Forensic Histopathological Recognition via a Context-Aware MIL Network Powered by Self-supervised Contrastive Learning

Chen Shen[1], Jun Zhang[4], Xinggong Liang[1], Zeyi Hao[1], Kehan Li[3], Fan Wang[3](✉), Zhenyuan Wang[1](✉), and Chunfeng Lian[2](✉)

[1] Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
wzy218@xjtu.edu.cn

[2] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi 710149, China
chunfeng.lian@xjtu.edu.cn

[3] Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
fan.wang@xjtu.edu.cn

[4] Tencent AI Lab, Shenzhen, China

Abstract. Forensic pathology is critical in analyzing death manner and time from the microscopic aspect to assist in the establishment of reliable factual bases for criminal investigation. In practice, even the manual differentiation between different postmortem organ tissues is challenging and relies on expertise, considering that changes like putrefaction and autolysis could significantly change typical histopathological appearance. Developing AI-based computational pathology techniques to assist forensic pathologists is practically meaningful, which requires reliable discriminative representation learning to capture tissues' fine-grained postmortem patterns. To this end, we propose a framework called FPath, in which a dedicated self-supervised contrastive learning strategy and a context-aware multiple-instance learning (MIL) block are designed to learn discriminative representations from postmortem histopathological images acquired at varying magnification scales. Our self-supervised learning step leverages multiple complementary contrastive losses and regularization terms to train a double-tier backbone for fine-grained and informative patch/instance embedding. Thereafter, the context-aware MIL adaptively distills from the local instances a holistic bag/image-level representation for the recognition task. On a large-scale database of $19,607$ experimental rat postmortem images and $3,378$ real-world human decedent images, our FPath led to state-of-the-art accuracy and promising cross-domain generalization in recognizing seven different postmortem tissues. The source code will be released on https://github.com/ladderlab-xjtu/forensic_pathology.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 528–538, 2023.
https://doi.org/10.1007/978-3-031-43987-2_51

Keywords: Forensic Pathology · Self-Supervised Learning · Multiple Instance Learning

1 Introduction

Computational pathology powered by artificial intelligence (AI) shows promising applications in various clinical studies [14,18], significantly easing the workload and promoting the development of clinical pathology. Inspired by such exciting progress, let's think step by step, so why not leverage advanced AI techniques to boost the research and applications in another important discipline, i.e., forensic pathology? Forensic pathology focuses on investigating the cause, manner, and time of (non-natural) deaths based on histopathological examinations of postmortem organ tissues [5]. As an indispensable part of the medicolegal autopsy, it provides critical evidence from the microscopic aspect to confirm, perfect, or refute macroscopic findings, establishing a reliable factual basis for future inferences [4]. Histopathological analysis in forensic pathology is challenging and time-consuming, since postmortem changes (e.g., putrefaction and autolysis) severely destroy tissues' typical image appearance, even making the manual differentiation between the tissues of different organs very difficult.

Although diverse deep-learning approaches have been proposed in clinical studies to process and analyze histopathological images [14,18], no similar work has yet in the forensic pathology community. The main reason could be threefold. **1)** Forensic and clinical pathology have distinct purposes. The former case analyzes the tissue images from multiple organs concurrently. In contrast, clinical diagnosis/prognosis usually focuses on one tissue type in one task [6]. **2)** Due to postmortem changes, histopathological images in forensic pathology have atypical appearances and more complex distributions than in clinical pathology, bringing additional challenges to deep representation learning [21,25]. **3)** Data in forensic pathology are more difficult to obtain and have relatively lower quality. Therefore, to deploy a reliable computational pathology system for forensic investigation, fine-grained discriminative representation learning from complex postmortem histopathological images is a very precondition.

In this paper, we introduce a deep computational pathology framework (*dubbed as* **FPath**) for forensic histopathological analysis. As shown in Fig. 1, FPath leverages the idea of self-supervised contrastive learning and multiple instance learning (MIL) to learn discriminative histopathological representations. Specifically, we propose a self-supervised contrastive learning strategy to learn a double-tier backbone network for fine-grained feature embedding of local image patches (i.e., instances in MIL). After that, a context-aware MIL block is designed, which adopts a self-attention mechanism to refine instance-level representations by aggregating contextual information, and then applies an adaptive-pooling operation to produce a holistic image-level representation for prediction. Our FPath performs efficient predictions without the need for tedious pre-processing (e.g., foreground extraction/segmentation). To the best of our knowledge, this paper is the first attempt that shows promising applications of advanced AI techniques (e.g., self-supervised contrastive learning) to forensic pathology.

The main technical contributions of our work are:

1) We design a double-tier backbone and a dedicated self-supervised learning strategy to capture discriminative instance-level histopathological patterns of postmortem organ tissues. The double-tier backbone combines CNN and transformer for local and non-local information fusion. To effectively train such a backbone to handle images acquired with varying microscopic magnifications, the dedicated self-supervised learning strategy leverages multiple complementary contrastive losses and regularization terms to concurrently maximize global and spatially fine-grained similarities between different views of the same instances/patches in an informative representation space.

2) We design a context-aware MIL branch to produce the bag-level discriminative representations for accurate and efficient postmortem histopathological recognition. Our MIL branch first refines instance embedding by leveraging a self-attention mechanism integrating positional embedding to model cross-patch associations for contextual information enhancement. Thereafter, an adaptive pooling operation is designed to learn deformable spatial attention to distill from contextually enhanced patch-level representations a holistic image-level representation for recognition.

3) Our FPath was applied to recognize postmortem organ tissues, a fundamental task in forensic pathology. To this end, we established a relatively large-scale multi-domain database consisting of an experimental rat postmortem dataset and a real-world human decedent dataset, each with 19, 607 and 3, 378 images acquired at a specific microscopic magnification (e.g., 5×, 10×, 20×, and 40×), respectively. On such a multi-domain database, our FPath led to promising cross-domain generalization and state-of-the-art accuracy in recognizing seven different postmortem organs.

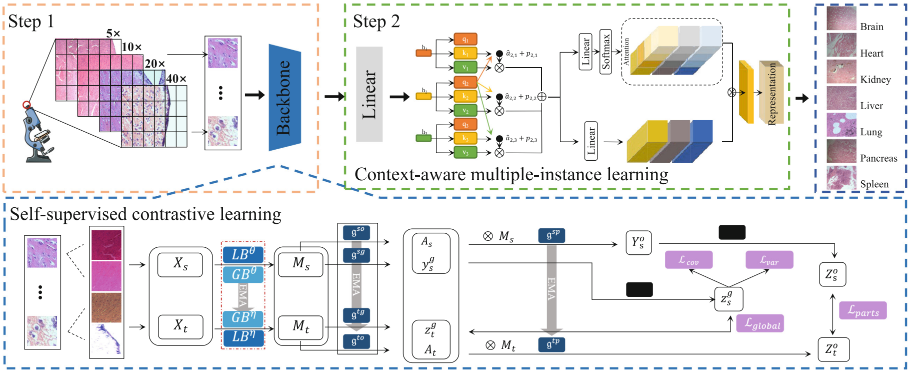

Fig. 1. Our FPath that consists of a self-supervised double-tier backbone (Step 1) and a context-aware MIL branch for postmortem recognition (Step 2).

2 Method

The schematic diagram of our FPath is shown in Fig. 1, which consists of two steps: **1)** Self-supervised contrastive learning of a double-tier backbone, and **2)** Context-aware multiple instance learning for postmortem tissue recognition.

2.1 Self-supervised Contrastive Patch Embedding

Double-Tier Backbone. Given patches from a postmortem histopathological image acquired at a specific magnification (i.e., 5×, 10×, 20×, or 40×), we adopt a backbone with a local branch (LB) and a global branch (GB) for instance/patch feature embedding. The LB is a ResNet50 [10] consisting of 16 successive bottlenecks, each with three convolutional layers with the kernel size of 1×1, 3×3, and 1×1, respectively. The GB is a Swin Transformer [17] that contains of a series of 12 window-based multi-head self-attention modules. Let an input patch be $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$. The corresponding feature embedding produced by the double-tier backbone will be $\mathbf{M} = \mathbf{M}_{\mathrm{LB}} \oplus \mathbf{M}_{\mathrm{GB}}$ $(\in \mathbb{R}^{h \times w \times C})$, where $\mathbf{M}_{\mathrm{LB}}$ and $\mathbf{M}_{\mathrm{GB}}$ denotes the representations from the LB and GB branch, respectively, and $\oplus$ stands for the channel-wise concatenation operation.

Self-supervised Contrastive Learning Strategy. We leverage the idea of self-supervised representation learning to establish the double-tier backbone. Referring to MoCo [9], our self-supervised learning is constructed by a teacher branch and a student branch. The student branch consists of six components, including a double-tier backbone (i.e., $\mathfrak{f}_\theta(\cdot)$), three projection layers (i.e., $\mathfrak{g}^{\mathrm{sg}}(\cdot)$, $\mathfrak{g}^{\mathrm{so}}(\cdot)$, $\mathfrak{g}^{\mathrm{sp}}(\cdot)$), and two prediction layers (i.e., $\mathfrak{p}^{\mathrm{sg}}(\cdot)$ and $\mathfrak{p}^{\mathrm{so}}(\cdot)$). The teacher branch contains four components, including a double-tier backbone $\mathfrak{f}_\eta(\cdot)$, and three projection layers (i.e., $\mathfrak{g}^{\mathrm{tg}}(\cdot)$, $\mathfrak{g}^{\mathrm{to}}(\cdot)$, and $\mathfrak{g}^{\mathrm{tp}}(\cdot)$). By feeding the two branches with different views of same patches, $\mathfrak{f}_\theta(\cdot)$ in the student branch (i.e., parameterized by θ) is trained via back-propagation to update $\mathfrak{f}_\eta(\cdot)$ in the teacher branch (i.e., parameterized by η) in a momentum-based moving average fashion, such as $\eta \leftarrow m \cdot \eta + (1 - m) \cdot \theta$, where $m = 0.99$ is the momentum parameter.

Another key issue that determines the quality of the embedding from such a self-supervised strategy is the formulation of respective contrastive loss functions and regularization terms. Accordingly, we design a thorough contrastive learning strategy to capture fine-grained discriminative patterns of postmortem tissues under varying microscopic magnifications. That is, let $\mathbf{X}_{\mathrm{s}}$ and $\mathbf{X}_{\mathrm{t}}$ be two different views of an image patch $\mathbf{X}$ generated by a random data augmentation process. Our contrastive learning strategy concurrently encourages the *global similarity* and *spatially fine-grained similarity* between the corresponding feature embedding $\mathbf{M}_{\mathrm{s}} = \mathfrak{f}_\theta(\mathbf{X}_{\mathrm{s}})$ and $\mathbf{M}_{\mathrm{t}} = \mathfrak{f}_\eta(\mathbf{X}_{\mathrm{t}})$ $(\in \mathbb{R}^{h \times w \times C})$. Also, *two regularization terms* are applied as auxiliary guidance to *protect the informativeness and avoid collapses* of the embedding learned by the backbone.

Specifically, the global similarity between $\mathbf{M}_{\mathrm{s}}$ and $\mathbf{M}_{\mathrm{t}}$ is encouraged by minimizing a general cosine contrastive loss, such as

$$\mathcal{L}_{\mathrm{global}} = 2 - 2 \cdot \frac{< \mathbf{z}_{\mathrm{s}}^{\mathrm{g}}, \mathbf{z}_{\mathrm{t}}^{\mathrm{g}} >}{||\mathbf{z}_{\mathrm{s}}^{\mathrm{g}}||_2 \cdot ||\mathbf{z}_{\mathrm{t}}^{\mathrm{g}}||_2}, \tag{1}$$

where $\mathbf{z}_{\mathrm{s}}^{\mathrm{g}} = \mathfrak{p}^{\mathrm{sg}}(\mathfrak{g}^{\mathrm{sg}}(\mathrm{GAP}(\mathbf{M}_{\mathrm{s}})))$ and $\mathbf{z}_{\mathrm{t}}^{\mathrm{g}} = \mathfrak{g}^{\mathrm{tg}}(\mathrm{GAP}(\mathbf{M}_{\mathrm{t}}))$, with $\mathrm{GAP}(\cdot)$ standing for the global average pooling that produces feature vectors.

In practice, forensic pathologists typically infer postmortem tissue type by evaluating the cellular compositions in multiple local regions. Accordingly, inspired by cross-view learning [11], we design a spatially fine-grained contrastive loss to explicitly encourage multi-parts similarity between $\mathbf{M}_{\mathrm{s}}$ and $\mathbf{M}_{\mathrm{t}}$. Assume $\mathbf{M}'_{\mathrm{s}}$ and $\mathbf{M}'_{\mathrm{t}}$ are two $(h \cdot w) \times C$ tensors flattened from $\mathbf{M}_{\mathrm{s}}$ and $\mathbf{M}_{\mathrm{t}}$ across the spatial dimension, respectively. They are further processed by $\mathfrak{g}^{\mathrm{so}}(\cdot)$ and $\mathfrak{g}^{\mathrm{to}}(\cdot)$ (followed by softmax normalization), respectively, to produce two $(h \cdot w) \times K$ attention matrices, i.e., $\mathbf{A}_{\mathrm{s}} = \mathfrak{g}^{\mathrm{so}}(\mathbf{M}'_{\mathrm{s}})$ and $\mathbf{A}_{\mathrm{t}} = \mathfrak{g}^{\mathrm{to}}(\mathbf{M}'_{\mathrm{t}})$, where K denotes the predefined number of parts. Thereafter, we aggregate the backbone representations in terms of the attention matrices to deduce multi-parts representations, i.e., $\mathbf{Z}_{\mathrm{s}}^{\mathrm{o}} = \mathfrak{p}^{\mathrm{so}}(\mathfrak{g}^{\mathrm{sp}}(\mathbf{A}_{\mathrm{s}}^T \otimes \mathbf{M}'_{\mathrm{s}}))$ and $\mathbf{Z}_{\mathrm{t}}^{\mathrm{o}} = \mathfrak{g}^{\mathrm{tp}}(\mathbf{A}_{\mathrm{t}}^T \otimes \mathbf{M}'_{\mathrm{t}})$, where $\otimes$ denotes tensor multiplication. Finally, the spatially fine-grained contrastive loss is quantified as

$$\mathcal{L}_{\mathrm{parts}} = \sum_{k=1}^{K} \left(2 - 2 \cdot \frac{< \mathbf{Z}_{\mathrm{s}}^{\mathrm{o}}[k,:], \mathbf{Z}_{\mathrm{t}}^{\mathrm{o}}[k,:] >}{||\mathbf{Z}_{\mathrm{s}}^{\mathrm{o}}[k,:]||_2 \cdot ||\mathbf{Z}_{\mathrm{t}}^{\mathrm{o}}[k,:]||_2} \right) \tag{2}$$

where $\mathbf{Z}^{\mathrm{o}}[k,:]$ denotes the kth part representation in $\mathbf{Z}^{\mathrm{o}} \in \mathbb{R}^{K \times D}$.

Besides, two additional regularization terms are further included to stabilize contrastive representation learning. Following [1], we penalize small changes between the global representations of different image patches across each feature dimension. Also, we encourage the global representations to be diverse/orthogonal across different feature dimensions. Let $\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}$ be a set of feature representations for an input mini-batch of patches in the student branch, and $\widetilde{\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}}$ and $\overline{\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}}$ denote their channel-wise variation and mean. The regularization terms are defined as

$$\mathcal{L}_{\mathrm{var}} = \frac{1}{D} \sum_{d=1}^{D} \max \left(0, 1 - \sqrt{\widetilde{\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}}[d] + \epsilon} \right) \tag{3}$$

$$\mathcal{L}_{\mathrm{cov}} = \frac{1}{D^2 - D} \sum_{i \neq j} \left(\left\{ (\mathbf{z}_{\mathrm{s}}^{\mathrm{g}} - \overline{\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}})^T (\mathbf{z}_{\mathrm{s}}^{\mathrm{g}} - \overline{\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}}) \right\} [i,j] \right)^2 \tag{4}$$

where ϵ is a small scalar to stabilize numerical computation, $\widetilde{\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}}[d]$ denotes the dth dimension of $\widetilde{\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}}$, and $\{(\mathbf{z}_{\mathrm{s}}^{\mathrm{g}} - \overline{\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}})^T (\mathbf{z}_{\mathrm{s}}^{\mathrm{g}} - \overline{\mathcal{Z}_{\mathrm{s}}^{\mathrm{g}}})\}[i,j]$ is the $[i,j]$th element in such a covariance matrix. According to [1], Eqs. (3) and (4) jointly encourage the diversity across patches and feature dimensions, thus protecting the informativeness and avoid collapse of self-supervised contrastive learning.

Overall, we combine Eqs. (1) to (4) as the final loss function to train the double-tier backbone, such as $\mathcal{L}_{\mathrm{all}} = \mathcal{L}_{\mathrm{global}} + \mathcal{L}_{\mathrm{parts}} + \gamma \mathcal{L}_{\mathrm{var}} + \lambda \mathcal{L}_{\mathrm{cov}}$, where γ and λ are two tuning parameters balancing different terms.

2.2 Context-Aware MIL

Given the patch/instance-level representations of a histopathological image from the double-tier backbone, we further design a context-aware MIL framework to aggregate their information for postmortem tissue recognition. Given patch embeddings of a Microscope image, our context-aware MIL part contains two main steps, i.e., a multi-head self-attention to refine each patch's feature and an adaptive pooling step to distill all patches' information.

In detail, we first adopt a multi-head self-attention (MSA) mechanism [19] integrating relative positional embedding to explicitly model cross-patch associations for contextual enhancement of the instance representations from the backbone. Let $\mathscr{Z} = \{\mathbf{z}_i\}_{i=1}^{I}$ be a set of the contextually enhanced instance embedding from an image. Thereafter, inspired by Deformable DETR [28], we further design an adaptive pooling operation, which is simple but effective to distill from $\mathscr{Z}$ a bag-level holistic representation for the classification purpose. Specifically, the bag-level holistic representation determined by the adaptive pooling is

$$\mathbf{z}_{\text{bag}} = \frac{1}{I}\sum_{i=1}^{I}(softmax(\mathfrak{h}_{\omega_1}(\mathbf{z}_i)) \circ \mathfrak{h}_{\omega_2}(\mathbf{z}_i)), \tag{5}$$

where $\mathfrak{h}_{\omega_1}(\cdot)$ and $\mathfrak{h}_{\omega_2}(\cdot)$ are two linear projections with the same number of output units, symbol $\circ$ denotes the Hadamard product between two tensors, and $softmax(\cdot)$ is performed across different instances to filter out uninformative patches and preserve discriminative patches in quantifying $\mathbf{z}_{\text{bag}}$ for classification.

3 Experiments

3.1 Data and Experimental Setup

Rat Postmortem Histopathology Dataset. Ninety Sprague-Dawley adult male rats were executed by the spinal cord dislocation and placed in a constant temperature and humidity environment for 6–8 h. The animal experiments were approved by the Laboratory Animal Care Committee of the anonymous institution. Seven organs, i.e., brain, heart, kidney, liver, lung, pancreas, and spleen, were removed and placed in the formalin solution. Briefly, paraffin sections of these organ tissues were stained with the H&E solution. The H&E-stained sections were then analyzed by three forensic pathologists, who used Lercai LAS EZ microscopes to record the areas according to their expertise. Overall, five to ten images were recorded from a section at each magnification (i.e., 5×, 10×, 20×, and 40×). Finally, we split the 90 rats as training, validation, and test sets of 60, 10, and 20 rats, respectively, each with $13,137$, $2,235$, and $4,325$ images.

Human Forensic Histopathology Dataset. The real forensic images were provided by the Forensic Judicial Expertise Center of the anonymous institution, after getting the informed consent of relatives. All procedures followed the requirements of local laws and institutional guidelines, and were approved and supervised by the Ethics Committee. A total of 32 decedents participated in this study. Four to six images were recorded at each of three magnifications (5×, 10×, and 20×) per H&E stained section. Similar to the rat dataset, the human dataset was selected from the same seven organs. Finally, the training, validation and test sets contain $1,691$ images, 628, and 1059 images, corresponding to 16, 6, and 10 different decedents, respectively.

Experimental Details. Notably, the double-tier backbone was self-supervised and learned on the rat training set for 100 epochs by setting the mini-batch size as 1024, with the parameters initialized by the ImageNet pre-trained models. The training data were augmented by a histopathology-oriented strategy by combining different kinds of staining jitters, random affine transformation, Gaussian blurring, resizing, etc. The image(patch) dimension in our implementation was 224*224. The tuning parameters γ and λ in $\mathcal{L}_{\text{all}}$ were set as 5 and 0.005, respectively. Thereafter, the MIL blocks on two different datasets were both trained by minimizing the cross-entropy loss for 20 epochs with the mini-batch size setting as 32. The experiments were conducted on three PCs with twenty NVIDIA GEFORCE RTX 3090 GPUs.

3.2 Results of Self-supervised Contrastive Learning

Our self-supervised double-tier backbone was compared with other state-of-the-art self-supervised learning approaches, including **balow twins** [27], **swin transformer (SSL)** [26], **TransPath** [23], **CTransPath** [24],**RetCCL** [22] and **MOCOV3** [3]. To evaluate the discriminative power of these competing methods, we adopted GAP to aggregate their instance representations from a whole image to train simple linear classifiers for the recognition of seven different organ tissues on both the rat and human datasets, with the test performance quantified in terms of four general classification metrics (i.e., **ACC**, **F1 score**, **MCC(Matthews Correlation Coefficient)**, and **Precision**). The corresponding results are summarized in Table 1, from which we can have two observations. *First*, our self-supervised double-tier backbone consistently outperformed all other competing methods in terms of all metrics on two datasets. *Second*, our method led to better generalization, as the backbone trained on the rat dataset shows promising performance on the challenging real-world human dataset (e.g., resulting in an ACC higher than 90%). These results suggest the effectiveness of our self-supervised learning strategy.

For a more detailed evaluation, we further conducted a series of ablation studies to evaluate the contributions of the contrastive losses (i.e., $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{parts}}$) and regularization strategy (i.e., $\mathcal{L}_{\text{var}} + \mathcal{L}_{\text{cov}}$). The corresponding results are summarized in Table 2, from which we can see that, given the baseline of $\mathcal{L}_{\text{global}}$, both the inclusion of the spatially fine-grained contrastive loss (i.e., $\mathcal{L}_{\text{parts}}$) and informativeness regularization (i.e., $\mathcal{L}_{\text{var}}$ and $\mathcal{L}_{\text{cov}}$) led to respective performance gains. These results further justify our self-supervised design.

Table 1. Linear classification results obtained by different self-supervised learning approaches on the rat and human testing sets, respectively.

Competing methods	Rat dataset				Human dataset			
	ACC	F1	MCC	Precision	ACC	F1	MCC	Precision
balow twins [27]	0.9232	0.9123	0.9076	0.9070	0.7306	0.7311	0.6854	0.7345
swin transformer(SSL) [26]	0.9450	0.9369	0.9299	0.9330	0.8079	0.8088	0.7758	0.8125
Transpath [23]	0.7351	0.7397	0.6958	0.7481	0.5838	0.5657	0.5264	0.6050
CTransPath [24]	0.9635	0.9610	0.9535	0.9596	0.8794	0.8799	0.8591	0.8842
RetCCL [22]	0.9794	0.9801	0.9768	0.9810	0.7796	0.7789	0.7448	0.7961
MOCOV3 [3]	0.9732	0.9738	0.9681	0.9745	0.8103	0.8124	0.7790	0.8187
Ours	**0.9831**	**0.9831**	**0.9796**	**0.9831**	**0.9049**	**0.9044**	**0.8886**	**0.9056**

Table 2. Ablation studies to evaluate the contributions of different self-supervised contrastive losses and regularization terms.

Loss functions				Rat dataset				Human dataset			
$\mathcal{L}_{\text{global}}$	$\mathcal{L}_{\text{parts}}$	$\mathcal{L}_{\text{var}}$	$\mathcal{L}_{\text{cov}}$	ACC	F1	MCC	Precision	ACC	F1	MCC	Precision
✓				0.9732	0.9713	0.9660	0.9689	0.8918	0.8913	0.8734	0.8935
✓	✓			0.9817	0.9819	0.9778	0.9822	0.8953	0.8956	0.8779	0.8981
✓		✓	✓	0.9793	0.9799	0.9757	0.9806	0.8978	0.8976	0.8802	0.8983
✓	✓	✓	✓	**0.9831**	**0.9831**	**0.9796**	**0.9831**	**0.9049**	**0.9044**	**0.8886**	**0.9056**

3.3 Results of Multiple-Instance Learning

Based upon the double-tier backbone learned on the rat training set, we compared our context-aware MIL with other MIL methods, including the gated attention-based approach (i.e., **AB-MIL** [12], **DSMIL** [15], **Transmil** [19] and **MSA** [2,16]) approaches with/without different positional embedding strategies, i.e., relative position embedding (**MSA-RP** [17]), learnable position embedding (**MSA-LP** [7]), and 2D sine-cosine position embedding (**MSA-SP** [8]). Notably, our approach used MSA-RP as the baseline, based on which an adaptive pooling operation is designed to produce the final bag-level representation. To check the efficacy of **adaptive pool**, we further conducted a corresponding set of ablation studies by replacing it with other operations, including **max pool**, and **soft pool** [20]. These comparison and ablations results are shown in Table 3, from which we can observe that our method led to the best results on both datasets, with relatively more significant improvements on the challenging human dataset. Also, compared with other pooling operations, the adaptive pool design brought consistent performance gains. These results suggest the efficacy of our context-aware MIL for postmortem tissue recognition.

In addition, we conducted LayerCAM-based analysis [13] to check the explainability and reliability of our postmortem histopathological recognition results. From the representative examples shown in Fig. 2, we can have an interesting observation that our method tends to focus on tissue-specific postmortem patterns at different microscopic scales. For example, the spatial attention maps reliably highlighted the meningeal structures of the brain tissue, the glomeruli in the kidney cortex, and the central vein area between the liver lobules. On

Table 3. Multiple-instance learning results obtained by the competing methods and our Context-Aware MIL with different pooling strategies.

Competing methods	Rat dataset				Human dataset			
	ACC	F1	MCC	Precision	ACC	F1	MCC	Precision
AB-MIL [12]	0.9815	0.9828	0.9793	0.9844	0.9011	0.9005	0.8838	0.9050
DSMIL [15]	0.9951	0.9948	0.9937	0.9945	0.9176	0.9166	0.9030	0.9170
Transmil [19]	0.9899	0.9888	0.9875	0.9878	0.8824	0.8813	0.8622	0.8821
MSA [2,16]	0.9875	0.9883	0.9861	0.9892	0.9082	0.9082	0.8921	0.9100
MSA-LP [7]	0.9879	0.9875	0.9853	0.9873	0.9097	0.9087	0.8945	0.9109
MSA-SP [8]	0.9851	0.9839	0.981	0.9832	0.8915	0.8905	0.8748	0.8948
MSA-RP [17]	0.9915	0.9915	0.9896	0.9916	0.9218	0.9213	0.9085	0.9218
Ours + Max pool	0.9910	0.9909	0.9888	0.9909	0.9144	0.9147	0.9001	0.9191
Ours + Soft pool [20]	0.9935	0.9929	0.9915	0.9924	0.9047	0.9023	0.8883	0.9056
Ours + Adaptive pool	**0.9956**	**0.9952**	**0.9943**	**0.9949**	**0.9229**	**0.9218**	**0.9093**	**0.9263**

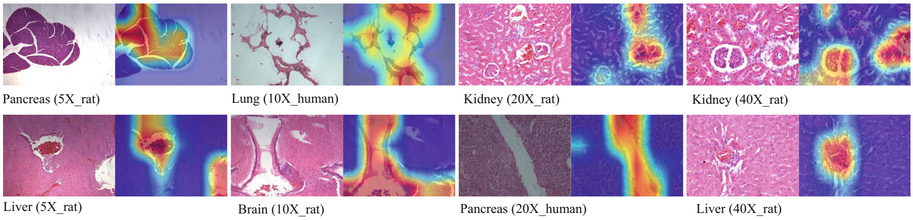

Fig. 2. Explainability analysis based on LayerCAM [13] for representative postmortem tissue images acquired at different microscopic scales.

the other hand, based on the pancreas example, we can see that our network can sensitively localize the pancreas glandular structure while filtering out the uninformative background in an end-to-end fashion, without the need for any pre-processing to segment first the foreground. These observations support our assumption that the proposed method is reliable and efficient in learning discriminative histopathological representations of postmortem organ tissues.

4 Conclusion

In this study, we have proposed a context-aware MIL framework powered by self-supervised contrastive learning to learn fine-grained discriminative representations for postmortem histopathological recognition. The dedicated self-supervised learning strategy concurrently maximizes multiple contrastive losses and regularization terms to deduce informative and discriminative instance embedding. Thereafter, the context-aware MIL framework adopts MSA followed by an adaptive pooling operation to distill from all instances a holistic bag/image-level representation. The experimental results on a relatively large-scale database suggest the state-of-the-art postmortem recognition performance of our method.

Funding Information. This work was supported in part by NSFC Grants (Nos. 62101430 & 62101431).

References

1. Bardes, A., Ponce, J., LeCun, Y.: VICReg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021)
2. Chen, R.J., et al.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144–16155 (2022)
3. Chen, X., Xie, S., He, K.: An Empirical Study of Training Self-Supervised Vision Transformers. arXiv e-prints (2021)
4. De La Grandmaison, G.L., Charlier, P., Durigon, M.: Usefulness of systematic histological examination in routine forensic autopsy. J. Forensic Sci. **55**(1), 85–88 (2010)
5. DiMaio, D., DiMaio, V.J.: Forensic Pathology. CRC Press, Boca Raton (2001)
6. Dolinak, D., Matshes, E., Lew, E.O.: Forensic Pathology: Principles and Practice. Elsevier, Amsterdam (2005)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum Contrast for Unsupervised Visual Representation Learning. arXiv e-prints (2019)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
11. Huang, L., You, S., Zheng, M., Wang, F., Qian, C., Yamasaki, T.: Learning where to learn in cross-view self-supervised learning. In: CVPR (2022)
12. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based Deep Multiple Instance Learning. arXiv e-prints (2018)
13. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: LayerCAM: exploring hierarchical class activation maps for localization. IEEE Trans. Image Process. **30**, 5875–5888 (2021)
14. Lee, Y., et al.: Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nat. Biomed. Eng. (2022)
15. Li, B., Li, Y., Eliceiri, K.W.: 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, pp. 14318–14328 (2021)
16. Li, H., et al.: DT-MIL: deformable transformer for multi-instance learning on histopathological image. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 206–216. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_20
17. Liu, Z., et al.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv e-prints (2021)
18. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. **5**(6), 555–570 (2021)

19. 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. Adv. Neural. Inf. Process. Syst. **34**, 2136–2147 (2021)
20. Stergiou, A., Poppe, R., Kalliatakis, G.: Refining activation downsampling with softpool. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10357–10366 (2021)
21. Wang, G., et al.: An emerging strategy for muscle evanescent trauma discrimination by spectroscopy and chemometrics. Int. J. Mol. Sci. **23**(21), 13489 (2022)
22. Wang, X., et al.: RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. Med. Image Anal. **83**, 102645 (2023)
23. Wang, X., et al.: TransPath: transformer-based self-supervised learning for histopathological image classification. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 186–195. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_18
24. Wang, X., et al.: Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. **81**, 102559 (2022)
25. Wu, H., et al.: Pathological and ATR-FTIR spectral changes of delayed splenic rupture and medical significance. Spectrochim. Acta. A Mol. Biomol. Spectrosc. **278**, 121286 (2022)
26. Xie, Z., et al.: Self-Supervised Learning with Swin Transformers. arXiv preprint arXiv:2105.04553 (2021)
27. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. arXiv preprint arXiv:2103.03230 (2021)
28. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv e-prints (2020)