scientific reports

Check for updates

OPEN

Human visual attention-inspired knowledge distillation underlying interpretable computational pathology

Muzhou Yu[1,8], Zihan Zhong[2,3,8], Xingang Zhou[4], Yuekun Wang[2], Tingyu Liang[2], Jiamin Chen[4], Hongmin Huang[3], Junhan Zhou[3], Dachun Zhao[5], Bo Lei[6]✉, Yu Wang[2]✉, Wenbin Ma[2]✉ & Kaisheng Ma[7]✉

Computational pathology leverages advanced deep-learning techniques to analyze medical images with high resolution. However, a trade-off exists between model lightweight, interpretability, and task performance in such real-world scenarios. Knowledge distillation (KD) is widely applied to compress deep learning models while preserving high performance. Despite this, deep learning-based KD often lacks interpretable design, leading to inaccurate attention to images. Inspired by human vision processing, we developed a human vision attention-inspired knowledge distillation (HVisKD) strategy that captures local and global patch relations to construct differentiated features. We employed it in pathological analysis to balance the tradeoff. HVisKD improves performance across various lightweight models in segmentation tasks. More importantly, the attention map of HVisKD showed promoted consistency with human expert-labeled segmentation. Furthermore, we examined HVisKD in a real-world intraoperative pathological diagnosis scenario and achieved an interpretable and fast analysis. Together, HVisKD offers a lightweight and interpretable strategy for computational pathology, aligning deep learning with brain-like information processing for more dependable output.

Keywords Computational pathology, Medical imaging, Bio-inspiration, Deep learning, Knowledge distillation

Computational pathology leverages advanced computational techniques, particularly artificial intelligence, to analyze high-quality whole-slide images (WSIs)[1,2], requiring both high performance and interpretability in real-world clinical diagnosis. Given the complexity and variety of pathological tissues, rapidly determining and precisely annotating subtypes on WSIs is crucial for biomarker discovery[3], therapeutic effect quantification[4] and intraoperative malignant tissue detection[5]. Commonly, this involves breaking down WSIs into smaller patches and employing deep convolutional neural networks for subtype segmentation[6]. However, the huge computational demands of these high-performing networks limit their practical clinical application.

Knowledge distillation (KD)[7,8] is a typical approach for developing a lightweight model (student model) by utilizing a pre-trained model (teacher model) to guide the learning process, thereby achieving the fast and precise inference[9–11] of the student model. The key aspect of KD is transferring important knowledge from the large well-trained teacher model to the light-weight student models during training. This transfer is typically achieved by aligning the logits[12,13], features[14,15], or data relations[16,17] between the two models. Although these strategies effectively improve learning performance in traditional computer vision, many studies often ignore the interpretability of the distillation process. It potentially causes inaccurate diagnosis and imprecise attention

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Shaanxi, China. [2]Department of Neurosurgery, Peking Union Medical College Hospital, Beijing, China. [3]Eight-Year Medical Doctor Program, China Academy of Medicine & Peking Union Medical College, Beijing, China. [4]Department of Pathology, Beijing Ditan Hospital, Capital Medical University, Beijing, China. [5]Department of Pathology, Peking Union Medical College Hospital, Beijing, China. [6]Beijing Academy of Artificial Intelligence, Beijing, China. [7]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. [8]Muzhou Yu and Zihan Zhong have contributed equally to this work. [9]Part of this work was carried out by Muzhou Yu during his internship at the Beijing Academy of Artificial Intelligence. ✉email: b.lei.2022@hotmail.com; ywang@pumch.cn; mawb2001@hotmail.com; kaishengthu@163.com

region on WSIs, which limits its application in practical pathology. Thus, balancing the preservation of task performance and improving interpretability during KD is still challenging for its application in real-world scenarios.

For human visual system, an effective mechanism to recognize key parts of a complete scene, focus precisely on specific areas, and integrate them for further processing is indispensable. Hierarchical attention mechanisms in the human visual system provide the foundation for balancing global information and local details[18,19]. Specifically, human vision exemplifies a complex interplay of global and local attentional processes, allowing for rapid and precise perception of visual information at both the scene and object levels. This capability is largely attributed to the hierarchical organization of visual attention, which seamlessly integrates broad contextual understanding with specific details[20,21]. Human pathologists, when examining histopathological slides, combine insights gained from both low-power and high-power magnifications to arrive at a final diagnosis. Low magnification provides an overview of tissue architecture and patterns, guiding focus to key areas, while high magnification examines fine cellular detail. This layered approach ensures thorough and accurate analysis (Fig. 1a). Based on this observation, we speculate that this biological visual process is analogous to the way

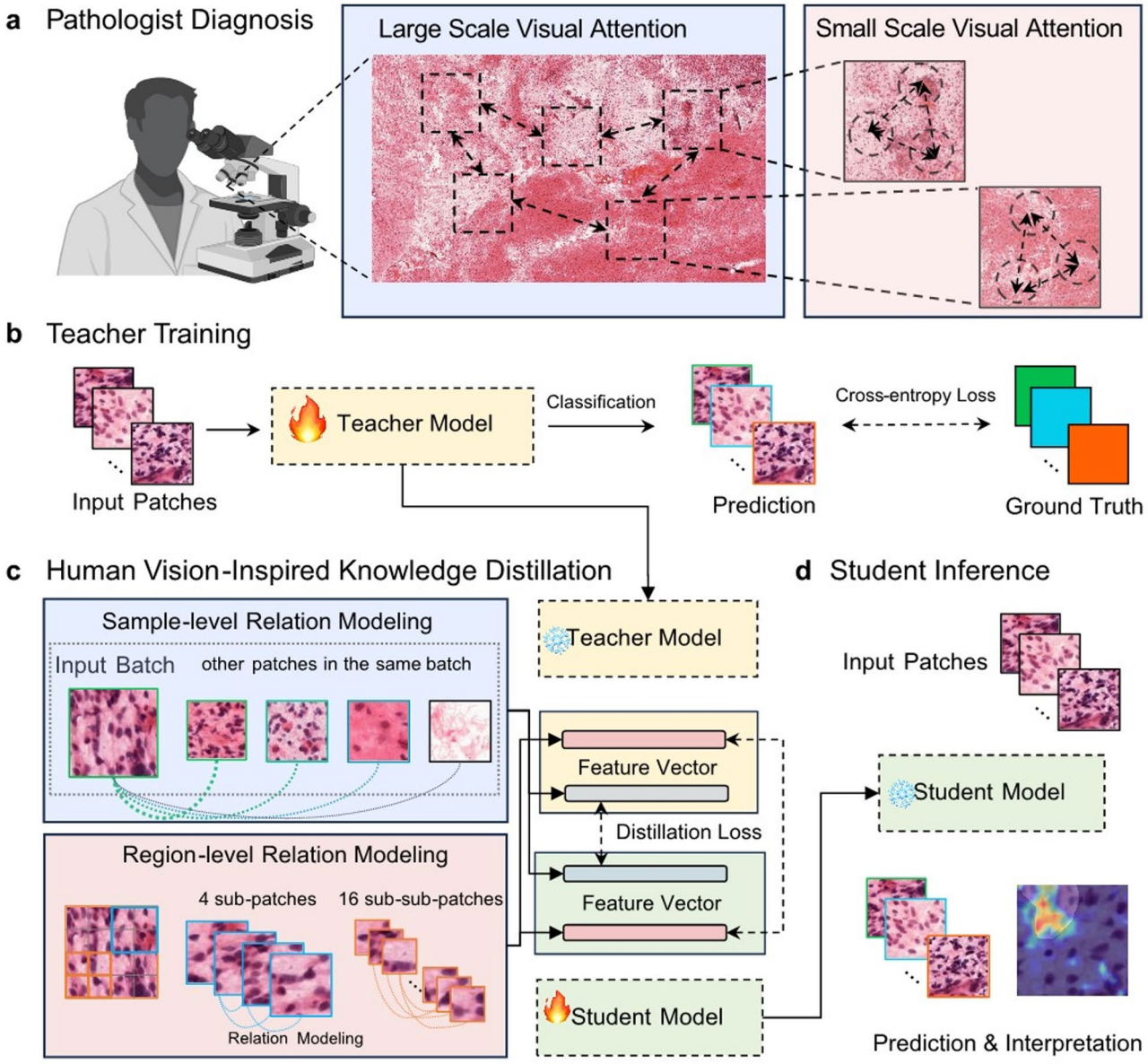


Fig. 1. The overview of proposed HVisKD for pathology WSI segmentation. (**a**) Human pathologists often combine insights gained from both low-power and high-power magnifications to arrive at a final diagnosis. (**b**) The pretraining process of the teacher model. (**c**) The distillation process of the student model via HVisKD. We construct both sample- and region- level relation among patches to transfer distinguished features from the teacher model to the student. (**d**) The inference process of the student model to predict the subtype of different patches and assemble them into a whole segmentation map.

doctors diagnose medical images, as both involve the process of hierarchical visual attention from various views to make informed decisions. Thus, we hope to incorporate this principal into knowledge distillation to achieve fast, efficient, and interpretable diagnosis in computational pathology.

In this work, we propose a novel algorithm, a human visual attention-inspired knowledge distillation (HVisKD). It constructs different visual levels of relation to allocate attentional knowledge, balancing the preservation of task performance and the interpretability of light models. Specifically, it builds long-range dependency among different patches at both the sample and region levels of pathological images, to allocate inter- and intra- relation. In this way, we can generate more differentiated features via relation modeling and transfer this knowledge to a light student model. With more effective information covering global perspectives and local details, the student model can present more precise attention on key parts of clinical symptoms. Given that our human vision-inspired design aligns naturally with the 2D spatial hierarchy of CNN features and that CNNs remain the predominant backbone in medical knowledge distillation, we focus our method on CNN-based architectures. At the same time, the limitations of directly applying HVisKD to transformer-based models[22–24] and potential strategies for extending it to ViTs are also described in detail in the Discussion section. We apply HVisKD to computational pathological analysis for efficient and interpretable visual evaluation in several datasets. Our approach not only improves KD performance across various light models but also aligns the attention focus with expert-labeled regions, demonstrating unprecedented interpretability.

Results

The pipeline of proposed HVisKD for pathology WSI segmentation

The model training and inference pipeline of HVisKD for pathology WSI segmentation is illustrated (Fig. 1). We pre-train our teacher model by inputting the patches from different WSI with category labels (Fig. 1b). The goal of the training is to enable the teacher model to classify the patches as accurately as possible. The cross-entropy loss is applied between the segmentation prediction and the ground truth label of each patch so as to optimize the teacher model.

Then, we construct discriminated features via relation modeling among feature space from sample-level and region-level, respectively (Fig. 1c). The sample-level relation modeling focuses on the relation of patches in a training batch that we construct the patch relation-aware features. In detail, we input a patch to obtain the feature from the model backbone, then we aggregate this feature with other features of other patches in a batch with weighted summation, where the weights are decided by the feature similarities between patches. Thus, we can obtain the patch relation-aware feature of the input patch where the feature representations are enhanced via consolidating similar features from other patches. Finally, we can enhance all the features of patches in a batch by mutually feature clustering, in which the features of different patches are discriminated. The region-level relation modeling emphasizes the relation of distinct regions in a patch. The region of tissue represents more class-specific information for model recognition compared to the surrounding regions. In order to make the class-relevant region more discriminated and represented, we construct the region-aware features. We divide the input patch into multiple small pieces and then build the pieces into sub-regions with various scales. Similarly to the patch relation-aware feature, we model the region relation-aware features via weighted fusing multi-scale regions within a feature map of a patch. The weights also depend on the similarities among regions. In this way, the features of regions with similar features can be mutually improved and thus enlarge the discrimination of dissimilar features.

Finally, we distill the two-level relation-aware feature from the teacher model to the student model, so as to achieve the purpose of making the features of the student discriminated, and the patch classification performance of the student model can be effectively improved (Fig. 1c). After distillation, the student model predicts the categories of patches, and then we assemble all the patches to obtain the segmentation map of WSI (Fig. 1d). In all, the design of HVisKD can be closer to the logic and principle of pathologists in processing pathological images, which conducts global and local relationship analysis, making KD method more interpretable and reliable in pathology analysis applications. For more detailed method description, please refer to our Method section.

HVisKD achieves superior performance on the ivyGAP pathology dataset

To validate HVisKD performance on pathology task, we collected 793 glioblastoma multiform (GBM) frozen section WSIs with tissue subtype annotations from ivyGAP[25] and tessellated each WSI to 75×75 μm patches to construct dataset for model training and validation (Fig. 2a and b).

To the best of our knowledge, we are the first to design model distillation of lightweight models for effective inference on the ivyGAP dataset. We evaluate the proposed HVisKD with the student model trained from scratch and a strong baseline, the original KD[7].

We show the Top-1 and Top-5 accuracy on 10 different teacher-student pairs (Table 1). The pairs include both similar and different teacher-student architectures. We report the mean ± standard deviation of our method HVisKD with the KD, based on results from ten independent runs. It is observed that our method HVisKD consistently surpasses the student and KD by a large margin in all model pairs. Meanwhile, we also visualize the improvement brought by HVisKD via point gap illustration (Fig. 2d), in which the performance of student trained from scratch is represented as diagonal line. In conclusion, the results indicate that HVisKD has better segmentation performance and generalization ability on the ivyGAP dataset.

To visualize the distinguishability, the area under the receiver operating characteristic curve (AUROC) alternatively proves HVisKD achieves better results (Fig. 2c, e and f). We present the AUROC results of five pairs, including VGG19-ShuffleV1, VGG19-MobileNetV2, VGG19-ShuffleNetV2, ResNet110-ResNet20, and ResNet32 × 4-ResNet8 × 4 (Fig. 2, Suppl. Fig. 1). Specifically, we choose the VGG19-ShuffleV1 pair as an example to indicate that our method can superiorly recognize tissue subtypes, averagely outperforming KD by 1.5% over

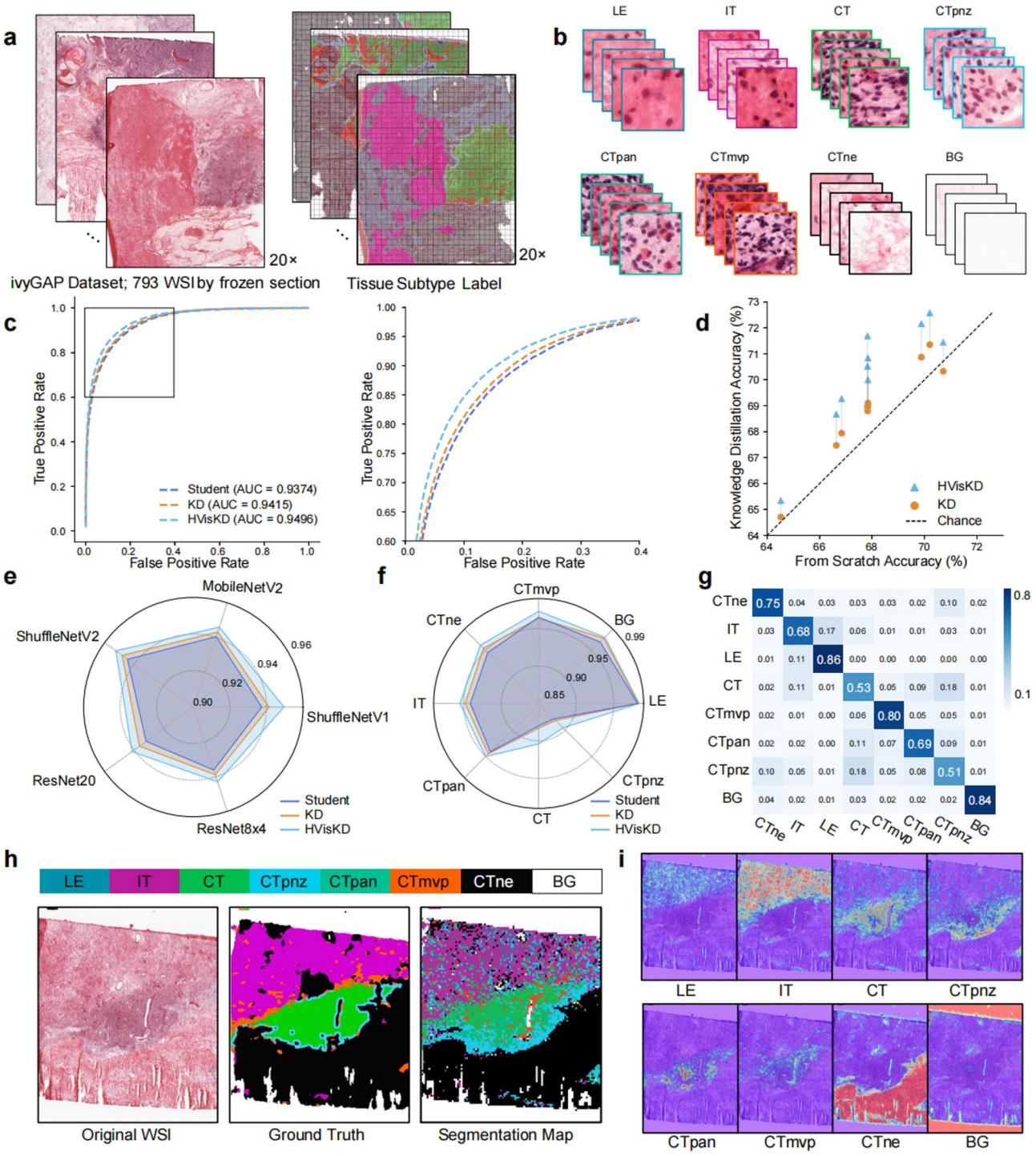


Fig. 2. Performance of HVisKD on the pathology ivyGAP dataset. (**a**) The ivyGAP dataset. (**b**) Glioblastoma tissue subtypes, including Leading Edge (LE), Infiltrating Tumor (IT), Cellular Tumor (CT), Pseudopalisading Cells Around Necrosis (CTpnz), Pushing Tumor Border (CTpan), Microvascular Proliferation (CTmvp), Tumor-Associated Necrosis (CTne), and Background (BG). (**c**) The AUROC curves of student model ShuffleNetV1 distilled from teacher model VGG19 are presented, where the results are averaged over all subtypes. We compare the results of HVisKD with KD and the student trained from scratch. (**d**) The illustration of accuracy improvement brought by HVisKD compared to KD. The chance line represents the performance of student trained from scratch. All the ten point pairs are from the ten teacher-student pairs in Table 1. (**e**) The comparison of AUROC values across different distillation pairs. (**f**) The comparison of AUROC values across different tissue subtypes. (**g**) The confusion matrix is applied to give a clearer visualization on the discriminability of HVisKD-distilled model across different subtypes. (**h**) The example segmentation results of WSI. It includes an original WSI, the segmentation map predicted by HVisKD, and the ground truth. (**i**) Besides, we show the attention map visualization of each subtype from the predicted segmentation map.

Teacher	Student	Teacher Top-1	Student Top-1	KD Top-1 (Mean ± SD)	HVisKD Top-1 (Mean ± SD)	Teacher Top-5	Student Top-5	KD Top-5 (Mean ± SD)	HVisKD Top-5 (Mean ± SD)
Resnet110	ResNet8	71.04	64.53	64.76 ± 0.05	65.32 ± 0.15	93.52	90.66	91.36 ± 0.26	91.08 ± 0.11
Resnet110	ResNet20	71.04	66.64	67.65 ± 0.17	68.72 ± 0.06	93.52	91.60	92.51 ± 0.09	92.83 ± 0.20
VGG19	VGG8	75.53	70.71	70.67 ± 0.36	71.23 ± 0.16	95.09	93.65	93.63 ± 0.13	93.71 ± 0.14
ResNet32 × 4	ResNet8 × 4	73.36	66.84	67.82 ± 0.51	69.15 ± 0.10	94.61	92.08	92.67 ± 0.15	92.77 ± 0.22
VGG19	ShuffleNetV2	75.53	69.88	71.15 ± 0.28	71.67 ± 0.41	95.09	93.55	94.02 ± 0.22	93.77 ± 0.38
VGG19	MobileNetV2	75.53	67.85	68.77 ± 0.25	69.72 ± 1.10	95.09	92.77	93.38 ± 0.28	93.42 ± 0.20
VGG19	ShuffleNetV1	75.53	67.83	69.14 ± 0.48	71.34 ± 0.24	95.09	92.48	93.22 ± 0.06	93.90 ± 0.06
Resnet32 × 4	ShuffleNetV1	73.36	67.83	68.67 ± 0.29	70.70 ± 0.26	94.61	92.48	93.05 ± 0.08	93.75 ± 0.42
Resnet32 × 4	MobileNetV2	73.36	67.85	68.61 ± 0.25	69.42 ± 0.35	94.61	92.77	93.05 ± 0.17	93.29 ± 0.08
Resnet32 × 4	ShuffleNetV2	73.36	70.20	70.59 ± 0.57	72.03 ± 0.47	94.61	93.68	93.84 ± 0.29	94.32 ± 0.06

Table 1. Top-1 and Top-5 test accuracy (%) of the student model on ivyGAP dataset among HVisKD, KD and the student model trained from scratch. 10 different teacher-student architecture pairs are performed, including 4 similar pairs and 6 different pairs (from top to bottom). We report the mean ± standard deviation of our method HVisKD with the KD, based on results from ten independent runs.

all subtypes (Fig. 2c). The AUROC results of each subtype is presented in the Supplementary (Suppl. Fig. 2b). For other pairs, we report the averaged results across all subtypes (Suppl. Fig. 2a) in the Supplementary. We also show the performance comparisons across different pairs and different tissue subtypes via radar chart (Fig. 2e and f) to demonstrate the generalization improvement of HVisKD. It's observed that each node and whole area of HVisKD can fully cover and exceed the original KD and the scratched student model. HVisKD reaches higher performance over all pairs and subtypes compared to other models.

Furthermore, we apply the confusion matrix to give a clearer visualization of how our model distinguishes between different subtypes (Fig. 2g, Suppl. Fig. 2c and d). HVisKD demonstrates superior performance in distinguishing confused subtypes. Notably, for the sub-confusion matrix containing CT-based subtypes, including CT, CTmvp, CTpan, and CTpnz, HVisKD significantly reduces the segmentation confusion between these subtypes. This demonstrates that the human-inspired relation modeling design for constructing and distilling distinguished features is beneficial for the student model to better distinguish different subtypes.

Apart from the quantification, we also present the visualization results of HVisKD on the ivyGAP tested WSIs (Fig. 2h and i). The example segmentation map of the WSI generated by the student model learned through HVisKD is generally consistent with the ground truth (Fig. 2h). The attention map for each tissue subtype also accurately focuses on the areas where the labels are located (Fig. 2i). Our proposed method HVisKD effectively enhances the lightweight model's capability to make precise predictions on pathology WSI, thereby assisting in accurate diagnoses in real-world scenarios.

We also design an additional experiment to further demonstrate the robustness in which we introduce the same image perturbations (Gaussian noise) to ivyGAP dataset (Suppl. Table 1). Our results show that student models trained with HVisKD exhibit significantly less performance degradation under noise, compared to both the student (non-distilled) and KD method.

We visualize the learning curves of the training and testing losses of KD and HVisKD over epochs. As shown in the figure (Suppl. Fig. 3), several observations can be made: (1) HVisKD shows higher loss variance due to intermediate feature supervision; (2) its loss curves converge more smoothly than KD; (3) it achieves greater overall loss reduction despite KD's faster initial drop; and (4) the smaller train-test gap suggests better generalization.

We also supplement the results from Table 1 with additional analysis (Suppl. Fig. 4). The results showed that our HVisKD method significantly improved the distilled model performance against others.

Moreover, to better demonstrate the outstanding performance of the proposed HVisKD compared current SOTA distillation methods, we also experiment on the standard computer vision datasets CIFAR100[26] and ImageNet[27] (Suppl. Tables 2 and 3, and Suppl. Fig. 5). The comparison SOTA methods are listed in Suppl. Experiment setting section[28–44]. The results show that HVisKD consistently present superior performance compared to other SOTA methods. We further conduct the ablation and sensitivity study of HVisKD to demonstrate the effectiveness and robustness (Suppl. Table 4, and Suppl. Figs. 6 and 7).

HVisKD emulates the visual attention of human pathologists

Considering the key concept of HVisKD inspired by the human visual process, we speculate that the HVisKD-distilled model can learn features comparable to those recognized by human pathologists. Microvascular proliferation, a key histopathological feature of GBM associated with poor prognosis[45,46], is classified as the CTmvp subtype in the ivyGAP dataset. Pathologists could identify microvascular regions from the background tumor tissue within patches. Thus, we engage pathologists with more than 10 years of experience in the neuro-oncology field to show their attention by labeling microvascular areas in CTmvp patches. The attention maps of HVisKD-distilled model, KD-distilled model, and non-distilled model (student model trained from scratch) were visualized by GradCAM[47] (Fig. 3a).

To quantitatively evaluate the alignment between model attention and pathologist attention, we compute the Dice coefficients[48] (Method) between pathologist annotations and GradCAM-derived heatmap. Among 70

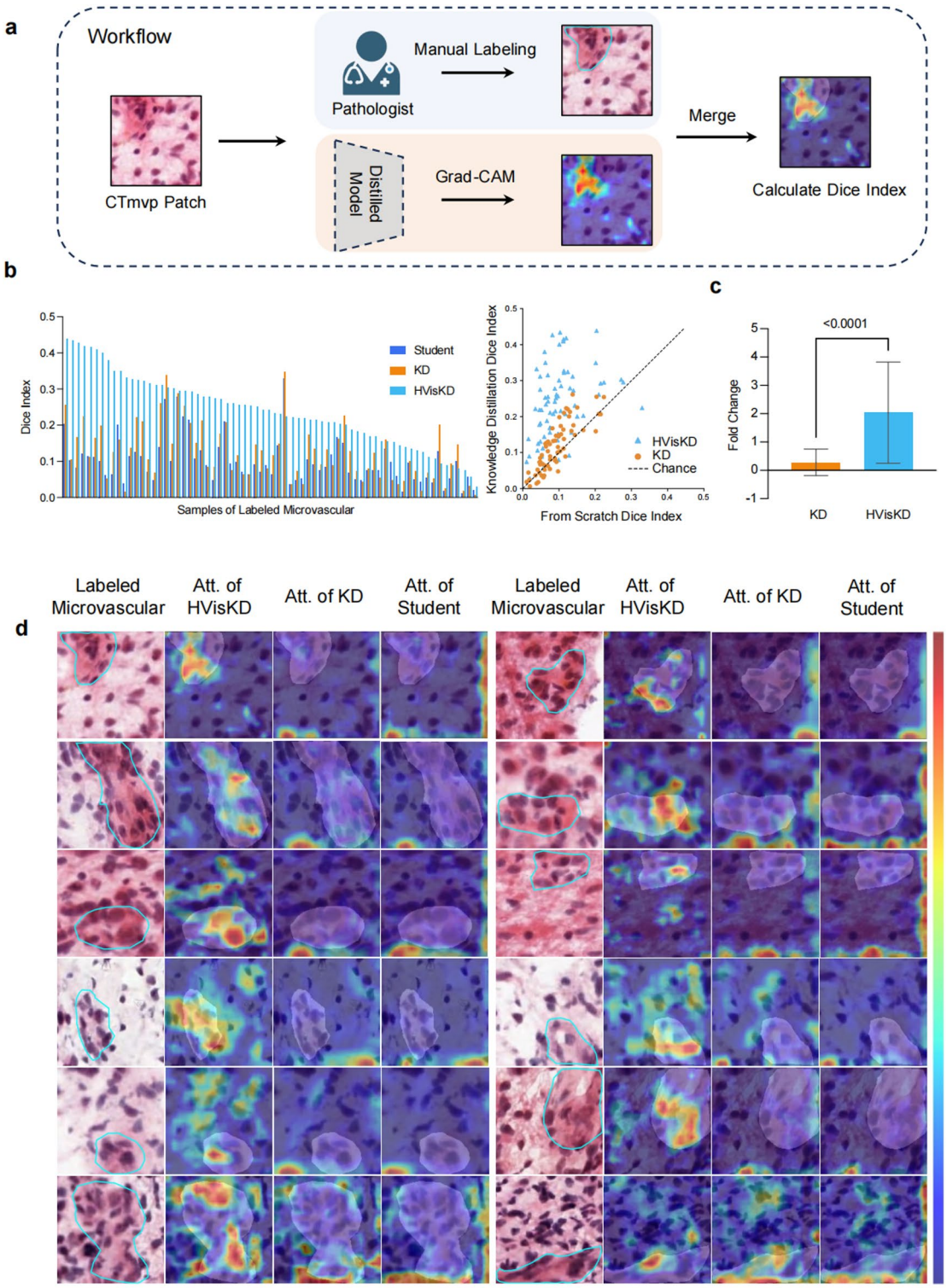


Fig. 3. HVisKD shows model attention consistency with pathologists. (**a**) Scheme for pathologist and model attention analysis workflow. (**b**) Overview of Dice coefficients between models' attentions and pathologist labeled microvascular area in CTmvp patches. We compare the results among HVisKD, KD and the student trained from scratch. (**c**) The HVisKD-distilled model performs significantly higher Dice coefficents than the KD-distilled model. (**d**) Visualization of pathologist labeled microvascular area (Labeled Microvascular, cyan polygon), and attention of HVisKD-distill model (Att. of HVisKD), KD-distill model (Att. of KD) and non-distilled model (Att. of Student). ns, not significant, $P > 0.05$; $^{****}P < 0.0001$; Paired two-tail t-test between each group (**b**).

patches, the HVisKD-distilled model demonstrates superior performance compared to both the KD-distilled model and the non-distilled model (Fig. 3b). Statistically, the HVisKD-distilled model exhibits significantly higher consistency ($P < 0.0001$) with the pathologist attention than KD-distilled models, while the difference between the KD-distilled model is not significant (Fig. 3c). Figure 3d illustrates the detailed distribution of attention for typical CTmvp patches. The HVisKD-distilled model's attention closely matches the pathologists' annotations, whereas the KD-distilled and non-distilled models' attention are often misaligned with pathologist-labeled regions, focusing instead on irrelevant background areas.

In conclusion, the HVisKD-distilled model, leveraging a human visual process-inspired design, exhibits pathologist-like attention, enhancing both interpretability and reliability.

Besides, we also present the GradCAM[47] attention map visualization on the standard natural images in artificial intelligence to prove its uniformity performance. The visualization is illustrated in the Supplementary (Suppl. Fig. 5c).

The HVisKD-distilled light model detects tumor boundaries efficiently and accurately

Surgical resection is the fundamental treatment for resectable GBM patients[49,50]. Plenty of clinical evidence indicates that whether the extent of resection (EOR) reaching the maximum safe resection is highly correlated with patients' prognosis[51–54]. The current intraoperative frozen section for EOR determination relies on the judgement of pathologist[55]. Portable and rapid deep-learning models have the potential to work on personal devices and assist human pathologists in making decisions.

We firstly demonstrate that our HVisKD can enhance light-weight model to better achieve the balance of inference cost and performance (Table 2). We report the model and time complexity and performance of different models, where the trials are simulated on practical hardware. Three different common hardware conditions are applied, that is desktop CPU (D-CPU), desktop GPU (D-GPU) and laptop CPU (L-CPU). Separately, we experiment the trials specifically on Intel Xeon E5-2650 v4 processor (D-CPU), a single NVIDIA GeForce RTX 2080 Ti (D-GPU) and Inter Core Ultra 9 185H (L-CPU). The prediction of pathology patches of evaluation are conducted, where we statistically compute the metrics above. It can be concluded that although the large parameter model has high accuracy, the exact model memory and time cost is not practical for application, especially hard for doctors' normal computer devices. In practice, when dealing with multiple pathology slides, it is time-sensitive and memory-limited, thus doctors need more efficient algorithm. Benefiting from our human visual-inspired distillation, which is similar to the view attention scenario of doctor's diagnosis of slides, the proposed HVisKD can achieve the better balance of complexity and performance, that the light-weight model has less memory and time cost meanwhile obtain well performance. Thus, we can provide an effective method HVisKD for real-world application of pathology slide prediction, that we hope it can be a powerful tool for doctors to assist them to better and effectively diagnose.

We analyzed two representative frozen section WSIs from in-house PUMCH-GBM cohort and public TCGA-GBM (method) to evaluate the consistency between distilled model predictions and pathologist-labeled normal tissue areas (Fig. 4). Pathologists meticulously labeled all normal brain tissue, which was then compared with the probability heatmaps derived from non-distilled, KD-distilled, and HVisKD-distilled models (Fig. 4a). The HVisKD-distilled model shows higher consistency with pathologists' labels compared to the non-distilled and KD-distilled models in both PUMCH-GBM cohort and TCGA-GBM (Fig. 4b and c). We analyze the distribution of probability on the normal-tumor tissue boundary (Fig. 4d, f, h and j). The HVisKD-distilled model's heatmap accurately focuses on the normal tissue. Further, we examine patches from both presumed normal and tumor tissue. Despite the tissue architecture being disrupted by the frozen section process, the model robustly differentiates between normal and tumor tissue (Fig. 4g and k). The quantitative results of dice coefficients also demonstrate the superiority of HVisKD-distilled model (Fig. 4e and i).

In summary, the above experiments suggest that HVisKD has the potential to have broad use in intraoperative EOR estimation for GBM and other cancer types.

Discussion

Contributions, interpretability, and limitations

With advances in computational pathology, deep learning has excelled in tasks like tissue subtype annotation[4,6], mutation prediction[56–58], tumor grading[59–61], prognosis[3,62,64]. Large language models[65,66] and foundation

Model	Model complexity ↓					Time complexity ↑	Acc. ↑ (HVisKD/Student)
	Params	Memory	MAdd	Flops	MemR + W	Infer speed (Patches/Sec.)	
ResNet20	280,024	42.39 M	1.82G	915.7 M	87.13 M	D-CPU:119 D-GPU:376 L-CPU:1.76	68.67/66.64
ResNet8 × 4	1,240,616	75.80 M	7.88G	3.95G	157.94 M	D-CPU:37 D-GPU:502 L-CPU:1.64	69.27/66.84
ShuffelNetV1	976,046	NA	292 M	865.82 M	NA	D-CPU:21 D-GPU:239 L-CPU:1.48	71.68/67.83
ShuffelNetV2	1,384,108	NA	524 M	996.47 M	NA	D-CPU:31 D-GPU:194 L-CPU:1.49	72.58/70.20
MobileNetV2	940,744	53.16 M	316.21 M	162.74 M	110.16 M	D-CPU:150 D-GPU:222 L-CPU:1.75	70.84/67.83

Table 2. Report on the model and time complexity and performance of several light-weight models. The trials simulate practical hardware conditions to process pathology slides, which are separately carried on three different hardware devices. D-CPU represents desktop CPU, D-GPU stands for desktop GPU and L-CPU refers to laptop CPU.

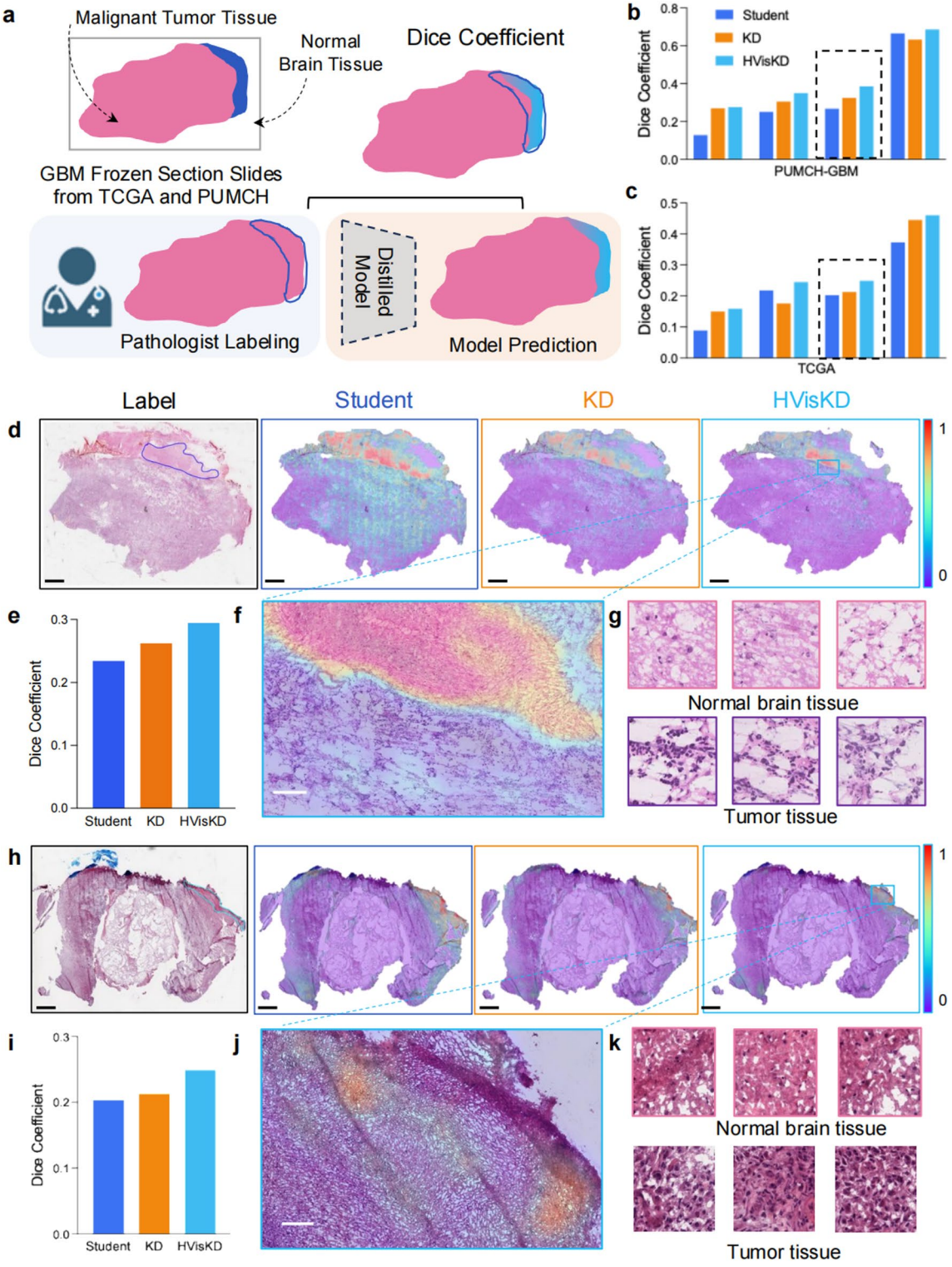


Fig. 4. HVisKD-distilled light model distinguished normal and tumor tissue comparable with pathologists on external cohort. (**a**) Scheme for evaluation the consistency between model segmentation results and pathologist labeling. (**b**, **c**) Dice coefficiencts for included frozen section WSIs from PUMCH-GBM cohort and TCGA-GBM. (**d**, **h**) Pathologists' labels (blue line) and predicted probability heatmap derived from non-distilled (Student), KD-distilled (KD), and HVisKD-distilled models (scale bar for 2 mm). (**e**, **i**) Dice coefficients between pathologists' labels and model-predicted probability heatmaps. (**f**, **j**) The model-predicted probability heatmaps distribution on the boundary of normal tissue (Scale bar for 200 μm). (**g**, **k**) Patches cropped from ROIs with high probability to be normal and tumor tissue.

models[5,67–70] show great potential but face deployment challenges due to high computational demands, especially in resource-limited settings and intraoperative pathology. Knowledge distillation (KD) offers a solution by compressing large models into smaller ones while preserving knowledge, enabling deployment on mobile devices and slide scanners.

We propose HVisKD, a novel KD method inspired by human vision, enhancing efficiency and interpretability in lightweight models. Experiments on standard vision datasets validate its reliability, and its application to GBM tissue subtype segmentation (using ivyGAP and PUMCH-GBM datasets) demonstrates superior distillation performance. HVisKD-distilled models generalize well to external frozen section WSIs, supporting fast, portable, and accurate intraoperative GBM detection.

Another major challenge in medical AI is model interpretability[71]. While methods like multi-instance learning and Transformers provide WSI-level heatmaps for mutation prediction or prognosis[56,60], patch-level attention for tissue classification remains underexplored. HVisKD mimics human pathologists by selectively attending to key regions, enhancing interpretability.

Limitations include focusing solely on tissue segmentation; future work should explore broader pathology tasks. Additionally, external validation was retrospective—prospective studies in intraoperative extent of resection (EOR) assessment are needed. CNNs remain the predominant choice in medical imaging and KD frameworks due to their proven efficacy under limited-data conditions[72–74] and their alignment with human vision-inspired hierarchical modeling. Accordingly, we focus on CNN-based architectures in this work. We next analyze the challenges of directly applying HVisKD to ViTs[22–24] and outline potential avenues for future research.

Extending HVisKD to transformer-based architectures

Transformer-based models, such as Vision Transformer (ViT)[22], DaViT[23], and CrossViT[24], have recently emerged as powerful alternatives, excelling at capturing global context through self-attention mechanisms. However, extending HVisKD to ViTs introduces additional challenges because ViT features are represented as sequences of patch tokens without a clear 2D spatial organization, making correspondence between teacher and student models non-trivial.

First, CNN and ViT features differ fundamentally in representation. CNN features provide explicit spatial hierarchy ((B,C,H,W)), which aligns naturally with our biologically inspired pyramid pooling design. In contrast, ViT features ((B,N,D)) lack inherent 2D structure, and tokens do not directly correspond to contiguous image regions. This discrepancy complicates region-level relation modeling.

Second, ViT tokens lack straightforward alignment between teacher and student models, introducing noise during relation distillation. To bridge these gaps, future work may explore strategies such as: (1) sample-level relation distillation via the CLS token, (2) token clustering or graph-based relation modeling to mimic coarse-to-fine perception, (3) multi-scale grouping for semantic dependency modeling, and (4) computationally efficient methods (e.g., sparse graphs, low-rank approximations, or top-k token selection).

These adaptations could enable relation-aware and human vision-inspired distillation for transformers, broadening the applicability of HVisKD beyond CNNs.

Method

HVisKD algorithm

Explanatory paragraph

To make the feature distillation process more interpretable and effective, we propose a human vision-inspired knowledge distillation (HVisKD) strategy. The core idea of HVisKD is to mimic how human vision perceives relationships among different visual patterns — not just focusing on individual features, but also understanding how these features relate to one another within and across samples. In our method, this is achieved by explicitly modeling sample-level and region-level relations. Specifically, we first compute a relation map that captures the similarity between samples (or regions) based on their visual features. This map highlights which samples or regions are more semantically related. We then use these relation maps to generate enhanced, relation-aware feature representations that integrate information from similar samples or regions. These relation-aware features, which encode both original features and their contextual dependencies, are then transferred from the teacher model to the student model through a dedicated distillation loss. This process helps the student model learn more discriminative and interpretable features, leading to improved generalization and performance, especially in lightweight settings. The following sections detail the formulation of these relation-aware features and the corresponding distillation losses.

S-HVisKD

We use $F \in \Re^{B\times C\times H\times W}$ to denote the feature from the backbone in a classification model, where B, C, H, W represents its batch size, channel number, height, and width, respectively. The feature F is obtained from a specific layer of the network, i.e.,$F = f^{(l)}(x)$, where $f^{(l)}$ denotes the function mapping up to layer , and x is the input patch. For each layer of the backbone model, we extract the corresponding feature maps and apply the same distillation process. To simplify the description, we illustrate the following procedure using the features from a single representative layer as an example.

We firstly feed the feature F into a global average pooling layer for spatial dimension reduction and then we reshape it to obtain the feature $F' \in \Re^{B\times C}$. After that, we perform a matrix multiplication between F' and its transpose, and then apply a normalization to calculate the sample relation map $R \in \Re^{B\times B}$:

$$R = Norm(F' \cdot F'^{T}),$$

where

$$R_{ij} = \frac{F'_i \cdot F'^T_j}{\sqrt{\sum_{j=1}^{B} (F'_i \cdot F'^T_j)^2}},$$

and we use subscript i and j to denote the i_{th} row and the j_{th} column of a matrix. Through the above calculation, we should note that R_{ij} represents the relation between the i_{th} sample and the j_{th} sample, and more similar of the two samples' features makes greater relation between them.

Then, we perform a matrix multiplication between R and F' to get a weighted sum of the sample features. Then we multiply the weighted sum by a scale parameter γ and perform an element-wise sum operation with the original feature map F' to obtain the final sample relation-aware feature $F^* \in \Re^{B\times C}$:

$$F^* = \gamma(R \cdot F') + F',$$

where

$$F^*_{ij} = \gamma(R_i \cdot F'_j) + F'_i,$$

Following[75], we initialize γ as 0 and take it as a part of the model parameters for collaborative training, so the value of this γ is continuously updated with the gradient back propagation in the training process. Thus, the sample relation-aware feature is a weighted sum of its original feature and features of all batch samples. This enhances similar samples while increasing the discrimination of dissimilar ones.

Then, we distill the sample relation-aware features from the teacher to the student. We use superscripts t and s to denote the teacher and the student, respectively, and thus the sample relation-aware features of the teacher and the student can be represented as F^{*t} and F^{*s}. Note that in knowledge distillation, the teacher always represents a network that performs well. Therefore, the feature relation of the teacher model is well-captured, and the obtained sample relation-aware features via the above process is well-discriminated. We distill these well-discriminated features from the teacher to the student can correct the weak feature discrimination of the student feature, to achieve the purpose of making the features of the student discriminated. The sample relation-aware feature distillation loss ℓ_{SFD} can be formulated as:

$$\ell_{SFD} = \ell_2(F^{*t}, F^{*s}),$$

where ℓ_2 is the L2 norm loss. Besides, in order to facilitate the sample relation-aware feature distillation, we further introduce the sample relation distillation loss ℓ_{SRD} to align the intermediate sample relation map of the teacher and the student:

$$\ell_{SRD} = \ell_2(R^t, R^s).$$

R-HVisKD

Biological motivation

Inspired by the hierarchical nature of human visual inspection, we designed R-HVisKD to mimic how pathologists interpret tissue slides using both low- and high-magnification views. At low magnification, global structures and spatial patterns guide initial attention, while high magnification reveals fine-grained cellular details for precise diagnosis. This multi-scale inspection process inherently models region-level relationships at different scales. In parallel, R-HVisKD employs a multi-level pyramid strategy to capture coarse-to-fine regional dependencies within feature maps. This design enables the model to integrate broad contextual information with local discriminative cues, thereby aligning computational feature learning with the biological mechanism of visual perception in pathology. Next, we describe the mathematical process in detail.

We first reduce the spatial dimension of a feature $f \in \Re^{C\times H\times W}$ to $f' \in \Re^{C\times 4\times 4}$ by an average pooling layer. Considering that the class-relevant regions are of various sizes in different images, we divide the feature f' into several sub-regions under a pyramid operation, of which the level of the pyramid operation represents one scale of the sub-region. For the coarsest level (Level 1), we consider the whole feature as one sub-region, and the following pyramid level (Level 2 and 3) splits the whole feature into several sub-regions of the same size. It should be noted that the size of each sub-region in different levels are varied. Then we perform the global average pooling for each sub-region in the pyramid level to obtain the output features ($f'_1 \in \Re^{C\times 1\times 1}$, $f'_2 \in \Re^{C\times 2\times 2}$ and $f'_3 \in \Re^{C\times 4\times 4}$). Note that the number of the levels in the pyramid operation and the scale of each sub-region in different levels can be modified. In this paper, considering the amount of calculation and storage, the pyramid operation contains three levels, whose size of sub-regions in each level is 1×1, 2×2, and 4×4, respectively.

Next, the similar operations in S-HVisKD are performed on the output features, f'_1, f'_2, and f'_3, respectively. We take the output feature $f'_2 \in \Re^{C\times 2\times 2}$ as an example. We firstly reshape it to $f'_2 \in \Re^{C\times 4}$ and transpose it to $f'_2 \in \Re^{4\times C}$. We perform a dot product between f'_2 and the reshape of itself, and then a normalization operation is applied to obtain the region relation map $r_2 \in \Re^{4\times 4}$:

$$r_2 = Norm(f'_2 \cdot f'^T_2),$$

$$r_{2,ij} = \frac{f'_{2,i} \cdot f'^T_{2,j}}{\sqrt{\sum_{j=1}^{4} (f'_{2,i} \cdot f'^T_{2,j})^2}},$$

and $r_{2,ij}$ represents the relation between the i_{th} sub-region and the j_{th} sub-region. We perform a dot product between r_2 and f_2' to get the result $f_2^{''} \in \Re^{4\times C}$, which is a weighted sum of the sub-regions' features within f_2'. For both f_1' and f_3', the above operations are performed.

Thus, we can obtain $f_1^{''} \in \Re^{1\times C}$, $f_2^{''} \in \Re^{4\times C}$ and $f_3^{''} \in \Re^{16\times C}$, respectively. Then, we reshape $f_1^{''}$, $f_2^{''}$ and $f_3^{''}$ to $\Re^{C\times 1\times 1}$, $\Re^{C\times 2\times 2}$ and $\Re^{C\times 4\times 4}$, respectively, and then upsample all of them to get the same size as the original feature $f' \in \Re^{C\times 4\times 4}$. After that, we multiply the summation of $f_1^{''}$, $f_2^{''}$ and $f_3^{''}$ by a scale parameter σ and perform an element-wise sum operation with f' to obtain the final region relation-aware feature $f^* \in \Re^{C\times 4\times 4}$. The process can be formulated as:

$$f^* = \sigma(r_1 \cdot f_1' + r_2 \cdot f_2' + r_3 \cdot f_3') + f',$$

where

$$f_{ij}^* = \sigma(r_{1,i} \cdot f_{1,j}' + r_{2,i} \cdot f_{2,j}' + r_{3,i} \cdot f_{3,j}') + f_i'.$$

Similarly, following[76], we initialize σ as 0 and take it as a part of the model parameters for collaborative training. Therefore, we can notice that the region relation-aware feature is a weighted fusing of multi-scale sub-regions within a feature map.

At last, we construct the region relation-aware feature distillation loss ℓ_{RFD} to transfer the region relation-aware feature from the teacher to the student:

$$\ell_{RFD} = \ell_2\left(f^{*t}, f^{*s}\right),$$

The region relation distillation loss is introduced to align the intermediate region relation map of the teacher and the student:

$$\ell_{RRD} = \ell_2(r^t, r^s).$$

Training loss

In our method, four distillation losses are constructed, and thus we introduce four hyper-parameters α_1, α_2 and β_1, β_2 to balance different distillation losses. Therefore, the overall distillation loss can be formulated as:

$$\ell_{overall} = \alpha_1\ell_{SFD} + \alpha_2\ell_{SRD} + \beta_1\ell_{RFD} + \beta_2\ell_{RRD}.$$

The overall distillation loss is model agnostic, which can be directly added to the original training loss of any model. In this paper, we empirically set $\alpha_1 = 0.05$, $\alpha_2 = 0.1$ and $\beta_1 = 0.01$, $\beta_2 = 0.1$. In the field of distillation, the training function commonly has cross-entropy classification loss $\ell_{cross-entropy}$ and the traditional knowledge distillation loss ℓ_{kd} [7], thus the final training loss can be formulated as:

$$\ell_{training} = \ell_{cross-entropy} + \ell_{kd} + \ell_{overall}.$$

Method statement

The study was approved by the Institutional review Board of Peking Union Medical College Hospital, and all experiments were performed in accordance with relevant guidelines and regulations.

Datasets and cohorts

IvyGAP dataset

We obtained 793 frozen section WSIs from 36 patients from the ivyGAP database. The WSIs were scanned at a resolution of 0.5 μm/pixel with a 20 × objective. For each WSI, the corresponding segmentation mask based on immunobiological stain (ISH) was available. To construct a dataset for GBM histopathological segmentation, we tessellated each WSI to patches and classified them into 8 classes (including LE, IT, CT, CTpan, CTpnz, CTmvp, CTne, and BG) referring to their localization on the mask. The size of each patch was set to 150 × 150 pixels. We ignored patches localized cross the boundary of tissue subtypes. For further model training and validation, we randomly selected 10,000 patches for each class and split them into a training set and validation set at a ratio of 7:3.

PUMCH-GBM cohort

We retrospectively collected 59 frozen section slides from 37 patients from Peking Union Medical College Hospital (PUMCH) to construct the PUMCH-GBM cohort. The WSIs were scanned at a resolution of 0.5 μm/pixel and saved as SVS format. The overall analysis was approved by the Institutional review Board of Peking Union Medical College Hospital. Informed consent was obtained from all subjects and/or their legal guardian(s).

TCGA-GBM cohort

We selectively downloaded 104 frozen section WSIs in The Cancer Genome Atlas Glioblastoma Multiforme (TCGA-GBM) database.

Model evaluation

Evaluation of segmentation performance

For image segmentation, we report Accuracy (ACC) as the proportion of correctly classified images and AUROC to measure class separability. ACC is calculated as the number of true positive and true negative results divided by the total number of cases (true positives, true negatives, false positives, and false negatives). AUROC represents the likelihood that a randomly chosen positive instance ranks higher than a negative one. A higher AUROC indicates better performance. For computational complexity, we evaluate Params, Memory, MAdds, MemR + W, Flops, and Inference Speed. Params denote the total model parameters, while Memory includes storage for weights and outputs. MAdds counts multiplication-addition operations per forward pass, MemR + W measures total memory reads/writes, and Flops quantifies floating-point operations per pass. Inference speed indicates the number of patches predicted per second. These metrics jointly assess both accuracy and efficiency of our distilled lightweight model.

Attention consistency between model and pathologist

We randomly selected 300 CTmvp patches from the validation set that was previously described. Two pathologists were involved in checking each patch and labeling the microvascular area (one for each patch) with Labelme software. Finally, 70 patches were labeled as binary matrix P with the shape of 150×150 and used for further analysis.

For model attention, we chose VGG19-ShuffleNetV1 pair. The HVisKD-distilled, KD-distilled and non-distilled ShuffleNetV1 were analyzed respectively. We used GradCAM to capture model attention heatmap $M \in [0, 1]$ for CTmvp class with size of 150×150. To quantify the similarity between pathologist and model attention, we calculated the Dice coefficients[48] for each patch:

$$Dice = \frac{2 \times \sum P * M}{\sum P + \sum M}$$

Paired t-tests between each two of HVisKD-distilled, KD-distilled and non-distilled models' Dice coefficients were performed using GraphPad Prism 10.

External test and tumor boundary detection

We constructed an external in-house PUMCH-GBM cohort and included TCGA-GBM as previously described. Considering the variation of stain procedure within and between datasets, we stain normalized the ivyGAP dataset with Macenko stain transfer and retrained the VGG19-ShuffleNetV1 teacher and student models with the same parameters as before.

To show the tumor boundary detection potential, two pathologists selected 4 WSI from the PUMCH-GBM cohort and 4 from TCGA-GBM which contain both normal brain tissue and malignant tumor tissue in the same slide. Then, they labeled all normal brain tissue area in detail with QuPath software. Each WSI was then tessellated and stain normalized for model segmentation. For each patch, the model returned the probabilities of three classes. We reconstructed the WSI-level heatmap based on patch coordinates. To get heatmap with higher resolution, we resized WSI-level heatmap to $100\times$ larger and smooth the heatmap with gaussian filter. Consistency between pathologist labeled normal brain tissue area and model predicted probability were calculated then as previous described.

Conclusion

In this study, we proposed HVisKD, a human visual attention-inspired knowledge distillation strategy, to address the challenge of balancing interpretability, model lightweight, and task performance in computational pathology. By modeling both global and local patch-level relations, HVisKD effectively transfers structured attention knowledge from teacher to student models, enabling enhanced feature differentiation and more precise focus on clinically relevant regions. Our method draws inspiration from the hierarchical attention mechanisms of the human visual system, mirroring how pathologists interpret WSIs through layered visual cues. Experimental results demonstrate that HVisKD consistently improves the performance of lightweight models across segmentation tasks while enhancing the interpretability of attention maps compared to conventional KD methods. These findings underscore the potential of bio-inspired designs in advancing deep learning applications for real-world clinical diagnostics, offering a practical and explainable solution for high-stakes medical decision-making.

Data availability

Ivy Glioblastoma Atlas Project is available at https://glioblastoma.alleninstitute.org, The TCGA frozen secti on whole-slide data and corresponding clinical information are available from NIH genomic data commons (https://portal.gdc.cancer.gov/projects/TCGA-GBM). CIFAR100 data is available at https://www.cs.toronto.edu/~kriz/cifar.html, and ImageNet data is available at https://www.image-net.org/download.php. The slide and patient data from PUMCH-GBM cohort is not provided due to ethic considerations.

Code availability

The code is available on GitHub repository at https://github.com/ChimesZ/RFD4Hist.

Received: 11 February 2025; Accepted: 27 October 2025
Published online: 26 November 2025

References

1. Song, A. H. et al. Artificial intelligence for digital and computational pathology. *Nat. Rev. Bioeng.* https://doi.org/10.1038/s44222-023-00096-8 (2023).
2. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: Enhancing cancer research and clinical oncology. *Nat. Cancer* **3**, 1026–1038 (2022).
3. Liang, J. et al. Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-023-00635-3 (2023).
4. Tolkach, Y. et al. Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: A retrospective algorithm development and validation study. *Lancet Digit. Health* **5**, e265–e275 (2023).
5. Hollon, T. *et al. Visual Foundation Models for Fast, Label-Free Detection of Diffuse Glioma Infiltration*. https://www.researchsquare.com/article/rs-4033133/v1 (2024) https://doi.org/10.21203/rs.3.rs-4033133/v1.
6. Kather, J. N. et al. Multi-class texture analysis in colorectal cancer histology. *Sci. Rep.* **6**, 27988 (2016).
7. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at http://arxiv.org/abs/1503.02531 (2015).
8. Ishtiaq, A., Mahmood, S., Anees, M. & Mumtaz, N. Model compression. Preprint at http://arxiv.org/abs/2105.10059 (2021).
9. Han, S., Mao, H. & Dally, W. J. Deep Compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. Preprint at http://arxiv.org/abs/1510.00149 (2016).
10. Chen, G., Choi, W., Yu, X., Han, T. & Chandraker, M. Learning efficient object detection models with knowledge distillation.
11. Bolya, D., Zhou, C., Xiao, F. & Lee, Y. J. YOLACT: Real-time instance segmentation. Preprint at http://arxiv.org/abs/1904.02689 (2019).
12. Zhang, Y., Xiang, T., Hospedales, T. M. & Lu, H. Deep mutual learning. Preprint at http://arxiv.org/abs/1706.00384 (2017).
13. Jin, Y., Wang, J. & Lin, D. Multi-level logit distillation. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 24276–24285 (IEEE, Vancouver, BC, Canada, 2023). https://doi.org/10.1109/CVPR52729.2023.02325.
14. Romero, A. *et al.* FitNets: Hints for Thin deep nets. Preprint at http://arxiv.org/abs/1412.6550 (2015).
15. Chen, D. *et al.* Knowledge distillation with the reused teacher classifier. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11923–11932 (IEEE, New Orleans, LA, USA, 2022). https://doi.org/10.1109/CVPR52688.2022.01163.
16. Park, W., Kim, D., Lu, Y. & Cho, M. Relational knowledge distillation. Preprint at http://arxiv.org/abs/1904.05068 (2019).
17. Heo, B., Lee, M., Yun, S. & Choi, J. Y. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. Preprint at http://arxiv.org/abs/1811.03233 (2018).
18. Lm, W. Determinants of attention to local and global features of visual forms. *J. Exp. Psychol. Hum. Percept. Perform.* **8**, 562 (1982).
19. Biederman, I. Perceiving real-world scenes. *Science* **177**, 77–80 (1972).
20. Brand, J. & Johnson, A. P. Attention to local and global levels of hierarchical Navon figures affects rapid scene categorization. *Front. Psychol.* **5**, 1274 (2014).
21. Shulman, G. L. & Wilson, J. Spatial frequency and selective attention to local and global information. *Perception* **16**, 89–101 (1987).
22. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).
23. Ding, M. *et al.* DaViT: Dual attention vision transformers. Preprint at https://doi.org/10.48550/arXiv.2204.03645 (2022).
24. Chen, C.-F., Fan, Q. & Panda, R. CrossViT: Cross-attention multi-scale vision transformer for image classification. Preprint at https://doi.org/10.48550/arXiv.2103.14899 (2021).
25. Puchalski, R. B. et al. An anatomic transcriptional atlas of human glioblastoma. *Science* **360**, 660–663 (2018).
26. Krizhevsky, A. Learning multiple layers of features from tiny images. in (2009).
27. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, Miami, FL, 2009). https://doi.org/10.1109/CVPR.2009.5206848.
28. Tian, Y., Krishnan, D. & Isola, P. Contrastive Representation Distillation. Preprint at http://arxiv.org/abs/1910.10699 (2022).
29. He, K., Zhang, X., Ren, S. & Sun, J. Identity Mappings in Deep Residual Networks. Preprint at http://arxiv.org/abs/1603.05027 (2016).
30. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated Residual Transformations for Deep Neural Networks. Preprint at http://arxiv.org/abs/1611.05431 (2017).
31. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Preprint at http://arxiv.org/abs/1801.04381 (2019).
32. Zhang, X., Zhou, X., Lin, M. & Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Preprint at http://arxiv.org/abs/1707.01083 (2017).
33. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at http://arxiv.org/abs/1409.1556 (2015).
34. Tung, F. & Mori, G. Similarity-preserving knowledge distillation. Preprint at http://arxiv.org/abs/1907.09682 (2019).
35. Xu, G., Liu, Z., Li, X. & Loy, C. C. Knowledge distillation meets self-supervision. Preprint at http://arxiv.org/abs/2006.07114 (2020).
36. Chen, L. et al. Wasserstein contrastive representation distillation. Preprint at http://arxiv.org/abs/2012.08674 (2021).
37. Peng, B. et al. Correlation congruence for knowledge distillation. Preprint at http://arxiv.org/abs/1904.01802 (2019).
38. Passalis, N. & Tefas, A. Learning deep representations with probabilistic knowledge transfer. Preprint at http://arxiv.org/abs/1803.10837 (2019).
39. Ahn, S., Hu, S. X., Damianou, A., Lawrence, N. D. & Dai, Z. Variational information distillation for knowledge transfer. Preprint at http://arxiv.org/abs/1904.05835 (2019).
40. Kim, J., Park, S. & Kwak, N. Paraphrasing complex network: network compression via factor transfer. Preprint at http://arxiv.org/abs/1802.04977 (2020).
41. Huang, Z. & Wang, N. Like what you like: Knowledge distill via neuron selectivity transfer. Preprint at http://arxiv.org/abs/1707.01219 (2017).
42. Ji, M., Heo, B. & Park, S. Show, attend and distill: Knowledge distillation via attention-based feature matching. Preprint at http://arxiv.org/abs/2102.02973 (2021).
43. Yim, J., Joo, D., Bae, J. & Kim, J. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 7130–7138 (IEEE, Honolulu, HI, 2017). https://doi.org/10.1109/CVPR.2017.754.
44. Zagoruyko, S. & Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. Preprint at http://arxiv.org/abs/1612.03928 (2016).
45. Tena-Suck, M. L., Celis-Lopez, M. A., Collado-Ortiz, M. A., Castillejos-Lopez, M. & Tenorio-Serralta, M. Glioblastoma multiforme and angiogenesis: A clinicopathological and immunohistochemistry approach. *J. Neurol. Res.* **5**, 199–206 (2015).
46. Das, S. & Marsden, P. A. Angiogenesis in Glioblastoma. *N. Engl. J. Med.* **369**, 1561–1563 (2013).
47. Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
48. Zou, K. H. et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad. Radiol.* **11**, 178–189 (2004).
49. Gerritsen, J. K. W. et al. Safe surgery for glioblastoma: Recent advances and modern challenges. *Neuro-Oncol. Pract.* **9**, 364–379 (2022).

50. Sales, A. H. A. et al. Surgical treatment of glioblastoma: State-of-the-art and future trends. *J. Clin. Med.* **11**, 5354 (2022).
51. Brown, T. J. et al. Association of the extent of resection with survival in glioblastoma: A systematic review and meta-analysis. *JAMA Oncol.* **2**, 1460 (2016).
52. Chaichana, K. L. et al. Establishing percent resection and residual volume thresholds affecting survival and recurrence for patients with newly diagnosed intracranial glioblastoma. *Neuro Oncol.* **16**, 113–122 (2014).
53. Bloch, O. et al. Impact of extent of resection for recurrent glioblastoma on overall survival: Clinical article. *JNS* **117**, 1032–1038 (2012).
54. Sanai, N., Polley, M.-Y., McDermott, M. W., Parsa, A. T. & Berger, M. S. An extent of resection threshold for newly diagnosed glioblastomas: Clinical article. *JNS* **115**, 3–8 (2011).
55. Somerset, H. L. & Kleinschmidt-DeMasters, B. K. Approach to the intraoperative consultation for neurosurgical specimens. *Adv. Anat. Pathol.* **18**, 446–449 (2011).
56. Wagner, S. J. et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell* **41**, 1650 (2023).
57. Niehues, J. M. et al. Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study. *Cell Rep. Med.* **4**, 100980 (2023).
58. Nasrallah, M. P. et al. Machine learning for cryosection pathology predicts the 2021 WHO classification of glioma. *Med* **4**, 526 (2023).
59. Pantanowitz, L. et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: A blinded clinical validation and deployment study. *Lancet Digit. Health* **2**, e407–e416 (2020).
60. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
61. Hollon, T. et al. Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. *Nat. Med.* **29**, 828. https://doi.org/10.1038/s41591-023-02252-4 (2023).
62. Zheng, Y., Carrillo-Perez, F., Pizurica, M., Heiland, D. H. & Gevaert, O. Spatial cellular architecture predicts prognosis in glioblastoma. *Nat. Commun.* **14**, 4122 (2023).
63. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865-878.e6 (2022).
64. Skrede, O.-J. et al. Deep learning for prediction of colorectal cancer outcome: A discovery and validation study. *Lancet* **395**, 350–360 (2020).
65. Xie, Q. *et al.* Me-LLaMA: Foundation large language models for medical applications. Preprint at https://doi.org/10.21203/rs.3.rs-4240043/v1 (2024).
66. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172. https://doi.org/10.1038/s41586-023-06291-2 (2023).
67. Wang, X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970. https://doi.org/10.1038/s41586-024-07894-z (2024).
68. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863. https://doi.org/10.1038/s41591-024-02856-4 (2024).
69. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924. https://doi.org/10.1038/s41591-024-03141-0 (2024).
70. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181. https://doi.org/10.1038/s41586-024-07441-w (2024).
71. Beam, A. L. et al. Artificial intelligence in medicine. *N. Engl. J. Med.* **388**, 1220–1221 (2023).
72. Kawadkar, K. Comparative analysis of vision transformers and convolutional neural networks for medical image classification. Preprint at https://doi.org/10.48550/arXiv.2507.21156 (2025).
73. Sepahvand, M. & Abdali-Mohammadi, F. Joint learning method with teacher–student knowledge distillation for on-device breast cancer image classification. *Comput. Biol. Med.* **155**, 106476 (2023).
74. Termritthikun, C., Umer, A., Suwanwimolkul, S., Xia, F. & Lee, I. Explainable knowledge distillation for on-device chest x-ray classification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **21**, 846–856 (2024).
75. Zhang, H., Goodfellow, I., Metaxas, D. & Odena, A. Self-Attention Generative Adversarial Networks. Preprint at http://arxiv.org/abs/1805.08318 (2019).
76. Zagoruyko, S. & Komodakis, N. Wide Residual Networks. Preprint at http://arxiv.org/abs/1605.07146 (2017).

Acknowledgements
We express our gratitude to all co-authors and datasets involved in this paper.

Author contributions
M.Y., Z.Z., B.L., Y.W., W.M., and K.M. conceived the project. M.Y. and Z.Z. designed the computational model. M.Y. and Z.Z. performed the theoretical analysis. M.Y. and Z.Z. performed the experiments and analyzed the data. M.Y., Z.Z., and B.L. wrote the paper. X.Z., Y.W., T.L., J.C., H.H, J.Z., and D.Z. participated in the investigation. B.L., Y.W., W.M., and K.M. supervised the project.

Funding
This work was supported by the National High Level Hospital Clinical Research Funding (2022-PUMCH-B-113), the National High Level Hospital Clinical Research Funding (2022-PUMCH-A-019) for Yu Wang, and the Institute for Interdisciplinary Information Core Technology (IIISCT).

Declarations

Competing interests
The authors declare that they have no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-26004-1.

Correspondence and requests for materials should be addressed to B.L., Y.W., W.M. or K.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025