CROSS-MODAL RELATIONAL DISTILLATION FOR WHITE-LIGHT ENDOSCOPIC GASTROINTESTINAL LESION CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

White-light imaging (WLI) is widely used for colorectal polyp assessment, whereas narrow-band imaging (NBI) reveals vascular and mucosal details but requires dedicated imaging equipment. Existing cross-modal distillation methods use verified WLI–NBI pairs for direct transfer or unpaired images for group-level supervision, leaving the relational information accessible through verified pairs underexplored. We propose Paired-anchor Cross-modal Transfer (PaCT), which uses verified pairs as shared identity anchors to transfer NBI relations to unpaired WLI images. PaCT builds an NBI teacher graph from paired and additional unpaired NBI images. For each unpaired WLI query, its affinities to WLI anchors are mapped through the verified identities to obtain a query-specific relational target from the NBI graph, without assuming a match to any unpaired NBI image. Under a common three-seed, five-fold evaluation protocol, PaCT improves mean accuracy, F1-score, and AUC over the best comparator for each metric by 2.48, 2.64, and 2.71 percentage points on PICCOLO, and by 1.93, 2.13, and 2.82 percentage points on PolypSet, respectively. PaCT thus extends the utility of verified pairs beyond paired examples while requiring only WLI at inference. Code is available at `https://anonymous.4open.science/r/PaCT-iclr2027-code/`.

1 INTRODUCTION

Colonoscopy plays a central role in colorectal cancer screening and in the assessment of suspicious lesions. In routine practice, however, image interpretation depends on specialist experience and imposes a substantial reading burden, with variation in the visual evidence available across examinations. These demands have motivated automated systems for polyp detection and classification (Zhang et al., 2017; Schmitz et al., 2022). White-light imaging (WLI) is the widely available modality used in routine colonoscopy, whereas narrow-band imaging (NBI) selectively enhances vascular and mucosal surface patterns that can provide additional diagnostic cues (Yue et al., 2023). Access to NBI is less widespread than to routine WLI (Wang et al., 2021b), making it valuable to exploit NBI during training while retaining a classifier that operates on WLI alone.

How can a WLI classifier learn from NBI without requiring it at inference? One direct answer is to use a verified WLI–NBI pair as the unit of transfer. The two views depict the same lesion, so the NBI view can provide a concrete cross-modal supervisory signal for its WLI counterpart. PolypsAlign performs domain alignment and contrastive learning between paired observations (Wang et al., 2021b); CPC-Trans enforces structured consistency between paired cross-modal representations (Ma et al., 2022); and ADD uses alignment-free dense interactions to transfer spatially distributed evidence between corresponding images (Hu et al., 2025). These methods demonstrate the value of instance-level correspondence: a known identity gives the learner a reliable route for deciding which NBI evidence belongs to which WLI observation.

The number of usable images, however, need not equal the number of verified cross-modal pairs. Establishing that two views depict the same lesion requires coordinated acquisition and curation, while clinical image collections can also contain WLI-only and NBI-only observations. Unpaired WLI images remain useful for ordinary class-supervised learning, but they do not have a corresponding NBI view from which to receive direct privileged supervision. Likewise, an unpaired NBI image

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

cannot be assigned to a particular WLI image without introducing an unverified match. Consequently, direct paired transfer confines cross-modal supervision to the paired subset, leaving an opportunity to extend NBI-derived supervision to WLI images outside that subset.

How can unpaired observations contribute to cross-modal learning? Auxiliary-modality learning provides a broader answer: a modality available during training can transfer knowledge to a target-modality model even when the auxiliary modality is absent at inference. Shen et al. study this setting with generalized curriculum distillation (Shen et al., 2023). In gastrointestinal endoscopy, PaGKD transfers NBI knowledge through class-conditioned prototypes and dense group relations, showing that WLI models can benefit from NBI without image-level pairing (Hu et al., 2026). We therefore focus on partially paired collections, where verified WLI–NBI identities coexist with unpaired WLI and NBI observations. Figure 1 summarizes this setting: paired observations provide strong cross-modal identities, whereas unpaired observations lack instance-level counterparts and expose weaker relationship structure.

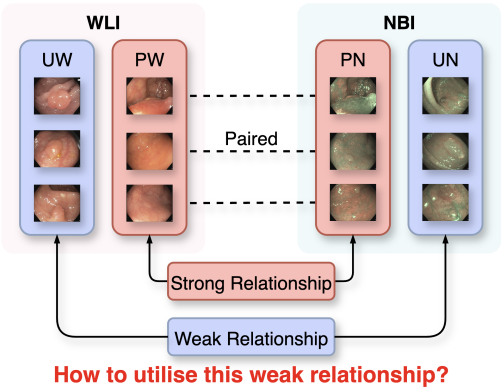

Figure 1: Motivation for PaCT.

These verified identities provide strong cross-modal connections through which additional observations can become useful to an individual WLI query. This motivates our central question: can paired identities serve as bridges for extending NBI-derived relational supervision to unpaired WLI images without fabricating cross-modal matches?

Our answer is Paired-anchor Cross-modal Transfer (PaCT). We use verified WLI–NBI pairs as shared anchor identities rather than restricting them to their own direct counterparts. A WLI query first expresses its relation to the paired WLI anchors. The shared identities transfer this query into the NBI anchor space, where the paired anchors are connected by relational structure built from both paired and unpaired NBI observations. Reading this structure produces a query-specific target distribution over the same known identities, which the WLI student learns together with its class objective. Thus, unpaired NBI contributes additional teacher-side relations, while the verified pairs preserve the semantic meaning of the cross-modal route. The construction transfers relational evidence without fabricating an NBI counterpart for any unpaired WLI image, and the trained predictor uses only WLI at deployment.

PaCT thereby extends the supervisory role of verified correspondences beyond the paired subset. Our experiments on PICCOLO and PolypSet evaluate the resulting WLI classification performance, while component ablations examine the contribution of relational transfer and the additional NBI observations.

Our contributions are summarized as follows.

- We study partially paired WLI–NBI classification through a new transfer route: verified pair identities serve as shared anchors that connect unpaired WLI queries to NBI relations without inventing cross-modal image matches.
- We propose PaCT, which builds a teacher graph from paired and unpaired NBI observations and reads its anchor-level relations through verified identities to produce query-specific distillation targets. The resulting classifier requires only WLI at inference.
- We evaluate PaCT on PICCOLO and PolypSet under a common three-seed, five-fold protocol. Relative to the best comparator for each metric, PaCT improves mean F1 by 2.64 and 2.13 percentage points and AUC by 2.71 and 2.82 points, respectively. Ablations separately support the roles of verified pair identities and additional unpaired NBI observations.

2 Related Work

2.1 WLI–NBI transfer with verified pairs

White-light imaging (WLI) is the deployment modality for endoscopic lesion classification, whereas narrow-band imaging (NBI) can provide complementary vascular and surface cues during training. This is an instance of learning with training-only auxiliary information: the learner can access NBI during training, while the deployed predictor operates on WLI alone (Vapnik & Vashist, 2009). In endoscopy, PolypsAlign trains an NBI teacher and a WLI student with domain alignment and contrastive learning, using verified WLI–NBI pairs for transfer (Wang et al., 2021b). CPC-Trans feeds a paired WLI–NBI example through a shared Transformer and aligns class tokens and spatial-attention responses (Ma et al., 2022). ADD handles spatial misalignment within corresponding WLI–NBI images by learning dense cross-domain affinities and filtering them with class activation maps (Hu et al., 2025). Thus, these methods treat a verified same-lesion pair as a unit for feature, attention, or dense evidence transfer. PaCT instead uses each verified identity as a reference through which NBI teacher neighborhoods can supervise an unpaired WLI query.

2.2 Pairing-free multimodal transfer

Pairing-free methods show how auxiliary modalities can strengthen a unimodal predictor when cross-modal identity is unavailable. Auxiliary-modality learning studies how an additional modality available during training can transfer knowledge to a target-modality model that operates without the auxiliary modality at inference. Shen et al. formulate this setting and develop generalized curriculum distillation as a general mechanism for such knowledge transfer (Shen et al., 2023). PaGKD transfers NBI knowledge from same-class groups with two complementary components: GKD-Pro extracts prototype representations with shared lesion-aware learned queries, while GKD-Den distills dense group relations from activation-derived maps (Hu et al., 2026). Better Together studies unpaired multimodal learning more generally: its UML framework alternates unpaired modalities through one shared model, allowing the representation of a target modality to benefit from auxiliary data without instance correspondence (Gupta et al., 2026). EvDistill provides a related cross-modal graph-distillation precedent in event vision, combining bidirectional modality reconstruction, distribution adaptation, and affinity-graph distillation for unpaired event data (Wang et al., 2021a). These approaches establish useful supervision regimes at the no-identity endpoint. PaCT addresses collections that retain both verified identities and additional unpaired NBI: identities determine how teacher-side relations are interpreted by WLI queries, while unpaired NBI enlarges the teacher-side topology.

2.3 Relational distillation and personalized propagation

Relational knowledge distillation transfers structure among examples rather than matching each teacher and student output independently; RKD instantiates this idea with distance-wise and angle-wise losses (Park et al., 2019). Graph Distillation transfers auxiliary multimodal information through graph structure in action videos with partially observed modalities (Luo et al., 2018). Personalized PageRank propagation supplies a fixed, adjustable-range mechanism for spreading information over a graph while retaining a node-specific starting distribution (Gasteiger et al., 2019).

PaCT combines these ingredients around partial WLI–NBI identity. It uses a frozen NBI teacher graph whose paired-anchor neighborhood is expanded by unpaired NBI nodes, while verified identities index how an unpaired WLI query's anchor affinities are interpreted on the NBI side. The student matches this query-specific teacher distribution through direct affinities to WLI anchors; the target retains class mass and relative preferences over individual anchors. The graph constructs fixed training targets; inference uses the WLI student alone. PaCT therefore transfers teacher-side structural information through known identities while allowing graph coverage to expand with additional NBI.

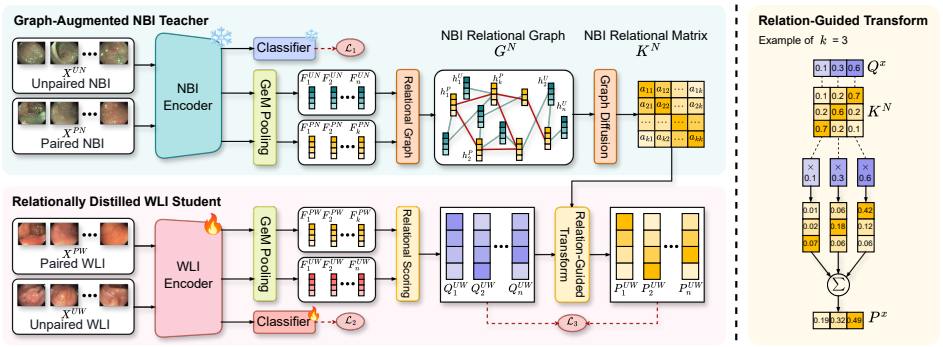

Figure 2: Overview of PaCT. The frozen NBI teacher builds a relational matrix from paired and unpaired observations. Verified pairs connect this structure to WLI student queries, yielding anchor-indexed distillation targets, as illustrated on the right. Snowflakes and flames denote frozen and trainable components, respectively.

3 Method

3.1 Preliminaries

We study colorectal polyp classification from WLI, using NBI during training only. Superscripts P/U denote paired/unpaired observations and W/N denote the modalities. The training images comprise

$$\begin{aligned} X^{PW} &= \{x_i^{PW}\}_{i=1}^{K}, \quad X^{PN} = \{x_i^{PN}\}_{i=1}^{K}, \\ X^{UW} &= \{x_j^{UW}\}_{j=1}^{N_W}, \quad X^{UN} = \{x_l^{UN}\}_{l=1}^{N_N}. \end{aligned} \tag{1}$$

Each verified pair (x_i^{PW}, x_i^{PN}) depicts the same lesion and shares label y_i. Unpaired images also have class labels, but their indices do not imply cross-modal correspondence.

An NBI encoder f_N is pretrained with cross-entropy on the NBI training images and then frozen. A paired-WLI classifier trained with cross-entropy provides θ_0, initializing both a frozen reference f_W^0 and the trainable student f_W^θ. We extract descriptors with GeM pooling (Radenović et al., 2019) ($\nu = 3$) and ℓ_2 normalization: F_i^{PN}, F_l^{UN} are NBI descriptors, and $F_i^{PW,0}$ are fixed WLI anchors. For an unpaired WLI image, $F_j^{UW,0}$ comes from the reference on a deterministic view, while $F_j^{UW,\theta}$ comes from the student on an augmented view. Descriptor implementation details appear in Appendix A.1.

3.2 Overview of PaCT

Figure 2 connects NBI relation construction to WLI learning. Paired and unpaired NBI form a graph whose diffusion produces a relation matrix K^N over paired anchors. Frozen WLI query affinities $Q_j^{UW,0}$ read this matrix through verified pair identities to obtain a fixed target P_j^{UW}; the student learns this target through its direct anchor affinities $Q_j^{UW,\theta}$. For compactness, the figure combines these frozen-reference and live-student scoring paths into one block; both use the same frozen WLI anchor bank. The illustrated losses are NBI pretraining cross-entropy $\mathcal{L}_1$, WLI cross-entropy $\mathcal{L}_2$, and unpaired-WLI relational distillation $\mathcal{L}_3$, defined below.

3.3 Graph-Augmented NBI Teacher

For each modality $m \in \{w, n\}$, descriptors are standardized coordinate-wise as $\widehat{F}^m = (F^m - \mu_m) \oslash \max(\sigma_m, \epsilon)$ using the paired-anchor mean and population standard deviation, with numerical floor $\epsilon > 0$. The RBF bandwidth β_m is the median positive k_{bw}-nearest-neighbor distance within that standardized anchor set. These statistics and bandwidths remain fixed, including when processing unpaired observations.

The NBI graph $\mathcal{G}^N$ contains $N = K + N_N$ distinct observations: paired nodes h_i^P and unpaired nodes h_l^U in the figure. With $\widehat{F}_a^N$ denoting node a's standardized descriptor, let $\mathcal{N}_{k_{\text{graph}}}(a)$ contain its k_{graph} nearest neighbors, excluding a. We construct

$$G_{ab} = \mathbf{1}[b \in \mathcal{N}_{k_{\text{graph}}}(a)] \exp\left(-\frac{\|\widehat{F}_a^N - \widehat{F}_b^N\|_2^2}{2\beta_n^2}\right), \qquad A = \max(G, G^\top) + I_N, \tag{2}$$

where the maximum is elementwise, and set $T = \operatorname{diag}(A\mathbf{1})^{-1}A$. Graph construction uses descriptors without class labels. Additional NBI can change local neighbors and multi-step paths between paired anchors.

Personalized PageRank (Gasteiger et al., 2019) aggregates these paths while retaining each starting anchor through restart. Let $\mathcal{A}_i$ index paired NBI anchor i and $E \in \mathbb{R}^{K\times N}$ have entries $E_{ia} = \mathbf{1}[a = \mathcal{A}_i]$. For $\rho \in (0, 1]$, define

$$S = \rho \sum_{t=0}^{\infty} (1-\rho)^t E T^t, \qquad S = \rho E + (1-\rho) S T. \tag{3}$$

Here S_{ia} is the endpoint probability of a walk starting at anchor i with geometrically distributed length. Intermediate nodes may be paired or unpaired. Restricting endpoints to paired anchors gives

$$K_{ij}^N = \frac{S_{i,\mathcal{A}_j}}{\sum_{l=1}^K S_{i,\mathcal{A}_l}}, \qquad K^N \mathbf{1} = \mathbf{1}. \tag{4}$$

Restart ensures a positive denominator. Each row is normalized *before* query mixing, so unpaired NBI changes the relations without changing the output identities. Appendix A.1 specifies the iterative computation.

3.4 Relational Scoring and Relation-Guided Transform

A WLI descriptor z is scored against the fixed paired-WLI anchors:

$$\text{Aff}_i(z) = \frac{\exp(-\|\widehat{z} - \widehat{F}_i^{PW,0}\|_2^2/(2\beta_w^2))}{\sum_{r=1}^K \exp(-\|\widehat{z} - \widehat{F}_r^{PW,0}\|_2^2/(2\beta_w^2))}, \tag{5}$$

using the fixed WLI standardization. The reference and student respectively produce $Q_j^{UW,0} = \text{Aff}(F_j^{UW,0})$ and $Q_j^{UW,\theta} = \text{Aff}(F_j^{UW,\theta})$, both row distributions over the same K anchors.

Let $M \in \{0, 1\}^{K\times K}$ be the permutation matrix mapping WLI anchor order to NBI anchor order through verified pairs; $M = I_K$ when the orders coincide. The relation-guided target is

$$P_j^{UW} = Q_j^{UW,0} M K^N M^\top. \tag{6}$$

The WLI weights determine how much to read from each NBI anchor's row in K^N, and $M^\top$ returns the mixture to WLI anchor order. P_j^{UW} is computed once from the canonical image and cached. It supplies query-specific NBI relations without inferring a match to an unpaired NBI image. At $\rho = 1$, $K^N = I_K$ and $P_j^{UW} = Q_j^{UW,0}$, recovering WLI anchor self-distillation.

3.5 Training Objectives and Inference

Let $s_N(x)$ and $s_W^\theta(x)$ denote classifier probabilities, distinct from the anchor distributions P, Q. For an NBI training batch B_N drawn from $X^{PN} \cup X^{UN}$, the teacher pretraining objective is

$$\mathcal{L}_1(B_N) = -\frac{1}{|B_N|} \sum_{(x,y)\in B_N} \log s_N(\widetilde{x})_y. \tag{7}$$

Here $\widetilde{x}$ denotes a training augmentation. After pretraining, the teacher is frozen. During student training, either a paired or an unpaired WLI batch B receives classification supervision:

$$\mathcal{L}_2(B) = -\frac{1}{|B|} \sum_{(x,y)\in B} \log s_W^\theta(\widetilde{x})_y. \tag{8}$$

For an unpaired WLI batch B_u, relational distillation additionally uses

$$\mathcal{L}_3(B_u) = \frac{1}{|B_u|} \sum_{(x_j^{UW}, y_j) \in B_u} \mathrm{KL}\Big(\mathrm{sg}(P_j^{UW}) \,\Big\|\, Q_j^{UW,\theta}\Big), \tag{9}$$

where sg stops gradients. Comparing distributions over the fixed anchor bank sends relational gradients through the live descriptor to the WLI backbone. This identity-level loss constrains both anchor-class mass and within-class identity preferences.

We use $k_{\mathrm{bw}} = k_{\mathrm{graph}} = 5$ and $\rho = 0.15$ for target construction. Each Stage-2 batch performs one optimizer step with

$$\mathcal{L}_{\mathrm{total}}(B) = \begin{cases} \mathcal{L}_2(B), & B \text{ is paired,} \\ \mathcal{L}_2(B) + \dfrac{\lambda}{\pi_u}\mathcal{L}_3(B), & B \text{ is unpaired.} \end{cases} \tag{10}$$

Here π_u is the fraction of unpaired batch updates, so λ/π_u accounts for the frequency of relational supervision. $\mathcal{L}_1$ is used only before student training. The reference, teacher, anchor bank, graph, and targets remain fixed; the classifier receives direct cross-entropy gradients, while the backbone receives both classification and relational gradients. Inference retains only the WLI backbone and GAP–linear classification head.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate how partially paired NBI supports WLI classification through a comparison with existing methods and controlled variants of PaCT. The study uses the three-class PICCOLO dataset (Sánchez-Peralta et al., 2020) and the binary PolypSet protocol (Li et al., 2021). PICCOLO contains three lesion categories, whereas the PolypSet protocol used here contains two classes. For each dataset, we organize the available WLI and NBI observations into verified WLI–NBI pairs and strict-unpaired pools. Detailed statistics are summarized in Table 1.

Table 1: Statistics of experimental datasets.

Dataset	Class	WLI–NBI pairs	Unpaired WLI	Unpaired NBI
PolypSet	Hyperplasia	63	200	200
	Adenoma	102	250	250
	Total	**165**	**450**	**450**
PICCOLO	Hyperplasia	249	155	35
	Adenoma	615	839	167
	Adenocarcinoma	156	106	47
	Total	**1020**	**1100**	**249**

Evaluation protocol and metrics. We perform five-fold cross-validation with patient/video-proxy (VP)-grouped splits, keeping all observations from each VP together to prevent leakage. Within each outer-training fold, 10% of VPs are reserved for validation; held-out WLI test data are used only for final evaluation. Split details and counts are provided in Appendix A.2.

We report accuracy (Acc), support-weighted F1-score (F1), and AUC (binary ROC-AUC for PolypSet and macro one-vs-rest ROC-AUC for PICCOLO). Results are reported as mean $\pm$ standard deviation across three seeds, with each seed's score averaged over the five outer folds.

4.2 Baselines

We compare PaCT with SSL-CPCD (Xu et al., 2024), FFCNet (Wang et al., 2022), CPC-Trans (Ma et al., 2022), PolypsAlign (Wang et al., 2021b), ADD (Hu et al., 2025), Better Together (Gupta et al., 2026), DLGNet (Wang et al., 2023), and PaGKD (Hu et al., 2026). NBI access follows each method's supervision: CPC-Trans, PolypsAlign, and ADD use verified pairs in D_p only, whereas methods with an unpaired-NBI branch may additionally use D_{unp}.

Table 2: Comparison of WLI classification performance on PICCOLO and PolypSet.

Dataset	Method	Acc (%)	F1 (%)	AUC (%)
PICCOLO	PolypsAlign† (Wang et al., 2021b)	69.48 ± 1.02	64.61 ± 2.45	67.92 ± 3.23
	CPC-Trans† (Ma et al., 2022)	73.55 ± 0.95	70.87 ± 0.90	74.53 ± 2.69
	FFCNet (Wang et al., 2022)	64.43 ± 1.64	59.13 ± 2.37	64.32 ± 4.23
	DLGNet† (Wang et al., 2023)	70.03 ± 0.90	64.43 ± 0.74	67.46 ± 1.26
	SSL-CPCD (Xu et al., 2024)	70.21 ± 1.06	67.28 ± 1.79	73.74 ± 3.37
	ADD (Hu et al., 2025)	68.87 ± 0.79	66.27 ± 0.49	74.32 ± 0.70
	Better Together (Gupta et al., 2026)	72.00 ± 1.58	67.69 ± 1.99	73.17 ± 1.28
	PaGKD (Hu et al., 2026)	75.65 ± 1.47	73.23 ± 1.09	75.52 ± 1.86
	PaCT (Ours)	$\mathbf{78.13 \pm 0.84}$	$\mathbf{75.87 \pm 0.67}$	$\mathbf{78.23 \pm 1.37}$
PolypSet	PolypsAlign† (Wang et al., 2021b)	82.35 ± 2.14	81.55 ± 2.56	84.23 ± 2.87
	CPC-Trans† (Ma et al., 2022)	87.28 ± 1.57	87.32 ± 1.54	86.15 ± 1.92
	FFCNet (Wang et al., 2022)	85.78 ± 2.71	85.71 ± 2.56	85.82 ± 3.45
	DLGNet† (Wang et al., 2023)	86.49 ± 0.65	86.60 ± 0.66	89.21 ± 2.87
	SSL-CPCD (Xu et al., 2024)	85.22 ± 1.27	85.25 ± 1.33	88.35 ± 2.44
	ADD (Hu et al., 2025)	87.41 ± 1.34	87.34 ± 1.89	86.78 ± 1.67
	Better Together (Gupta et al., 2026)	83.73 ± 0.65	83.55 ± 1.02	83.05 ± 5.53
	PaGKD (Hu et al., 2026)	88.69 ± 1.26	88.74 ± 1.21	87.46 ± 0.98
	PaCT (Ours)	$\mathbf{90.62 \pm 1.17}$	$\mathbf{90.87 \pm 1.44}$	$\mathbf{92.03 \pm 1.95}$

Note: Results are reported as the mean ± standard deviation over three random seeds. indicates the highest value (including ties), while indicates the second-highest value within each dataset and metric column. † indicates methods that use additional lesion-location annotations.

4.3 Comparative Experiment Results

All values in Table 2 come from our fold-specific reruns under the three-seed protocol described above, rather than from the original publications or an earlier comparison table. We keep the fold splits, preprocessing, WLI initialization and training pool, optimization budget, and metric implementation fixed across methods. When checkpoint selection or configurable settings are required, we apply the same inner-validation rule to each method.

Across the three-seed evaluation, PaCT reaches 78.13% accuracy, 75.87% F1, and 78.23% AUC on PICCOLO, and 90.62%, 90.87%, and 92.03% on PolypSet, respectively. These results establish a consistent advantage for the proposed paired-anchor transfer under the common WLI-only inference setting.

4.4 Ablation Study

We present the ablation as a staged construction with compact labels that make the active supervision explicit. M0 denotes CE continuation. M1 denotes anchor distillation using the $\rho = 1.00$ setting, M2 adds the paired-only NBI graph, and M3 denotes the full PaCT configuration with unpaired NBI graph. All variants retain the same WLI training pool, paired-data initialization, teacher, outer-fold partition, and Stage-2 budget. Each variant follows the same inner-validation checkpoint-selection rule; test-set results are reported only after the variant is fixed. Table 3 reports the staged comparisons using Accuracy, F1, and AUC.

The progression from M0 to M3 exposes the contribution of increasingly rich cross-modal supervision while keeping the evaluation protocol fixed. Macro-F1 is not reported in this ablation table.

Effect of verified pair identities. To test whether verified correspondence provides information beyond class labels, we replace the pair mapping with within-class cyclic or random reassignment while keeping the NBI graph and training configuration fixed. As shown in Table 4, PaCT outperforms both controls in F1 and AUC on PICCOLO and PolypSet, with F1 gains over Cyclic of 1.87 and 3.58 percentage points, respectively.

Table 3: Staged component ablations of PaCT on the PICCOLO and PolypSet datasets.

Setting	Components			PICCOLO (%)			PolypSet (%)		
	Anchor distillation	Paired NBI graph	Unpaired NBI nodes	Acc	F1	AUC	Acc	F1	AUC
M0	–	–	–	70.49	68.43	71.05	80.01	80.02	82.75
M1	✓	–	–	72.46	69.85	73.42	83.96	83.75	86.04
M2	✓	✓	–	77.33	74.99	76.66	87.36	87.34	88.34
M3	✓	✓	✓	**78.13**	**75.87**	**78.23**	**90.62**	**90.87**	**92.03**

Note: Checkmarks denote active components. denotes the best result (bold), and denotes the second-best result within each dataset and metric column.

Table 4: Effect of verified WLI–NBI correspondence on relational distillation.

Pair identity	PICCOLO (%)			PolypSet (%)		
	Acc	F1	AUC	Acc	F1	AUC
Cyclic	76.15	74.00	76.21	87.63	87.29	88.64
Random	75.91	73.80	77.51	87.25	87.20	86.99
PaCT	**78.13**	**75.87**	**78.23**	**90.62**	**90.87**	**92.03**

Note: denotes the highest value, and denotes the second-highest value within each dataset and metric column.

Effect of additional unpaired NBI data. We vary the fraction of unpaired NBI images included in the teacher graph from 0% to 100%, while keeping the paired anchors, frozen encoders, WLI training data, and optimization settings fixed. Figure 3 reports the resulting Accuracy and AUC. Incorporating 25% of the unpaired NBI pool improves Accuracy by 0.72 and 1.61 percentage points and AUC by 0.87 and 1.48 points on PICCOLO and PolypSet, respectively. On PICCOLO, this subset reaches 78.05% Accuracy and 77.53% AUC, while the full pool reaches 78.13% and 78.23%. On PolypSet, the updated 25% configuration reaches 88.97% Accuracy and 89.82% AUC, and the full pool reaches 90.62% and 92.03%, respectively. Both the 25% and full-pool configurations improve Accuracy and AUC over the paired-only graph on both datasets, showing that PaCT can exploit additional NBI observations through a fixed set of paired anchors.

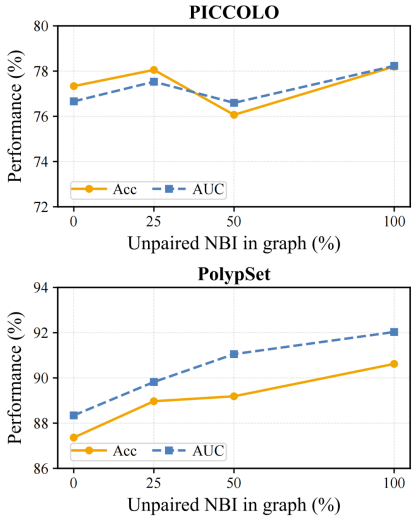

Figure 3: Experiment of unpaired NBI availability.

Hyperparameter sensitivity. Figure 5 presents the F1 scores under different values of the graph-construction neighborhood size k_{graph} (left) and the restart probability ρ (right). In each sweep, only the displayed hyperparameter is varied; for every value of k_{graph}, the graph certificate and query targets are regenerated, while the remaining training protocol is kept fixed. Candidate configurations are compared using F1 on the corresponding inner-validation split. The complete tabulated results for the k_{graph} and ρ sweeps are provided in Appendix A.4 and Appendix A.5, respectively.

Qualitative case study. To complement the quantitative comparisons, Figure 4 presents representative response maps on four WLI images. The GT column shows the annotated lesion regions, while the remaining columns show the corresponding model responses. These examples provide a qualitative view of how the methods distribute their responses around the annotated lesions under WLI-only inference.

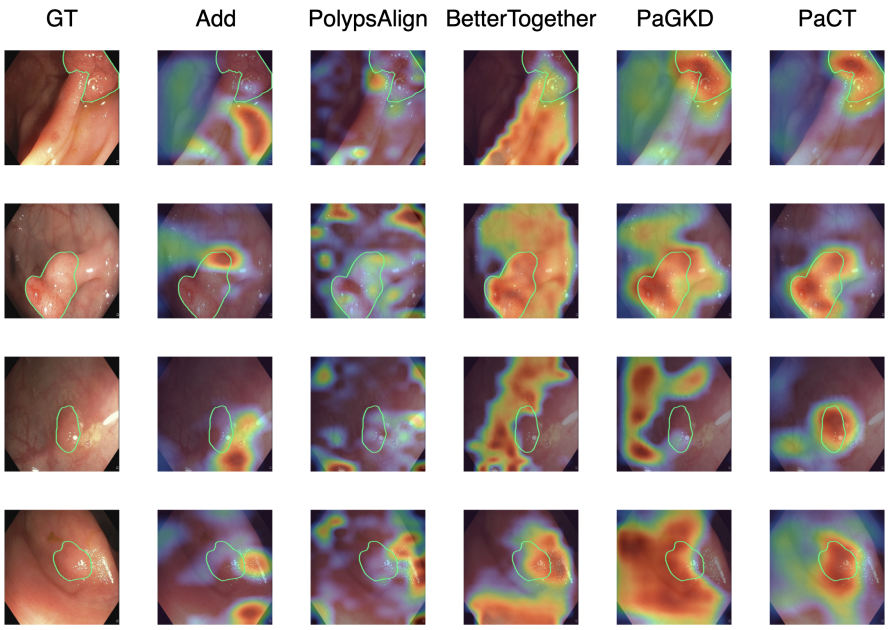

Figure 4: Qualitative comparison of model response maps on representative WLI images. GT denotes the annotated lesion regions.

4.5 Implementation Details

We fix $k_{\text{graph}} = 5$ and $\rho = 0.15$ based on the inner-validation procedure described above, before evaluating the held-out test splits, and use these values as the default configuration in the main experiments. All experiments were implemented in Python 3.10 with PyTorch 2.2.0 and run on a workstation equipped with four NVIDIA RTX A6000 GPUs (48 GB each), dual AMD EPYC 7763 CPUs, and 1 TB of system memory. Further architectural, training, and evaluation details are provided in Appendix A.6, while the input configurations and computational-cost measurements are reported in Appendix A.3.

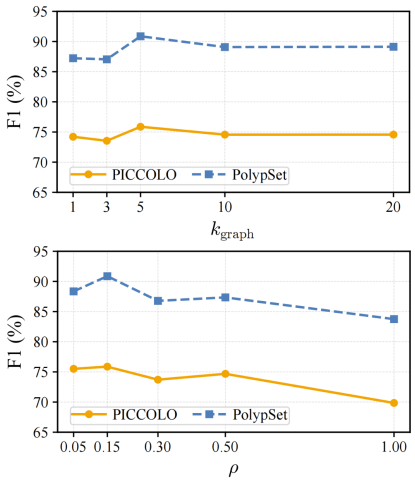

Figure 5: Ablation on hyperparameters: k_{graph} and ρ.

5 Conclusion

In this paper, we presented PaCT, a cross-modal relational distillation framework for improving white-light endoscopic image classification under partial WLI–NBI correspondence. We studied how verified WLI–NBI correspondences can extend NBI-derived supervision to WLI images without paired NBI counterparts. PaCT treats each verified pair as a shared identity anchor: affinities from an unpaired WLI query read an NBI teacher graph built from paired and unpaired NBI observations, yielding a query-specific relational target for the WLI student. This construction uses additional NBI images without assigning them unverified WLI matches and retains WLI-only inference. Under a common three-seed, five-fold evaluation on PICCOLO and PolypSet, PaCT improves mean accuracy, F1, and AUC over the best comparator for each metric on both datasets. Controlled ablations of anchor identities and unpaired NBI availability support the distinct contributions of verified correspondence and expanded teacher-side relations. Within the evaluated colorectal polyp classification setting, PaCT thus shows that verified pairs can serve as reusable anchors, extending NBI relational supervision to WLI images outside the paired subset.

AI USE STATEMENT

We used large language models (LLMs) solely for polishing the writing, such as improving sentence fluency and fixing grammar issues. No content generation, idea creation, or experimental analysis was conducted by LLMs.

ETHICS STATEMENT

For the PICCOLO dataset only, images and anonymized data from patients included in this study were provided by PICCOLO database from the Basque Biobank. The experiments in this paper use the publicly available PICCOLO and PolypSet endoscopic image datasets. We use these datasets in accordance with their original terms and cite the corresponding dataset publications. This study does not involve new data collection, participant recruitment, or intervention, and we do not attempt to identify individuals or process direct personal identifiers.

REPRODUCIBILITY STATEMENT

The PaCT architecture and training objective are described in Section 3. Dataset protocols and evaluation metrics are specified in Section 4. Appendix A.2 provides fold-specific partition statistics, and the fixed training settings are given in Appendix A.6. Computational-cost measurements and sensitivity analyses are reported in Appendix A.3, Appendix A.4, and Appendix A.5. Code is available at this repository: `https://anonymous.4open.science/r/PaCT-iclr2027-code/`.

REFERENCES

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1gL-2A9Ym`.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=5OIgg5YkC3`.

Qiang Hu, Qimei Wang, Jia Chen, Xuantao Ji, Mei Liu, Qiang Li, and Zhiwei Wang. Holistic white-light polyp classification via alignment-free dense distillation of auxiliary optical chromoendoscopy. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15970 of *Lecture Notes in Computer Science*, pp. 251–261. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05141-7_25.

Qiang Hu, Qimei Wang, Yingjie Guo, Qiang Li, and Zhiwei Wang. Pairing-free group-level knowledge distillation for robust gastrointestinal lesion classification in white-light endoscopy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4905–4913, 2026. doi: 10.1609/aaai.v40i6.42494.

Kaidong Li, Mohammad I. Fathan, Krushi Patel, Tianxiao Zhang, Cuncong Zhong, Ajay Bansal, Amit Rastogi, Jean S. Wang, and Guanghui Wang. Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. *PLOS ONE*, 16(8):e0255809, 2021. doi: 10.1371/journal.pone.0255809.

Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision*, pp. 174–192, 2018. doi: 10.1007/978-3-030-01264-9_11.

Weijie Ma, Ye Zhu, Ruimao Zhang, Jie Yang, Yiwen Hu, Zhen Li, and Li Xiang. Toward clinically assisted colorectal polyp recognition via structured cross-modal representation consistency. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 141–150, 2022. doi: 10.1007/978-3-031-16437-8_14.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019. doi: 10.1109/TPAMI.2018.2846566.

Luisa F. Sánchez-Peralta, J. Blas Pagador, Artzai Picón, Ángel José Calderón, Francisco Polo, Nagore Andraka, Roberto Bilbao, Ben Glover, Cristina L. Saratxaga, and Francisco M. Sánchez-Margallo. PICCOLO white-light and narrow-band imaging colonoscopic dataset: A performance comparative of models and datasets. *Applied Sciences*, 10(23):8501, 2020. doi: 10.3390/app10238501.

Rüdiger Schmitz, Rene Werner, Alessandro Repici, Raf Bisschops, Alexander Meining, Michael Zornow, Helmut Messmann, Cesare Hassan, Prateek Sharma, and Thomas Rösch. Artificial intelligence in GI endoscopy: Stumbling blocks, gold standards and the role of endoscopy societies. *Gut*, 71(3):451–454, 2022. doi: 10.1136/gutjnl-2020-323115.

Yu Shen, Xijun Wang, Peng Gao, and Ming Lin. Auxiliary modality learning with generalized curriculum distillation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 31057–31076. PMLR, 2023. URL https://proceedings.mlr.press/v202/shen23f.html.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042.

Kai-Ni Wang, Yuting He, Shuaishuai Zhuang, Juzheng Miao, Xiaopu He, Ping Zhou, Guanyu Yang, Guang-Quan Zhou, and Shuo Li. FFCNet: Fourier transform-based frequency learning and complex convolutional network for colon disease classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 78–87, 2022. doi: 10.1007/978-3-031-16437-8_8.

Kai-Ni Wang, Shuaishuai Zhuang, Qi-Yong Ran, Ping Zhou, Jie Hua, Guang-Quan Zhou, and Xiaopu He. Dlgnet: A dual-branch lesion-aware network with the supervised gaussian mixture model for colon lesions classification in colonoscopy images. *Medical Image Analysis*, 87:102832, 2023.

Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. EvDistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 608–619, 2021a. doi: 10.1109/CVPR46437.2021.00067.

Qin Wang, Hui Che, Weizhen Ding, Li Xiang, Guanbin Li, Zhen Li, and Shuguang Cui. Colorectal polyp classification from white-light colonoscopy images via domain alignment. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 24–32, 2021b. doi: 10.1007/978-3-030-87234-2_3.

Ziang Xu, Jens Rittscher, and Sharib Ali. Ssl-cpcd: Self-supervised learning with composite pretext-class discrimination for improved generalisability in endoscopic image analysis. *IEEE Transactions on Medical Imaging*, 43(12):4105–4119, 2024. doi: 10.1109/TMI.2024.3411933.

Guanghui Yue, Guibin Zhuo, Siying Li, Tianwei Zhou, Jingfeng Du, Weiqing Yan, Jingwen Hou, Weide Liu, and Tianfu Wang. Benchmarking polyp segmentation methods in narrow-band imaging colonoscopy images. *IEEE Journal of Biomedical and Health Informatics*, 27(7):3360–3371, 2023. doi: 10.1109/JBHI.2023.3270724.

R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. W. Lau, and C. C. Y. Poon. Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. *IEEE Journal of Biomedical and Health Informatics*, 21(1):41–47, 2017. doi: 10.1109/JBHI.2016.2635662.

A SUPPLEMENTARY MATERIAL

A.1 ALGORITHMIC DETAILS

Algorithms 1 and 2 summarize the offline construction of fixed relational targets and the subsequent Stage-2 WLI optimization used by PaCT.

Initialization. The NBI teacher f_N is pretrained on $X^{PN} \cup X^{UN}$ using $\mathcal{L}_1$. A paired-WLI cross-entropy model initializes the frozen reference f_W^0 and the student f_W^θ. All offline descriptors use deterministic views. The paired anchor lists follow the orders encoded by the verified permutation M; NBI anchors precede unpaired NBI nodes.

Algorithm 1 PaCT: offline NBI relations and WLI target construction

Input: Paired sets X^{PW}, X^{PN}; unpaired sets X^{UW}, X^{UN}; verified mapping M
Input: Frozen WLI reference f_W^0 and NBI teacher f_N
Input: $k_{\text{graph}} = k_{\text{bw}} = 5$, $\rho = 0.15$, $\delta = 10^{-6}$, $J_{\max} = 100$, $\epsilon = 10^{-6}$
Output: Fixed WLI descriptors $F^{PW,0}$, statistics $(\mu_w, \sigma_w, \beta_w)$, target cache $\mathcal{C}_P$

A. Extract canonical descriptors and fit anchor statistics
Set both encoders to evaluation mode; disable gradients for offline operations
$F_i^{PW,0} \leftarrow \text{Desc}(f_W^0(x_i^{PW}))$; $F_i^{PN} \leftarrow \text{Desc}(f_N(x_i^{PN}))$ for $i = 1, \ldots, K$
$\mu_w \leftarrow \text{mean}(F^{PW,0})$; $\sigma_w \leftarrow \text{std}_{\text{pop}}(F^{PW,0})$
$\mu_n \leftarrow \text{mean}(F^{PN})$; $\sigma_n \leftarrow \text{std}_{\text{pop}}(F^{PN})$
$\widehat{F}^{PW,0} \leftarrow \text{Std}(F^{PW,0}; \mu_w, \sigma_w)$; $\widehat{F}^{PN} \leftarrow \text{Std}(F^{PN}; \mu_n, \sigma_n)$
$\beta_w \leftarrow \textsc{Bandwidth}(\widehat{F}^{PW,0}, k_{\text{bw}})$; $\beta_n \leftarrow \textsc{Bandwidth}(\widehat{F}^{PN}, k_{\text{bw}})$
$F_l^{UN} \leftarrow \text{Desc}(f_N(x_l^{UN}))$; $\widehat{F}_l^{UN} \leftarrow \text{Std}(F_l^{UN}; \mu_n, \sigma_n)$
$V \leftarrow [\widehat{F}^{PN}; \widehat{F}^{UN}] \in \mathbb{R}^{N \times d}$, where $N = K + N_N$
B. Construct the NBI graph and propagate from paired anchors
$G \leftarrow 0_{N \times N}$
for $a = 1, \ldots, N$ **do**
  $\mathcal{N}(a) \leftarrow k_{\text{graph}}$ nearest neighbors of V_a, excluding a
  for $b \in \mathcal{N}(a)$ **do**
    $G_{ab} \leftarrow \exp(-\|V_a - V_b\|_2^2/(2\beta_n^2))$
  end for
end for
$A \leftarrow \max(G, G^\top) + I_N$; $T \leftarrow \text{RowNorm}(A)$ ▷ Elementwise maximum
$E \leftarrow [I_K \; 0_{K \times (N-K)}]$; $S \leftarrow E$
for $t = 1, \ldots, J_{\max}$ **do**
  $S_{\text{next}} \leftarrow \rho E + (1 - \rho) S T$
  $\Delta \leftarrow \max_i \sum_a |(S_{\text{next}})_{ia} - S_{ia}|$; $S \leftarrow S_{\text{next}}$
  if $\Delta \le \delta$ **then**
    break
  end if
end for
$K^N \leftarrow \text{RowNorm}(S_{:,1:K})$ ▷ Condition on paired endpoints, row by row
C. Read NBI relations with frozen WLI query weights
$\mathcal{C}_P \leftarrow \varnothing$
for $j = 1, \ldots, N_W$ **do**
  $F_j^{UW,0} \leftarrow \text{Desc}(f_W^0(x_j^{UW}))$ ▷ Canonical view
  $Q_j^{UW,0} \leftarrow \text{Aff}(F_j^{UW,0})$
  $P_j^{UW} \leftarrow \text{RowNorm}(Q_j^{UW,0} M K^N M^\top)$
  Cache $\mathcal{C}_P[j] \leftarrow \text{sg}(P_j^{UW})$
end for
return $F^{PW,0}, (\mu_w, \sigma_w, \beta_w), \mathcal{C}_P$

Algorithm 2 PaCT: Stage-2 WLI training and WLI-only inference

Input: Labeled loaders for X^{PW} and X^{UW}; R training epochs
Input: Stage-1 backbone/head parameters (θ_0, ψ_0); fixed outputs of Algorithm 1
Input: Nominal distillation weight λ; training augmentation $\mathcal{T}$
Output: Trained WLI backbone f_W^θ and classification head C_ψ

$(\theta, \psi) \leftarrow (\theta_0, \psi_0)$; initialize a fresh Adam optimizer
$\pi_u \leftarrow$ fraction of scheduled Stage-2 updates assigned to the unpaired route
for $e = 1, \dots, R$ **do**
  Set the student to training mode; keep all cached quantities fixed
  for each route-tagged batch (r, B) in the interleaved paired/unpaired schedule **do**
    Clear optimizer gradients
    $\widetilde{x} \leftarrow \mathcal{T}(x)$; $H_x \leftarrow f_W^\theta(\widetilde{x})$ for each $(x, y) \in B$
    $s_W^\theta(\widetilde{x}) \leftarrow \mathrm{softmax}(C_\psi(\mathrm{GAP}(H_x)))$
    $\mathcal{L}_2(B) \leftarrow -|B|^{-1} \sum_{(x,y)\in B} \log s_W^\theta(\widetilde{x})_y$
    $\mathcal{L}_{\mathrm{total}}(B) \leftarrow \mathcal{L}_2(B)$
    if $r =$ unpaired **then**
      $F_j^{UW,\theta} \leftarrow \mathrm{Desc}(H_{x_j^{UW}})$ for each $(x_j^{UW}, y_j) \in B$
      $Q_j^{UW,\theta} \leftarrow \mathrm{Aff}(F_j^{UW,\theta})$; $P_j^{UW} \leftarrow \mathcal{C}_P[j]$
      $\mathcal{L}_3(B) \leftarrow |B|^{-1} \sum_{(x_j^{UW}, y_j)\in B} D_{\mathrm{KL}}(\mathrm{sg}(P_j^{UW}) \| Q_j^{UW,\theta})$
      $\mathcal{L}_{\mathrm{total}}(B) \leftarrow \mathcal{L}_2(B) + (\lambda/\pi_u)\mathcal{L}_3(B)$
    end if
    Backpropagate $\mathcal{L}_{\mathrm{total}}(B)$; update (θ, ψ) with one optimizer step
  end for
end for
return (f_W^θ, C_ψ) ▷ Checkpoint selection follows the evaluation protocol

Deployment: a deterministic WLI view, without teacher or anchors

function PREDICT(x; f_W^θ, C_ψ)
  return $\arg\max_c [C_\psi(\mathrm{GAP}(f_W^\theta(x)))]_c$
end function

Descriptor and fixed-coordinate operators. For a feature map $H \in \mathbb{R}^{d\times L}$, the descriptor is

$$g_r(H) = \left(\frac{1}{L}\sum_{t=1}^{L} \max(H_{rt}, \epsilon)^3\right)^{1/3}, \qquad \mathrm{Desc}(H) = \frac{g(H)}{\max(\|g(H)\|_2, \epsilon)},$$

$$\mathrm{Std}(z; \mu, \sigma) = (z - \mu) \oslash \max(\sigma, \epsilon), \qquad \mathrm{RowNorm}(X)_{ij} = \frac{X_{ij}}{\sum_l X_{il}}.$$

Statistics are channel-wise and fitted on paired anchors only. For arrays, Std acts on each descriptor row. The operator $\mathrm{Aff}(z)$ takes an unstandardized GeM–ℓ_2 descriptor, applies the fixed WLI statistics, and normalizes over the fixed WLI anchors:

$$\mathrm{Aff}_i(z) = \mathrm{softmax}_i\left(-\frac{\|\,\mathrm{Std}(z; \mu_w, \sigma_w) - \widehat{F}_i^{PW,0}\|_2^2}{2\beta_w^2}\right).$$

$\textsc{Bandwidth}(B, k_{\mathrm{bw}})$ is the median of strictly positive Euclidean distances among each anchor's k_{bw} nearest other anchors. The arrays $F^{PW,0}$ and F^{PN} stack the respective descriptors; V stacks standardized NBI descriptors with paired anchors first. The cache $\mathcal{C}_P$ stores one target per unpaired WLI training-image index j.

Gradient flow and numerical evaluation. The teacher uses $\mathcal{L}_1$ only during pretraining. During Stage 2, $\mathcal{L}_2$ updates the WLI backbone and classification head, whereas $\mathcal{L}_3$ updates the backbone through GeM and affinities. Targets, anchors, statistics, and the teacher remain fixed. Evaluate affinities with log-sum-exp stabilization and floor logarithm arguments at 10^{-8}. The final row normalization in Algorithm 1 compensates for numerical roundoff. At $\rho = 1$, $K^N = I_K$ and $P_j^{UW} = Q_j^{UW,0}$, yielding WLI anchor self-distillation.

Table 5: Sensitivity to the restart probability ρ.

Dataset	ρ	Accuracy (%)	F1 (%)	AUC (%)
PICCOLO	0.05	<u>77.77</u>	<u>75.51</u>	76.83
	0.15	**78.13**	**75.87**	<u>78.23</u>
	0.30	75.72	73.71	**79.12**
	0.50	76.83	74.68	77.07
	1.00	72.46	69.85	73.42
PolypSet	0.05	<u>88.54</u>	<u>88.35</u>	88.76
	0.15	**90.62**	**90.87**	**92.03**
	0.30	86.69	86.78	88.98
	0.50	87.34	87.35	<u>90.51</u>
	1.00	83.96	83.75	86.04

Note: Bold $\rho = 0.15$ indicates the default setting. The highest and second-highest values within each dataset and metric column are shown in bold and underlined, respectively.

A.2 Fold protocol

Fold construction. We construct all partitions using the dataset-provided VP identifier, which serves as a patient/video proxy. Each VP is treated as an indivisible group: all paired WLI–NBI images, unpaired WLI images, and unpaired NBI images associated with the same VP are assigned to the same partition. Outer folds are constructed such that no VP or annotated lesion is shared across folds. The subsequent 90%/10% training–inner-validation split is performed over VPs rather than individual images and follows the same no-overlap constraint. Thus, no image from a VP or annotated lesion used for training or checkpoint selection appears in the corresponding held-out test partition.

A.3 Computational cost

Table 6 reports the parameter count, GFLOPs, and measured training and inference latency of all compared methods. The input configurations follow the original implementation of each method. FFCNet uses a $6 \times 400 \times 400$ Fourier mosaic, while DLGNet receives two $3 \times 400 \times 400$ images per sample. SSL-CPCD, CPC-Trans, and Better Together use $3 \times 224 \times 224$ inputs, whereas ADD, PolypsAlign, PaGKD, and PaCT use $3 \times 448 \times 448$ inputs. CPC-Trans is evaluated with a ViT-S/16 encoder; its reported parameter count, GFLOPs, and latency therefore correspond to the ViT-S/16 configuration.

Measurement protocol. All GPU measurements were collected on an otherwise idle NVIDIA RTX A6000 GPU with 48 GB of memory. Training time is measured for one steady-state update using batch size 24. Better Together processes 24 WLI images and 24 NBI images at each update, while DLGNet processes two input images per sample. Inference time is measured with batch size 1 and includes only the forward pass on inputs already in memory; disk loading and image preprocessing are excluded. Better Together was measured in its official PyTorch 2.9 environment, whereas all other methods were measured with PyTorch 2.2.

Offline graph construction is excluded from the training-time measurement. For PaCT, Stage-2 graph construction, certificate generation, and target precomputation are performed outside the timed iterations. The reported training time includes steady-state DataLoader waiting, data transfer, forward computation, loss computation, backpropagation, and the optimizer update, but excludes validation, checkpoint saving, worker startup, and the one-time memory-bank initialization of SSL-CPCD. Therefore, the training-time column represents per-step update latency rather than complete end-to-end training time.

A.4 Graph-neighborhood sensitivity

We vary only the graph-construction neighborhood size k_{graph} under the locked training protocol. For each value of k_{graph}, we regenerate the NBI graph, adjacency normalization A, transition matrix

Table 6: Computational complexity and latency of the compared methods. Parameter counts, GFLOPs, and measured training and inference latencies are reported for the configurations used in the comparison.

Method	**Params (M)**	**GFLOPs**	**Training time (ms/step)**	**Inference time (ms/image)**
FFCNet	47.05	52.87	619.9	28.61
ADD	23.78	16.49	506.7	4.64
SSL-CPCD	23.51	**4.11**	221.4	4.58
CPC-Trans	21.67	4.61	**101.1**	**3.90**
DLGNet	48.08	26.58	325.9	10.61
Better Together	**21.63**	6.14	148.7	4.01
PolypsAlign	23.51	16.44	344.5	4.53
PaGKD	23.78	16.49	1139.3	4.44
PaCT	23.78	16.49	212.0	4.52

Note: denotes the lowest value, and denotes the second-lowest value in each column.

Table 7: Sensitivity to the graph-construction neighborhood size k_{graph}. Only k_{graph} varies. For each value of k_{graph}, all graph-derived quantities and targets are regenerated; no quantities from $k_{\text{graph}} = 5$ are reused.

Dataset	k_{graph}	Accuracy (%)	F1 (%)	AUC (%)
PICCOLO	1	76.60	74.21	75.89
	3	75.68	73.54	75.59
	5	**78.13**	**75.87**	78.23
	10	76.55	74.56	**79.94**
	20	76.75	74.57	76.90
PolypSet	1	87.34	87.23	89.95
	3	87.34	87.05	86.85
	5	**90.62**	**90.87**	92.03
	10	89.20	89.08	**93.00**
	20	89.18	89.13	87.67

Note: Bold $k_{\text{graph}} = 5$ indicates the default setting. denotes the highest value, and denotes the second-highest value within each dataset and metric column.

T, PPR diffusion matrix S, NBI kernel K^N, and the resulting targets P_j^{UW} under the same locked protocol. No graph certificate or target generated with $k_{\text{graph}} = 5$ is reused.

A.5 Restart-probability sensitivity

We vary only the restart probability ρ under the same locked training protocol used for the graph-neighborhood sweep. The resulting Accuracy, F1, and AUC values are reported in Table 8.

A.6 Implementation Details

All comparisons use the same fold-specific paired cross-entropy initialization, WLI training pool, and Stage-2 optimization budget. Each NBI-dependent method uses a fold-specific frozen NBI teacher trained with the supervision allowed by its formulation. PaGKD and PaCT use the complete available NBI training pool, whereas the pair-only methods CPC-Trans, PolypsAlign, and ADD train their NBI teachers using paired NBI images in D_p only. Each teacher is frozen before Stage 2.

Stage 2 runs for 100 epochs with batch size 24 and Adam, using a learning rate of 10^{-4} and weight decay of 10^{-8}. Paired and unpaired batches each perform one optimizer step. PaCT uses $k_{\text{graph}} = 5$ and $\rho = 0.15$, with distillation weights of 0.06 on PICCOLO and 7.0 on PolypSet. The paired route uses cross-entropy, while the unpaired route uses cross-entropy together with the PaCT identity-level KL loss and the unpaired-update correction in Equation (10). Unless otherwise stated, these settings

Table 8: Sensitivity to the restart probability ρ.

Dataset	ρ	Accuracy (%)	F1 (%)	AUC (%)
PICCOLO	0.05	77.77	75.51	76.83
	0.15	**78.13**	**75.87**	78.23
	0.30	75.72	73.71	**79.12**
	0.50	76.83	74.68	77.07
	1.00	72.46	69.85	73.42
PolypSet	0.05	88.54	88.35	88.76
	0.15	**90.62**	**90.87**	**92.03**
	0.30	86.69	86.78	88.98
	0.50	87.34	87.35	90.51
	1.00	83.96	83.75	86.04

Note: Bold $\rho = 0.15$ indicates the default setting. denotes the highest value, and denotes the second-highest value within each dataset and metric column.

are fixed across folds and seeds. For each fold and seed, checkpoints and configurable default settings are selected using the inner-validation split and its F1 score; the held-out test split is not used for selection. For the main results, scores are first averaged over the five folds within each seed, followed by the mean and standard deviation over the three seed-level means.