Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TopoNet: A Topology-Guided Dual-Branch Framework for Medical Image Classification

Jin Sun
Shanghai University
Nan Gao
nangao@shu.edu.cn
Shanghai University
Wenqi Tao
Renji Hospital, Shanghai Jiao Tong University School of Medicine
Zaifang Zhang
Shanghai University
Chenghao Shao
Shanghai Changzheng Hospital

Research Article

Keywords: Volumetric medical image classification, Persistent homology, Confidence bands, Topological structures, Multimodal fusion

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9928087/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TopoNet: A Topology-Guided Dual-Branch Framework for Medical Image Classification

Jin Sun[1], Nan Gao[1,5*], Wenqi Tao[2†], Zaifang Zhang[3], Chenghao Shao[4]

[1*]School of Science, Shanghai University, 200444, Shanghai, China.
[2]Renji Hospital, Shanghai Jiao Tong University School of Medicine, 200127, Shanghai, China.
[3]School of Mechatronic Engineering and Automation, Shanghai University, 200444, Shanghai, China.
[4]Department of Pancreatic-biliary Surgery, Shanghai Changzheng Hospital, 200003, Shanghai, China.
[5]New touch Center for Mathematics of Shanghai University, 200444, Shanghai, China.

*Corresponding author(s). E-mail(s): nangao@shu.edu.cn;
Contributing authors: sgolds@shu.edu.cn; tao_wenqi@163.com;
zaifangzhang@shu.edu.cn; shaochenghao_czyy@163.com;
†These authors contributed equally to this work and are co-corresponding authors.

Abstract

Volumetric medical image classification remains challenging due to limited robustness and interpretability, largely because modern models based on CNNs and Transformers primarily rely on appearance-driven cues and struggle to explicitly capture global anatomical topology, such as connectivity, loops, and cavities. To address this limitation, we propose TopoNet, a topology-guided dual-branch framework that integrates multiscale topological structures into volumetric representation learning. Specifically, TopoNet computes cubical persistent homology to obtain persistence diagrams and distills robust topological signals from background noise via statistical confidence bands, enabling adaptive and noise-robust filtering without relying on rigid thresholds. Crucially, to overcome the spatial information loss inherent in traditional vectorization methods, the selected persistence features are mapped back to the voxel domain.

This process restores spatial context to generate spatially registered topological structures, forming an explicit topology modality aligned with the input images and providing spatially grounded structural cues that improve interpretability. The image and topology modalities are encoded in parallel and fused via stage-wise cross-branch attention to facilitate complementary feature interactions across scales. We validate TopoNet on two datasets, where it consistently outperforms strong baselines across key metrics, achieving $\mathbf{0.8197 \pm 0.0587}$ in Accuracy and $\mathbf{0.8123 \pm 0.1127}$ in AUC on the MSD Task07 Pancreas dataset, and $\mathbf{0.9333 \pm 0.0079}$ in Accuracy and $\mathbf{0.9267 \pm 0.0437}$ in AUC on the BraTS 2018 dataset. Overall, these results indicate that topology-guided multimodal fusion provides a promising avenue for improving performance, robustness, and clinical interpretability in volumetric medical image classification. The code is available at https://github.com/hbqdcq/TopoNet.

Keywords: Volumetric medical image classification, Persistent homology, Confidence bands, Topological structures, Multimodal fusion

1 Introduction

Medical image classification is central to clinical decision-making, as the accurate interpretation of radiological and tomographic data is essential for early disease diagnosis, prognostic evaluation, and the optimization of therapeutic interventions [1]. Despite rapid progress, achieving robust and generalizable performance remains challenging, especially for volumetric medical imaging, due to intrinsic heterogeneity, high dimensionality, and noise. These factors often limit the reliability of purely appearance-driven learning pipelines and can lead to inconsistent performance across patients and acquisition settings.

In response, several studies have successfully employed Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in various medical imaging classification tasks. Studies such as [2, 3] have utilized CNN-based architectures to enhance diagnostic performance across diverse modalities, demonstrating the effectiveness of convolutional feature extraction in clinical contexts. Meanwhile, ViT-based frameworks have gained increasing attention, with works like [4, 5] highlighting their potential for disease classification, particularly in neurological and oncological domains. To further improve performance and generalizability, several studies [6–8] have proposed hybrid architectures that integrate CNNs with Transformers, aiming to combine the complementary strengths of local convolutional patterns and global attention mechanisms. However, these pixel-level architectures primarily model texture and contextual cues and often provide insufficient representation of intrinsic anatomical topology in medical images. Unlike natural images, medical images frequently exhibit prominent anatomical topology. For example, in the diagnosis of pancreatic ductal adenocarcinoma (PDAC), healthy pancreatic ducts typically appear as continuous tubular structures. In contrast, PDAC is frequently associated with abrupt narrowing, fragmentation, or abnormal dilation of the duct, resulting in substantial topology-related changes, including altered connectivity and loop-like patterns [9]. Conventional

architectures inherently lack the mathematical guarantees to distinguish between a fragmented duct and a continuous one based purely on local appearance, thereby limiting both their diagnostic accuracy and interpretability. Therefore, incorporating topological analysis into classification frameworks offers a promising direction. It provides complementary, clinically aligned structural information and improves both predictive robustness and interpretability.

Topological data analysis (TDA) is a collection of mathematical and computational methodologies designed to construct and extract topological features embedded within complex data [10, 11]. Among these techniques, persistent homology has emerged as a foundational and widely adopted approach [12], offering a robust and theoretically grounded framework for capturing structural features across multiple spatial scales, from local to global. Briefly, persistent homology tracks the evolution of topological features across multi-scale spatial resolutions. Instead of relying on fixed geometric shapes, it characterizes data through fundamental structural invariants: connected components (H_0), loops (H_1), and voids (H_2). In medical imaging, H_0 features may correspond to discrete lesion candidates or tissue components, H_1 features can reflect ring-like or boundary-related patterns, and H_2 features often indicate internal cavities or disrupted tissue organization. Fig. 1 offers an intuitive visual summary of PDAC-relevant topological patterns captured by persistent homology in pancreatic imaging.

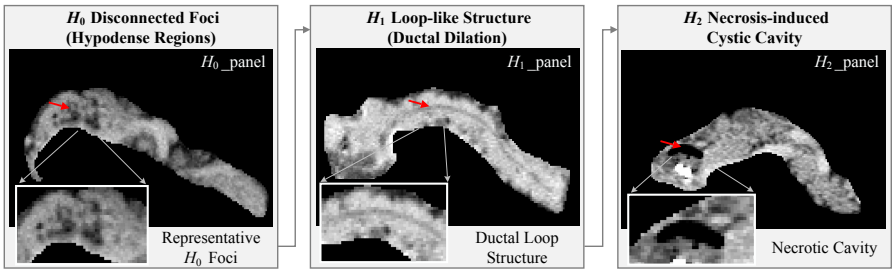


Fig. 1 The topological structure in PDAC. From left to right, red arrows highlight representative topology-related patterns within the pancreatic region: H_0 disconnected low-attenuation foci corresponding to multiple isolated hypodense regions, H_1 loop-like ductal/boundary structures associated with ductal dilation, and H_2 a necrosis-induced cystic cavity forming an enclosed void.

Persistent homology has demonstrated notable success across a wide range of biomedical applications, particularly in the analysis of medical imaging data. Representative tasks include predicting pathological complete response [13, 14], lesion detection and classification [15–19], biomedical network analysis [20, 21], and image segmentation [22, 23]. Despite these achievements, a critical challenge in applying persistent homology to medical imaging lies in the presence of substantial topological noise (e.g., spurious loops arising from intensity inhomogeneities or scanning artifacts) [24], which can obscure clinically meaningful structures. The extraction of significant topological features under such conditions is inherently nontrivial. To address this issue, most existing studies [16, 19] employ threshold-based filtering strategies that retain only a subset of features based on persistence values. However, such approaches

can be sensitive to case-to-case variability and acquisition differences, risking discarding subtle but clinically relevant topological patterns and reducing robustness under distribution shifts. Furthermore, a fundamental limitation of current approaches is the "spatial detachment" problem. Many studies [13, 17] integrate topological features by vectorizing them (e.g., persistence images) and concatenating them with neural network features. This process inherently discards the spatial coordinates of topological features. Consequently, the network learns global structural complexity but loses the ability to localize where the specific lesion or ductal disruption occurs. This lack of explicit interaction between topological descriptors and the hierarchical spatial representations hinders the exploitation of their complementarity. As a result, the model's ability to make comprehensive, interpretable, and structure-aware decisions in clinical scenarios is markedly constrained.

To address these limitations, we propose TopoNet, a topology-guided dual-branch framework for volumetric medical image classification. Unlike previous works that treat topology as a global descriptor, we introduce a mechanism to restore topological features to their spatial origins. Specifically, we compute cubical persistent homology and derives persistence diagrams on cubical complexes. It then selects statistically significant persistence pairs via confidence bands, enabling adaptive and robust topology selection under substantial topological noise. We then perform a dimension-aware topology-to-voxel mapping that leverages cubical persistence to localize the selected persistence pairs and rasterize $H_0/H_1/H_2$ structures into spatially registered binary topology masks, yielding an explicit topology modality complementary to the raw image. Finally, the image and topology modalities are encoded in parallel and coupled through an attention-based cross-branch fusion mechanism. Topology-derived structural context is then injected into hierarchical features in a stage-wise manner, which stabilizes cross-modal learning, mitigates representational mismatch, and enables the joint capture of fine-grained appearance cues and global topological features.

In summary, our work has four main contributions:

(1) We propose a statistically grounded topology extraction framework that identifies significant persistent homology features on cubical complexes via confidence bands, effectively separating robust clinical signals from background topological noise without relying on rigid thresholds.
(2) We bridge the gap between abstract topological summaries and grid-based deep learning by developing a dimension-aware topology-to-voxel mapping pipeline. This extracts topological structures across multiple homology dimensions ($H_0/H_1/H_2$) and converts them into spatially aligned, voxel-wise topology representations.
(3) We design TopoNet, a topology-guided dual-branch architecture that encodes the original image and the topology modality in parallel, facilitating explicit cross-modal interactions to guide the network's attention toward structurally significant regions.
(4) We conduct extensive experiments on two datasets, showing consistent improvements over strong baselines. In particular, TopoNet achieves highly competitive performance across key metrics on both datasets, demonstrating the effectiveness and generalizability of the proposed topology-guided learning paradigm.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method. Section 4 reports the experimental results and performance evaluation on public datasets. Section 5 concludes the paper.

2 Related work

2.1 Deep learning methods

Recent advances in deep learning, particularly those based on CNNs and ViTs, have significantly improved performance in medical image classification tasks. State-of-the-art approaches have evolved from pure convolutional networks to hybrid architectures that aim to balance local texture extraction with global contextual reasoning.

In the realm of CNN-based approaches, researchers initially focused on enhancing feature hierarchies through transfer learning and multi-task frameworks. Maqsood et al. [2] and Akter et al. [3] demonstrated that integrating auxiliary tasks, such as segmentation, can significantly boost classification accuracy by enforcing shape awareness within the feature encoders. However, to overcome the intrinsic locality bias of convolutions, recent attention has shifted toward ViTs and hybrid designs. For instance, Huang and Qiu [4] developed MC-ViT to leverage stochastic sampling for global context, while hybrid architectures such as DSMT-Net [5], Eff-CTM [6], CTransCNN [7], and StoHisNet [8] were proposed to bridge the gap between CNNs and Transformers. By integrating attention mechanisms with convolutional layers, these methods mitigate the trade-off between local detail preservation and long-range dependency modeling, demonstrating strong generalization across datasets ranging from gastric histopathology to endoscopic ultrasound.

Despite these advances, a major limitation persists: the features extracted by CNNs and ViTs are primarily sensitive to local intensity variations and texture-level descriptors. Consequently, they often fail to capture essential global and topological anatomical structures, such as connected components, loops, and cavities, that are critical for accurate clinical interpretation. This limitation can hinder model performance, particularly in tasks requiring structural reasoning and topological awareness.

2.2 Topological deep learning methods

Persistent homology, a method for identifying topological features across multiple spatial scales, has achieved remarkable success in various medical analysis tasks. Early works, such as [16] and [18], pioneered the use of handcrafted topological descriptors. Asaad et al. [16] utilized barcode vectorizations for mammogram analysis, while Stolz et al. [18] introduced topological lenses to quantify vascular connectivity. While foundational, these methods rely heavily on manual feature engineering and are often sensitive to parameter selection, limiting their end-to-end learnability.

To address these constraints, recent efforts have sought to integrate TDA directly into deep learning frameworks, broadly categorizing into regularization-based and feature-fusion approaches. For regularization, Du et al. [14] leveraged knowledge distillation to transfer topological knowledge into neural networks, improving response

prediction accuracy. For feature fusion, Peng et al. [19] and Hu et al. [25] investigated hybrid streams where topological features are computed and fused with visual features. Similarly, Wang et al. [13] introduced TopoTxR to guide attention mechanisms using topological biomarkers. However, two critical bottlenecks remain in these state-of-the-art integration methods: (1) Sensitivity to Noise: Significant features are often selected via rigid thresholding, which may discard subtle but clinically relevant structures in noisy medical scans. (2) The "Spatial Detachment" Problem: Most existing methods [17, 25] inherently vectorize topological features (e.g., into Betti curves or persistence images) to form global descriptors. This vectorization process collapses the spatial dimensions, effectively divorcing topological information from its original coordinates. Consequently, while the network learns global structural complexity, it loses the ability to localize where the specific topological anomaly (e.g., a broken duct) occurred, leading to a representational mismatch between spatially-aware deep features and spatially-agnostic topological vectors. To address these issues, we propose a unified framework that restores spatial context to topological features via a dual-branch architecture coupled with an attention-based fusion mechanism.

3 Method

In this paper, we propose a topology-guided framework for medical image classification. The key idea is to incorporate persistent homology as an explicit structural prior into deep representation learning, and to enhance the guiding effect of topology features on image feature extraction through attention mechanisms.

For clarity, the remainder of this section is organized as follows. Section 3.1 briefly reviews the fundamentals of persistent homology and its construction on volumetric data. Section 3.2 presents a confidence band-based strategy for selecting significant topological features. This strategy adaptively retains statistically salient persistence pairs from the persistence diagram, resulting in more robust and reliable topological structures. Section 3.3 describes the overall design of the proposed topology-guided dual-branch network, TopoNet. It details how the two modalities are encoded in parallel and how cross-branch interactions are implemented. In particular, Section 3.3.1 introduces the procedure for mapping the selected persistence pairs back to voxel space to construct topology-derived representations that are strictly spatially aligned with the input images. Building on these components, Section 3.3.2 provides a systematic description of the complete TopoNet architecture and its end-to-end training paradigm.

3.1 Persistent homology

In this section, we present a brief introduction to the fundamentals of persistent homology within the context of algebraic topology. For a more detailed and rigorous exposition, the readers are referred to [11]. Persistent homology provides a robust, multi-scale framework for identifying and quantifying topological structures that remain stable across varying spatial scales, thereby offering meaningful insights into the intrinsic organization of complex data.

Let $f : \Omega \to \mathbb{R}$ denote a scalar field defined on an image domain Ω (e.g., voxel intensity or density). For any threshold $t \in \mathbb{R}$, the associated sublevel set is defined as:

$$S_t = \{x \in \Omega \mid f(x) \leq t\} \tag{1}$$

A filtration is a nested sequence of sublevel sets:

$$S_{t_0} \subseteq S_{t_1} \subseteq \cdots \subseteq S_{t_n} \tag{2}$$

where $t_0 < t_1 < \cdots < t_n$. For $i \leq j$, the inclusion $S_{t_i} \to S_{t_j}$ induces homomorphisms $H_k(S_{t_i}) \to H_k(S_{t_j})$. The resulting persistence module summarizes how k-dimensional topological features are created and destroyed across the filtration. For each threshold t_i, topological features can be characterized by the homology groups $\{H_k(S_{t_i})\}$, which encode k-dimensional cycles at the corresponding scale. In particular, H_0 summarizes connected components (connectivity), H_1 captures loop-like or tunnel-like structures (e.g., the interior of a ring or the handle of a mug), and H_2 represents enclosed voids or cavities (e.g., the hollow interior of a sphere). In this work, we focus on $k \in \{0, 1, 2\}$, which captures the most interpretable structures in 3D medical volumes; higher-dimensional homology is rarely used in our setting. For volumetric images, these sublevel sets can be instantiated on a cubical complex defined over the voxel grid, so that S_t corresponds to a subcomplex that is spatially aligned with the original volume.

Persistent homology further tracks how these features evolve along the filtration. A topological feature is said to be **born** at threshold t_i when it first appears, and **dies** at threshold t_j when it disappears (or merges into an older feature). This event is recorded as a persistence pair (t_i, t_j), which corresponds to a point (b, d) in the persistence diagram with $b = t_i$ on the x-axis and $d = t_j$ on the y-axis. The quantity $d - b$ measures the persistence of the feature: features with larger persistence are more likely to reflect stable and salient structures, whereas short-lived features may arise from minor fluctuations or noise. In subsequent sections, we introduce principled criteria to select statistically meaningful persistence pairs for our task.

Fig. 2 summarizes the H_1 persistent homology of the synthetic image f. As the threshold increases, each loop is summarized by a birth value b and a death value d, which form a persistence pair (b, d). Its lifetime is quantified by $d - b$, and more persistent loops correspond to more prominent structures.

3.2 Extraction of statistically significant topological features guided by confidence bands

In medical image analysis, a central challenge associated with persistent homology is effectively distinguishing between meaningful topological features related to pathological or anatomical structures and irrelevant noise within the persistence diagram, thereby improving diagnostic performance. Most existing studies address this issue via manually tuned thresholds to filter out low-persistence features. However, such strategies have several notable limitations. First, manual threshold selection is inherently subjective and heavily reliant on intuition, which hinders full automation, reduces

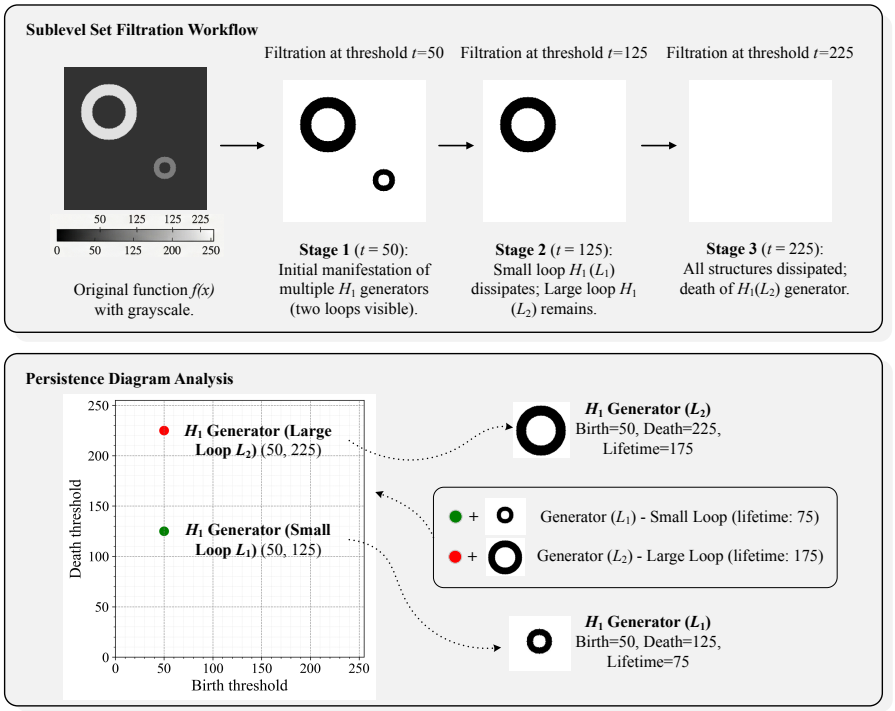


Fig. 2 Illustration of one-dimensional persistent homology (H_1) under a sublevel set filtration. **Top**: The sublevel set filtration workflow for an input image f containing a dark background of intensity 50 and two bright annular boundaries with intensities 125 (small loop L_1) and 225 (large loop L_2). We construct sublevel sets $S_t = \{x \mid f(x) \leq t\}$ at thresholds $t = 50, 125, 225$. At $t = 50$, the background enters first while the bright annuli are still excluded, yielding two H_1 generators (loops) associated with the two rings. As t increases to 125, the small annulus enters S_t and the corresponding H_1 cycle (L_1) is filled in by 2-cells and thus dies; at $t = 225$, the large cycle (L_2) dies analogously. **Bottom**: The corresponding persistence diagram analysis, where the points $(50, 125)$ and $(50, 225)$ represent the birth-death pairs of the small and large loops, respectively.

cross-study consistency, and can undermine the stability and reproducibility of the resulting representations. Second, fixed thresholding is often insensitive to the intrinsic topology of the underlying data distribution, increasing the likelihood of discarding subtle yet clinically relevant anatomical or pathological signals. Consequently, diagnostic sensitivity and overall accuracy may be compromised. In this section, we first present the confidence-band framework in a general point-cloud setting. We then apply the corresponding stability guarantees to cubical complexes for volumetric images.

In reference [26], the authors propose an effective method for distinguishing topological noise from genuine topological signal. For completeness, let X and Y be two compact subsets of a Euclidean space, and let η denote a filtration operator (e.g., Rips or Alpha complex) that maps a geometric object to its persistence diagram, denoted by $Dgm_k(\eta(\cdot))$. Here Dgm_k denotes the persistence diagram in homology dimension k. The stability theorem for persistent homology guarantees that, for any two compact sets:

$$d_B(Dgm_k(\eta(X)), Dgm_k(\eta(Y))) \leq C d_H(X, Y) \tag{3}$$

where $d_B(\cdot,\cdot)$ is the bottleneck distance between persistence diagrams, $d_H(\cdot,\cdot)$ is the Hausdorff distance between sets, and C is a filtration-dependent stability constant (e.g., $C = 1$ for the Alpha complex filtration, $C = 2$ for the Rips filtration). Now consider an i.i.d. sample X_n of size n drawn from a distribution P supported on X. To construct a $(1-\alpha)$-level confidence band for the true persistence diagram $Dgm_k(\eta(X))$, one first establishes:

$$P(d_H(X, X_n) \leq c_\alpha) \geq 1 - \alpha \tag{4}$$

That is, with probability at least $1-\alpha$, the Hausdorff deviation between the empirical point cloud and the true support is bounded by c_α. Combining (4) with the stability bound in (3) yields:

$$P(d_B(Dgm_k(\eta(X)), Dgm_k(\eta(X_n))) \leq Cc_\alpha) \geq 1 - \alpha \tag{5}$$

We define the persistence diagram bandwidth as the bottleneck radius $\varepsilon_\alpha := Cc_\alpha$. The resulting $(1-\alpha)$-confidence set for the true persistence diagram is:

$$\ell_{1-\alpha} = \{Dgm \mid d_B(Dgm, Dgm_k(\eta(X_n))) \leq \varepsilon_\alpha\} \tag{6}$$

Consequently, $\ell_{1-\alpha}$ contains the true diagram $Dgm_k(\eta(X))$ with probability at least $1-\alpha$. Under the ℓ_∞ metric used in the bottleneck distance, matching a point (b, d) to the diagonal incurs cost $dist_\infty((b, d), \Delta) = (d - b)/2$. Therefore, a feature is regarded as statistically significant if it cannot be absorbed by the confidence band around the diagonal, i.e.,

$$d - b > 2\varepsilon_\alpha \tag{7}$$

Pairs with $d - b < 2\varepsilon_\alpha$ are treated as likely topological noise. In practice, c_α is estimated via a subsampling procedure and taken as an empirical quantile of Hausdorff deviations. Specifically, we repeatedly draw subsamples of size m and compute the one-sided Hausdorff deviation from the held-out set to each subsample. Let $\{D^{(b)}\}_{b=1}^{B}$ denote the resulting maxima over B subsamples; we set:

$$\hat{c}_\alpha = 2 \cdot Quantile_{1-\alpha}(\{D^{(b)}\}_{b=1}^{B}) \tag{8}$$

where the factor 2 follows the standard scaling in the subsampling-based bound. We then obtain $\hat{\varepsilon}_\alpha = C\hat{c}_\alpha$ and apply $d - b > 2\hat{\varepsilon}_\alpha$ to select significant persistence pairs. Overall, this framework provides a principled and statistically grounded mechanism for assigning confidence to topological features in persistence diagrams. Further theoretical details and stability proofs are provided in [26].

As illustrated in Fig. 3, we generated two noisy concentric-circle point clouds by sampling 1500 points from a circle of radius 0.4 and 3000 points from a circle of radius 0.8, with each point perturbed by i.i.d. uniform noise in $[-0.02, 0.02]$ added to both coordinates. We then applied a subsampling-based method to obtain a $(1 - \alpha) = 0.9$ upper bound on the Hausdorff distance between the empirical point cloud and its underlying true support, which in turn determines the persistence diagram bandwidth ε_α. Features that lie within the semi-transparent gray band of height $2\varepsilon_\alpha$ above the diagonal $y = x$ are treated as topological noise, since their persistence is too small to

rise above the uncertainty induced by sampling variability. In contrast, points outside this band, most notably the two dominant H_1 features corresponding to the inner and outer rings, are well separated from the diagonal and are therefore deemed statistically significant. This example demonstrates that the subsampling-derived confidence band effectively distinguishes genuine loop structures from spurious topological artifacts.

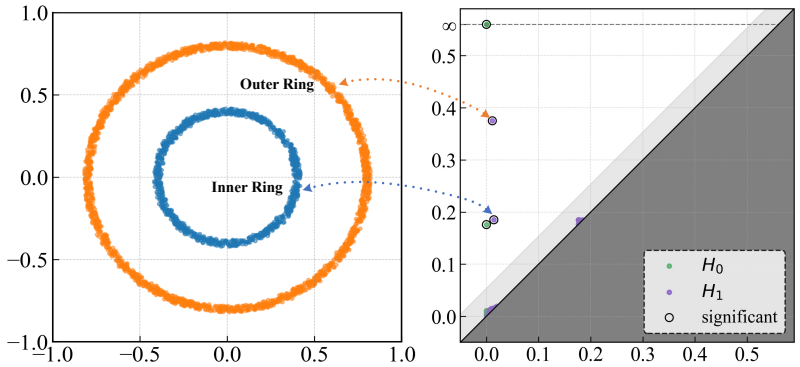


Fig. 3 Persistent homology on a noisy concentric-circle point cloud. **Left:** Two simulated noisy concentric rings, with the inner ring shown in blue and the outer ring shown in orange. **Right:** Persistence diagram for homology dimensions H_0 and H_1. The semi-transparent gray region of height $2\varepsilon_\alpha$ above the diagonal $y = x$ visualizes the $(1 - \alpha) = 0.9$ confidence band parameterized by ε_α. Features with persistence $d - b > 2\varepsilon_\alpha$ are highlighted by black-edged hollow circles to indicate statistical significance. The essential feature with infinite death time is shown at the top boundary as a gray dashed line for visualization.

Unlike Rips or Alpha complexes, which are constructed from point-cloud filtrations and admit confidence-band constructions based on Hausdorff perturbations, cubical complexes are induced directly by voxel-wise scalar fields (e.g., intensity or density functions). Consequently, confidence-band methods designed for simplicial complexes are not directly applicable in voxel-based medical imaging. To bridge this gap, we propose a strategy compatible with cubical complexes to quantify diagram-level uncertainty in persistence diagrams and to identify statistically significant topological features.

Let $F : \Omega \subset \mathbb{Z}^3 \to \mathbb{R}$ denote a voxel function defined on a 3D grid Ω. We simulate incomplete observations by subsampling the voxel field and reconstructing a complete surrogate function. Specifically, for $b = 1, ..., B$, we generate an identically distributed random binary mask $M^{(b)} \in \{0, 1\}^{|\Omega|}$ that retains each voxel with probability p (restricted to a region of interest to focus on anatomically relevant structures). This produces a partially observed volume $F^{(b)}_{obs}$, where a subset of voxel values is treated as missing. A deterministic reconstruction operator $\mathfrak{R}(\cdot)$, such as constant-value filling, grayscale morphological dilation, or interpolation, is then applied to obtain a complete surrogate field $\tilde{F}^{(b)} = \mathfrak{R}(F^{(b)}_{obs})$. In our implementation under a sublevel-set filtration, dropped voxels are imputed by a value slightly larger than $\max(F)$ so that they enter the filtration last.

The subsampling and reconstruction procedure is repeated B times to characterize variability induced by incomplete sampling. We quantify uncertainty either at the function level or directly at the persistence-diagram level. For function-level uncertainty, we compute the reconstruction error under the supremum norm,

$$\delta^{(b)} = \|F - \tilde{F}^{(b)}\|_\infty = \max_{x \in \Omega} |F(x) - \tilde{F}^{(b)}(x)| \tag{9}$$

and let $\hat{\delta}_{1-\alpha}$ denote the $(1-\alpha)$-quantile of $\{\delta^{(b)}\}_{b=1}^{B}$. We then define the confidence bandwidth as:

$$\varepsilon_\alpha = \hat{\delta}_{1-\alpha} \tag{10}$$

Accordingly, statistical significance is determined by whether a persistence pair remains separable from the diagonal under the ℓ_∞ bottleneck metric; we retain only features satisfying $d - b > 2\varepsilon_\alpha$.

In practice, to tightly align with discrete cubical implementations and avoid estimating additional filtration-dependent constants, we adopt a diagram-level estimate by directly computing bottleneck deviations,

$$\Delta^{(b)} = d_B(Dgm_k(F), Dgm_k(\tilde{F}^{(b)})) \tag{11}$$

and setting $\hat{\Delta}_{1-\alpha}$ as the $(1-\alpha)$-quantile of $\{\Delta^{(b)}\}$, with $\varepsilon_\alpha = \hat{\Delta}_{1-\alpha}$. This yields an empirical confidence bandwidth that is filtration-consistent and implementation-aware.

According to the stability of cubical persistence, the bottleneck distance between $Dgm_k(F)$ and $Dgm_k(\tilde{F}^{(b)})$ is controlled by the perturbation magnitude of the underlying voxel function. Consequently, any feature $(b, d) \in Dgm_k(F)$ with lifespan $d - b \leq 2\varepsilon_\alpha$ is statistically indistinguishable from voxel-level subsampling and reconstruction noise and is regarded as unstable or spurious. Conversely, features satisfying $d - b > 2\varepsilon_\alpha$ are retained as statistically significant topological signals. This criterion is applied independently to $k \in \{0, 1, 2\}$ to extract confidence-filtered connected components, loops, and cavities, which are subsequently localized back to the voxel domain and used as topology-derived inputs in our framework.

3.3 Topology-guided dual-branch network

We propose TopoNet, a topology-guided dual-branch framework for medical image classification that integrates complementary image information with explicit topological structures. This section is organized into two parts. First, we describe topological structure extraction via persistent homology. Specifically, we construct a cubical-complex sublevel set filtration to compute H_0 (connected components), H_1 (loops), and H_2 (cavities). Guided by the criteria in Section 3.2, we further select statistically significant topological features, and map the persistence-derived structures back to the voxel space to form spatially aligned topological tissue structures consistent with the input images. Second, we detail the TopoNet architecture, describing how the topology modality is encoded in parallel with the image modality, and how cross-branch interaction modules are designed to fuse topology-aware representations for robust prediction.

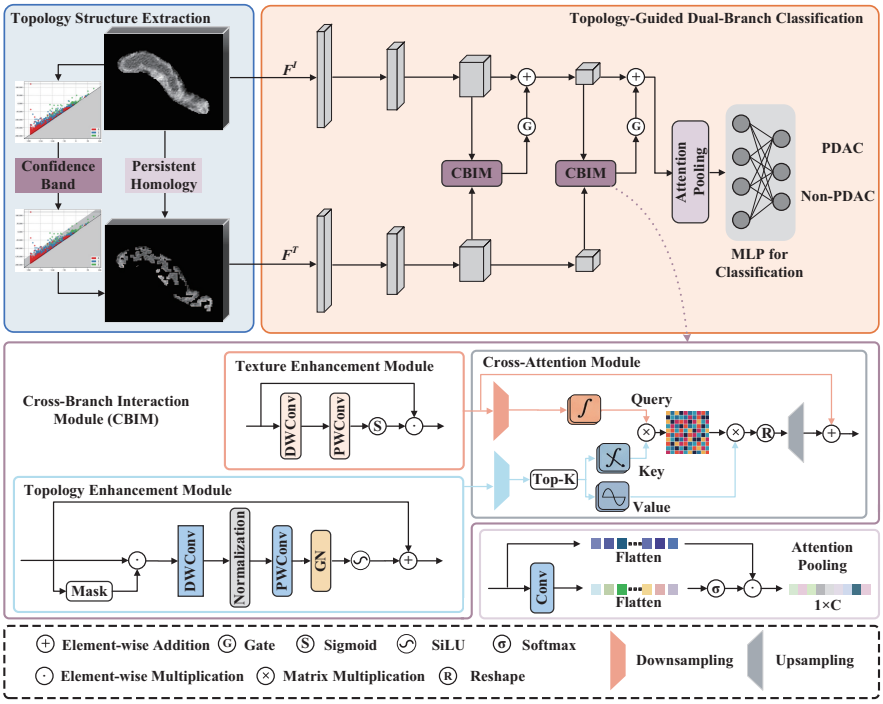


Fig. 4 The overview of the proposed TopoNet framework. The **top** panel consists of two parts. The left part illustrates topological structure extraction, where persistent homology together with a confidence band is used to identify statistically significant topological structures. The right part shows the topology-guided dual-branch classification network, which comprises an image branch and a topology branch for parallel feature encoding and topology-guided fusion. The **bottom** panel details the architecture of the cross-branch interaction module (CBIM) (including the texture enhancement module, the topology enhancement module, and the cross-attention module), as well as the attention pooling used for feature aggregation before the MLP head.

3.3.1 Topology Structure Extraction via Cubical Persistence

We explicitly encode topological structure into volumetric representation learning by computing persistent homology on cubical complex and extracting voxel-wise topological structures from the input intensity field. Specifically, a sublevel set filtration is constructed to compute the persistence of H_0, H_1, and H_2. Let $f : \Omega \to \mathbb{R}$ denote the 3D intensity field, and let K be the associated cubical complex. We represent the input volume on a regular cubical grid and assign filtration values on lattice sites as $val(v) = f(v)$. Higher-dimensional cells inherit filtration values via the standard cubical rule, defined as the maximum over their incident vertices, which ensures inclusion consistency across sublevel sets. The filtration is defined as:

$$K_t = \{\sigma \in K \mid val(\sigma) \leq t\} \tag{12}$$

where $val(\cdot)$ denotes the cubical filtration value of a cell σ, derived from image intensities. Following Section 3.2, statistically salient persistence pairs are identified

using confidence bands and subsequently mapped back to the image domain, yielding topology-derived tissue structures that are spatially aligned with the original volume. These topology-informed structures provide interpretable, scale-aware geometric constraints and serve as an auxiliary modality for our topology-guided dual-branch network. In the following, we describe the extraction procedures for topology-derived structures induced by H_0, H_1, and H_2, respectively.

H_0: Seed-guided connected component mask extraction from persistence pairs. Under this filtration, H_0 characterizes the birth and merging of connected components. Each H_0 persistence pair (b_i, d_i) indicates that a component is born at $t = b_i$ and dies at $t = d_i$ when it merges into an older component according to the elder rule. After identifying statistically significant persistence pairs via the confidence band, we localize the spatial support of each selected component by taking its birth seed s_i, provided by the creator index (i.e., the birth voxel that creates the corresponding H_0 class). We then extract the seed-connected component immediately prior to the corresponding merge event. We determine the target connected component using a pre-merge threshold:

$$t_i^{high} = d_i - \delta \tag{13}$$

together with a birth-side stabilization threshold:

$$t_i^{low} = b_i + \varepsilon \tag{14}$$

specifically, we sample the connected component containing the seed s_i at t_i^{high}, while t_i^{low} serves to stabilize the birth-side thresholding against discretization and tie effects. Unless otherwise stated, we set $\delta = \varepsilon = 10^{-3}$. Let $V_t = \{v \in \Omega \mid f(v) \leq t\}$ denote the sublevel set lattice sites at threshold t. We define the voxel-wise connected-component mask as:

$$M_i = CC(V_{t_i^{high}}, s_i) = \{x \in V_{t_i^{high}} \mid x \leftrightarrow s_i\} \tag{15}$$

where $\leftrightarrow$ denotes voxel connectivity under a chosen adjacency (e.g., 6-neighborhood). The resulting connected set is directly interpreted as a binary voxel mask on the input grid.

H_1: Persistent 1-cycle extraction via matrix reduction. Complementary to region-level connectivity, H_1 captures loop-like structures. Within the same cubical filtration to guarantee filtration monotonicity and inclusion consistency, each cell σ is assigned a non-decreasing filtration value using the standard cubical rule:

$$val(v) = f(v), val(e) = \max\{f(v_1), f(v_2)\}, val(q) = \max\{f(v_1), f(v_2), f(v_3), f(v_4)\} \tag{16}$$

where v, e, and q denote vertices, edges, and square faces, respectively. We obtain a deterministic filtration order by sorting cells by $val(\sigma)$, with a fixed tie-breaking rule (e.g., by cell dimension and lexicographic index) to handle equal $val(\sigma)$ values.

Over the field $\mathbb{Z}_2$, we define the chain groups and boundary operators:

$$\partial_1 : C_1 \to C_0, \partial_2 : C_2 \to C_1 \tag{17}$$

The boundary matrices are implemented in a column-set form: each edge column $\partial_1(e)$ contains its two incident vertices, and each face column $\partial_2(q)$ contains its four

boundary edges (orientations are irrelevant over $\mathbb{Z}_2$). Applying the standard persistent homology matrix reduction with respect to the above filtration order yields the set of H_1 persistence pairs. For each pair (b_i, d_i), a representative 1-cycle is reconstructed by tracking the column operations throughout the reduction procedure, and is returned as cycle-edge-indices, i.e., the set of supporting edge indices forming a discrete loop. Subsequently, we rasterize each representative 1-cycle into a binary loop mask. Specifically, for each persistence pair (b_i, d_i), we take the edge indices returned by cycle-edge-indices and map them onto the voxel lattice by marking voxels adjacent to the corresponding edges, resulting in a sparse loop mask. To obtain a more stable and thickness-aware structure, we optionally apply a morphological dilation with radius $r = 1$, thereby forming a tubular neighborhood around the loop. Here r is measured in voxels on the native grid. Finally, we aggregate the masks associated with all salient persistence pairs selected by the confidence band criterion to produce the H_1 loop mask.

H_2: Destroyer seeded cavity identification and shell mask construction. In 3D cubical persistent homology, an H_2 class corresponds to a bounded cavity enclosed by the object. Such cavities often align with clinically meaningful structures in medical imaging (e.g., cystic lumens or necrotic voids), and the associated representative cycle can be interpreted as the closed shell surface that bounds the cavity. In a sublevel set filtration, the birth of an H_2 class indicates the formation of an enclosed void space, whereas its death occurs when the cavity is filled by the inclusion of a 3-cell (cube), causing the void to vanish.

To localize the image-domain support of a specific H_2 feature, we adopt a seeded strategy analogous to the H_0 extraction procedure. Specifically, after enumerating all interior cavities in the sublevel set, we use the destroyer 3-cell coordinate associated with the selected persistence pair to uniquely identify the target cavity. The destroyer voxel serves as a seed for geodesic region growing within the complement of the sublevel set object, yielding a cavity mask that is precisely aligned with the current H_2 class. To obtain a stable voxel-wise approximation of the representative 2-cycle, we further construct the object boundary adjacent to the extracted cavity using morphological operators, and optionally refine it by seeded connectivity constrained within this boundary using the neighborhood of the destroyer coordinate. The resulting cavity shell mask provides a robust discrete surrogate of the closed bounding surface, capturing the topology-defining geometry more faithfully than the cavity interior alone.

Finally, the above procedures yield binary topology masks corresponding to H_0, H_1, and H_2. These masks are combined by union to form a unified topology representation (with the specific fusion strategy determined through ablation studies), which is then used to extract spatially aligned regions from the original image volume while preserving the associated voxel intensity information, thereby completing the topological structure extraction stage.

3.3.2 The architecture of TopoNet

As illustrated in Fig. 4, we propose TopoNet, a dual-branch 3D architecture for volumetric medical image classification that jointly leverages image information and

topological structural features. The overall framework comprises three stages: **(1)** mono-modal feature extraction, **(2)** dual-modal feature fusion, and **(3)** classification head.

1) Mono-modal feature extraction. Given an input image $I \in \mathbb{R}^{D\times H\times W\times C}$ and a topology-related volume $T \in \mathbb{R}^{D\times H\times W\times C}$, we employ two independent 3D DenseNet121 branches, ψ_{img} and ψ_{topo}, to extract modality-specific multi-scale representations:

$$F_i^I = \psi_{img}(I;\theta_I), F_i^T = \psi_{topo}(T;\theta_T), i \in \{1,2,3,4\} \tag{18}$$

where $F_i^I, F_i^T \in \mathbb{R}^{d_i\times h_i\times w_i\times c_i}$ denotes the stage-wise feature maps (corresponding to layer 1–4 in the implementation), and θ_I and θ_T are the learnable parameters of the image branch and the topology branch, respectively. In our 3D DenseNet121 instantiation, the channel dimensions follow $c_i = \{256, 512, 1024, 2048\}$. Notably, after projecting the persistence-derived topological structures back into voxel space, the topology modality is typically highly sparse: responses are concentrated in a small subset of voxels, whereas most voxels remain near zero. To enhance robustness under such sparse activations, we replace all Batch Normalization layers in the topology branch with Group Normalization (GN). We use a fixed group number (e.g., $G = 32$) for GN across the topology branch. By eliminating reliance on mini-batch statistics (mean and variance), GN yields more stable feature normalization and facilitates more reliable optimization in volumetric training settings.

2) Dual-modal feature fusion. To effectively aggregate the complementary information from the two modalities, we introduce, at stage 3 and 4 (i.e., a halfway fusion paradigm), a **cross-branch interaction module (CBIM)** that consists of three components: a texture enhancement module, a topology enhancement module, and a cross-attention module.

First, on the image branch, we employ a lightweight **texture enhancement module** to strengthen local texture features with minimal computational overhead. Conventional spatial attention mechanisms typically incur substantial memory cost; moreover, their strong emphasis on global context may under-model local spatial variations, thereby weakening fine-grained detail representations. To alleviate this limitation, we adopt a spatial gating scheme based on depth-wise convolutions that approximates spatial self-attention using only convolutions. As illustrated in Fig. 4, the texture enhancement module is defined as:

$$W = \sigma(PW(DW(F^I))) \tag{19}$$

$$F^{\tilde{I}} = F^I \odot (1 + W) \tag{20}$$

where $DW(\cdot)$ denotes a $3\times 3\times 3$ depth-wise convolution, $PW(\cdot)$ denotes a $1\times 1\times 1$ point-wise convolution, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication with broadcasting over channels. This module acts as a residual spatial amplifier that selectively boosts texture-salient regions. In this way, the texture enhancement module strengthens locally discriminative appearance information on the image branch before it is fed into the subsequent cross-branch fusion.

Next, complementary to the texture enhancement module on the image branch, we further refine the topology features by employing a sparsity-aware **topology enhancement module**. Given a topology feature map F^T, we first construct a binary validity mask M from non-zero activations and restrict the subsequent depth-wise convolution to these valid regions. The resulting responses are then normalized by the number of valid voxels within each convolutional kernel, yielding a sparsity-aware local aggregation that is robust to variations in activation density. Finally, a point-wise $1 \times 1 \times 1$ convolution, followed by GN and a SiLU nonlinearity, is applied and added back to the input in a residual fashion. This design prevents background zeros from contaminating topology representations and enables stable, multi-scale aggregation of topology features on their sparse spatial support. Formally, the topology enhancement module is defined as:

$$M = I(|F^T|_1 > 0) \tag{21}$$

$$F = DW(F^T \odot M)/(conv(M) + \varepsilon) \tag{22}$$

$$F^{\tilde{T}} = F^T + SiLU(GN(PW(F))) \tag{23}$$

where M is the binary validity mask derived from non-zero activations, $DW(\cdot)$ and $PW(\cdot)$ denote depth-wise and point-wise convolutions, $conv(\cdot)$ is a convolution with an all-ones kernel used to count valid voxels for normalization, $GN(\cdot)$ denotes Group Normalization, $SiLU(\cdot)$ is the SiLU activation function, $\odot$ denotes element-wise multiplication, and ε is a small constant for numerical stability.

Finally, we perform cross-modal fusion using a cross-attention operator $\Phi_i(\cdot)$ that conditions image features on topology-guided context:

$$F_i^{I \leftarrow T} = \Phi_i(F_i^{\tilde{I}}, F_i^{\tilde{T}}; \theta_f), i \in \{3, 4\} \tag{24}$$

here, $\Phi_i(\cdot)$ corresponds to the implemented **cross-attention module** at stage i, where image features are used to construct queries and topology features provide keys and values to modulate the image representations. Let $F_i^{\tilde{I}}$ and $F_i^{\tilde{T}}$ denote the image and topology feature maps after the texture enhancement module and the topology enhancement module, respectively. Both feature maps are flattened into token sequences $T_I, T_T \in \mathbb{R}^{N \times C}$. The topology tokens are then projected into value and key spaces, while the image tokens are projected into the query space, as defined:

$$Q_I = T_I W^Q, K_T = T_T W^K, V_T = T_T W^V \tag{25}$$

where W^Q, W^K and W^V are learnable projection matrices. Notably, to control the computational cost of 3D attention and to focus on salient topology tokens, we apply token compression and selection before computing attention. Specifically, we downsample $F_i^{\tilde{I}}$, $F_i^{\tilde{T}}$ to a fixed token grid via max pooling (e.g., 4^3 at stage 3 and 3^3 at stage 4), and select a fixed proportion of Top-K_i topology tokens according to their l_2 energy, which are further used to generate T_I and T_T. By default, we set $(K_3, K_4) = (0.5, 0.5)$.

Based on the remaining tokens, we build the cross-modal correlation matrix via dot-product attention and obtain the topology-conditioned image representation:

$$Z_I = softmax(\frac{Q_I K_T^T}{\sqrt{D_K}}) \cdot V_T \tag{26}$$

where D_K is the key dimension, and the $softmax$ is applied row-wise so that each image token aggregates information from the most relevant topology tokens. The resulting fused tokens are then mapped back to the 3D feature space, and the fused representation is written into the image branch. Concretely, we reshape Z_I and upsample it back to the original stage resolution using trilinear interpolation, after which it is added to $F_i^{\tilde{I}}$ through a residual connection. It is defined as:

$$F_i^{T\to I} = Up_{tri}(reshape(Z_I)) + F_i^{\tilde{I}} \tag{27}$$

where $Up_{tri}(\cdot)$ denotes the trilinear interpolation-based upsampling operator. Next, instead of using simple additive fusion, concatenation followed by a $1 \times 1 \times 1$ convolution, or ungated residual fusion, we adopt a gated residual scheme at stages 3–4. For each fusion stage $i \in \{3, 4\}$, we first compute a learnable scalar gate:

$$g_i = \sigma(\theta_i) \tag{28}$$

and then update the image features as:

$$\hat{F}_i^I = F_i^I + g_i F_i^{T\to I} \tag{29}$$

where F_i^I denotes the pre-fusion image feature map at stage i, $F_i^{T\to I}$ is the topology-guided fused feature produced by the cross-attention, θ_i is a trainable parameter, and $\sigma(\cdot)$ denotes the sigmoid function. This gated residual formulation allows the network to adaptively control the strength of cross-branch interaction: when topology representations are unreliable, the model can effectively fall back to the image-only pathway, whereas informative fused features are progressively injected as g_i increases during training.

3) Classification head with attention pooling. After the final fusion stage, we obtain the top-level fused representation $\hat{F}_4^I$. In volumetric medical images, a substantial portion of voxels typically corresponds to background with near-zero responses. As a result, standard global average pooling (GAP), which aggregates all voxels with equal weights, may dilute discriminative foreground information by overwhelming background activations. To address this issue, we draw inspiration from [27] and adopt an attention-based pooling operator to perform adaptive, content-aware aggregation over the feature, thereby highlighting informative spatial locations while suppressing irrelevant background contributions. Specifically, a $1\times1\times1$ convolution produces a spatial attention logit map. After flattening the spatial dimensions, a $softmax$ function is applied to yield normalized attention weights over all voxels. The pooled representation is then computed as the attention-weighted sum of voxel features, producing a

compact global descriptor. It is defined as:

$$\alpha = softmax(flat(conv(\hat{F}_4^I))) \tag{30}$$

$$z = \sum_{n=1}^{N} \alpha_n f_n \tag{31}$$

where N denotes the total number of voxels in the 3D feature volume after flattening, with $N = D \times H \times W$. Let $f_n \in \mathbb{R}^C$ denote the feature vector at the n-th voxel location, and $\alpha \in \mathbb{R}^N$ denote the normalized attention weights. The resulting pooled representation z is subsequently fed into a lightweight classifier:

$$y = \varphi_{cls}(z; \theta_c) \tag{32}$$

where θ_c are the learnable parameters of classifier, and $\varphi_{cls}(\cdot)$ is an MLP with dropout regularization, which outputs the final logits y for downstream prediction.

3.4 Loss function

In the proposed method, we optimize the classification objective using the categorical cross-entropy (CE) loss, which measures the discrepancy between the ground-truth label distribution p and the predicted class-probability distribution q. It is defined as:

$$L_{CE} = -\sum_{i=1}^{C} p_i \log(q_i) \tag{33}$$

where C is the number of classes, p_i denotes the ground-truth label for class i, and q_i is the predicted probability for class i.

4 Experiments

4.1 Datasets

The MSD Task07 Pancreas dataset [28] consists of portal-venous phase, contrast-enhanced abdominal CT scans acquired at Memorial Sloan Kettering Cancer Center (MSKCC), comprising 420 three-dimensional volumes. As reported in [28], the cohort includes patients who underwent surgical resection for pancreatic masses, with pathological diagnoses spanning intraductal papillary mucinous neoplasm (IPMN), pancreatic neuroendocrine tumors (PNET), and pancreatic ductal adenocarcinoma (PDAC). The official release provides voxel-level annotations of the pancreatic parenchyma and pancreatic lesions to support segmentation tasks. For this study, an experienced chief physician specializing in pancreaticobiliary surgery re-reviewed the 281 publicly labeled training cases and curated two subsets: 82 PDAC cases and 62 IPMN cases. Based on this curated cohort, we formulated a binary classification task to determine whether each case corresponds to PDAC (PDAC versus non-PDAC). Notably, the

curated cohort is relatively small (only 144 cases). Moreover, 3D volumetric classification is high-dimensional and prone to overfitting, making it challenging to achieve stable performance with strong generalizability on this dataset.

The Brain Tumor Segmentation Challenge (BraTS 2018) [29–31] is one of the most widely used benchmark datasets and challenge suites for multimodal brain tumor MRI segmentation in medical imaging. It was originally established to evaluate tumor segmentation algorithms and has been progressively extended to related tasks such as tumor classification and prognosis prediction. The BraTS 2018 cohort comprises 285 glioma subjects, including 210 high-grade gliomas (HGG) and 75 low-grade gliomas (LGG), providing a reliable basis for glioma grading and classification studies using public data. For each subject, four 3D MRI modalities are available: T1, post-contrast T1-weighted imaging (T1-Gado), T2, and fluid-attenuated inversion recovery (FLAIR). Manual segmentations were generated by 1–4 annotators following a unified annotation protocol and were subsequently reviewed and approved by experienced neuroradiologists. The annotations delineate enhancing tumor, peritumoral edema, and the necrotic and non-enhancing tumor core. To facilitate cross-subject comparison and robust model training, all images undergo standardized preprocessing, including registration to a common template, resampling to 1 mm^3 isotropic resolution, and skull stripping. Building on this dataset setting, we further evaluate the generalization capability of our method by performing binary glioma grading (HGG versus LGG) using the T1-Gado sequence as input.

4.2 Evaluation metrics

In medical image classification tasks, relying solely on a single evaluation metric is insufficient for comprehensively assessing model performance. Therefore, this study utilizes multiple evaluation metrics to ensure robust and reliable evaluation. Specifically, three widely recognized metrics, including Accuracy, Recall, and F1-score, are employed to quantitatively evaluate the effectiveness of different classification methods. These metrics are formally defined as follows:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{34}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{35}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{36}$$

$$F1 = \frac{2 \times \mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \tag{37}$$

where TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively. In addition, we further evaluate model performance using the image-level area under the receiver operating characteristic curve (AUC).

4.3 Implementation details

We implement TopoNet using PyTorch 2.8 and the MONAI [32]. All models are trained on an NVIDIA GeForce RTX 5090 GPU with 32 GB of memory. We employ the AdamW [33] optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-2}. The maximum number of training epochs is set to 200. To improve training stability and convergence, we adopt a warm-up cosine annealing learning rate schedule: during the first 8 epochs, the learning rate is linearly warmed up from 0 to the initial value; thereafter, cosine annealing is applied for the remaining epochs, decaying the learning rate to a minimum of 1×10^{-6}.

For the MSD Task07 Pancreas dataset, we first resample all volumes to an isotropic spacing of 1mm^3. To reduce irrelevant background interference, we leverage the segmentation annotations to extract the pancreas region as the region of interest (ROI). We expand the ROI boundary by 5 voxels to obtain the minimal 3D bounding box that encloses the ROI, crop the corresponding subvolume, and then uniformly resize it to $96 \times 96 \times 96$. The resulting input is then intensity-normalized and fed into the network for training. For the BraTS 2018 dataset, we define the ROI as the union of three tumor subregions, namely the enhancing tumor, peritumoral edema, and the necrotic and non-enhancing tumor core, and apply the same preprocessing pipeline: cropping with a 5-voxel padding, resizing to $96 \times 96 \times 96$, followed by intensity normalization before network training.

For topological feature extraction, we compute persistent homology using the GUDHI library [34]. On MSD Task07 Pancreas dataset, we employ sublevel filtration to emphasize topological patterns associated with low-intensity structures. In contrast, for the BraTS 2018 dataset, tumor-related regions on contrast-enhanced T1 (T1-Gado) typically exhibit high signal intensity, whereas sublevel filtration inherently emphasizes low-valued structures. Therefore, we adopt a bright-structure oriented filtration strategy. Specifically, we invert the input prior to applying sublevel filtration, enabling a more effective characterization of the topology of high-intensity enhancing regions. To extract statistically significant topological structures, we perform $B = 1000$ subsampling trials within the ROI, randomly retaining a fraction of $p = 0.9$ ROI voxels per trial to quantify topological uncertainty and determine the $(1 - \alpha) = 0.9$ confidence-band width.

Model training is conducted using 5-fold cross-validation, and we report the mean $\pm$ standard deviation across the five validation folds. In addition, we employ data augmentation strategies including random rotations and random flips.

4.4 Comparison with other methods

We compared TopoNet with eight image classification baselines, including ResNet50 [35], DenseNet121 [36], SENet154 [37], SEResNet50 [37], SEResNeXt50 [37], EfficientNetB5 [38], ViT [39], and Swin Transformer [40]. To ensure a fair comparison, all competing methods were implemented and trained under the MONAI framework using their official codebases. All models were trained in the same environment without any pretraining or fine-tuning.

4.4.1 Comparison on the MSD Task07 Pancreas dataset

Table 1 provides a quantitative comparison between representative CNN-based architectures and Transformer-based models. As shown, the proposed TopoNet achieves the best overall performance across most key evaluation metrics (Accuracy 0.8197±0.0587, F1-score 0.8443 ± 0.0406, AUC 0.8123 ± 0.1127). While DenseNet121 obtains a higher mean Recall, TopoNet yields a highly competitive Recall (0.8412 ± 0.0310) with significantly lower variance compared to DenseNet121 (± 0.1194), thereby exhibiting a more balanced and stable advantage particularly in Accuracy, F1-score, and AUC.

Compared with conventional CNN baselines, TopoNet yields consistent and notable improvements on multiple metrics, indicating that explicitly incorporating topological structural priors offers crucial complementary cues beyond appearance-driven representation learning and leads to more discriminative volumetric feature representations. Meanwhile, the dual-branch design enables parallel modeling of texture information and topology-derived structural features, and facilitates effective cross-branch interactions for complementary fusion, thereby enhancing robustness to anatomical variability, lesion morphology changes, and imaging noise. Notably, although pure Transformer-based models are inherently capable of global representation learning and long-range dependency modeling, their performance is often constrained in data-scarce settings, where a higher sample complexity is required to fully exploit attention mechanisms. These results further support the view that topology-guided learning provides an effective and competitive pathway for improving small-sample 3D medical image classification performance.

Table 1 Quantitative comparison of the proposed method with representative CNN- and Transformer-based approaches for volumetric classification on the MSD Task07 Pancreas dataset. Best results are in **bold**.

Type	Method	Accuracy ↑	Recall ↑	F1-score ↑	AUC ↑
CNN-based	ResNet50 [35]	0.7493 ± 0.0705	0.7779 ± 0.1655	0.7724 ± 0.0843	0.7603 ± 0.0978
	DenseNet121 [36]	0.7773 ± 0.0492	**0.8684 ± 0.1194**	0.8142 ± 0.0459	0.7660 ± 0.0632
	SENet154 [37]	0.7633 ± 0.0820	0.7949 ± 0.1462	0.7876 ± 0.0836	0.7342 ± 0.1304
	SEResNet50 [37]	0.7567 ± 0.0629	0.8272 ± 0.1347	0.7918 ± 0.0614	0.7714 ± 0.0635
	SEResNeXt50 [37]	0.7704 ± 0.0438	0.7684 ± 0.0873	0.7916 ± 0.0341	0.7694 ± 0.0779
	EfficientNetB5 [38]	0.6943 ± 0.0463	0.7919 ± 0.1660	0.7409 ± 0.0543	0.6737 ± 0.1106
Transformer-based	ViT [39]	0.7012 ± 0.0214	0.8191 ± 0.0610	0.7593 ± 0.0142	0.6486 ± 0.0922
	Swin Transformer [40]	0.7288 ± 0.0329	0.8147 ± 0.0445	0.7721 ± 0.0177	0.7295 ± 0.1096
	TopoNet (ours)	**0.8197 ± 0.0587**	0.8412 ± 0.0310	**0.8443 ± 0.0406**	**0.8123 ± 0.1127**

4.4.2 Comparison on the BraTS 2018 dataset

To demonstrate the generalization and robustness of our TopoNet on MRI modalities, we conduct volumetric glioma grading on the BraTS 2018 dataset using the contrast-enhanced T1 modality. As summarized in Table 2, TopoNet achieves the best overall

performance, reaching 0.9333 ± 0.0079 in Accuracy, 0.9561 ± 0.0053 in F1-score, and 0.9267 ± 0.0437 in AUC, outperforming representative CNN and Transformer baselines. Its Recall (0.9857 ± 0.0130) matches the highest performance achieved by the best competing method (0.9857 ± 0.0117, SEResNeXt50). Notably, while maintaining high sensitivity to lesions, TopoNet shows substantially lower variability in key metrics like Accuracy ($\pm$ 0.0079) and F1-score ($\pm$ 0.0053) compared to other baselines, indicating more stable predictions across folds. Overall, these results indicate that TopoNet generalizes well to MRI modalities and further validate that incorporating topology-guided representations can simultaneously enhance discriminability and robustness for glioma grading.

Table 2 Quantitative comparison of the proposed method with representative CNN- and Transformer-based approaches for volumetric classification on the BraTS 2018 dataset. Best results are in **bold**.

Type	Method	Accuracy ↑	Recall ↑	F1-score ↑	AUC ↑
CNN-based	ResNet50 [35]	0.9053 ± 0.0179	0.9667 ± 0.0323	0.9379 ± 0.0101	0.9095 ± 0.0340
	DenseNet121 [36]	0.9228 ± 0.0238	0.9810 ± 0.0095	0.9495 ± 0.0148	0.9171 ± 0.0384
	SENet154 [37]	0.8842 ± 0.0238	0.9714 ± 0.0233	0.9254 ± 0.0139	0.8171 ± 0.0667
	SEResNet50 [37]	0.8947 ± 0.0333	0.9571 ± 0.0278	0.9309 ± 0.0205	0.8581 ± 0.0727
	SEResNeXt50 [37]	0.9018 ± 0.0263	**0.9857 ± 0.0117**	0.9369 ± 0.0155	0.8676 ± 0.0530
	EfficientNetB5 [38]	0.8667 ± 0.0306	0.9810 ± 0.0233	0.9159 ± 0.0171	0.8206 ± 0.0788
Transformer-based	ViT [39]	0.8351 ± 0.0409	0.9714 ± 0.0278	0.8975 ± 0.0211	0.7427 ± 0.0861
	Swin Transformer [40]	0.8807 ± 0.0205	0.9524 ± 0.0261	0.9218 ± 0.0121	0.8873 ± 0.0651
	TopoNet (ours)	**0.9333 ± 0.0079**	0.9857 ± 0.0130	**0.9561 ± 0.0053**	**0.9267 ± 0.0437**

4.5 Ablation studies

In this section, we conduct a two-part ablation study to systematically evaluate the contribution of each component within our framework. The first part investigates the topology extraction pipeline, assessing the effect of persistent-homology-derived topological structures. The second part examines the architectural design of TopoNet, including the dual-branch encoding scheme and the cross-modal fusion mechanism. All experiments are performed on the MSD Task07 Pancreas dataset under identical training configurations to ensure fair comparison.

4.5.1 Ablation on topological structure extraction

1) Effects of confidence band-based topology retention. To evaluate how topological structure retention strategies affect classification performance, we compare two schemes: (i) fixed-threshold selection, which retains 20%–70% of the topological structures, and (ii) an adaptive selection based on statistical significance using a confidence band (ours). For a controlled comparison, we extract only H_0 topological structures in all settings for this experiment. The results are summarized in Table 3. Overall, the

fixed-threshold strategy exhibits a clear non-monotonic behavior: when the threshold is too low (e.g., 20%), insufficient topological cues are preserved, leading to inferior Accuracy and F1-score; when the threshold becomes too high (e.g., 70%), more noisy or unstable structures may be introduced, resulting in degraded overall performance and increased variance in metrics such as Recall and F1-score, which indicates reduced model stability. In contrast, the confidence-band selection (ours) achieves the best performance in Accuracy, F1-score and AUC. Notably, while the 30% threshold yields the highest Recall, it suffers from a severe degradation in AUC (0.6868); our confidence-band selection maintains a highly competitive and more stable Recall (0.8662 ± 0.0450) while preventing such catastrophic performance trade-offs. Overall, the confidence band based adaptive selection can preserve statistically salient and stable topological structures without manual threshold tuning, thereby improving the discriminative quality of the learned representations and enhancing overall classification performance and robustness.

Table 3 Ablation study of topology-structure retention criteria using H_0 topological structures (fixed percentile thresholds versus confidence-band selection). Best results are in **bold**.

	Accuracy ↑	Recall ↑	F1-score ↑	AUC ↑
20%	0.7436 ± 0.0446	0.8059 ± 0.1016	0.7808 ± 0.0370	0.7154 ± 0.0742
30%	0.7567 ± 0.0332	**0.9272 ± 0.0572**	0.8133 ± 0.0148	0.6868 ± 0.0365
40%	0.7712 ± 0.0333	0.8522 ± 0.0860	0.8084 ± 0.0302	0.7334 ± 0.0600
50%	0.7357 ± 0.0315	0.8544 ± 0.1333	0.7838 ± 0.0305	0.7526 ± 0.0984
60%	0.7502 ± 0.0388	0.9007 ± 0.0754	0.8036 ± 0.0314	0.7612 ± 0.0372
70%	0.7079 ± 0.0391	0.8272 ± 0.1273	0.7591 ± 0.0513	0.6735 ± 0.0277
confidence band (ours)	**0.7985 ± 0.0342**	0.8662 ± 0.0450	**0.8310 ± 0.0234**	**0.7773 ± 0.0834**

2) Effects of different topological structure combinations. To systematically assess the contributions of different topological structures to pancreatic cancer classification, we conducted an ablation study with seven topology-input configurations. Specifically, the topology branch was fed with H_0 (connected components), H_1 (loops), H_2 (cavities), their pairwise combinations ($H_0 + H_1, H_0 + H_2, H_1 + H_2$), and the full combination $H_0 + H_1 + H_2$ (ours), as summarized in Table 4.

Compared with the image-only DenseNet121 baseline (Table 1), introducing a single homology dimension consistently improves Accuracy and AUC, but does not yield consistent improvements in Recall. In contrast, fusing multi-dimensional topological structures generally leads to clearer and more consistent performance gains, and $H_1 + H_2$ achieves the highest AUC, suggesting that loop- and cavity-related structures provide particularly discriminative signals. Further incorporating H_0 yields the final configuration, which delivers the best overall performance in Accuracy and F1-score and exhibits improved cross-fold stability.

These findings are consistent with the radiological characteristics of pancreatic ductal adenocarcinoma (PDAC) on CT. PDAC often presents with hypodense necrotic or

cystic components, ductal dilatation, and irregular gland boundaries. From a topological perspective, these patterns can be naturally mapped to separated low-attenuation connected regions (H_0), enclosure-induced loop structures (H_1), and volumetric voids such as cystic or necrotic cavities (H_2). Overall, the three homology dimensions provide complementary information: H_1 and H_2 primarily enhance discriminability, whereas H_0 further contributes global connectivity and spatial distribution cues, resulting in stronger and more robust overall performance. Therefore, we adopt $H_0 + H_1 + H_2$ as the default topology input to the topology branch in subsequent experiments.

Table 4 Ablation of topology-structure combinations for pancreatic cancer classification. Different subsets of $\{H_0, H_1, H_2\}$ are fed to the topology branch. Best results are in **bold**.

Topological structures	H_0	H_1	H_2	Accuracy ↑	Recall ↑	F1-score ↑	AUC ↑
H_0	✓	-	-	0.7985 ± 0.0342	**0.8662 ± 0.0450**	0.8310 ± 0.0234	0.7773 ± 0.0834
H_1	-	✓	-	0.7988 ± 0.0451	0.7779 ± 0.1109	0.8102 ± 0.0583	0.8072 ± 0.0591
H_2	-	-	✓	0.7988 ± 0.0699	0.8662 ± 0.0718	0.8310 ± 0.0556	0.7807 ± 0.1018
$H_0 + H_1$	✓	✓	-	0.8062 ± 0.0734	0.8544 ± 0.0469	0.8375 ± 0.0477	0.7961 ± 0.1101
$H_0 + H_2$	✓	-	✓	0.7640 ± 0.0545	0.8301 ± 0.1317	0.7995 ± 0.0344	0.7711 ± 0.0900
$H_1 + H_2$	-	✓	✓	0.7985 ± 0.0596	0.8390 ± 0.1298	0.8203 ± 0.0711	**0.8300 ± 0.1020**
$H_0 + H_1 + H_2$ (ours)	✓	✓	✓	**0.8197 ± 0.0587**	0.8412 ± 0.0310	**0.8443 ± 0.0406**	0.8123 ± 0.1127

3) Effects of intensity inversion in topology extraction. To evaluate the effect of intensity inversion in preprocessing (i.e., whether the topology extractor should emphasize bright or dark structures) on topological structure extraction and downstream classification, we conduct two strictly controlled experiments. After intensity normalization, the volumetric intensity field is denoted as $f \in [0, 1]$. In the first setting, we construct a sublevel set filtration directly on f to extract persistent homology structures (No inversion, ours). In the second setting, following TopoTxR [13], we invert the intensity field as $f_{inv} = 1 - f$ and perform sublevel set filtration on f_{inv}, which is semantically equivalent to applying a superlevel set filtration on the original f.

As shown in Table 5, the no-inversion setting yields substantial improvements in mean performance across all four metrics. Although it introduces a higher variance in Accuracy and AUC, it achieves a notably higher and more stable sensitivity (Recall: 0.8412 ± 0.0310 versus 0.7794 ± 0.0945), indicating that sublevel set filtration on f provides more discriminative topological information for pancreatic cancer classification. This result aligns with a key principle of persistent homology: the choice of filtration function determines the anatomical and pathological semantics of the extracted topological structures. In pancreatic CT, clinically informative cues are often dominated by low-density dark regions, such as necrotic cavities, ductal dilatation, and hypo perfused areas, whose spatial distribution and enclosure patterns are closely associated with tumor growth and heterogeneity. Therefore, sublevel set filtration on f naturally highlights and captures the topology of these dark structures. In contrast,

intensity inversion shifts the focus toward high-density bright regions (e.g., glandular parenchyma and contrast-enhanced areas), which tend to be less stable and less consistently discriminative for this task, thereby diminishing the overall performance.

Table 5 Ablation on intensity inversion in topology extraction. We compare sublevel set filtration on f with filtration on $f_{inv} = 1 - f$ (equivalently, superlevel set filtration on f). All settings are identical. Best results are in **bold**.

Setting	Invert	Filtration	Accuracy ↑	Recall ↑	F1-score ↑	AUC ↑
No inversion (ours)	×	Sublevel on f	**0.8197 ± 0.0587**	**0.8412 ± 0.0310**	**0.8443 ± 0.0406**	**0.8123 ± 0.1127**
Intensity inversion	✓	Sublevel on $1 - f$	0.7569 ± 0.0379	0.7794 ± 0.0945	0.7821 ± 0.0467	0.7471 ± 0.0378

Notably, for tasks where enhanced bright regions constitute the primary semantic signal, such as TopoTxR [13] for predicting treatment response in breast cancer, intensity inversion is a reasonable and effective choice to emphasize bright-structure topology. This further suggests that topology-oriented preprocessing should be tailored to task-specific imaging semantics. Based on the above analysis and empirical evidence, we adopt no inversion as the default topology extraction strategy for pancreatic cancer classification.

4.5.2 Ablation on network architecture

1) Comparisons with different backbones. To verify the generality and cross-backbone robustness of TopoNet, we replace the dual-branch encoders with ResNet50 [35], SENet154 [37], EfficientNetB5 [38], Swin Transformer [40], and DenseNet121 [36], while keeping the topology extraction pipeline and the cross-modal fusion design strictly identical. As shown in Table 6, integrating our topology-guided framework consistently improves the comprehensive classification performance (specifically Accuracy and F1-score) across all tested baseline backbones. The + ours variants maintain a strong and stable performance range across different architectures (Accuracy: 0.7493–0.8197; AUC: 0.7219–0.8123), indicating that the proposed topology-prior modeling and inter-branch interaction are largely insensitive to the choice of backbone and thus exhibit excellent plug-and-play transferability.

Moreover, DenseNet121 + ours achieves the best overall results in Accuracy (0.8197 ± 0.0587), F1-score (0.8443 ± 0.0406), and AUC (0.8123 ± 0.1127). Meanwhile, SENet154 + ours attains the highest Recall (0.8912 ± 0.0705). Notably, although DenseNet121 + ours does not yield the absolute highest Recall (0.8412), it maintains a highly competitive sensitivity with a much lower variance (± 0.0310 vs. ± 0.0705), ensuring a more reliable clinical diagnosis. Overall, these results confirm that TopoNet can be effectively applied to a variety of mainstream CNN and Transformer backbones, demonstrating robust cross-architecture generalization capability.

2) Effects of CBIM on cross-modal representation learning. To systematically assess how the enhancement direction and order within CBIM affect TopoNet, we

Table 6 Ablation study on TopoNet with different backbones. Best results are in **bold**.

Method	Accuracy ↑	Recall ↑	F1-score ↑	AUC ↑
ResNet50 [35]	0.7493 ± 0.0705	0.7779 ± 0.1655	0.7724 ± 0.0843	0.7603 ± 0.0978
ResNet50 [35] + ours	**0.7502 ± 0.0495**	**0.7801 ± 0.1001**	**0.7780 ± 0.0546**	**0.7705 ± 0.0515**
SENet154 [37]	0.7633 ± 0.0820	0.7949 ± 0.1462	0.7876 ± 0.0836	0.7342 ± 0.1304
SENet154 [37] + ours	**0.7638 ± 0.0150**	**0.8912 ± 0.0705**	**0.8109 ± 0.0065**	**0.7590 ± 0.0476**
EfficientNetB5 [38]	0.6943 ± 0.0463	**0.7919 ± 0.1660**	0.7409 ± 0.0543	0.6737 ± 0.1106
EfficientNetB5 [38] + ours	**0.7776 ± 0.0477**	0.7713 ± 0.1415	**0.7942 ± 0.0489**	**0.7889 ± 0.1027**
Swin Transformer [40]	0.7288 ± 0.0329	0.8147 ± 0.0445	0.7721 ± 0.0177	**0.7295 ± 0.1096**
Swin Transformer [40] + ours	**0.7493 ± 0.0634**	**0.8537 ± 0.1012**	**0.7929 ± 0.0596**	0.7219 ± 0.0747
DenseNet121 [36]	0.7773 ± 0.0492	**0.8684 ± 0.1194**	0.8142 ± 0.0459	0.7660 ± 0.0632
DenseNet121 [36] + ours	**0.8197 ± 0.0587**	0.8412 ± 0.0310	**0.8443 ± 0.0406**	**0.8123 ± 0.1127**

design four strictly controlled variants. All variants share the same backbone and training configuration. Only the direction/order of cross-modal interaction and the output feature type are varied. Specifically, as shown in Fig. 5: (a) $T \to I$ (ours) directly leverages topological priors to guide image feature enhancement (topology-guided image enhancement) and outputs the enhanced $F^{T\to I}$, which is used as the input to the deeper network. (b) $I \to T$ enhances topology features F^T conditioned on image features F^I (image-guided topology enhancement), yielding the enhanced $F^{I\to T}$. (c) $I \leftrightarrow T$ performs bidirectional cross-modal enhancement, producing enhanced $F^{T\to I}$ and $F^{I\to T}$, which are then fused as $F^{T\to I} + F^{I\to T}$. (d) $I \to T \to I$ adopts a sequential strategy: it first updates topology representations via $I \to T$, and then uses the updated topology features to enhance image representations via $T \to I$, finally outputting the enhanced $F^{T\to I}$.

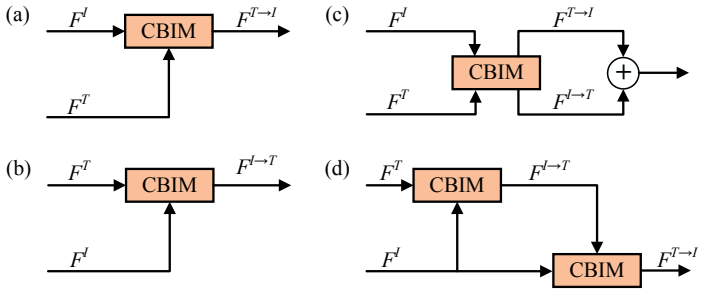


Fig. 5 Cross-modal enhancement designs of CBIM. (a) $T \to I$ (ours): topology-guided image enhancement, outputting $F^{T\to I}$. (b) $I \to T$: image-guided topology enhancement, outputting $F^{I\to T}$. (c) $I \leftrightarrow T$: bidirectional enhancement producing $F^{T\to I}$ and $F^{I\to T}$, which are fused by summation. (d) $I \to T \to I$: sequential enhancement that first performs $I \to T$ and then $T \to I$, outputting $F^{T\to I}$.

As shown in Table 7, $T \to I$ (ours) achieves the best performance on most key metrics (Accuracy 0.8197 ± 0.0587, Recall 0.8412 ± 0.0310, and F1-score 0.8443 ± 0.0406).

This suggests that, for this task, using topological structural priors to constrain and refine image representations helps the model learn more discriminative and stable features, leading to consistent improvements in overall classification performance. Notably, although the bidirectional design $I \leftrightarrow T$ employs fused outputs $F^{T\to I}+F^{I\to T}$ and achieves a higher AUC (0.8431 ± 0.0449), it underperforms $T \to I$ (ours) across Accuracy, Recall, and F1-score. This suggests that the gains of our method are not driven by increased capacity from naive feature concatenation, but rather stem from effective selection and enhancement of image features under topological guidance, enabling stronger generalization with a less redundant output. Overall, these results confirm that the direction and order of cross-modal interaction are critical to TopoNet performance, and further demonstrate the effectiveness of the proposed enhancement strategy for volumetric classification.

Table 7 Ablation study of CBIM cross-modal enhancement designs. Best results are highlighted in **bold**.

Cross-modal design	Output feature	Accuracy ↑	Recall ↑	F1-score ↑	AUC ↑
$I \to T$	$F^{I\to T}$	0.7919 ± 0.0480	0.8059 ± 0.0889	0.8150 ± 0.0385	0.8010 ± 0.0684
$I \leftrightarrow T$	$F^{I\to T} + F^{T\to I}$	0.7988 ± 0.0113	0.8406 ± 0.0813	0.8253 ± 0.0194	**0.8431 ± 0.0449**
$I \to T \to I$	$F^{T\to I}$	0.7640 ± 0.0587	0.8176 ± 0.1087	0.7994 ± 0.0282	0.7756 ± 0.0502
$T \to I$ (ours)	$F^{T\to I}$	**0.8197 ± 0.0587**	**0.8412 ± 0.0310**	**0.8443 ± 0.0406**	0.8123 ± 0.1127

3) Discussion on different input modalities. To investigate the impact of different input-modality configurations on TopoNet representation learning, and to further assess whether complementary modality fusion yields consistent gains, we construct five input settings: single-modality *img*, single-modality *topo*, homogeneous dual-branch *img/img*, homogeneous dual-branch *topo/topo*, and complementary dual-modality *img/topo* (ours). For fair comparison, the image branch is activated for the *img* input, whereas the topology branch is activated for the *topo* input. For dual-branch configurations, modality-specific enhancement modules are applied within their corresponding branches (e.g., the texture enhancement module for the image branch and the topology enhancement module for the topology branch), thereby avoiding confounding factors introduced by changes in network capacity or additional components.

As shown in Table 8, the complementary *img/topo* fusion achieves the best performance in Accuracy, F1-score, and AUC. Although the *img/img* setting yields a higher mean Recall, it suffers from significantly higher variance (± 0.0712) and lower overall accuracy, suggesting a tendency to over-predict positives. In contrast, our *img/topo* fusion maintains a highly competitive and much more stable Recall (0.8412 ± 0.0310), demonstrating that cross-modality interaction substantially strengthens discriminative representations and predictive stability. In contrast, homogeneous dual-branch settings (*img/img* or *topo/topo*) provide only limited improvements, suggesting that duplicated modalities introduce redundancy and fail to form effective complementarity. Overall, image feature (texture or intensity) and topology-derived structural

priors are inherently complementary; their integration better captures lesion phenotype variations, achieving a more robust balance between missed detections and false alarms, and consequently improving overall classification performance. This ablation study quantitatively confirms that cross-modal complementary fusion is a primary contributor to the performance gains of TopoNet.

Table 8 Ablation study on different input modality. Best results are in **bold**.

Input modalities	Accuracy ↑	Recall ↑	F1-score ↑	AUC ↑
img	0.7773 ± 0.0492	0.8684 ± 0.1194	0.8142 ± 0.0459	0.7660 ± 0.0632
topo	0.7426 ± 0.0760	0.7574 ± 0.1142	0.7680 ± 0.0700	0.7035 ± 0.0800
img/img	0.7852 ± 0.0441	**0.8904 ± 0.0712**	0.8259 ± 0.0290	0.7644 ± 0.0572
topo/topo	0.7357 ± 0.0492	0.7941 ± 0.1048	0.7723 ± 0.0442	0.7441 ± 0.0679
img/topo (ours)	**0.8197 ± 0.0587**	0.8412 ± 0.0310	**0.8443 ± 0.0406**	**0.8123 ± 0.1127**

4) Importance of components and stage-wise connections. We validate the effectiveness of each architectural component and the stage-wise connection strategy in TopoNet through comprehensive ablation experiments, with results reported in Table 9. Here, TEM, TpEM, CAM, and AP denote the texture enhancement module, topology enhancement module, cross-attention module, and attention pooling, respectively. Unless otherwise specified, all experimental settings are kept identical, and we use the full model with TEM + TpEM + CAM + AP and fusion at Stage 3 + 4 (ours) as the reference configuration.

In the component ablation study (fusion stages fixed to Stage 3 + 4), removing any single module generally degrades the overall balance of classification performance, indicating that each component contributes meaningfully. In particular, discarding TEM leads to the most pronounced performance drop in Recall (0.7662 ± 0.1283), highlighting the importance of texture enhancement for maintaining high sensitivity to lesions; although its AUC marginally increases to 0.8225, such a severe loss in Recall is highly undesirable for clinical diagnosis. Removing CAM yields the largest decrease in Accuracy (0.7712), F1-score (0.7961), and AUC (0.7745), demonstrating that cross-attention is the most crucial component for cross-modal alignment and effective information interaction. Removing TpEM also results in clear decreases in Accuracy and F1-score, confirming that topology enhancement provides complementary structural priors beyond appearance cues. Finally, removing AP degrades Accuracy and F1-score, suggesting that attentive aggregation is critical for stable decision-making.

In the stage-connection ablation (components fixed to full TEM + TpEM + CAM + AP), we compare three fusion schemes: Stage 4 only, Stage 3 + 4 (ours), and Stage 2 + 3 + 4. The results show that Stage 3 + 4 achieves the best overall stability, specifically delivering the highest Accuracy and F1-score. Restricting fusion to Stage 4 yields a higher mean Recall (0.8669), but it introduces extreme variance (± 0.1195) and weakens mid-to-high-level semantic interaction, limiting overall benefits in F1-score and AUC. In contrast, enabling fusion as early as Stage 2 + 3 + 4 leads to a

clear performance drop, suggesting that overly early cross-modal fusion may introduce noise or disrupt the hierarchical abstraction of features. Overall, Table 9 quantitatively confirms that the synergistic design of TEM/TpEM/CAM/AP and cross-modal interaction at mid-to-high stages (Stage 3 + 4) are key factors driving the superior performance of TopoNet.

Table 9 Importance of components and stage-wise connections. Best results are in **bold**.

Components				Fusion stages	Metrics			
TEM	TpEM	CAM	AP		Accuracy ↑	Recall ↑	F1-score ↑	AUC ↑
(I) Component ablation (fusion stages fixed to Stage 3 + 4)								
-	✓	✓	✓	3+4	0.7990 ± 0.0444	0.7662 ± 0.1283	0.8072 ± 0.0600	**0.8225 ± 0.0891**
✓	-	✓	✓	3+4	0.7919 ± 0.0570	0.8191 ± 0.1062	0.8159 ± 0.0519	0.7908 ± 0.0480
✓	✓	-	✓	3+4	0.7712 ± 0.0333	0.7912 ± 0.0850	0.7961 ± 0.0359	0.7745 ± 0.0588
✓	✓	✓	-	3+4	0.7852 ± 0.0581	0.8662 ± 0.0450	0.8219 ± 0.0457	0.7961 ± 0.0493
(II) Stage connection ablation (components fixed to full TEM + TpEM + CAM + AP)								
✓	✓	✓	✓	2+3+4	0.8057 ± 0.0556	0.7824 ± 0.1033	0.8215 ± 0.0417	0.8104 ± 0.1001
✓	✓	✓	✓	3+4 (ours)	**0.8197 ± 0.0587**	0.8412 ± 0.0310	**0.8443 ± 0.0406**	0.8123 ± 0.1127
✓	✓	✓	✓	4	0.8123 ± 0.0422	**0.8669 ± 0.1195**	0.8162 ± 0.0454	0.8039 ± 0.0691

4.6 Visualization

We employ Gradient-weighted Class Activation Mapping (Grad-CAM) [41] to visualize the evidence driving PDAC versus non-PDAC decisions and to localize the most contributive regions. Grad-CAM computes channel-wise importance by globally averaging the gradients of the target class score with respect to a selected convolutional feature map, and then forms a class-discriminative activation map via a weighted sum of feature channels. We apply Grad-CAM after the last stage (stage 4) of the image-branch backbone to obtain activation maps that are closely coupled to the final prediction. In Fig. 6, column 1 shows the image-branch input (pancreatic ROI), and column 2 shows the topology-branch input, namely statistically significant topological structures extracted via persistent homology and confidence-band filtering. Columns 3–5 present Grad-CAM heatmaps from DenseNet121 [36], SEResNeXt50 [37], and SENet154 [37], respectively, while column 6 corresponds to TopoNet. We highlight two challenging cases in which all three strong baselines misclassify PDAC as non-PDAC, whereas TopoNet correctly predicts PDAC, enabling a direct comparison of where different models attend.

For case 1 (first row), baseline activations are relatively diffuse, spreading across broad parenchymal regions or along organ boundaries, and they do not consistently emphasize duct-related structural information. In contrast, TopoNet exhibits a more coherent, structurally oriented high-response band within the pancreas, which aligns well with the salient topological structure. Notably, the extracted topological structure delineates a tubular pattern traversing the pancreas, consistent with loop-like patterns

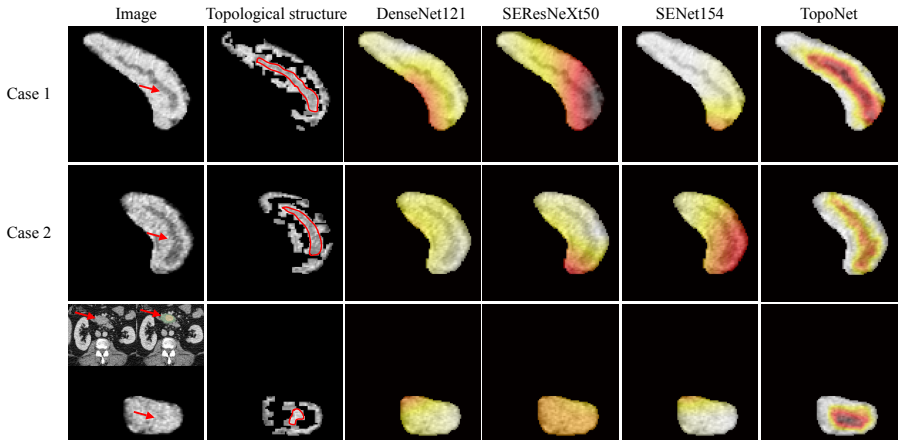


Fig. 6 Grad-CAM comparisons across methods. Column 1 shows the pancreatic ROI, where red arrows mark radiologically suspicious regions. Column 2 shows the corresponding significant topological structures, where red contours outline a dominant duct-like structures (associated with H_1) and tumor lesion-related structures. Columns 3–6 present Grad-CAM heatmaps from DenseNet121, SEResNeXt50, SENet154, and TopoNet, respectively. In both cases, TopoNet yields more coherent activations that better align with the topological structures. The bottom row additionally includes a tumor-focused example with CT and annotation for reference.

captured by the first homology group H_1, which often manifests as duct enclosure and connectivity in volumetric data. This is clinically plausible because PDAC commonly induces duct-related alterations on CT, including main pancreatic duct dilatation, abrupt ductal cutoff, and distal pancreatic atrophy. By explicitly incorporating such structural priors, TopoNet is better positioned to capture disease-relevant duct morphology, reducing errors when appearance contrast is subtle or confounded by noise. A similar phenomenon is observed in case 2 (second row), where baseline heatmaps remain locally responsive but unstable and fail to form attention patterns consistent with the duct trajectory. TopoNet produces more contiguous and spatially consistent activations that agree with the topological structures, suggesting improved robustness under inter-case variability (e.g., anatomical differences and heterogeneous textures). In the tumor lesion region of case 2 (third row), we additionally show the raw CT image and the corresponding pancreas and tumor segmentation masks (top-left: raw CT; top-right: pancreas and tumor labels) to assess lesion coverage. The baselines show weaker or off-target responses around the tumor, whereas TopoNet more reliably highlights regions proximal to the annotated lesion, while remaining consistent with topology-emphasized structures.

Overall, Grad-CAM provides qualitative evidence consistent with the quantitative gains: TopoNet shifts attention from diffuse texture-driven cues toward more disease-relevant structural patterns, exhibiting strong agreement with salient topological structures (particularly duct-like H_1 patterns). These observations support that topology-guided dual-branch learning supplies stable structural priors, reduces reliance on incidental textures, and enhances discriminability, robustness, and interpretability for PDAC classification.

5 Conclusion

In conclusion, we presented TopoNet, a topology-guided dual-branch framework for volumetric medical image classification. Rather than treating topology as a global statistic, our approach fundamentally integrates explicit structural cues into the deep learning pipeline. By leveraging confidence-band selection and a novel dimension-aware topology-to-voxel mapping, TopoNet effectively resolves the "spatial detachment" issue of traditional TDA methods, identifying and localizing statistically significant features (H_0 / H_1 / H_2) even under substantial noise. Through an attention-based fusion mechanism, the framework promotes effective cross-modal interactions, mitigating the representational mismatch between appearance and topology. Extensive experiments demonstrate that TopoNet achieves consistent improvements over baselines, validating the effectiveness of restoring spatial context to topological analysis for robust 3D representation learning.

Despite these promising results, several limitations offer avenues for future research. First, our current design treats homology dimensions largely independently. Future work could employ Graph Neural Networks (GNNs) or hypergraphs to explicitly model the synergistic interactions between loops (H_1) and cavities (H_2), which may jointly characterize complex anatomical signatures. Second, as is common in medical imaging, performance is constrained by data availability. Broader validation on larger, multi-center cohorts is warranted to rigorously assess generalizability across diverse scanner protocols and domain shifts. Moving forward, we aim to investigate scalable, data-centric strategies to further enhance the interpretability and trustworthiness of topology-guided AI in clinical practice.

Funding. This work was supported by the National Natural Science Foundation of China [12271333].

Author contribution. [Jin Sun]: Conceptualization of this study, Methodology, Software, Writing-Original Draft. [Zaifang Zhang]: Conceptualization of this study, Writing-Review & Editing, Supervision. [Wenqi Tao]: Validation, Investigation & Editing. [Chenghao Shao]: Conceptualization of this study, Supervision. [Nan Gao]: Conceptualization of this study, Writing-Review

Data availability. The MSD Pancreas dataset is publicly available at http://medicaldecathlon.com/. The BraTS 2018 dataset can be requested via the CBICA Image Processing Portal at https://www.med.upenn.edu/cbica/brats/. The source code for TopoNet is available at https://github.com/hbqdcq/TopoNet.

Declarations

- **Competing interests** The authors declare that they have no conflicts of interest.
- **Ethics approval** This article does not contain any studies with human participants or animals performed by any of the authors.
- **Consent to participate** The authors have approved this article before submission, including the names and order of authors.
- **Consent for publication** The authors agreed with the content and gave explicit consent to submit.

References

[1] Lambin P, Leijenaar RT, Deist TM, et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology 14(12):749–762

[2] Maqsood S, Damaševičius R (2023) Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare. Neural networks 160:238–258

[3] Akter A, Nosheen N, Ahmed S, et al (2024) Robust clinical applicable cnn and u-net based algorithm for mri classification and segmentation for brain tumor. Expert Systems with Applications 238:122347

[4] Huang F, Qiu A, Initiative ADN, et al (2024) Ensemble vision transformer for dementia diagnosis. IEEE Journal of Biomedical and Health Informatics 28(9):5551–5561

[5] Li J, Zhang P, Wang T, et al (2023) Dsmt-net: dual self-supervised multi-operator transformation for multi-source endoscopic ultrasound diagnosis. IEEE Transactions on Medical Imaging 43(1):64–75

[6] Liu S, Wang L, Yue W (2024) An efficient medical image classification network based on multi-branch cnn, token grouping transformer and mixer mlp. Applied Soft Computing 153:111323

[7] Wu X, Feng Y, Xu H, et al (2023) Ctranscnn: Combining transformer and cnn in multilabel medical image classification. Knowledge-Based Systems 281:111030

[8] Fu B, Zhang M, He J, et al (2022) Stohisnet: A hybrid multi-classification model with cnn and transformer for gastric pathology images. Computer Methods and Programs in Biomedicine 221:106924

[9] Al-Hawary MM, Francis IR, Chari ST, et al (2014) Pancreatic ductal adenocarcinoma radiology reporting template: consensus statement of the society of abdominal radiology and the american pancreatic association. Radiology 270(1):248–260

[10] Carlsson G (2009) Topology and data. Bulletin of the American Mathematical Society 46(2):255–308

[11] Edelsbrunner H, Harer J (2010) Computational topology: an introduction. American Mathematical Soc.

[12] Edelsbrunner H, Harer J, et al (2008) Persistent homology-a survey. Contemporary mathematics 453(26):257–282

[13] Wang F, Kapse S, Liu S, et al (2021) Topotxr: a topological biomarker for predicting treatment response in breast cancer. In: International Conference on Information Processing in Medical Imaging, Springer, pp 386–397

[14] Du S, Lao Q, Kang Q, et al (2022) Distilling knowledge from topological representations for pathological complete response prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 56–65

[15] Saadat-Yazdi A, Andreeva R, Sarkar R (2021) Topological detection of alzheimer's disease using betti curves. In: International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, Springer, pp 119–128

[16] Asaad A, Ali D, Majeed T, et al (2022) Persistent homology for breast tumor classification using mammogram scans. Mathematics 10(21):4039

[17] Vandaele R, Mukherjee P, Selby HM, et al (2023) Topological data analysis of thoracic radiographic images shows improved radiomics-based lung tumor histology prediction. Patterns 4(1)

[18] Stolz BJ, Kaeppler J, Markelc B, et al (2022) Multiscale topology characterizes dynamic tumor vascular networks. Science Advances 8(23):eabm2456

[19] Peng Y, Wang H, Sonka M, et al (2024) Phg-net: Persistent homology guided medical image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 7583–7592

[20] Songdechakraiwut T, Krause BM, Banks MI, et al (2023) Wasserstein distance-preserving vector space of persistent homology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 277–286

[21] Lee H, Kang H, Chung MK, et al (2012) Persistent brain network homology from the perspective of dendrogram. IEEE transactions on medical imaging 31(12):2267–2277

[22] Clough JR, Byrne N, Oksuz I, et al (2020) A topological loss function for deep-learning based image segmentation using persistent homology. IEEE transactions on pattern analysis and machine intelligence 44(12):8766–8778

[23] Hu X, Li F, Samaras D, et al (2019) Topology-preserving deep image segmentation. Advances in neural information processing systems 32

[24] Dey TK, Li K, Sun J, et al (2008) Computing geometry-aware handle and tunnel loops in 3d models. In: ACM SIGGRAPH 2008 papers. p 1–9

[25] Hu CS, Lawson A, Chen JS, et al (2021) Toporesnet: A hybrid deep learning architecture and its application to skin lesion classification. Mathematics 9(22):2924

[26] Fasy BT, Lecci F, Rinaldo A, et al (2014) Confidence sets for persistence diagrams. The Annals of Statistics 42(6):2301–2339

[27] Qiu S (2018) Global weighted average pooling bridges pixel-level localization and image-level classification. arXiv preprint arXiv:180908264

[28] Simpson AL, Antonelli M, Bakas S, et al (2019) A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:190209063

[29] Menze BH, Jakab A, Bauer S, et al (2014) The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging 34(10):1993–2024

[30] Bakas S, Akbari H, Sotiras A, et al (2017) Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4(1):1–13

[31] Bakas S, Reyes M, Jakab A, et al (2018) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:181102629

[32] Cardoso MJ, Li W, Brown R, et al (2022) Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:221102701

[33] Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:171105101

[34] Maria C, Boissonnat JD, Glisse M, et al (2014) The gudhi library: Simplicial complexes and persistent homology. In: International congress on mathematical software, Springer, pp 167–174

[35] He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

[36] Huang G, Liu Z, Van Der Maaten L, et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708

[37] Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141

[38] Tan M, Le QE, et al (2019) Rethinking model scaling for convolutional neural networks. In: Proceedings of the International conference on machine learning, Long Beach, CA, USA

[39] Dosovitskiy A (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929

[40] Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022

[41] Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626