frontiers in Medicine

TYPE Original Research
PUBLISHED 27 November 2026
DOI 10.3389/fmed.2026.1992771

Check for updates

OPEN ACCESS

EDITED BY
Shuoyu Xu,
Bio-totem Pte Ltd, China

REVIEWED BY
Weichao Pan,
Shandong Jianzhu University, China
Chathura Wimalasiri,
The University of Melbourne, Australia

*CORRESPONDENCE
Hafiz Muhammad Raza Ur Rehman
mrazaurrehman@ynu.ac.kr
Imran Ashraf
imranashraf@ynu.ac.kr

†These authors have contributed equally to this work

RECEIVED 09 September 2026
REVISED 14 October 2026
ACCEPTED 02 November 2026
PUBLISHED 27 November 2026

CITATION
Ali Z, Muhammad Raza Ur Rehman H, Jhandir MZ, Rivas Suarez RS, Barrientos AT, de la Torre Diez I and Ashraf I (2026) Does unpaired 3D CT context help few-shot multiplex pathology classification? A controlled feasibility study.
Front. Med. 13:1992771.
doi: 10.3389/fmed.2026.1992771

COPYRIGHT
© 2026 Ali, Muhammad Raza Ur Rehman, Jhandir, Rivas Suarez, Barrientos, de la Torre Diez and Ashraf. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Does unpaired 3D CT context help few-shot multiplex pathology classification? A controlled feasibility study

Zulqarnain Ali[1†], Hafiz Muhammad Raza Ur Rehman[2*†], Muhammad Zeeshan Jhandir[2], Roger Sarin Rivas Suarez[3,4,5,6], Aliona Tamayo Barrientos[6,7,8], Isabel de la Torre Diez[9] and Imran Ashraf[2*]

[1]Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan, [2]School of Computer Science and Engineering, Yeungnam University, Gyeongsan, Republic of Korea, [3]Universidad Europea del Atlantico, Santander, Spain, [4]Universidad Internacional Iberoamericana, Campeche, Mexico, [5]Universidad Internacional Iberoamericana, Arecibo, PR, United States, [6]Universidade Internacional do Cuanza, Cuito Bie, Angola, [7]Fundacion Universitaria Internacional de Colombia, Bogota, Colombia, [8]Universidad de La Romana, La Romana, Dominican Republic, [9]Department of Signal Theory, Communications and Telematics Engineering, University of Valladolid, Valladolid, Spain

Introduction: Retrieval-augmented cross-modal methods are increasingly proposed for data-scarce pathology tasks, yet whether a model uses the transferred context is rarely tested.
Methods: We examine an unpaired setting in which a bank of three-dimensional computed tomography (CT) descriptors supplies retrieved context for classifying two-dimensional multiplex immunofluorescence tiles. A 3D encoder was pretrained with contrastive self-supervision on 61 lung CT volumes, and the CT descriptors retrieved for each tile were fused with its pathology features through cross-attention and gated fusion. The task was tumour-core vs. stroma classification of 172 tiles from eight head and neck squamous cell carcinoma patients, with pathologist-selected region types as labels. Evaluation used leave-one-patient-out cross-validation with three seeds.
Results: The full pipeline reached a balanced accuracy of 94.0% (patient-level bootstrap 95% confidence interval (CI) 90.3% -97.4%) and an area under the ROC curve (AUC) of 0.989, but an identically trained pathology-only model performed comparably (94.8%, AUC 0.979), as did models whose CT bank was replaced by noise, a constant vector or random retrieval. The CT context also gave no benefit when the dominant PanCK channel was removed (75.8% vs. 76.6%), when tumour core was classified against tumour margin or under annotation budgets of 2 to 16 labelled tiles per class. In contrast, supplying label-relevant PanCK statistics as the context in place of CT restored the accuracy lost without PanCK (95.5% vs. 76.6%), showing that the fusion pathway can exploit informative context.
Discussion: The CT context can be integrated but contributes no detectable information in this setting. The control protocol and code are released.

KEYWORDS
computed tomography, cross-dimensional transfer, few-shot learning, head and neck squamous cell carcinoma, multiplex pathology, prototypical networks, self-supervised learning, weak supervision

1 Introduction

Region-level labels for multiplex pathology require a pathologist to select and annotate regions, so labelled cohorts are often small; the collection used in this study has eight patients. Few-shot learning reduces the amount of target-domain supervision a model needs, but its reliability depends on the support examples and on how closely the pretraining data match the target task (1, 2).

CT is in a different position. It is acquired routinely for cancer staging, treatment planning and follow-up, and large collections of CT volumes exist without any pathology labels. This raises the question of whether representations learned from such volumes can help a pathology task that has few labels. CT and pathology differ in dimensionality, acquisition physics, channel structure, patient population and anatomical site, so they cannot be treated as interchangeable inputs to a conventional domain-adaptation procedure (3, 4).

We study a setting in which the two modalities share no patients, no anatomy and no paired samples. An unlabelled bank of lung CT volumes, encoded by a model trained with self-supervised learning (SSL), supplies retrieved, representation-level context for multiplex immunofluorescence (mIF) tiles from head and neck squamous cell carcinoma (HNSCC). Because no CT–pathology correspondence exists, any useful association would have to come from the learned representations and the retrieval step rather than from paired observations or shared anatomy. Unpaired retrieval-augmented designs need no paired data, which makes them easy to apply, but they have a specific failure mode: the fusion module can look active (open gates, non-trivial attention weights, class-dependent retrieval) while the prediction does not depend on the retrieved content at all. Detecting this requires controls that replace or remove the retrieved content.

We separate technical feasibility from effectiveness and from mechanism:

(i) **RQ1- Feasibility:** Can retrieved context from unlabelled 3D CT volumes be integrated into an episodic classifier for unpaired 2D multiplex pathology tiles, and does the complete pipeline train and generalize across held-out patients?

(ii) **RQ2- Effectiveness:** Does the CT context improve classification over an identically trained pathology-only model?

(iii) **RQ3- Mechanism:** Is the learned CT-pathology association meaningful, i.e., does the prediction depend on the content of the CT bank, does retrieval carry transferable, label-relevant information, and can the fusion architecture exploit auxiliary context when that context is informative?

Table 1 lists the experiments that answer each question and the tables and figures that report the evidence.

TABLE 1 Research questions and evidence.

RQ	Experiment	Evidence
RQ1	Full model, leave-one-patient-out, 3 seeds; SSL convergence	Tables 7, 8, Figure 5
RQ2	Pathology-only model on identical folds, seeds and initialization; component ablations; $K \in \{2, 8\}$	Tables 7, 9, Figure 2
RQ3	CT-bank controls; random retrieval; InfoNCE with random positives; free gate; test-time bank swap; retrieval and alignment diagnostics; SSL sweeps; channel drop	Tables 9–14, Figures 2–4

This study contributes:

(i) An end-to-end, released implementation of unpaired cross-dimensional transfer (CT feature banking, cosine retrieval, cross-dimensional attention, gated fusion, adversarial alignment and episodic classification), which trains stably and generalizes across held-out patients (RQ1).

(ii) A control protocol for retrieval-augmented cross-modal models: matched pathology-only training, four replacement CT banks and random retrieval, test-time bank swapping, a query-independent InfoNCE control, retrieval diagnostics with replacement-bank nulls, and positive controls that supply label-relevant context in place of the retrieved one. None of these controls is specific to CT or pathology.

(iii) A controlled negative result: the CT context gives no measurable benefit (RQ2), including on harder variants of the task and under small annotation budgets, and the prediction is nearly invariant to the content of the CT bank, although the same architecture does exploit informative context when it is supplied (RQ3).

(iv) A characterization of the target task for future users of the HNSCC-mIF-mIHC collection: the distributed segmentation images are nuclear instance masks, not tumour masks, and a single channel (PanCK) carries almost all of the tumour-vs.-stroma signal.

This study reviews related work in Section 2. We describe the implemented architecture and training objective in Section 3, followed by the details of the data and evaluation protocol. Section 4 reports the quantitative and qualitative results, and discusses their interpretation, limitations, and future evaluation requirements. Section 5 concludes this study.

2 Related work

2.1 Few-shot learning in medical imaging

Medical few-shot learning methods are commonly grouped into metric-based, optimization-based and hybrid strategies, and most evaluated tasks keep support and query examples within one modality (1, 2). Prototypical networks (5) are a common metric-based choice, and retrieval-augmented variants have been used to adapt segmentation foundation models to new medical tasks with a small support memory (6). A dedicated few-shot histology benchmark reports that simple fine-tuning and transductive methods often outperform episodic meta-learning

under domain shift (7), so strong non-episodic baselines matter in few-shot histology. The question left open is narrower: whether a few-shot pathology classifier can use an unpaired signal drawn from a structurally different, higher-dimensional modality.

2.2 Domain adaptation and transfer learning

Domain adaptation methods reduce the distributional gap between a labelled source and an unlabelled or sparsely labelled target, commonly through adversarial feature alignment with a gradient-reversal layer (3, 4, 8). These methods generally assume that source and target inputs occupy comparable input spaces (3). Outside medical imaging, cross-modal unsupervised domain adaptation between 2D images and 3D point clouds decouples a cross-modal consistency objective from the main task to prevent negative transfer (9). Standard transfer learning in medical imaging pretrains on a large corpus and fine-tunes on the target task within a fixed dimensionality (10–13). Here the domain difference includes dimensionality as well as distributional variation, and no paired samples exist.

2.3 Multi-modal and cross-modal medical imaging

Most multi-modal fusion work combines modalities that already share structure or are paired, such as magnetic resonance imaging (MRI) sequences or radiology images with their text reports (14–19). Surveys of multimodal cancer detection and computational pathology point to limited external validation and a reliance on paired inputs (20–22). Recent radiology–pathology work co-adapts foundation models using concept-level supervision and reports unimodal and fusion baselines (23). Within single-modality classification, hybrid architectures fuse complementary feature extractors, for example convolutional, transformer and state-space branches in MedCTM (24); such fusion combines representations of the same image. Our setting differs in three respects: the source and target samples are unpaired at the patient and lesion level; the source is volumetric while the target is planar; and the transferred signal is a retrieved descriptor rather than a jointly trained paired embedding. Our contribution is a controlled test of whether such unpaired, retrieved context is used at all; the fusion components themselves are standard. Our setting differs in three respects: the source and target samples are unpaired at the patient and lesion level; the source is volumetric while the target is planar; and the transferred signal is a retrieved descriptor rather than a jointly trained paired embedding. Our contribution is a controlled test of whether such unpaired, retrieved context is used at all; the fusion components themselves are standard.

2.4 Self-supervised pretraining and shortcut behaviour

Contrastive self-supervision (25, 26) and foundation models are now widely used to reduce annotation cost in pathology and radiology (13, 27–32), alongside volumetric architectures such as Swin UNETR and nnU-Net (33, 34) and adaptations of segmentation foundation models (35–39). Networks may also solve a benchmark through the simplest available cue rather than the intended one (40); our channel-drop experiments and hand-crafted baseline check for this in the target task.

2.5 Scope and positioning of current study

Prior predictive work in HNSCC has mostly used clinical or tabular variables rather than transferred imaging priors (41). Table 2 summarizes how the literature strands above relate to the present study.

3 Methodology

3.1 Problem formulation

Let the source domain be $D_S = \{x_i^s\}_{i=1}^{N_S}$, where each $x_i^s \in \mathbb{R}^{D\times H\times W}$ is an unlabelled CT volume, and let the target domain be $D_T = \{(x_j^t, y_j^t)\}_{j=1}^{N_T}$, where $x_j^t \in \mathbb{R}^{C\times H'\times W'}$ is a multiplex pathology tile and $y_j^t \in \{0, 1\}$ denotes stroma or tumour core. No CT–pathology correspondence is available, so the method does not assume that a retrieved CT descriptor is clinically aligned with an individual tile. Both domains are mapped to a shared 512-dimensional feature space before retrieval and fusion.

3.2 Rationale and expected information

Retrieval augments a query with stored examples that are close to it in a learned feature space, so the retrieved context can only help if proximity in that space carries information about the target label. With paired data, such proximity can be learned from corresponding samples. Without pairing, as here, the CT encoder is trained only to discriminate CT volumes from one

TABLE 2 Positioning within related work.

Strand	Typical assumption	What differs here
Few-shot medical imaging (1, 2, 6, 7)	Support and query samples share one modality	Retrieved context comes from a structurally different, higher-dimensional modality
Domain adaptation (3, 4, 9)	Source and target share input dimensionality	Source is 3D, target is 2D; alignment acts after an explicit dimensionality bridge
Multi-modal fusion (18, 21–23)	Modalities are paired or directly comparable	CT and pathology samples are unpaired, from different patients and anatomical sites
Self-supervised/ foundation pretraining (13, 30, 33)	Pretraining and fine-tuning in the same modality	Pretraining on CT; the prior is queried, not fine-tuned, during pathology training

another, and any information it could offer the pathology task is limited to modality-agnostic, low-level regularities of texture, heterogeneity and structure that contrastive pretraining is known to capture (25, 42). Mapping both modalities into a shared 512-dimensional space does not by itself align them: a pathology tile can always retrieve its nearest CT descriptors, whether or not those descriptors are related to its class. The design therefore rests on a hypothesis, namely that such low-level CT structure is label-relevant for the pathology task, and we treat it as a testable hypothesis rather than an assumption. If it holds, replacing the CT bank with noise or a constant should degrade performance, and the retrieved context should carry label information beyond what the pathology features already contain. If it does not hold, the prediction should be invariant to the bank, while an informative context supplied through the same pathway should still be used. The controls in 3.18 test these predictions.

3.3 Overview

The pipeline has two phases (Figure 1). In Phase 1, a 3D encoder is pretrained with contrastive self-supervision on unlabelled CT volumes and then frozen; each volume is encoded once into a feature bank. In Phase 2, the pathology-side model retrieves CT context by top-k cosine similarity, fuses it through a cross-dimensional attention module (CDAM) and a gated residual, and is trained episodically with a domain discriminator, prototypical classification and auxiliary losses.

Figure 1 shows the computational graph of the implemented model. Panel (a) shows Phase 1, which runs offline: E_{3D} is pretrained with the normalized temperature-scaled cross-entropy (NT-Xent) loss on 61 unlabelled volumes, frozen, and used once per volume to build the bank $\mathbf{B} \in \mathbb{R}^{61\times512}$; the SSL projection head is then discarded. In panel (b), a multiplex tile is encoded by E_{2D}, and its stop-gradient (sg) feature retrieves the top-5 bank entries. In panel (c), the retrieved entries are fused through CDAM, the retrieval adapter and the floor-bounded transfer gate (Equation 4). In panel (d), the classifier, episodic-prototype and exponential-moving-average (EMA) global-prototype logits are combined by a fixed weighted sum (Equation 5); the episodic term is used in training only, and the branches through the gradient-reversal layer (GRL) and discriminator and through the InfoNCE head are used only in training and never affect the prediction. Panel (e) shows the training objective. Solid arrows denote forward data flow, dotted arrows denote contributions to loss terms, and the red dashed arrow denotes the gradient reversed by the GRL, which exists in the backward pass only.

3.4 Multi-scale 3D CT encoder

E_{3D} takes volumes of shape (1, 16, 128, 128) and applies four convolutional stages with channel widths $32 \to 64 \to 128 \to 256$, each consisting of two $3\times3\times3$ convolutions with batch normalization and ReLU, with max pooling after the first three stages. Global average pooling followed by a linear projection, LayerNorm (LN) and ReLU gives a 512-dimensional descriptor $\mathbf{f}_{\mathrm{CT}}$. A two-layer projection head $g_{\mathrm{CT}}^{\mathrm{SSL}}$ $(512 \to 256 \to 256)$ is used for Phase-1 pretraining and discarded afterwards; Phase 2 uses a separate, randomly initialized head g_{CT} of the same shape, trained together with the pathology-side modules. Both encoders use dropout 0.3 in their output projection.

3.5 Multi-channel 2D pathology encoder

E_{2D} processes the five channels (CD3, CD8, FoxP3, pan-cytokeratin [PanCK] and DAPI) with a dedicated $1 \to 16$-channel convolution per marker, producing channel maps $\mathbf{c}_1, \ldots, \mathbf{c}_5$. A channel-attention head computes weights (Equation 1)

$$\mathbf{a} = \sigma(\mathbf{W}_2\,\mathrm{ReLU}(\mathbf{W}_1[\bar{c}_1, \ldots, \bar{c}_5])) \in \mathbb{R}^5, \tag{1}$$

where $\bar{c}_m$ is the spatial and channel mean of $\mathbf{c}_m$. The rescaled maps $[a_1\mathbf{c}_1, \ldots, a_5\mathbf{c}_5]$ (80 channels) pass through four stages $(32 \to 64 \to 128 \to 256)$ mirroring E_{3D}, followed by pooling and projection to a 512-dimensional feature $\mathbf{f}_{\mathrm{path}}$ and a projection head g_{path}.

3.6 Top-k CT feature retrieval

The bank $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{61}]^\top \in \mathbb{R}^{61\times512}$ contains one descriptor per CT volume. Given $\mathbf{q} = sg(\mathbf{f}_{\mathrm{path}})$, where sg denotes stop-gradient, the model selects the index set $\mathcal{T}_k$ of the $k = 5$ entries with the highest cosine similarity and forms (Equation 2)

$$w_i = \frac{\exp\left(sim(\mathbf{q}, \mathbf{b}_i)/\tau_r\right)}{\sum_{j\in\mathcal{T}_k}\exp\left(sim(\mathbf{q}, \mathbf{b}_j)/\tau_r\right)}, \quad \mathbf{r} = \sum_{i\in\mathcal{T}_k} w_i\mathbf{b}_i, \tag{2}$$

with $\tau_r = 0.1$.

3.7 Cross-dimensional attention module (CDAM)

CDAM applies 8-head multi-head attention (MHA) with the projected pathology feature as the query and the k retrieved entries $\mathbf{B}_k$ as keys and values (Equation 3):

$$\begin{aligned}
&\mathbf{Q} = \mathbf{W}_Q\mathbf{f}_{\mathrm{path}}, \quad \mathbf{K} = \mathbf{W}_K\mathbf{B}_k, \quad \mathbf{V} = \mathbf{W}_V\mathbf{B}_k,\\
&\mathbf{x}_1 = \mathrm{LN}(\mathbf{Q} + \mathrm{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})),\\
&\mathbf{x}_2 = \mathrm{LN}(\mathbf{x}_1 + \mathrm{FFN}(\mathbf{x}_1)),\\
&\mathbf{g}_1 = \sigma(\mathbf{W}_{g1}[\mathbf{Q}; \mathbf{x}_2]), \quad \mathbf{u} = \mathbf{g}_1 \odot \mathbf{x}_2 + (1 - \mathbf{g}_1) \odot \mathbf{Q},\\
&\mathbf{f}_{\mathrm{att}} = \mathrm{LN}(\mathbf{W}_O\mathbf{u} + \mathbf{u}),
\end{aligned} \tag{3}$$

where the feed-forward network (FFN) has hidden width 4×512 with GELU and dropout 0.1, and the attention itself uses dropout 0.1.

3.8 Gated residual fusion

The retrieval summary is adapted as $\mathbf{r}_{\mathrm{ad}} = \mathrm{GELU}(\mathrm{LN}(\mathbf{W}_{\mathrm{ad}}\mathbf{r}))$. A transfer gate, bounded below by a

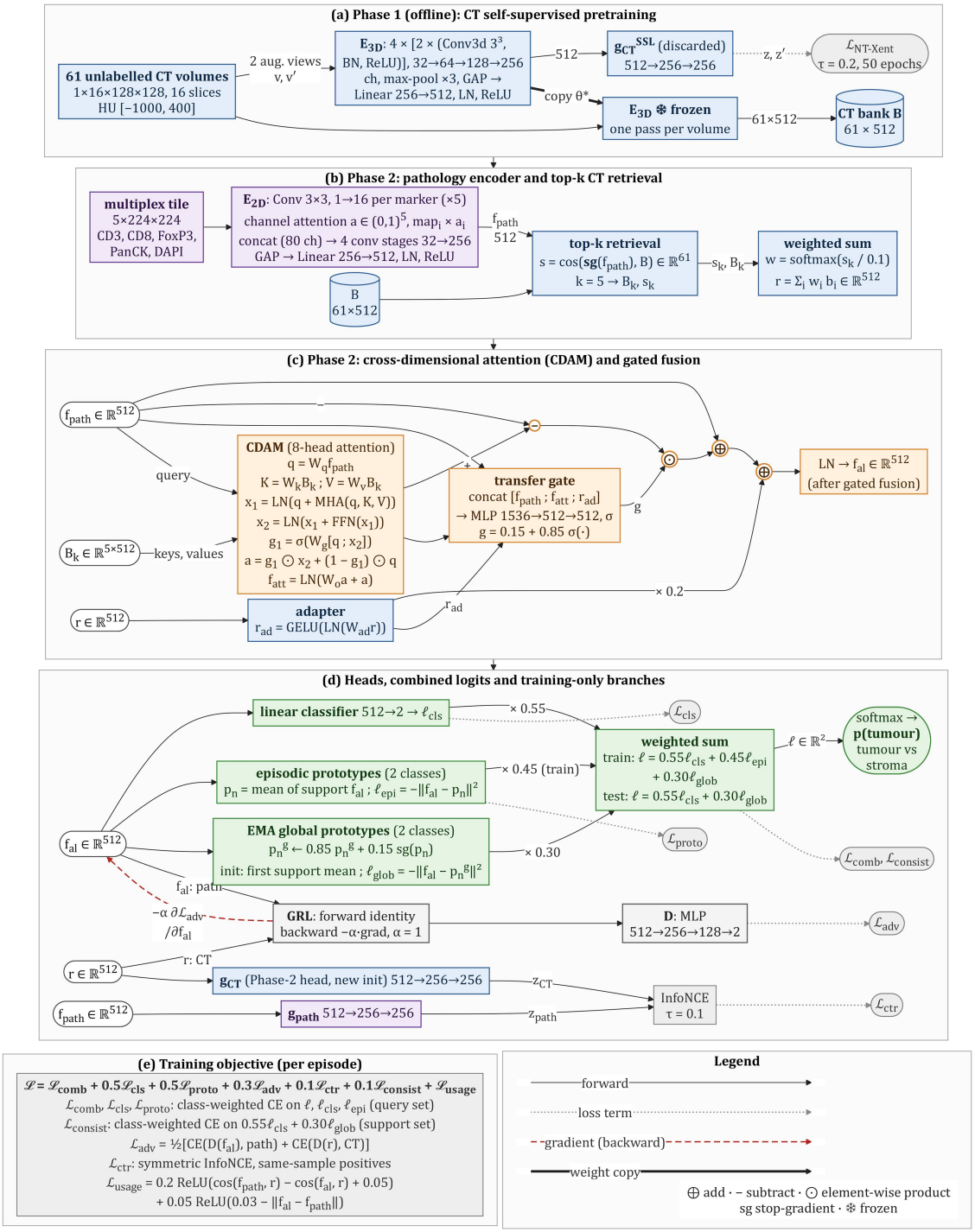

FIGURE 1
Architecture of the proposed cross-dimensional transfer model. A pathology feature retrieves whole-volume CT descriptors, which are fused through CDAM and a gated residual connection before episodic classification. **(a)** Phase 1: SSL pretraining of the 3D CT encoder, CT bank. **(b)** Phase 2: pathology encoder, top-k retrieval. **(c)** CDAM and gated fusion. **(d)** Heads, combined logits, training-only branches. **(e)** Per-episode training objective.

floor $\gamma_g = 0.15$, controls how much attended context is injected:

$$\begin{aligned} \mathbf{g} &= \gamma_g + (1-\gamma_g)\sigma\Big(\mathrm{MLP}([\mathbf{f}_{\mathrm{path}}; \mathbf{f}_{\mathrm{att}}; \mathbf{r}_{\mathrm{ad}}])\Big), \\ \mathbf{f}_{\mathrm{al}} &= \mathrm{LN}\Big(\mathbf{f}_{\mathrm{path}} + \mathbf{g} \odot (\mathbf{f}_{\mathrm{att}} - \mathbf{f}_{\mathrm{path}}) + \lambda_{\mathrm{res}}\mathbf{r}_{\mathrm{ad}}\Big), \end{aligned} \tag{4}$$

where MLP is a two-layer perceptron, $\lambda_{\mathrm{res}} = 0.20$, and the final LayerNorm has no learned parameters. The gate floor and the usage penalty defined below keep the transfer pathway open; an ablation removes both.

3.9 Adversarial domain alignment

A GRL (8) is the identity in the forward pass and multiplies the gradient by $-\alpha$ ($\alpha = 1$) in the backward pass.

A three-layer MLP discriminator D ($512 \to 256 \to 128 \to 2$) receives $\mathbf{f}_{\text{al}}$ (label: pathology) and $\mathbf{r}$ (label: CT) through the GRL and is trained with the unweighted cross-entropy $\mathcal{L}_{\text{adv}} = \frac{1}{2}[\text{CE}(D(\mathbf{f}_{\text{al}}), \text{path}) + \text{CE}(D(\mathbf{r}), \text{CT})]$, so the pathology branch is pushed toward features that D cannot separate from CT context. D contributes only to $\mathcal{L}_{\text{adv}}$ and never to the prediction.

3.10 Classification logits

Three sets of logits are computed from $\mathbf{f}_{\text{al}}$. The linear classifier gives $\ell^{\text{cls}} = \mathbf{W}_c\mathbf{f}_{\text{al}} + \mathbf{b}_c$. For an N-way K-shot episode with support set $\mathcal{S}$, episodic prototypes $\mathbf{p}_n = \frac{1}{|\mathcal{S}_n|}\sum_{i\in\mathcal{S}_n}\mathbf{f}_{\text{al},i}$ give $\ell_n^{\text{epi}} = -\mathbf{f}_{\text{al}} - \mathbf{p}_n 2^2$ (5). Global prototypes are initialized to the first support mean of each class and then maintained by an EMA, $\mathbf{p}_n^{\text{glob}} \leftarrow \beta\,\mathbf{p}_n^{\text{glob}} + (1-\beta)\,\mathbf{p}_n$ with $\beta = 0.85$, updated with each episode's (detached) support features before that episode's logits are computed; they are saved with the selected checkpoint and fixed at test time. They give $\ell_n^{\text{glob}} = -\mathbf{f}_{\text{al}} - \mathbf{p}_n^{\text{glob}} 2^2$. The combined logits are a fixed weighted sum (Equation 5):

$$\begin{aligned} \ell^{\text{train}} &= 0.55\,\ell^{\text{cls}} + 0.45\,\ell^{\text{epi}} + 0.30\,\ell^{\text{glob}}, \\ \ell^{\text{test}} &= 0.55\,\ell^{\text{cls}} + 0.30\,\ell^{\text{glob}}. \end{aligned} \tag{5}$$

No temperature or normalization is applied to the distance logits, and the weights are not constrained to sum to one. At test time no support set is available, so the episodic term is omitted and $p(\text{tumour}) = softmax(\ell^{\text{test}})_1$. Because squared distances between 512-dimensional layer-normalized features are large, the prototype term dominates the magnitude of ℓ (quantified in Section 4.4).

3.11 Training objective

With class-weighted cross-entropy CE_w (weights inversely proportional to training class frequencies), the loss per episode (Equations 6, 7) is

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{comb}} &+ 0.5\,\mathcal{L}_{\text{cls}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \\ &+ \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}} + \lambda_{\text{cons}}\mathcal{L}_{\text{consist}} + \mathcal{L}_{\text{usage}}, \end{aligned} \tag{6}$$

$$\begin{aligned} \mathcal{L}_{\text{comb}} &= \text{CE}_w(\ell_{\mathcal{Q}}^{\text{train}}, y), \quad \mathcal{L}_{\text{cls}} = \text{CE}_w(\ell_{\mathcal{Q}}^{\text{cls}}, y), \\ \mathcal{L}_{\text{proto}} &= \text{CE}_w(\ell_{\mathcal{Q}}^{\text{epi}}, y), \\ \mathcal{L}_{\text{consist}} &= \text{CE}_w(0.55\,\ell_{\mathcal{S}}^{\text{cls}} + 0.30\,\ell_{\mathcal{S}}^{\text{glob}}, y), \end{aligned} \tag{7}$$

where subscripts $\mathcal{Q}$ and $\mathcal{S}$ denote query and support tiles, $\lambda_{\text{proto}} = 0.5$, $\lambda_{\text{adv}} = 0.3$, $\lambda_{\text{ctr}} = 0.1$ and $\lambda_{\text{cons}} = 0.1$. The transfer-usage penalty, applied to query and support features and summed, (Equation 8) is

$$\begin{aligned} \mathcal{L}_{\text{usage}} = 0.20\,\text{ReLU}\Big(&\cos(\mathbf{f}_{\text{path}}, \mathbf{r}) - \cos(\mathbf{f}_{\text{al}}, \mathbf{r}) + 0.05\Big) \\ &+ 0.05\,\text{ReLU}\Big(0.03 - \mathbf{f}_{\text{al}} - \mathbf{f}_{\text{path}}2\Big) \end{aligned} \tag{8}$$

3.12 InfoNCE positives and a query-independent control

$\mathcal{L}_{\text{ctr}}$ is a symmetric InfoNCE loss (26) with temperature 0.1 between $g_{\text{path}}(\mathbf{f}_{\text{path}})$ and the Phase-2 head $g_{\text{CT}}(\mathbf{r})$ over the query batch; the positive for each query is its own retrieved summary $\mathbf{r}$ and the other queries' summaries are negatives. Because $\mathbf{r}$ is itself selected by the query, the loss reinforces the query's own retrieval rather than an independently defined correspondence. We test this with a control in which the CT side of each positive pair is a bank entry drawn uniformly at random, independently of the query (Section 3.18).

3.13 Training procedure

Phase 1 pretrains E_{3D} and $g_{\text{CT}}^{\text{SSL}}$ for 50 epochs with the NT-Xent loss (25) (temperature 0.2, batch size 8 with the last incomplete batch dropped, AdamW, constant learning rate 10^{-4}, gradient-norm clipping at 1.0) on pairs of augmented views (axis flips, intensity scaling and shift, Gaussian noise). Phase 2 trains all modules except the frozen E_{3D} jointly (including the discriminator, gate, adapter and g_{CT}; gradient-norm clipping at 1.0) for up to 30 epochs with early stopping (patience 10) on a composite validation score ($0.5\times$ balanced accuracy $+0.3\times$ macro-F1 $+0.2\times$ accuracy). Each episode samples K support and 5 query tiles per class; support and query tiles are encoded in one batch. The checkpoint with the best validation score is evaluated on the held-out patient. Algorithm 1 summarizes the procedure.

Algorithm 1 Training and evaluation protocol.

1: Pretrain E_{3D}, g_{CT}^{SSL} with NT-Xent on CT volumes (per seed); freeze E_{3D}
2: Encode every volume once: bank $\mathbf{B} \in \mathbb{R}^{61\times512}$ (or a control bank)
3: for each held-out patient $f \in \{1, \ldots, 8\}$ and seed $s \in \{0, 1, 2\}$ do
4: Validation patient ← next patient in rotation; train on the other six
5: for each epoch (max. 30) do
6: for each episode do
7: Sample K support and 5 query tiles per class; augment
8: Retrieve top- k bank entries (stop-gradient); fuse via CDAM and Equation 4
9: Update all pathology-side modules by minimizing Equation 6
10: end for
11: Evaluate on the validation patient; keep best checkpoint; early stop
12: end for
13: Predict the held-out patient's tiles with Equation 5 (test form)
14: end for
15: Pool out-of-fold predictions; patient-level cluster bootstrap for intervals

3.14 Experimental setup

3.14.1 Datasets

(i) **Source domain: LungCT-Diagnosis.** The source data are 61 diagnostic contrast-enhanced thoracic CT series (one per patient) from The Cancer Imaging Archive (TCIA),[1] collected to study CT descriptors of lung adenocarcinoma. The series have 24–150 axial slices (mean 76.8) with slice thicknesses of 2.5–6 mm. Diagnostic labels are not used.

(ii) **Target domain: HNSCC-mIF-mIHC-comparison.** The target data are multiplex pathology images from eight HNSCC patients (`Case1`–`Case8`), also from TCIA.[2] The same tumour sections were stained with mIF and then restained with multiplex immunohistochemistry (mIHC) and co-registered (43). For each patient, an experienced pathologist selected nine regions of interest on both the mIF and the mIHC images: three in the tumour core (T1–T3), three at the tumour margin (M1–M3) and three in the adjacent stroma (S1–S3) (43). Each region is subdivided into up to four 512 × 512 patches, giving 268 tiles in the release (of at most 288). Files are named `Case[patient]_[T/M/S][1-3]_[patch]_[marker]`; patient identity is therefore the `CaseN` prefix, and all tiles of a patient are always kept in the same partition. We use the five mIF channels CD3, CD8, FoxP3, PanCK and DAPI.

Table 3 gives the number of tiles per patient and region type, where T, M and S denote tumour core, margin and stroma. The task uses tumour core (T, positive) vs. stroma (S, negative), and margin (M) tiles are excluded, so the last column lists the tiles used. Table 4 compares the source and target data as used in this study.

TABLE 3 Pathology tiles per patient.

Patient	T	M	S	All tiles	Used (T/S)
Case1	12	12	5	29	12/5
Case2	12	12	12	36	12/12
Case3	11	12	4	27	11/4
Case4	12	12	9	33	12/9
Case5	12	12	12	36	12/12
Case6	12	12	12	36	12/12
Case7	11	12	12	35	11/12
Case8	12	12	12	36	12/12
Total	94	96	78	268	94/78

TABLE 4 Source and target data compared.

Characteristic	LungCT-diagnosis	HNSCC-mIF-mIHC
Modality	Computed tomography	Multiplex immunofluorescence
Dimensionality	3D volumetric	2D planar
Samples used	61 volumes (61 patients)	172 tiles (8 patients)
Native size	512 × 512 × (24--150)	512 × 512 px
Model input	16 × 128 × 128, 1 channel (HU)	224 × 224, 5 channels
Label usage	None (self-supervised)	Region type: T vs. S

3.15 Target labels and label audit

Each tile is labelled with the region type chosen by the pathologist in the source study (43): tumour core (T) as positive and stroma (S) as negative. Margin tiles (M), which by construction contain both tissue types, are excluded from this main task. This gives 94 tumour and 78 stroma tiles, and every patient contributes both classes.

The labels are region-level, not per-pixel annotations, and are not computed from the image channels given to the model. Because the regions were selected by inspecting the mIF and mIHC images, however, tumour-core regions are by construction regions rich in the epithelial marker PanCK, which is relevant for the channel results in Section 4.9. As a harder secondary task, we also classify tumour core (positive) against tumour margin (negative), giving 94 and 96 tiles; margin regions contain tumour cells as well, so this task depends less on the presence of the epithelial marker. The collection also distributes a segmentation image per tile, produced by the data contributors with an interactive deep-learning whole-cell/nuclear segmentation tool (43). On inspection, these images contain nuclear instance masks drawn in two colours (filled nuclei and their one-pixel boundaries) that coincide with DAPI-stained nuclei; they are not tumour masks. Because these masks are distributed alongside the tiles, a mask-coverage label is an easy alternative to derive, but it would measure nuclear density rather than tumour content. With a 5% boundary-pixel threshold (close to the median fraction of 0.049, so the tiles split roughly in half), such a label disagrees with the region type for 55 of 94 tumour-core and 18 of 78 stroma tiles, and on the 172 T/S tiles it can be predicted from the mean DAPI intensity alone with an AUC of 0.899, i.e., from an input channel of the model. We use region type as the only label, and the segmentation images are not used. The model's inputs are masked with a label-blind region derived from the 75th percentile of the DAPI channel (Section 3.16); this operation uses no label information.

Table 5 summarizes this audit. Its upper part compares the mask label with the region type for all 268 tiles. Its lower part gives the AUC of a single channel's mean intensity on the 172 T/S tiles, for the region-type label used here and for the mask label; these AUCs involve no learning and are computed on the images before DAPI masking.

[1] https://www.cancerimagingarchive.net/collection/lungct-diagnosis/, DOI: 10.7937/K9/TCIA.2015.A6V7JIWX.

[2] https://www.cancerimagingarchive.net/collection/hnscc-mif-mihc-comparison/, DOI: 10.7937/TCIA.2020.T90F-WB82.

TABLE 5 Label and channel audit.

Region type	Mask label = 0	Mask label = 1
Tumour core (T)	55	39
Margin (M)	26	70
Stroma (S)	60	18
Channel mean	AUC, T vs. S	AUC, mask label
CD3	0.620	0.625
CD8	0.580	0.468
FoxP3	0.459	0.507
PanCK	0.998	0.578
DAPI	0.489	0.899

3.16 Preprocessing

CT. Stored values are converted to Hounsfield units (HU) using the DICOM rescale slope and intercept. Slices are ordered by their z-position, and 16 slices are sampled uniformly over the whole series. Values are windowed to $[-1,000, 400]$ HU, scaled to $[0, 1]$ and resized to 128×128. *Pathology.* Each channel is resized to 224×224 and scaled to $[0, 1]$. A label-blind region mask is derived from the 75th percentile of the DAPI channel (with an intensity fallback when the mask is too sparse) and applied to all channels. Training augmentation applies horizontal/vertical flips, random 90° rotations, per-channel intensity scaling and Gaussian noise.

3.17 Cross-validation protocol and statistics

With eight patients, we use leave-one-patient-out cross-validation, holding out each patient once. The next patient in a fixed rotation serves as the validation patient for checkpoint selection, and the remaining six patients are used for training. Every condition uses the same folds, the same three seeds (0, 1, 2) and the same model initialization for a given seed and fold, so comparisons are paired. For each condition and seed, out-of-fold predictions of all 172 tiles are pooled to compute accuracy, balanced accuracy, macro-F1, the area under the receiver operating characteristic (ROC) curve (AUC), Cohen's κ, the Matthews correlation coefficient (MCC), sensitivity, specificity, the expected calibration error (ECE, 10 bins) (44) and the Brier score. We report the mean and standard deviation over seeds. The 95% confidence intervals (CIs) come from a patient-level cluster bootstrap (45) (2,000 resamples of the eight patients, with the metric averaged over seeds in each resample). Differences between two conditions (Δ) are computed on the same bootstrap resamples and seeds, which gives paired intervals. No correction for multiple comparisons is applied.

The episodic support size K describes the composition of each training episode, not the annotation budget: episodes are drawn repeatedly from all labelled tiles of the six training patients, which amounts to 124–136 unique labelled tiles per fold (mean 129), and the validation patient contributes 15–24 further labelled tiles for checkpoint selection. To measure performance under strictly limited annotation, the annotation-budget conditions below restrict the training pool itself.

The harder tasks, positive controls, annotation budgets, selection variants, preprocessing variants, three-seed channel-drop runs and extended bank-swap measurements form a second set of runs with its own reference full and pathology-only models (balanced accuracy 94.8% and 94.2%, AUC 0.989 and 0.979), so that every comparison is made within the same set of runs; its CT encoders were pretrained anew with the same settings. With eight patients, these intervals are wide, and we interpret an interval containing zero as no detectable difference, not as evidence of equivalence. With eight patients these intervals are wide, and we interpret an interval containing zero as no detectable difference, not as evidence of equivalence.

3.18 Experimental conditions

Every condition below, except for the test-time bank swap, is trained from scratch on the same folds as the full model:

- **Pathology-only comparator**: no retrieval, CDAM, adversarial or contrastive terms; the same encoder, prototypical heads, episodes and selection.
- **CT-bank controls**: the bank is replaced by descriptors from an untrained (randomly initialized) CT encoder; by Gaussian noise with the per-dimension mean and standard deviation of the real bank; by the real bank with each entry's dimensions randomly permuted; or by a constant bank (every entry equal to the bank mean). A further control keeps the real bank but retrieves k entries at random.
- **Circularity and forced-use controls**: InfoNCE with query-independent random bank entries as positives; removal of the gate floor and the usage penalty ($\gamma_g = 0$, $\mathcal{L}_{\text{usage}} = 0$), so the model is free to ignore the context.
- **Component ablations**: without the fusion pathway (CDAM, adapter, gate, residual and usage penalty removed, so the fused feature is $\mathbf{f}_{\text{path}}$; CT then enters only through $\mathcal{L}_{\text{adv}}$ and InfoNCE), without the adversarial loss, without InfoNCE, and without prototypes (classifier logits only).
- **Test-time bank swap**: the trained full models are re-evaluated with each control bank and with random retrieval, without retraining. Besides the predicted probabilities, we record the change of the weighted logit margins, the cosine similarity and Euclidean distance between the fused features before and after the swap, the fraction of decisions that flip, Cohen's κ between the decisions, and the change of each patient's balanced accuracy.
- **Support size**: $K = 8$ for the full and pathology-only models, trained independently of the $K = 2$ models.
- **SSL sensitivity**: downstream training with encoders from 0 (random initialization), 10, 25, 50 and 100 pretraining epochs at temperature 0.2 (snapshots of one 100-epoch run), and with temperatures 0.05, 0.1 and 0.5 at 50 epochs (seed 0).

- **Harder tasks**: PanCK removed from the input (full model, pathology-only and a Gaussian-noise bank), and tumour core vs. margin (full model and pathology-only).
- **Positive controls**: with PanCK removed from the input, the retrieved CT context is replaced by a context vector that is known to be label-relevant. In the *PanCK-context* control this vector is a fixed random projection to 512 dimensions of a standardized descriptor of each tile's own PanCK intensities (a 16-bin histogram, mean, standard deviation, fraction above 0.2, median and 90th percentile, computed without labels). In the *oracle* control it is a fixed random vector per class plus per-tile noise; this context encodes the label and serves only as a capacity check. Both are fed through the same fusion pathway as the CT context.
- **Annotation budgets**: the training pool is restricted to 2, 4, 8 or 16 randomly chosen labelled tiles per class (drawn per seed and fold), for the full model and the pathology-only model.
- **Checkpoint selection**: the validation patient is the previous rather than the next patient in the rotation, or no validation is used and the model after the last epoch (30) is kept; for both the full and pathology-only models.
- **Channel reliance**: each input channel set to zero in training and testing (three seeds). The DAPI-derived region mask is still applied in these runs, so the DAPI-drop condition removes the DAPI intensities but not the mask. Preprocessing without the mask is evaluated separately for the full model, the pathology-only model and the DAPI-drop condition. A hand-crafted baseline, ℓ_2-regularized logistic regression ($C = 1$) on per-channel mean, standard deviation and fraction of pixels above 0.2 (15 features, computed before DAPI masking), is evaluated with the same leave-one-patient-out folds, also for the tumour-vs.-margin task and without the PanCK statistics.

3.19 Retrieval and alignment diagnostics

On the out-of-fold predictions of the full model we compute the following.

(i) **Bank usage**: the number of bank entries ever retrieved, the normalized entropy of retrieval counts, and the share of retrievals taken by the five most-used entries.

(ii) **Class dependence of retrieval**: the mutual information between the label and the top-1 retrieved entry, with a 1000-permutation p-value. Because retrieval is computed from pathology features that already encode the class, a label permutation is not the relevant null. We therefore compare the mutual information with that of models trained with the replacement banks of 3.18, which contain no CT information.

(iii) **Per-patient diagnostics**: for each held-out patient, the overlap (Jaccard index) of the bank entries retrieved for its tumour and stroma tiles, the within-patient mutual information, and the context-probe AUC.

(iv) **Transferable label information**: the AUC of a leave-one-patient-out logistic probe ($C = 0.1$, trained on the seven other patients) that predicts the label from the retrieved context $\mathbf{r}$ alone, with a 100-permutation p-value.

(v) **Stability**: the Jaccard overlap of each tile's retrieved set across seeds.

(vi) **Domain separability**: the balanced accuracy of a logistic classifier separating bank entries from pathology features (5-fold cross-validation), and a linear maximum mean discrepancy (MMD) (46) between normalized features, both before and after fusion.

(vii) **Representation similarity**: linear centred kernel alignment (CKA) (47) between pathology features and their retrieved context.

(viii) **Usage of the fusion module**: the normalized attention entropy over the k retrieved entries, the mean transfer-gate value, and the median absolute contribution of each (weighted) logit term to the tumour-vs.-stroma margin. Because the out-of-fold features of different patients come from different fold models, the pooled separability, CKA and t-SNE (48) analyses are descriptive. For the SSL encoder we record the loss, the view-retrieval top-1 accuracy (whether the second augmented view of each volume is the nearest neighbour of the first among the 61 volumes), the effective rank of the bank, and the mean pairwise cosine similarity (42).

3.20 Implementation and hyperparameters

Table 6 lists the configuration. Training used two NVIDIA Tesla T4 GPUs with mixed-precision (fp16, with gradient scaling) computation in the pathology encoder; distances, prototypes and losses are computed in fp32. The main experiment set (464 training runs and seven SSL pretraining runs) took about 3 h of wall-clock time with two jobs sharing each GPU, and the second set (696 training runs and three SSL pretraining runs) about 3.7 h; one seed (eight folds) took about 12 min for the full model and 10–11 minutes for the retrained controls. Non-deterministic cuDNN kernels were enabled for speed, so runs with the same seed are reproducible up to floating-point non-determinism.

4 Results

4.1 Feasibility and main comparison (RQ1, RQ2)

Table 7 reports pooled out-of-fold results. The complete pipeline trained stably on all folds and reached a balanced accuracy of 94.0% and an AUC of 0.989. The pathology-only model, trained identically but without any CT context, reached 94.8% and 0.979. The paired differences were small and their intervals contained zero: Δ balanced accuracy (pathology-only minus full) +0.7pp (−3.1 to +5.0) and Δ AUC −0.010 (−0.029

to +0.006). With $K = 8$ support tiles per class the two models were again indistinguishable (95.2% vs. 95.1%). A logistic regression on 15 hand-crafted channel-intensity statistics numerically exceeded both networks (98.1%, AUC 0.996), although its paired interval against the full model also contained zero. Performance was high for every held-out patient (Table 8). The lowest balanced accuracies were for Case1 and Case4 (84.7% and 89.4%), and every patient had a mean AUC of at least 0.963. Pooled over seeds, the full model's errors were mostly missed tumour tiles (24 false negatives vs. 8 false positives out of 516 tile predictions).

In Table 7, all metrics are computed from the pooled out-of-fold predictions of the 172 tiles and are given as mean ± SD over three seeds, with 95% CIs from the patient-level cluster bootstrap (Section 3.17). Sensitivity and specificity are for the tumour class, and the confusion counts (TN/FP/FN/TP) are summed over the three seeds (516 predictions). The channel-statistics baseline is deterministic and was therefore evaluated in one run. Table 8 gives the held-out results of the full model ($K = 2$) for each patient, as mean ± SD over seeds.

TABLE 6 Hyperparameter configuration.

Category	Parameter	Value
Data	CT input	$16 \times 128 \times 128$, HU $[-1,000, 400]$
	Pathology input	224×224, 5 channels
Phase 1 (SSL)	Epochs/batch size	50/8
	Optimizer	AdamW, lr 10^{-4}, wd 10^{-4}
	NT-Xent temperature	0.2
	Augmentation strength	0.5
Phase 2	Max. epochs/patience	30/10
	Optimizer	AdamW, lr 10^{-4}, wd 10^{-4}, clip 1.0
	LR schedule	Cosine, warm restarts ($T_0 = 20$)
	Episodes per epoch	31–34 ($K = 2$)/20 ($K = 8$)
Episodes	N-way/K-shot/queries	2/2 (8)/5 per class
	Global-prototype momentum β	0.85
	Logit weights (cls, epi, glob)	0.55, 0.45, 0.30
Retrieval/ CDAM	Top-k/temperature τ_r	5/0.1
	Attention heads	8
	Gate floor γ_g/residual λ_{res}	0.15/0.20
	Usage penalty (weights, margin, shift)	0.20, 0.05; 0.05; 0.03
Loss weights	λ_{proto}, λ_{adv}	0.5, 0.3
	λ_{ctr} (InfoNCE, $\tau = 0.1$), λ_{cons}	0.1, 0.1
Evaluation	CV	Leave-one-patient-out (8 folds)
	Seeds/bootstrap resamples	0, 1, 2/2,000 (patient level)

4.2 CT-bank controls and ablations (RQ2, RQ3)

Table 9 and Figure 2 compare every control and ablation with the full model. Replacing the SSL bank with descriptors from an untrained encoder, with Gaussian noise or with a constant vector, or retrieving entries at random, did not reduce performance. All balanced accuracies were between 94.5% and 96.0%, and every Δ interval contained zero except one: a small

TABLE 8 Per-patient held-out results.

Patient	Tiles (T/S)	Bal. acc. (%)	AUC
Case 1	17 (12/5)	84.7 ± 16.8	1.000 ± 0.000
Case 2	24 (12/12)	100.0 ± 0.0	1.000 ± 0.000
Case 3	15 (11/4)	95.5 ± 0.0	1.000 ± 0.000
Case 4	21 (12/9)	89.4 ± 6.3	0.963 ± 0.064
Case 5	24 (12/12)	100.0 ± 0.0	1.000 ± 0.000
Case 6	24 (12/12)	93.1 ± 12.0	0.984 ± 0.028
Case 7	23 (11/12)	92.4 ± 5.2	0.995 ± 0.004
Case 8	24 (12/12)	97.2 ± 2.4	0.998 ± 0.004

TABLE 7 Leave-one-patient-out results.

Model	Bal. acc. (%)	95% CI	AUC	95% CI	MCC	Sens./Spec. (%)	ECE	Brier	TN/FP/ FN/TP
Full model (CT context), $K = 2$	94.0 ± 1.2	[90.3, 97.4]	0.989 ± 0.001	[0.975, 0.998]	0.878 ± 0.026	91.5/96.6	0.062	0.058	226/8/24/258
Pathology-only, $K = 2$	94.8 ± 0.6	[90.7, 98.4]	0.979 ± 0.006	[0.952, 0.998]	0.893 ± 0.012	92.6/97.0	0.054	0.053	227/7/21/261
Full model, $K = 8$	95.2 ± 2.5	[92.9, 97.8]	0.987 ± 0.010	[0.971, 0.998]	0.903 ± 0.048	95.0/95.3	0.048	0.046	223/11/14/268
Pathology-only, $K = 8$	95.1 ± 1.1	[91.2, 98.9]	0.984 ± 0.011	[0.968, 0.999]	0.900 ± 0.020	93.6/96.6	0.051	0.050	226/8/18/264
Channel statistics + logistic regression	98.1	[95.1, 100.0]	0.996	[0.983, 1.000]	0.965	100.0/96.2	0.053	0.019	75/3/0/94

AUC reduction for the dimension-shuffled bank (−0.021, CI −0.040 to −0.001) without a change in balanced accuracy. This is the only one of the 24 paired intervals in Table 9 that excludes zero; with no multiplicity correction, one such interval is consistent with chance. Using query-independent random positives in the InfoNCE term, removing the gate floor and usage penalty, and removing the fusion pathway, the adversarial loss, InfoNCE or the prototypes likewise produced no detectable change.

Each condition in Table 9 was trained with three seeds on the same folds as the full model. The Δ columns give the condition minus the full model, with the paired 95% patient-bootstrap CI in brackets. Figure 2 plots the paired differences from the full model for all controls, and ablations as points, with 95% patient-level bootstrap intervals as lines, for balanced accuracy (a) and AUC (b); intervals that exclude zero are drawn in vermillion.

4.3 Dependence of predictions on CT content (RQ3)

Swapping the bank of the trained full models at test time left their predictions essentially unchanged on average (Table 10). The mean absolute change in the predicted tumour probability was 0.008–0.015, and balanced accuracy stayed between 93.6% and 94.0%. Together with the retrained controls, these results indicate that the prediction depends at most weakly on which

TABLE 9 Controls and ablations.

Condition	Bal. acc. (%)	AUC	Δ bal. acc. (pp)	Δ AUC
Full model (SSL bank, top-k)	94.0 ± 1.2	0.989 ± 0.001	–	–
Pathology-only	94.8 ± 0.6	0.979 ± 0.006	+0.7 [−3.1, +5.0]	−0.010 [−0.029, +0.006]
Bank: untrained CT encoder	95.2 ± 1.5	0.980 ± 0.018	+1.1 [−3.6, +6.2]	−0.009 [−0.031, +0.007]
Bank: Gaussian noise	95.2 ± 0.4	0.989 ± 0.007	+1.2 [−1.9, +4.6]	0.000 [−0.009, +0.009]
Bank: dimension-shuffled	94.5 ± 2.4	0.968 ± 0.030	+0.4 [−3.0, +3.5]	−0.021 [−0.040, −0.001]
Bank: constant (mean)	95.8 ± 1.7	0.988 ± 0.006	+1.8 [−2.1, +6.1]	−0.001 [−0.018, +0.015]
Random retrieval	96.0 ± 0.4	0.994 ± 0.003	+2.0 [−0.5, +4.9]	+0.004 [−0.002, +0.017]
InfoNCE: random positives	95.3 ± 1.2	0.990 ± 0.010	+1.3 [−2.8, +6.3]	+0.001 [−0.012, +0.013]
No gate floor/usage penalty	95.2 ± 1.8	0.986 ± 0.011	+1.2 [−2.8, +5.8]	−0.003 [−0.013, +0.007]
−Fusion (CDAM, gate)	95.5 ± 2.2	0.991 ± 0.007	+1.4 [−0.4, +3.6]	+0.002 [−0.008, +0.013]
−Adversarial	95.7 ± 4.1	0.985 ± 0.016	+1.7 [−1.4, +5.2]	−0.005 [−0.017, +0.007]
−InfoNCE	94.1 ± 2.1	0.976 ± 0.011	0.0 [−3.9, +3.8]	−0.013 [−0.037, +0.006]
−Prototypes	95.0 ± 1.8	0.991 ± 0.005	+1.0 [−2.4, +5.6]	+0.002 [−0.006, +0.014]

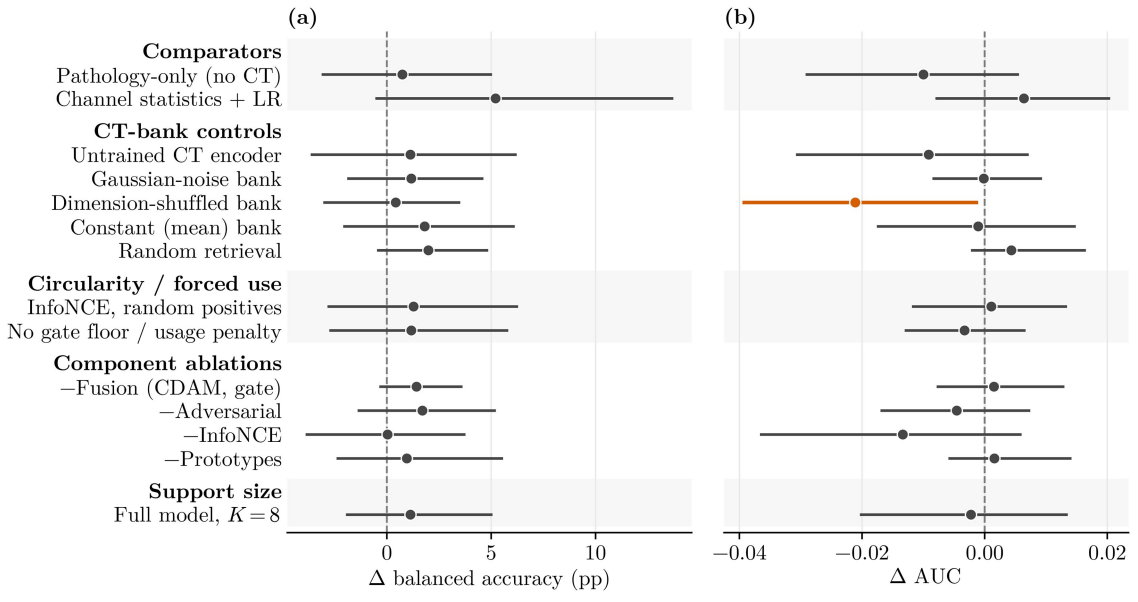

FIGURE 2
Paired differences from the full model. **(a)** Change in balanced accuracy. **(b)** Change in AUC.

descriptors are retrieved, including when they are noise or a constant vector rather than CT descriptors.

In Table 10, the full models were trained with the SSL bank and evaluated with a replacement bank; values are mean ± SD over seeds, and $|\Delta p|$ is the mean absolute change of the predicted tumour probability.

Because the predicted probabilities are saturated, small probability changes could hide larger internal changes. Table 11 therefore reports the swap at the level of logits, features and decisions for the reference runs of the second set. Swapping the bank changed the total weighted logit margin by 2.3–2.6 on average, i.e., 5%–6% of its typical magnitude, and left the fused features nearly unchanged in direction (mean cosine similarity 0.983–0.993, minimum per fold 0.970). In Euclidean terms, the fused features moved by 0.32–0.52 times the median distance between two tiles of the same held-out patient, so the context does shift the representation, but along directions that do not affect the decision: only 2 to 7 of 516 decisions flipped (Cohen's κ 0.973–0.992 between the decisions before and after the swap), and the balanced accuracy of individual patients changed in 2 to 4 of the 24 patient-seed combinations, by at most 12.5pp.

In Table 11, $|\Delta\text{margin}|$ is the mean absolute change of the total weighted logit margin (in parentheses relative to the mean absolute margin before the swap), feature cos. is the mean cosine similarity between the fused features before and after the swap, and rel. shift is their mean Euclidean distance divided by the median distance between fused features of two tiles of the same held-out patient; flips and κ are pooled over three seeds and eight folds.

4.4 Retrieval and alignment diagnostics (RQ3)

Table 12 summarizes the diagnostics of the full model, computed out of fold as mean ± SD over seeds. The fusion pathway was open (mean transfer gate 0.62), and attention over the five retrieved entries was moderately non-uniform (normalized entropy 0.93; mean maximum weight 0.33 vs. 0.20 for uniform weights). On average 46, (40–49 across seeds) of the 61 bank entries were retrieved at least once, with the five most-used entries taking 31% of retrievals. The top-1 retrieved entry carried 0.32 nats of mutual information with the label (label-permutation $p \leq 0.001$). But this apparent class dependence does not indicate CT information. Retrieval is computed from pathology features that already encode the class, and models trained with banks that contain no CT information showed the same dependence (Table 13): 0.24 nats for the untrained encoder, 0.33 for Gaussian noise, 0.38 for the dimension-shuffled bank and 0.31 for random retrieval. Within patients, the entries retrieved for tumour and stroma tiles overlapped only partly (Jaccard index 0.22–0.50 per patient), consistent with class-specific pathology features rather than class-specific CT content. Three further results show that retrieval carries no transferable label information. First, a logistic probe trained on the retrieved context of the seven other patients and tested on the held-out patient did not exceed chance (AUC 0.36, permutation $p = 0.98$; per patient 0.09–0.65). Second, the retrieved sets were almost entirely different between seeds (Jaccard 0.07). Third, the CT bank and the pathology features remained perfectly linearly separable before and after fusion

TABLE 10 Test-time bank swap.

Evaluation bank	Bal. acc. (%)	AUC	$\lvert\Delta p\rvert$
SSL bank (as trained)	94.0 ± 1.2	0.989 ± 0.001	–
Untrained CT encoder	94.0 ± 1.7	0.985 ± 0.004	0.015
Gaussian noise	93.6 ± 0.7	0.986 ± 0.005	0.008
Dimension-shuffled	93.8 ± 0.9	0.986 ± 0.005	0.008
Constant (mean)	93.6 ± 0.7	0.986 ± 0.005	0.008
Random retrieval	93.8 ± 0.9	0.987 ± 0.005	0.010

TABLE 12 Retrieval and alignment diagnostics.

Diagnostic	Value
Bank entries ever retrieved (of 61; range)	46.0 ± 5.2 (40–49)
Normalized retrieval entropy	0.842 ± 0.023
Share of top-5 entries	0.310 ± 0.030
Label–top-1 mutual information (nats)/perm. p	0.315 ± 0.016/≤0.001
Context-only probe AUC/perm. p	0.358 ± 0.050/0.98
Retrieval Jaccard across seeds	0.068
Domain classifier bal. acc., CT vs. raw/fused	1.000/1.000
Linear MMD, CT vs. raw/fused	0.254/0.457
Linear CKA, raw/fused vs. context	0.366/0.334
Attention entropy/log k	0.927 ± 0.041
Mean max. attention weight ($1/k = 0.2$)	0.333 ± 0.043
Mean transfer gate	0.618 ± 0.006
Median |weighted margin|: classifier/global prototypes	2.64/30.5
Decisions agreeing in sign: global-prototype/classifier term	100%/97.3%

TABLE 11 Bank swap: logits, features, decisions.

Evaluation bank	$\lvert\Delta p\rvert$	$\lvert\Delta\text{margin}\rvert$ (rel.)	Feature cos.	Rel. shift	Flips/516	κ
Untrained CT encoder	0.003	2.37 (0.051)	0.983	0.52	2	0.992
Gaussian noise	0.010	2.30 (0.059)	0.992	0.36	7	0.973
Dimension-shuffled	0.010	2.56 (0.064)	0.989	0.42	7	0.973
Constant (mean)	0.009	2.28 (0.057)	0.993	0.32	6	0.977
Random retrieval	0.007	2.41 (0.056)	0.992	0.34	5	0.981

(domain-classifier balanced accuracy 1.00), and the linear MMD increased rather than decreased after fusion (0.25 to 0.46), so the adversarial term did not align the domains.

The logit magnitudes point to a separate property of the model: the weighted global-prototype term had a median absolute margin of 30.5 against 2.6 for the classifier term, and it agreed in sign with every decision (the classifier term with 97.3%).

Figure 3a shows how often each of the 61 CT bank entries is retrieved (out-of-fold, seed 0), with a dashed line marking uniform usage; Figures 3b, c show the distributions of the normalized attention entropy and the mean transfer-gate value over all out-of-fold tiles and seeds.

TABLE 13 Retrieval class dependence: nulls.

Bank (training and test)	MI (nats)
SSL CT bank (full model)	0.315 ± 0.016
Untrained CT encoder	0.244 ± 0.051
Gaussian noise	0.326 ± 0.025
Dimension-shuffled	0.379 ± 0.059
Random retrieval	0.311 ± 0.000

Figure 4 is a t-SNE projection of the CT bank (grey) and the out-of-fold pathology features of seed 0, before (a) and after (b) gated fusion, with tumour-core tiles in vermillion and stroma tiles in blue. The CT bank forms a separate cluster, and after fusion the tumour and stroma tiles of each cluster lie side by side. Because each held-out patient is embedded by its own fold model, the pathology clusters largely correspond to folds, so the projection is qualitative; the quantitative measures are those in Table 12.

In Table 13, the mutual information is computed on the pooled out-of-fold predictions of each model (mean ± SD over three seeds). The constant bank is not listed because all of its entries are identical, so the retrieved entry carries no information by construction.

4.5 SSL convergence and sensitivity (RQ1, RQ3)

The view-retrieval criterion saturated within the 50-epoch schedule (Figure 5). At temperature 0.2 the view-retrieval top-1 accuracy (chance 0.016) rose from 0.21 at initialization to 0.66 after 10 epochs, 0.92 after 25 and 1.00 after 50 (0.97–1.00 for

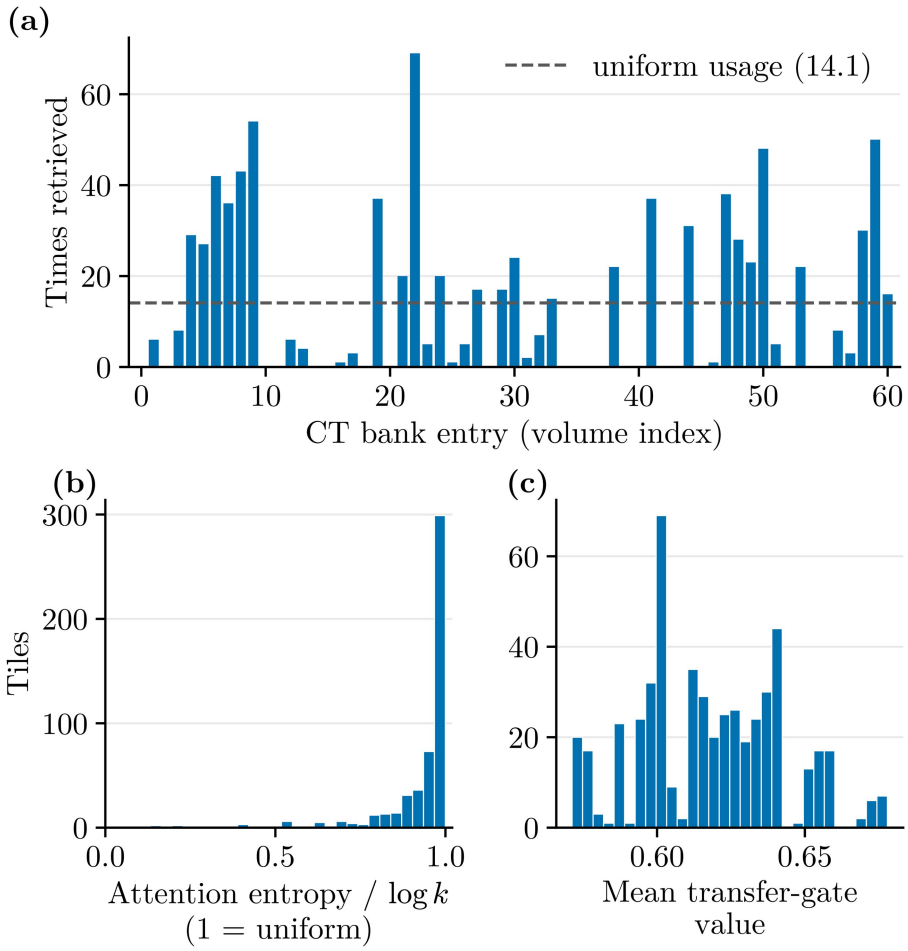

FIGURE 3
Bank usage, attention and gating. **(a)** Number of times each CT bank entry was retrieved. **(b)** Distribution of tiles by normalized attention entropy. **(c)** Distribution of tiles by mean transfer-gate value.

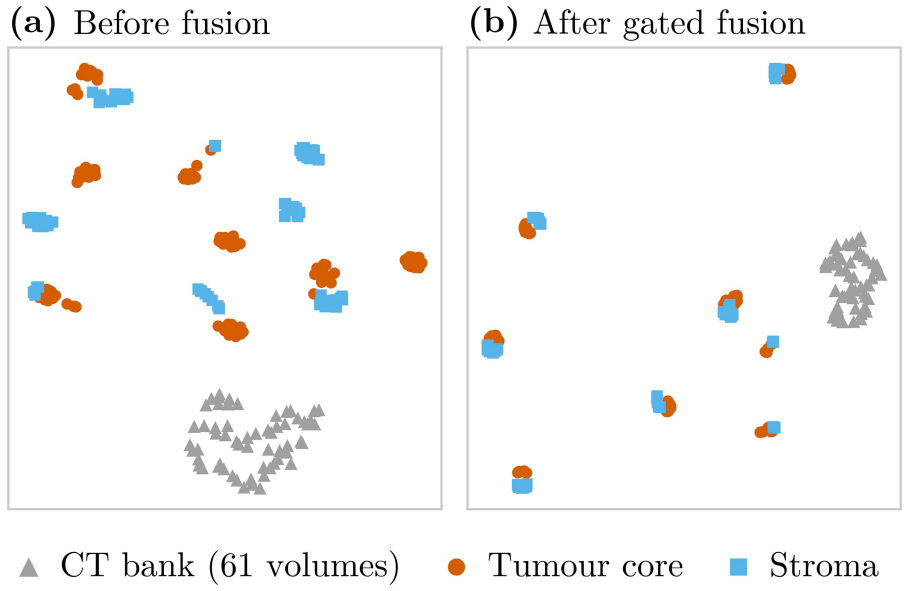

FIGURE 4
t-SNE feature projection. **(a)** Before fusion. **(b)** After gated fusion.

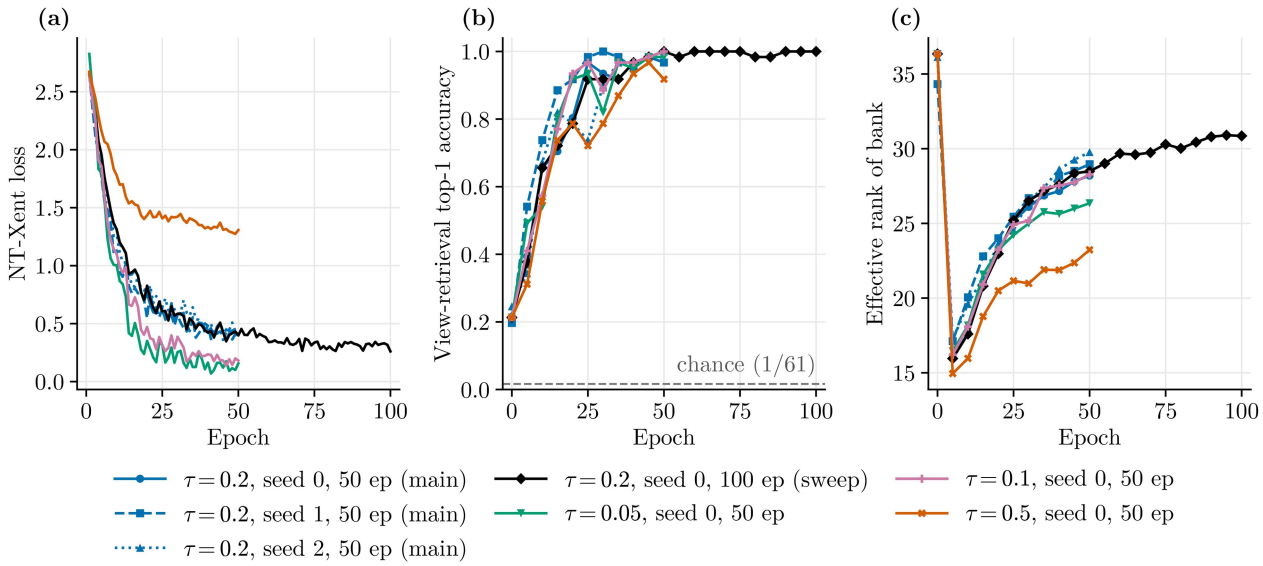

FIGURE 5
SSL pretraining convergence. **(a)** NT-Xent loss. **(b)** View-retrieval top-1 accuracy. **(c)** Effective rank of the CT bank, over training epochs.

the three main encoders), and remained at or above 0.98 up to 100 epochs, while the loss (0.40 to 0.26) and effective rank (28.5 to 30.9) continued to improve slowly between 50 and 100 epochs.

The effective rank of the bank first dropped and then recovered to 28–31, and the mean pairwise cosine similarity fell from 0.999 (random initialization) to 0.41–0.45, so the representation did not collapse. The temperature mattered for the representation: $\tau = 0.5$ gave a lower effective rank (23.2) and view-retrieval accuracy (0.92) than $\tau \leq 0.2$ (loss values are not comparable across temperatures). None of these differences reached the downstream task, however (Table 14). Balanced accuracy varied between 91.5% and 96.1% and AUC between 0.984 and 0.996 across all settings, including the untrained encoder. This range is comparable to the seed-to-seed variation of the full model (92.9%–95.3%).

4.6 Harder tasks and annotation budgets (RQ2)

Because PanCK nearly separates tumour core from stroma, the main task leaves little room for any auxiliary signal. We therefore repeated the comparison where there is room for improvement (Table 15, Figure 6). With PanCK removed from the input,

TABLE 14 SSL sensitivity analysis.

Epochs	τ	Loss	View top-1	Eff. rank	Bal. acc. (%)	AUC
0	0.2	–	0.213	36.3	94.8	0.988
10	0.2	1.313	0.656	17.6	95.8	0.996
25	0.2	0.606	0.918	25.2	94.7	0.990
50	0.2	0.398	1.000	28.5	92.0	0.996
100	0.2	0.260	1.000	30.9	95.7	0.996
50	0.05	0.157	0.984	26.3	95.7	0.991
50	0.1	0.181	1.000	28.3	96.1	0.989
50 (main)	0.2	0.407	1.000	28.2	92.9	0.990
50	0.5	1.307	0.918	23.2	91.5	0.984

TABLE 15 Harder tasks, budgets and positive controls.

Setting	Model	Bal. acc. (%)	AUC	Δ bal. acc. (pp)	Δ AUC
PanCK removed	Pathology-only	76.6 ± 4.8	0.829 ± 0.036	reference	reference
	Full model (CT bank)	75.8 ± 1.1	0.805 ± 0.016	−0.9 [−6.1, + 6.3]	−0.024 [−0.060, + 0.026]
	Gaussian-noise bank	78.1 ± 5.0	0.828 ± 0.028	+1.4 [−1.7, + 5.8]	−0.001 [−0.035, + 0.033]
	Channel statistics + LR	75.6	0.830	–	–
PanCK removed, positive controls	PanCK-statistics context	95.5 ± 0.6	0.990 ± 0.007	+18.9 [+10.5, + 27.5]	+0.161 [+0.051, + 0.269]
	Oracle context	100.0 ± 0.0	1.000 ± 0.000	+23.4 [+13.7, + 32.0]	+0.171 [+0.056, + 0.278]
Tumour core vs. margin	Pathology-only	77.9 ± 4.5	0.840 ± 0.039	reference	reference
	Full model (CT bank)	76.5 ± 2.5	0.844 ± 0.024	−1.4 [−3.8, + 1.2]	+0.004 [−0.022, + 0.027]
	Channel statistics + LR	77.9	0.878	–	–
Annotation budget (full vs. path.-only)	2 tiles per class	66.0/68.1	0.718/0.771	−2.1 [−6.8, + 3.8]	−0.053 [−0.085, + 0.001]
	4 tiles per class	81.6/80.7	0.870/0.867	+0.9 [−5.7, + 8.2]	+0.003 [−0.083, + 0.086]
	8 tiles per class	92.6/93.7	0.964/0.979	−1.1 [−2.9, + 0.4]	−0.015 [−0.034, + 0.004]
	16 tiles per class	94.1/94.6	0.982/0.983	−0.5 [−2.9, + 2.0]	−0.001 [−0.011, + 0.008]

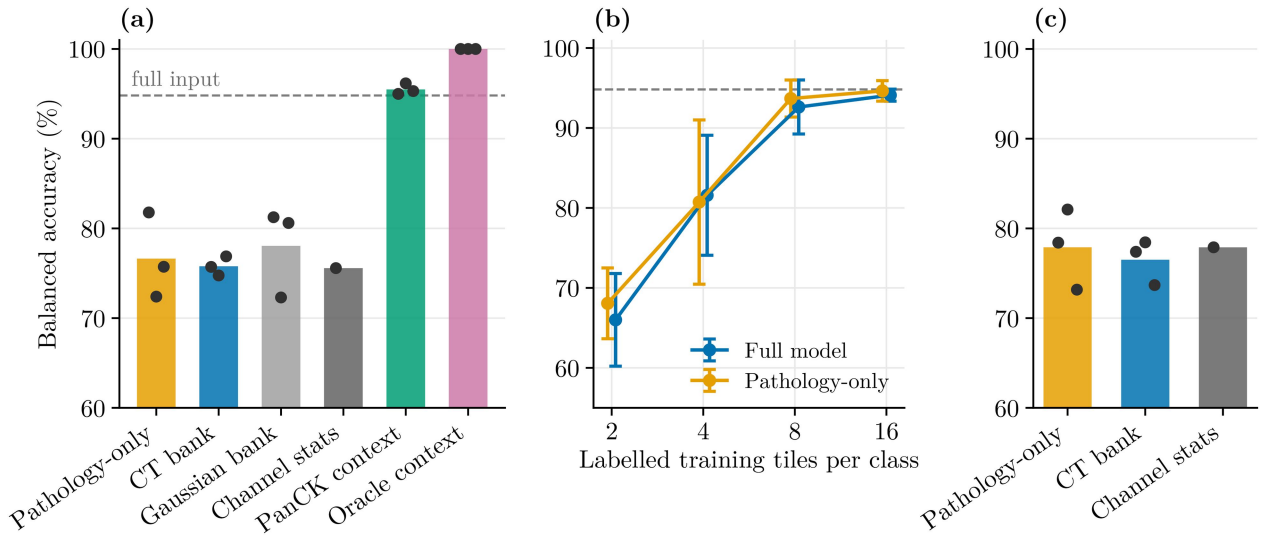

FIGURE 6
Harder tasks and positive controls. **(a)** Balanced accuracy for six input conditions. **(b)** Balanced accuracy versus labelled training tiles per class. **(c)** Balanced accuracy for three input conditions.

balanced accuracy fell to 75.8% for the full model and 76.6% for the pathology-only model; the CT context did not help (Δ −0.9pp, CI −6.1 to +6.3; Δ AUC −0.024, −0.060 to +0.026), and a Gaussian-noise bank performed as well as the CT bank (78.1%). On tumour core vs. margin, the full model reached 76.5% and the pathology-only model 77.9% (Δ −1.4pp, −3.8 to +1.2), and a logistic regression on channel statistics reached 77.9%. Under strictly limited annotation, performance rose from 66%–68% with 2 labelled tiles per class to 94%–95% with 16, for both models, and the CT context gave no benefit at any budget (all intervals contain zero; Table 15).

In Table 15, values are mean ± SD over three seeds of the pooled out-of-fold predictions, and Δ is the paired difference from the pathology-only model of the same setting, with 95% patient-level bootstrap CIs; for the budgets it is the full model minus the pathology-only model at the same budget. The channel-statistics baselines are deterministic (one run). Figure 6 shows the per-seed balanced accuracies of the PanCK-removed regime with the positive controls (a), the annotation-budget curves (b) and the tumour-vs.-margin task (c); the dashed line marks the reference full model with all channels.

4.7 Positive controls (RQ3)

The controls above show that the model does not use the CT context, but not whether it could use an informative context at all. With PanCK removed from the input, supplying each tile's own PanCK intensity statistics as the context, through the same fusion pathway with retrieval bypassed, raised balanced accuracy from 76.6% to 95.5% (Δ +18.9pp, CI +10.5 to +27.5; AUC 0.829 to 0.990), essentially recovering the accuracy of the model that receives PanCK as an input channel (94.8%). The oracle context reached 100% (Table 15, Figure 6a). The fusion architecture can therefore exploit an auxiliary context when that context carries label information; the absence of an effect of the CT context reflects the content of the CT bank, not an inability of the architecture to use context.

4.8 Sensitivity to patients, checkpoint selection and preprocessing

With eight patients, a single patient could dominate the comparison. Leaving out each patient in turn changed the difference in balanced accuracy between the full and pathology-only models only within −1.9 to +1.0pp in the main analysis (−0.2 to +1.3pp in the reference runs of the second set), and the difference in AUC stayed between +0.002 and +0.016. Per patient, the difference ranged from −12.5pp (Case1) to +6.9pp (Case8) with no consistent sign. Checkpoint selection did not produce a stable advantage either (Table 16): with the previous patient as the validation patient, the full model was 2.5pp better than the pathology-only model (CI +0.1 to +5.6), whereas without any validation (model after the last epoch) it was 1.5pp worse (−3.1 to +0.1). Given the number of comparisons and the opposite signs, we do not interpret either difference as an effect of the CT context.

TABLE 16 Selection, preprocessing and channels.

Comparison	Bal. acc. (%)	Δ bal. acc. (pp)
Full vs. path.-only, validation = next	94.8/94.2	+0.6 [−1.2, + 2.7]
Full vs. path.-only, validation = previous	95.1/92.6	+2.5 [+0.1, + 5.6]
Full vs. path.-only, last epoch	95.3/96.8	−1.5 [−3.1, + 0.1]
Full vs. path.-only, no DAPI mask	91.4/95.8	−4.4 [−9.0, − 0.8]
No DAPI mask: −DAPI vs. full	96.7/91.4	+5.3 [+2.5, + 8.5]
−CD3 vs. full	94.1/94.8	−0.7 [−4.7, + 2.6]
−CD8 vs. full	96.6/94.8	+1.7 [−1.3, + 5.4]
−FoxP3 vs. full	92.9/94.8	−2.0 [−4.9, + 0.5]
−PanCK vs. full	75.8/94.8	−19.0 [−29.6, − 8.7]
−DAPI vs. full	97.2/94.8	+2.3 [+0.8, + 4.0]

In Table 16, all rows come from the second set of runs (three seeds each); balanced accuracies are given for the first and second model of each comparison, and Δ is their paired difference with a 95% patient-level bootstrap CI.

4.9 Channel reliance and calibration (RQ3)

Removing PanCK was the only channel intervention with a large effect (Table 16): over three seeds, balanced accuracy fell by 19.0pp (CI −29.6 to −8.7) and AUC by 0.184 (−0.310 to −0.062), to 75.8% and 0.805. Removing CD3, CD8 or FoxP3 changed balanced accuracy by −2.0 to +1.7pp, with intervals containing zero, and removing the DAPI intensities slightly improved it (+2.3pp, +0.8 to +4.0). Without the DAPI-derived region mask, removing DAPI improved the full model further (+5.3pp, +2.5 to +8.5), so the reliance on PanCK does not depend on the mask, and DAPI is not needed with or without it; without the mask, the full model was also 4.4pp below the pathology-only model (−9.0 to −0.8), the only setting in which the CT context was detectably harmful. Consistent with these results, the mean PanCK intensity alone separates tumour core from stroma with an AUC of 0.998 (Table 5). Because PanCK marks epithelial cells, this reliance is biologically expected; it also means that the main task is close to a single-marker intensity decision. The out-of-fold ECE of the four networks in Table 7 was 0.048–0.062 (Brier 0.046–0.058), but the predicted probabilities were saturated: 502 of the full model's 516 pooled predictions had a confidence of at least 0.95, whereas its accuracy was 0.94. The networks are therefore moderately overconfident, and no post hoc calibration was applied.

Figure 7a shows the out-of-fold ROC curves, pooled over seeds, of the full model, the pathology-only model and the channel-statistics baseline. The AUCs in its legend are computed on the predictions pooled over the three seeds (0.987 for the full model), whereas Table 7 reports the mean of the per-seed AUCs (0.989). Figure 7b shows the reliability

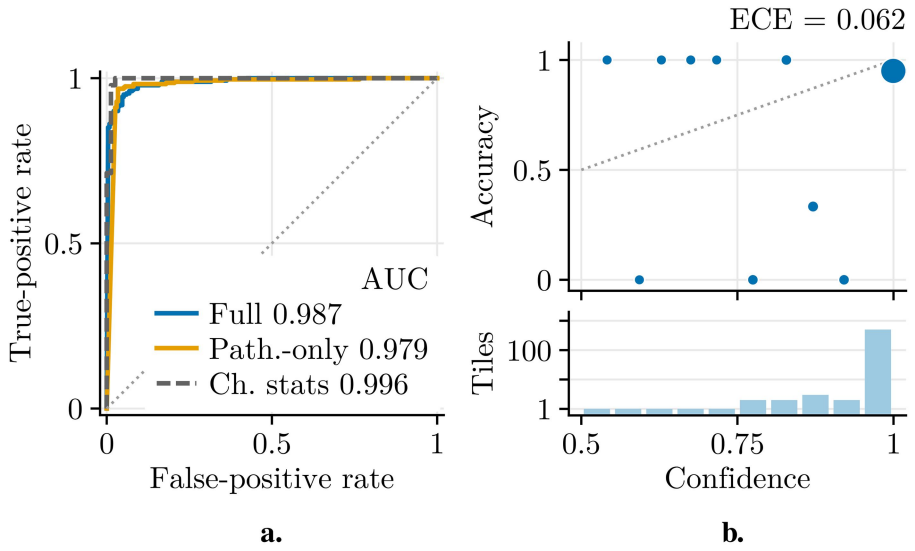

FIGURE 7
ROC curves and reliability diagram. **(a)** ROC curves **(b)** Reliability diagram.

diagram of the full model with the number of predictions in each confidence bin.

5 Discussion

5.1 Feasibility is not effectiveness

The results support RQ1: the pipeline trains on every fold and reaches an out-of-fold AUC of 0.989. They do not support a benefit of the CT context (RQ2), either on the main task or where there is room for improvement: without PanCK, on tumour core vs. margin and with 2 to 16 labelled tiles per class. They also do not support a meaningful CT–pathology association (RQ3), while the positive controls show that the same architecture uses an informative context when one is supplied. The accuracy of the complete pipeline comes from the pathology branch, and within that branch mostly from the PanCK channel, as the channel-drop experiment and the channel-statistics baseline show. Because no published benchmark exists for this dataset pair, the relevant reference points are the matched pathology-only model and the channel-statistics baseline rather than results reported on other data.

5.2 Possible reasons the CT context is not used

The positive controls narrow down the possible explanations: the fusion pathway, the gate and the classifier can propagate label information from an auxiliary context, so the null result is not caused by an architecture that ignores context in general. Three observations point to the content of the CT context instead. First, the context offers nothing stable. Each bank entry summarizes an entire lung CT volume, and without paired data there is no mechanism that would make a particular volume informative about a particular pathology tile. Retrieval appears class-dependent but noise banks show the same dependence, so it derives from the pathology features rather than from the CT content, and the retrieved set changes almost completely with the seed. Second, the architecture allows the context to be neutralized. The gate floor and usage penalty force the fused feature to differ from the raw pathology feature, but they do not force the difference to carry information; removing them did not change performance. The dominance of the prototype-distance logits is not the explanation, since the classifier-only ablation behaved the same. Third, the target task leaves little room for improvement. A near-deterministic single-marker cue is available, which is a setting in which any auxiliary signal can be expected to be redundant (40).

5.3 Information a CT encoder could transfer

In this setting, CT can at best supply modality-agnostic, low-level regularities of texture, heterogeneity and structure learned from volumetric data. It cannot supply patient- or lesion-specific information, because the domains are unpaired. The SSL diagnostics show that the encoder does learn a non-trivial, instance-discriminative representation of the CT volumes (view-retrieval accuracy 0.97–1.00 for the main encoders, no collapse). Yet an encoder with random weights served the pathology task equally well. This indicates that whatever structure the encoder learned is either not relevant to tumour-vs.-stroma discrimination in multiplex immunofluorescence, or not accessible through whole-volume descriptors and cosine retrieval. A contribution from the CT pathway would plausibly require finer-grained (slice- or patch-level) tokens, paired radiology–pathology data, or a target task that the pathology signal alone does not solve.

5.4 Implications for retrieval-augmented cross-modal methods

The fusion module in this study showed the usual signs of activity: open gates, moderately non-uniform attention and class-dependent retrieval. Yet the prediction was nearly invariant to the retrieved content. Attention maps, t-SNE plots, gate values and class-dependent retrieval are therefore not enough to claim transfer. We recommend that retrieval-augmented and unpaired cross-modal methods report at least a matched no-context comparator, a noise or constant bank, and a test-time bank swap, complemented by a positive control that shows the architecture can use informative context. These controls are inexpensive: in our case the test-time swap required no retraining, and each retrained control took 10–11 min per seed on a shared T4.

5.5 Limitations

(i) **Small cohort:** The target data come from eight patients. The confidence intervals are correspondingly wide (e.g., roughly $\pm$4pp for balanced-accuracy differences), so a small benefit or harm of the CT context cannot be excluded; as noted in Section 3.17, an interval containing zero indicates no detectable difference, not equivalence.

(ii) **Labels and task difficulty:** Labels are the pathologist-selected region types, not tile-level annotations. Region types are coarse, and some tiles may contain mixed tissue. PanCK nearly separates tumour core from stroma; the harder settings (PanCK removed, tumour core vs. margin, small annotation budgets) leave more room for an auxiliary signal, but they use the same eight patients and the same region-level labels. Other targets, such as immune phenotypes, were not tested because they would need labels that the collection does not provide.

(iii) **Single source–target pair and design choices:** We tested one CT collection, one pathology collection, whole-volume descriptors, one retrieval scheme and one fusion architecture. The negative result applies to this configuration and does not rule out cross-modal transfer with paired data, finer-grained tokens or other architectures.

(iv) **Protocol:** Validation for checkpoint selection used a single patient per fold, and the selected epochs were early (mean about 5); the selection variants show that this choice can move the full-vs.-pathology-only difference by a few percentage points in either direction. The SSL sweeps used one seed. The PanCK-context positive control derives its context from the same image rather than from a separate modality, so it shows what the architecture can use, not what an independent auxiliary source would provide. GPU kernels were not bit-deterministic. The hand-crafted baseline used untuned default regularization. No external validation cohort was available.

5.6 Outlook

The most informative next step is a setting in which a CT signal could plausibly help: paired radiology–pathology data from the same patients, or a target task that pathology intensities alone do not solve. Slice- or patch-level CT tokens, stronger pretrained volumetric encoders, and fusion designs in which unused context is penalized rather than neutralized (e.g., by predicting pathology from context alone) should be tested with the control protocol introduced here. External validation on independent HNSCC cohorts and expert tile-level annotation would strengthen any positive claim. Finally, the networks are moderately overconfident; calibration-aware training objectives that combine temperature scaling with entropy regularization (49), or post hoc calibration fitted on separate patients, could be evaluated within the same protocol.

6 Conclusion

Integrating retrieved context from an unpaired 3D CT bank into a few-shot multiplex-pathology classifier is technically straightforward and yields a well-performing model, but in this setting the performance comes from the pathology branch: matched pathology-only training, bank replacement and test-time bank swapping all leave predictions essentially unchanged, including when PanCK is removed, on tumour core vs. margin and under small annotation budgets. The same architecture does use an informative context when one is supplied, so the CT context contributes no detectable information here because of its content, and signs of fusion activity are not evidence of transfer. We recommend that retrieval-augmented and unpaired cross-modal methods report matched no-context comparators, bank-replacement controls and test-time bank swaps; the released code runs all three.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

This study is a secondary analysis of two public, de-identified image collections distributed by The Cancer Imaging Archive (TCIA): LungCT-Diagnosis (human thoracic CT) and HNSCC-mIF-mIHC-comparison (human head and neck cancer tissue). The data were collected by the contributing institutions under their own approvals and are released by TCIA for research use under its data-use terms, without identifiable information. The author had no access to identifiable data and no interaction with participants; under these conditions, secondary research use of public de-identified data does not require additional ethical approval or informed consent.

Author contributions

ZA: Conceptualization, Data curation, Writing – original draft. HR: Conceptualization, Formal analysis, Writing – original

draft. MJ: Data curation, Methodology, Writing – original draft. RS: Funding acquisition, Investigation, Writing – original draft. AB: Investigation, Software, Writing – original draft. ID: Project administration, Visualization, Writing – original draft. IA: Supervision, Validation, Writing – review & editing.

Funding

The author(s) declared financial support was received for this work and/or its publication. This study is funded by the European University of Atlantic.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Nayem J, Hasan SS, Amina N, Das B, Ali MS, Ahsan MM, et al. Few shot learning for medical imaging: a comparative analysis of methodologies and formal mathematical framework. *arXiv* [Preprint]. *arXiv:2305.04401* (2023).

2. Pachetti E, Colantonio S. A systematic review of few-shot learning in medical imaging. *arXiv* [Preprint]. *arXiv:2309.11433* (2023).

3. Guan H, Liu M. Domain adaptation for medical image analysis: a survey. *IEEE Trans Biomed Eng.* (2022) 69:1173–85. doi: 10.1109/TBME.2021.3117407

4. Li Y, Fan Y. Medical image segmentation with domain adaptation: a survey. *arXiv* [Preprint]. *arXiv:2311.01702* (2023).

5. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In *Advances in neural information processing systems* (2017), 30, 4077–87.

6. Zhao L, Chen X, Chen EZ, Liu Y, Chen T, Sun S. Retrieval-augmented few-shot medical image segmentation with foundation models. *arXiv* [Preprint]. *arXiv:2408.08813* (2024).

7. Shakeri F, Boudiaf M, Mohammadi S, Sheth I, Havaei M, Ben Ayed I, et al. FHIST: a benchmark for few-shot classification of histological images. *arXiv* [Preprint]. *arXiv:2206.00092* (2022).

8. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. *J Mach Learn Res.* (2016) 17:1–35.

9. Jaritz M, Vu T-H, de Charette R, Wirbel É, Pérez P. xMUDA: cross-modal unsupervised domain adaptation for 3D semantic segmentation. *arXiv* [Preprint]. *arXiv:1911.12676* (2020).

10. Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, et al. Advances in medical image analysis with vision transformers: a comprehensive review. *arXiv* [Preprint]. *arXiv:2301.03505* (2023).

11. He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, et al. Transformers in medical image analysis: a review. *arXiv* [Preprint]. *arXiv:2202.12165* (2022).

12. Paschali M, Chen Z, Blankemeier L, Varma M, Youssef A, Bluethgen C, et al. Foundation models in radiology: what, how, when, why and why not *arXiv* [Preprint]. *arXiv:2411.18730* (2024).

13. Wang W-C, Ahn E, Feng D, Kim J. A review of predictive and contrastive self-supervised learning for medical images. *arXiv* [Preprint]. *arXiv:2302.05043* (2023).

14. Han Q, Liu J, Qin Z, Zheng Z. Integrating MedCLIP and cross-modal fusion for automatic radiology report generation. *arXiv* [Preprint]. *arXiv:2412.07141* (2024).

15. Liu X, Yang G, Luo Y, Mao J, Zhang X, Gao M, et al. Expert-level vision-language foundation model for real-world radiology and comprehensive evaluation. *arXiv* [Preprint]. *arXiv:2409.16183* (2024).

16. Tao Y, Ma L, Yu J, Zhang H. Memory-based cross-modal semantic alignment network for radiology report generation. *arXiv* [Preprint]. *arXiv:2404.00588* (2024).

17. Wang J, Zhu L, Yu X, Bhalerao A, He Y. Improving medical visual representation learning with pathological-level cross-modal alignment and correlation exploration. *arXiv* [Preprint]. *arXiv:2506.10573* (2025).

18. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. *arXiv* [Preprint]. *arXiv:2210.10163* (2022).

19. Zambrano Chaves JM, Huang S-C, Xu Y, Xu H, Usuyama N, Zhang S, et al. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. *arXiv* [Preprint]. *arXiv:2403.08002* (2024).

20. Jennings C, Broad A, Godson L, Clarke E, Westhead D, Treanor D. Machine learning-based multimodal prognostic models integrating pathology images and high-throughput omic data for overall survival prediction in cancer: a systematic review. *arXiv* [Preprint]. *arXiv:2507.16876* (2025).

21. Li D, Wan G, Wu X, Wu X, Chen X, He Y, et al. Multi-modal foundation models for computational pathology: a survey. *arXiv* [Preprint]. *arXiv:2503.09091* (2025).

22. Tian Y, Xu Z, Ma Y, Ding W, Wang R, Gao Z, et al. Survey on deep learning in multimodal medical imaging for cancer detection. *arXiv* [Preprint]. *arXiv:2312.01573* (2023).

23. Chen Y, Huang Y, Wang F, Yeung M, Jiang Y, Wang S, et al. Bridging radiology and pathology foundation models via concept-based multimodal co-adaptation. In: *International conference on learning representations* (2026).

24. Pan W, Kang J, Wang X, Lv C, Zhang X, Liu X. MedCTM: a CNN-transformer-mamba hybrid network for medical image classification. *Inf Process Manag.* (2026) 63:104867. doi: 10.1016/j.ipm.2026.104867

25. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th international conference on machine learning, Proceedings of Machine Learning Research* (2020), 119, 1597–607.

26. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. *arXiv* [Preprint]. *arXiv:1807.03748* (2018).

27. El Nahhas OSM, Wölflein G, Ligero M, Lenz T, van Treeck M, Khader F, et al. Joint multi-task learning improves weakly-supervised biomarker prediction in computational pathology. *arXiv* [Preprint]. *arXiv:2403.03891* (2024).

28. Lenz T, El Nahhas OSM, Ligero M, Kather JN. Reducing self-supervised learning complexity improves weakly-supervised classification performance in computational pathology. *arXiv* [Preprint]. *arXiv:2403.04558* (2024).

29. Li D, Wan G, Wu X, Wu X, Nirmal AJ, Lian CG, et al. A survey on computational pathology foundation models: datasets, adaptation strategies, and evaluation tasks. *arXiv* [Preprint]. *arXiv:2501.15724* (2025).

30. Neidlinger P, El Nahhas OSM, Muti HS, Lenz T, Hoffmeister M, Brenner H, et al. Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. *arXiv* [Preprint]. *arXiv:2408.15823* (2024).

31. Nielsen M, Wenderoth L, Sentker T, Werner R. Self-supervision for medical image classification: state-of-the-art performance with about 100 labeled training samples per class. *arXiv* [Preprint]. *arXiv:2304.05163* (2023).

32. Sharma A, Liu B, Rantalainen M. WEEP: a method for spatial interpretation of weakly supervised CNN models in computational pathology. *arXiv* [Preprint]. *arXiv:2403.15238* (2024).

33. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. *arXiv* [Preprint]. *arXiv:2201.01266* (2022).

34. Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, et al. nnU-Net revisited: a call for rigorous validation in 3D medical image segmentation. *arXiv* [Preprint]. *arXiv:2404.09556* (2024).

35. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *arXiv* [Preprint]. *arXiv:2304.12306* (2023).

36. Roy S, Wald T, Koehler G, Rokuss MR, Disch N, Holzschuh J, et al. SAM.MD: zero-shot medical image segmentation capabilities of the segment anything model. *arXiv* [Preprint]. *arXiv:2304.05396* (2023).

37. Shi P, Qiu J, Abaxi SMD, Wei H, Lo FP-W, Yuan W. Generalist vision foundation models for medical imaging: a case study of segment anything model on zero-shot medical segmentation. *arXiv* [Preprint]. *arXiv:2304.12637* (2023).

38. Wei X, Cao J, Jin Y, Lu M, Wang G, Zhang S. I-MedSAM: implicit medical image segmentation with segment anything. *arXiv* [Preprint]. *arXiv:2311.17081* (2023).

39. Wu J, Ji W, Liu Y, Fu H, Xu M, Xu Y, et al. Medical SAM adapter: adapting segment anything model for medical image segmentation. *arXiv* [Preprint]. *arXiv:2304.12620* (2023).

40. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. *Nat Mach Intell.* (2020) 2:665–73. doi: 10.1038/s42256-020-00257-z

41. Dhariwal N, Giridharan A. Towards precision oncology: predicting mortality and relapse-free survival in head and neck cancer using clinical data. *arXiv* [Preprint]. *arXiv:2502.11200* (2025).

42. Wang T, Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: *Proceedings of the 37th international conference on machine learning* (2020), *Proceedings of machine learning research*, 119, 9929–39.

43. Ghahremani P, Marino J, Hernandez-Prera J, de la Iglesia JV, Slebos RJC, Chung CH, et al. An AI-ready multiplex staining dataset for reproducible and accurate characterization of tumor immune microenvironment. In: *Medical image computing and computer assisted intervention—MICCAI 2023* (Springer) (2023), *Lecture notes in computer science*, 14225, 704–13. doi: 10.1007/978-3-031-43987-2_68.

44. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *Proceedings of the 34th international conference on machine learning* (2017), *Proceedings of machine learning research*, 70, 1321–30.

45. Field CA, Welsh AH. Bootstrapping clustered data. *J R Stat Soc Series B Stat Methodol.* (2007) 69:369–90. doi: 10.1111/j.1467-9868.2007.00593.x

46. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. *J Mach Learn Res.* (2012) 13:723–73.

47. Kornblith S, Norouzi M, Lee H, Hinton G. Similarity of neural network representations revisited. In: *Proceedings of the 36th international conference on machine learning* (2019), *Proceedings of machine learning research*, 97, 3519–29.

48. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res.* (2008) 9:2579–605.

49. Pan W. TET loss: a temperature-entropy calibrated transfer loss for reliable medical image classification. *J Imaging Inform Med.* (2026):1–13. Online ahead of print. doi: 10.1007/s10278-025-01816-9