Supplementary Material for: TopoCL: Topological Contrastive Learning for Medical Imaging

Guangyu Meng, Pengfei Gu, Peixian Liang, John P. Lalor, Erin Wolf Chambers, Danny Z. Chen

This supplementary material provides additional details and experimental results to support the main paper. It is organized as follows:

- Section 1: Complete details on topology-aware augmentation design and per-dataset configurations
- Section 2: Full implementation details and training hyperparameters for all methods
- Section 3: Expert gating analysis for all contrastive learning methods
- Section 4: Additional experimental results and statistical analysis

1. Topology-Aware Augmentation Design

In the main paper (Section 3.2), we introduce topology-aware augmentations that control topological perturbations using relative bottleneck distance. Here we provide complete details on the augmentation configurations and the experimental methodology for determining optimal d_B^{rel} ranges.

1.1. Experimental Methodology

Table 4 in the main paper demonstrates that topology-aware augmentations with weak-strong pairs on SAM-extracted ROIs achieve optimal performance. Here we provide the complete ablation study that determined the optimal d_B^{rel} ranges (5-15% weak, 15-25% strong) mentioned in Section 3.2.

To determine optimal d_B^{rel} ranges, we conduct systematic experiments using MoCo-v3+TopoCL across three representative datasets: OrganSMNIST (abdominal CT), ISIC2019 (dermoscopy), and Kvasir (endoscopy). We evaluate d_B^{rel} range configurations: (1) Config 1 (0-10%, 10-20%): Conservative ranges testing whether minimal perturbations suffice; (2) Config 2 (5-15%, 15-25%): Our proposed ranges balancing structural preservation and diversity; (3) Config 3 (10-20%, 20-30%): Aggressive ranges testing whether stronger perturbations improve diversity; (4) Config 4 (0-20%, 20-40%): Very wide ranges as baseline.

For each configuration, we train MoCo-v3+TopoCL for 150 epochs followed by 100 epochs of linear probe evaluation. All experiments use ResNet-50, batch size 256, AdamW optimizer (learning rate 3×10^{-4}), and cosine annealing. We report mean accuracy and standard deviation over five independent runs.

Table 1. Impact of d_B^{rel} range selection on linear probe accuracy (%) using MoCo-v3+TopoCL.

Config	d_B^{rel} Range (%)		Linear Probe Accuracy (%)		
	Weak	Strong	OrganS	ISIC	Kvasir
1	0-10	10-20	$77.48_{\pm 0.51}$	$75.88_{\pm 0.39}$	$89.49_{\pm 0.44}$
2	**5-15**	**15-25**	$\mathbf{80.58}_{\pm 0.09}$	$\mathbf{78.44}_{\pm 0.83}$	$\mathbf{91.17}_{\pm 0.83}$
3	10-20	20-30	$80.01_{\pm 0.11}$	$77.92_{\pm 0.45}$	$90.76_{\pm 0.38}$
4	0-20	20-40	$73.61_{\pm 0.78}$	$71.56_{\pm 1.03}$	$86.55_{\pm 0.95}$

1.2. Results and Analysis

Table 1 presents the impact of different d_B^{rel} range configurations on linear probe accuracy.

Config 2 (5-15%, 15-25%) consistently achieves the best performance across all three datasets. Config 1 provides insufficient topological diversity, limiting robust representation learning. Config 3 applies excessive perturbations that destroy diagnostically relevant structures such as lesion boundaries or organ connectivity. Config 4 with very wide ranges exhibits high variance as extreme augmentations no longer preserve semantic identity. The consistent superiority of Config 2 validates our choice of 5-15% weak and 15-25% strong ranges, striking optimal balance between structural preservation and diversity.

1.3. Per-Dataset Augmentation Configurations

Table 2 lists representative augmentation operation combinations and parameter ranges for achieving optimal d_B^{rel} ranges on each dataset. During training, we randomly select from multiple validated configurations (not exhaustively listed) and sample operation parameters from specified ranges.

Implementation Details: During training, operation parameters are uniformly sampled from ranges in Table 2. For example, when applying topology-weak augmentation on OrganSMNIST, we apply Flip, sample Gaussian noise $\sigma \sim \text{Uniform}[0.04, 0.12]$, and sample brightness $\beta \sim$

Table 2. Representative augmentation operations and parameter ranges used to achieve optimal d_B^{rel} ranges for each dataset.

Dataset	Topology-Weak (5-15%)	Topology-Strong (15-25%)
PathMNIST	Flip, GaussianNoise ($\sigma \in [0.03, 0.08]$), Contrast ($\alpha \in [0.05, 0.15]$)	Flip, GaussianNoise ($\sigma \in [0.12, 0.20]$), GaussianBlur ($\sigma \in [1.5, 2.5]$), Contrast ($\alpha \in [0.15, 0.30]$)
OCTMNIST	Flip, Rotation ($\theta \in [5°, 15°]$), Contrast ($\alpha \in [0.08, 0.18]$)	Flip, GaussianBlur ($\sigma \in [2.0, 3.0]$), Dilation (kernel=3), Brightness ($\beta \in [-0.15, 0.15]$)
OrganSMNIST	Flip, GaussianNoise ($\sigma \in [0.04, 0.12]$), Brightness ($\beta \in [-0.10, 0.10]$)	Flip, GaussianNoise ($\sigma \in [0.13, 0.23]$), Contrast ($\alpha \in [0.20, 0.35]$), GaussianBlur ($\sigma \in [1.5, 2.5]$)
ISIC2019	Flip, GaussianNoise ($\sigma \in [0.05, 0.15]$), Contrast ($\alpha \in [0.10, 0.20]$)	Flip, GaussianNoise ($\sigma \in [0.15, 0.25]$), Erosion (kernel $\in [3, 5]$), Contrast ($\alpha \in [0.20, 0.35]$)
Kvasir	Flip, GaussianNoise ($\sigma \in [0.04, 0.10]$), Brightness ($\beta \in [-0.08, 0.08]$)	Flip, GaussianNoise ($\sigma \in [0.15, 0.22]$), GaussianBlur ($\sigma \in [1.5, 2.5]$), Dilation (kernel=3)

Uniform$[-0.10, 0.10]$. This stochastic sampling ensures augmentations fall within target d_B^{rel} ranges while providing diversity across training iterations. Different modalities require tailored strategies: OrganSMNIST (abdominal CT) uses subtle noise and brightness to preserve organ boundaries; ISIC2019 (dermoscopy) tolerates morphological operations like erosion due to natural boundary variations; Kvasir (endoscopy) balances noise and blur for mucosal texture; OCTMNIST (retinal OCT) uses rotation with controlled blurring for boundary-sensitive structures; PathMNIST (histopathology) uses subtle noise and contrast for cellular structure preservation.

1.4. Augmentation Operation Specifications

All operations are implemented using standard image processing libraries:

- **Flip**: Horizontal and/or vertical flip with 50% probability each.
- **Rotation**: Random rotation within specified angle range using bilinear interpolation.
- **GaussianNoise**: Additive Gaussian noise with standard deviation σ, applied as $I' = I + \mathcal{N}(0, \sigma^2)$.
- **GaussianBlur**: Gaussian blur with kernel size $k = \lceil 6\sigma \rceil + 1$ (ensuring odd kernel size).
- **Contrast**: Linear contrast adjustment with factor α, applied as $I' = \alpha \cdot (I - \mu) + \mu$ where μ is mean intensity.
- **Brightness**: Additive brightness adjustment with factor β, applied as $I' = \text{clip}(I + \beta, 0, 1)$.
- **Erosion**: Binary morphological erosion with square kernel of specified size, applied after Otsu thresholding.
- **Dilation**: Binary morphological dilation with square kernel of specified size, applied after Otsu thresholding.

For ROI extraction, we apply SAM-ViT-H [6] to automatically identify foreground regions containing anatomically relevant structures. Persistence diagrams are computed on these ROIs using GUDHI [4] with sublevel set filtration based on grayscale intensity values. During training, augmentations are applied to full images, but d_B^{rel} is implicitly controlled through validated parameter ranges in Table 2 rather than computed online.

2. Complete Implementation Details

This section provides key implementation details for reproducibility. We follow standard configurations for all baseline contrastive learning methods [1–3, 5, 7], and describe TopoCL-specific components below.

2.1. TopoCL Architecture

Hierarchical Topology Encoder: The PH Encoder uses 4 FC layers (4→64→128→256→384) to process top-k persistent features ($k_{H_0} = 48$, $k_{H_1} = 96$). Self-attention and bidirectional cross-attention employ 4 heads with dimension 384 and weighted residuals ($\lambda_0 = \lambda_1 = 0.5$). Features are aggregated via max and mean pooling (6×384→2304), followed by a 3-layer projection MLP (2304→768→512→256).

MoE Fusion Module: Visual (ResNet-50 output: 2048-dim) and topological features (256-dim) are projected to 256 dimensions via separate 3-layer MLPs. Five expert networks (each 512→384→256→512) process concatenated features. A multi-gating network (512→256→128→5 + Softmax) computes sample-specific weights. The final projection MLP (512→384→256→256) produces fused representations for the contrastive learning objective.

2.2. Training Configuration

All methods use ResNet-50 as the visual encoder, batch size 256, AdamW optimizer (learning rate 3×10^{-4}, weight decay 0.05), cosine annealing with 10-epoch linear warmup, and 150 pretraining + 100 linear probe epochs. For linear

probing, we freeze encoder weights and train a linear classifier using SGD (momentum 0.9, learning rate 0.1 with cosine decay) for 100 epochs without data augmentation.

Implementation: Experiments run on NVIDIA H100 (80GB) GPUs using PyTorch 2.0 and CUDA 11.8. SAM-ViT-H [6] extracts ROIs, and GUDHI [4] computes persistence diagrams. We use five random seeds (42, 123, 456, 789, 1024) for all reported results. Code is available at https://github.com/gm3g11/TopoCL.

3. Expert Gating Analysis for All Methods

In the main paper Section 4.5, we present expert gating analysis for BYOL+TopoCL. Here we provide complete gating patterns for the remaining four contrastive learning methods: SimCLR, MoCo-v3, DINO, and Barlow Twins.

Figure 1 reveals distinct gating patterns across methods and datasets, demonstrating TopoCL's adaptive fusion strategy.

Method-Specific Patterns. Each contrastive learning method exhibits characteristic expert preferences aligned with its learning objective. SimCLR shows relatively balanced expert utilization across all five experts, with no single expert dominating, consistent with its symmetric contrastive loss that benefits from diverse fusion strategies. MoCo-v3 demonstrates more pronounced dataset-specific variation, with certain datasets strongly preferring gated blending while others favor concatenation or cross-attention experts, indicating that momentum-based methods adaptively select fusion strategies based on image characteristics.

DINO exhibits the most distinctive gating pattern, with notably elevated cross-attention expert weights across multiple datasets. This strong preference for cross-modal interaction aligns with DINO's self-distillation objective that emphasizes learning rich cross-view relationships. The consistently high cross-attention activation suggests that self-distillation benefits particularly from explicit visual-topological feature interaction rather than simple concatenation. Barlow Twins displays more uniform expert weight distributions, with all five experts receiving moderate activation. This balanced pattern is consistent with its redundancy-reduction objective that encourages diverse, decorrelated feature representations.

Dataset-Specific Patterns. Certain datasets consistently activate specific experts across multiple methods. ISIC2019 shows elevated topology-only expert weights in several methods, confirming that persistent homology effectively captures discriminative skin lesion boundary features for dermoscopy classification. The topology-only preference is most pronounced in BYOL (as shown in the main paper) but remains visible across other methods, validating the importance of boundary topology for this modality.

PathMNIST consistently activates cross-attention and gated experts across methods, indicating that histopathology images benefit from explicit integration of visual appearance and topological structure to capture cellular organization patterns. OCTMNIST and OrganSMNIST exhibit more method-dependent patterns, with expert preferences varying substantially between SimCLR, MoCo-v3, and DINO. This suggests that retinal OCT and abdominal CT modalities require different fusion strategies depending on the contrastive learning objective. Kvasir demonstrates relatively stable expert distributions across methods, with balanced activation of concatenation and visual-only experts, indicating that endoscopy images may rely more heavily on visual texture features with modest topological augmentation.

Key Observations. Three important patterns emerge from cross-method analysis. First, topology-only experts generally receive lower weights compared to fusion-based experts (concatenation, gated, cross-attention) across most method-dataset combinations. This indicates that topological features are most effective when integrated with visual features rather than used in isolation, validating our MoE fusion design. Second, fusion-based experts collectively dominate the gating weights across all methods, with their combined weights substantially exceeding those of single-modality experts. This confirms that adaptive integration of visual and topological modalities outperforms approaches that rely on a single feature type.

Third, the substantial variation in expert preferences across methods demonstrates that our MoE architecture successfully adapts to different contrastive learning objectives. The multi-gating network learns these method-specific and dataset-specific preferences without supervision, automatically determining optimal fusion strategies for each combination. This adaptability explains how TopoCL achieves consistent performance improvements across diverse methods (Table 3 in main paper): rather than imposing a fixed fusion strategy, the MoE module learns to combine features in ways tailored to each method's learning dynamics and each dataset's structural characteristics.

4. Additional Experimental Results

4.1. Complete Statistical Significance Analysis

Table 3 provides complete p-values for all 50 comparisons (5 methods × 5 datasets × 2 metrics) from the main paper Table 3, assessed using paired t-tests across five independent runs.

Summary: Of 50 total comparisons, 43 (86%) achieve statistical significance at $p < 0.05$, with 80% (8/10) of dataset-averaged metrics reaching p¡0.001. Notably, all accuracy improvements on PathMNIST, OrganSMNIST, and OCTMNIST achieve $p < 0.01$, demonstrating particularly strong and consistent benefits on datasets with structured

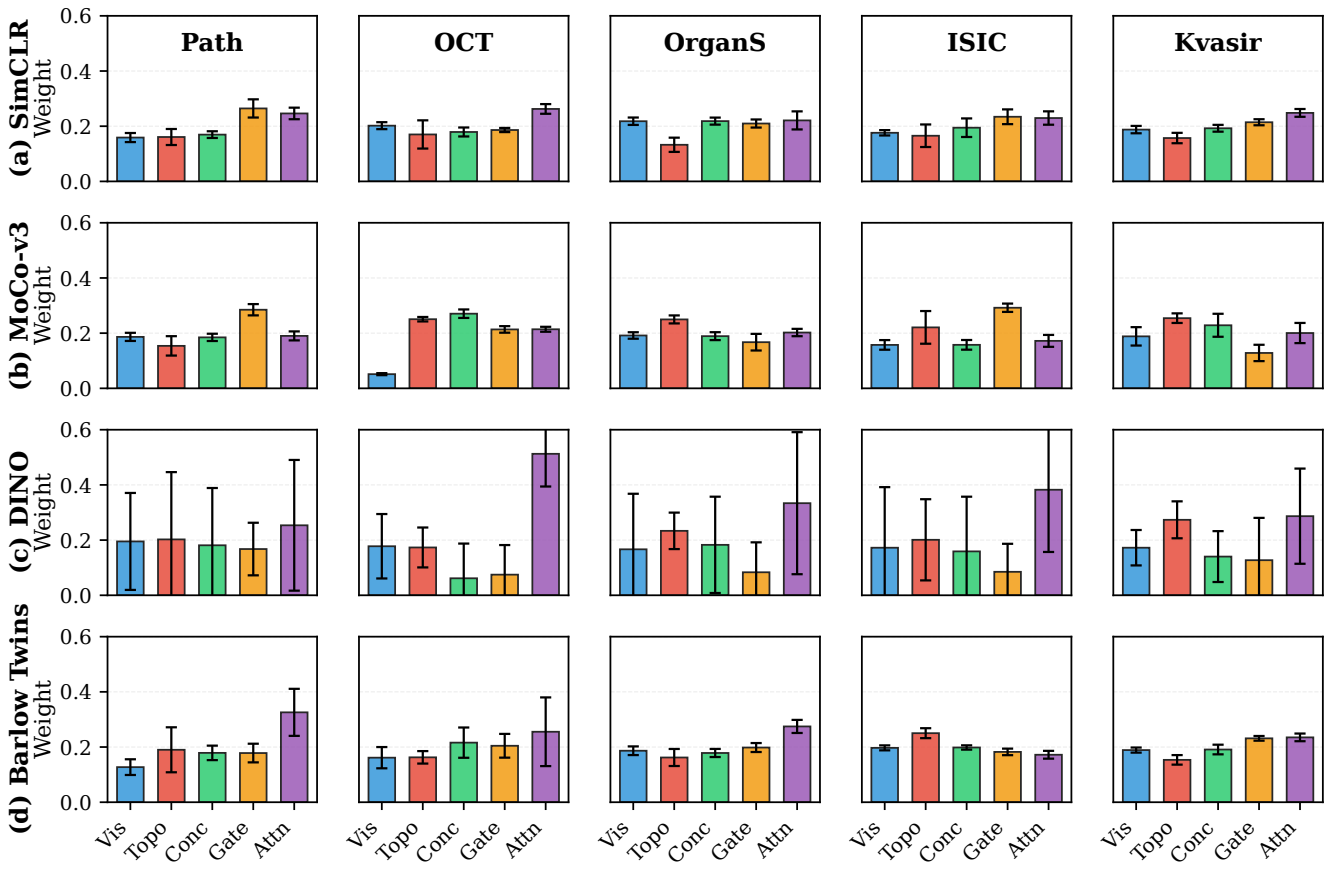

Figure 1. Expert gating analysis for TopoCL integrated with (a) SimCLR, (b) MoCo-v3, (c) DINO, and (d) Barlow Twins across five datasets (Path, OCT, OrganS, ISIC, Kvasir from left to right). BYOL results are shown in the main paper (Figure 5). Error bars show standard deviation across test samples. The five experts are: visual-only (Vis), topology-only (Topo), concatenation (Conc), gated blending (Gate), and cross-attention (Attn). Method-specific gating preferences demonstrate the adaptive nature of our MoE fusion module across diverse contrastive learning objectives.

Table 3. Complete p-values for all TopoCL improvements over baseline methods. Values < 0.001 are shown as <0.001. Bold indicates $p < 0.05$.

Method	Metric	p-value				
		Path	OrganS	OCT	ISIC	Kvasir
SimCLR	ACC	**0.042**	**<0.001**	**<0.001**	**0.008**	**0.003**
	AUC	0.112	**0.019**	**0.006**	**0.004**	**0.021**
MoCo-v3	ACC	**0.003**	**<0.001**	**<0.001**	**0.037**	**0.022**
	AUC	**0.007**	**0.023**	**<0.001**	0.385	0.158
BYOL	ACC	**<0.001**	**<0.001**	**<0.001**	0.097	**0.005**
	AUC	0.241	**0.002**	**0.018**	0.174	**0.031**
DINO	ACC	**<0.001**	**<0.001**	**<0.001**	**0.029**	**0.004**
	AUC	**0.003**	**<0.001**	**<0.001**	**0.002**	**0.009**
Barlow	ACC	**<0.001**	**<0.001**	**<0.001**	**0.006**	**0.008**
	AUC	0.076	**<0.001**	**0.019**	**0.002**	**0.034**

anatomical features. The few non-significant results (7/50) primarily occur on AUC metrics for high-performing baselines (e.g., MoCo-v3 on Kvasir: 97.37% baseline), where further improvements face saturation effects. This comprehensive statistical evidence validates TopoCL's effectiveness across diverse contrastive learning methods and medical imaging modalities.

4.2. Ablation on Number of Persistent Features

Table 4 ablates the number of persistent features (k_{H_0} for connected components, k_{H_1} for holes) used in the Hierarchical Topology Encoder. Our default configuration ($k_{H_0} = 48$, $k_{H_1} = 96$) balances expressiveness and efficiency.

Performance saturates at our default configuration, indicating that top-48 H_0 and top-96 H_1 features capture the most persistent topological structures. Additional features primarily represent noise or unstable topological signals that do not improve representation learning. Larger configurations also increase computational cost without benefit: (144,288) requires 3× memory and 2.1× training time compared to (48,96) while achieving lower accuracy. This validates our choice based on preliminary analysis of per-

Table 4. Ablation on number of persistent features using MoCo-v3+TopoCL on PathMNIST. ACC (%) reported.

k_{H_0}	k_{H_1}	**Accuracy**
24	48	$93.82_{\pm 0.31}$
48	**96**	$\mathbf{94.55}_{\pm 0.20}$
96	192	$94.48_{\pm 0.28}$
144	288	$94.42_{\pm 0.34}$

sistence diagram distributions across all datasets.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pages 1597–1607, 2020.

[3] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[4] Pawel Dlotko. Cubical complex. In *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.11.0 edition, 2025.

[5] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap Your Own Latent: A new approach to self-supervised learning. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 21271–21284, 2020.

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[7] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the International Conference on Machine Learning*, pages 12310–12320, 2021.