ResGAT: A Residual Graph Attention Network for Cancer Subtype Classification in Whole Slide Images

Zhenhan Lin[1] ZHENHAN.LIN@VANDERBILT.EDU
Hao Tong[2] TONGH@ALUMNI.WFU.EDU
Yunfei Hu[1] YUNFEI.HU@VANDERBILT.EDU
Xianyong Sean Gui[2] XGUI@WAKEHEALTH.EDU
Jeanne Shen[3] JEANNES@STANFORD.EDU
Byrne Lee [4] BYRNELEE@STANFORD.EDU
Lu Zhang [5] ERICLUZHANG@HKBU.EDU.HK
Daniel Moyer [1] DANIEL.MOYER@VANDERBILT.EDU
Mu Zhou[6] MUZHOU1@GMAIL.COM
Xin Maizie Zhou[1,7,*] MAIZIE.ZHOU@VANDERBILT.EDU
Konstantinos Votanopoulos[2,*] KVOTANOP@WAKEHEALTH.EDU
[1] *Department of Computer Science, Vanderbilt University, Nashville, TN, United States*
[2] *Department of General Surgery, Wake Forest University, Winston-Salem, NC, United States*
[3] *Department of Pathology, Stanford University School of Medicine, Palo Alto, CA, United States*
[4] *Department of Surgery, Stanford University, Palo Alto, CA, United States*
[5] *Department of Computer Science, Hong Kong Baptist University, Hong Kong*
[6] *Department of Computer Science, Rutgers University, New Brunswick, NJ, United States*
[7] *Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, United States*

Editors: Under Review for MIDL 2026

Abstract

Multiple instance learning (MIL) provides a weakly supervised framework for whole slide image (WSI) classification, enabling slide-level prediction from gigapixel images with only slide-level labels. However, WSI subtype classification in realistic settings is still challenging. In this work, we propose ResGAT, a residual graph attention framework that operates on patch graphs and models representations with stacked residual graph attention blocks. ResGAT is evaluated on binary subtype classification task across a rare, class-imbalanced appendiceal cancer cohort and two public TCGA datasets. It outperforms SOTA MIL baselines on the appendiceal cancer cohort and remains competitive on the TCGA datasets. We further assess cross-site generalization via few-shot adaptation under source shift, showing that ResGAT adapts effectively to new domains with limited labels. An ablation study is provided to assess the effectiveness of key architectural components of our method.

Keywords: whole slide image classification, multiple instance learning, residual graph attention framework, cross-site generalization

1. Introduction

As histopathology digitization becomes routine, incorporating computational models into diagnostic workflows is becoming increasingly feasible (Hanna et al., 2019; Kumar et al., 2020; Yilmaz et al., 2024). These computational models provide slide-level classification results together with interpretable justifications, promoting consistent decisions and transparent verification (Tizhoosh and Pantanowitz, 2018; Yilmaz et al., 2024). This is particularly valuable for rare diseases, where expert diagnosticians are scarce. However, a fundamental challenge lies in the gigapixel scale of whole-slide images (WSIs), which prevents them from being processed as a single image. In practice, the standard approach involves tiling tissue regions into thousands of patches, formulating the task as a Multiple Instance Learning (MIL) problem.

The evolution of MIL for WSI classification has shifted from simple feature pooling to sophisticated context modeling. Initial frameworks adopted static aggregation strategies, such as max-pooling (Campanella et al., 2019) and mean-pooling. While computationally efficient, these methods often lose critical contextual information by focusing only on the extreme feature or diluting signals through averaging. The introduction of Attention-based MIL (ABMIL) (Ilse et al., 2018) marked a pivotal advancement by using trainable weights to rank instances. Subsequent research has sought to address overfitting and attention concentration through advanced strategies: pseudo-bag augmentation and feature distillation methods like DTFD-MIL (Zhang et al., 2022); and attention-challenging frameworks such as ACMIL (Zhang et al., 2024) and MHIM (Tang et al., 2023) that mitigate attention concentration by suppressing high-confidence instances to encourage the discovery of comprehensive diagnostic patterns. Despite these improvements, the attention mechanisms often treat instances as independent and identically distributed (i.i.d.). To explicitly capture inter-instance correlations, recent sequence-based works like TransMIL (Shao et al., 2021) and the Mamba-based architecture (Yang et al., 2024) leverage self-attention and selective scan mechanisms to explicitly model long-range dependencies, marking a paradigm shift towards correlated feature learning.

Running parallel to sequence-based advancements, Graph Neural Networks (GNNs) have emerged as a distinct paradigm focused on explicitly encoding the structural topology of the tissue. By representing patches as nodes and their interactions as edges, these methods avoid flattening the spatial structure into a sequence. Early implementations employed k-nearest neighbor (KNN) algorithms to construct spatial graphs, demonstrating that explicitly modeling local neighborhoods enhances diagnostic accuracy (Chen et al., 2021; Zheng et al., 2022). Subsequent research has explored more intricate graph constructions, including hierarchical formulations for multi-resolution reasoning (Hou et al., 2022) and heterogeneous graphs that distinguish between different tissue components (Chan et al., 2023). However, the ”over-smoothing” phenomenon (Chen et al., 2020) is challenging for graph-base MIL approaches. Stacking multiple message passing layers induce node representations to become homogenized, losing the discriminative power essential for classification. This degradation poses an obstacle in realistic clinical settings, which are characterized by extreme heterogeneity in tissue scale. In such diverse scenarios, applying standard readout functions to homogenized features yields inconsistent diagnostic profiles across varying graph sizes will harm the reliability required for clinical deployment.

In this work, we propose a residual graph attention network (ResGAT), a weakly supervised MIL framework tailored for whole slide image subtype classification. ResGAT processes hybrid k-NN patch graphs with stacked residual graph attention blocks, where each block combines a multi-head graph attention branch with a parallel linear projection path. This design preserves individual patch information while updating node features via graph attention, yielding representations that support effective slide-level aggregation across graphs of varying sizes. Through comprehensive evaluation, our model achieves superior performance on a rare, class-imbalanced appendiceal cancer cohort and remains competitive on public TCGA benchmarks. We also introduce a benchmarking protocol to assess cross-site generalization and few-shot adaptation, demonstrating that ResGAT maintains strong performance when labeled data are limited in new domains. Ablation analyses confirm that the hybrid graph connectivity strategy and graph normalization contribute positively to performance and training stability. Furthermore, our framework supports qualitative interpretation with prediction-related heatmaps that can aid diagnostic review.

2. Method

2.1. Problem Formulation

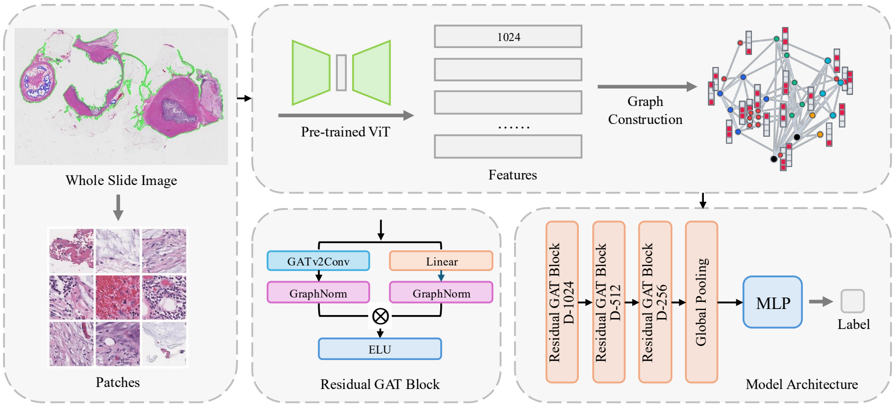

Figure 1: Overview of the ResGAT pipeline for WSI classification. The pipeline consists of three main stages: patch feature extraction, hybrid k-NN patch graph construction, and graph-based slide-level prediction.

We treat each WSI as a bag of patch embeddings under the multiple instance learning (MIL) setting. Given a slide s, we extract tissue patches at a fixed magnification and encode each patch into a feature vector $\mathbf{x}_i \in \mathbb{R}^D$ using a pretrained encoder. This yields a set

$$\mathcal{B}_s = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$

with a bag-level label $y_s \in \{0, 1\}$ indicating the cancer subtype. Our goal is to learn a permutation-invariant function $f_\theta : \mathcal{B}_s \mapsto y_s$ for subtype classification.

Following prior graph-based MIL methods, we represent each slide as a patch graph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, which enables the model to incorporate spatial relationships and feature-level associations among patches. Each node $v_i \in \mathcal{V}_s$ corresponds to a patch embedding $\mathbf{x}_i$, and edges in $\mathcal{E}_s$ are constructed based on both spatial proximity and feature similarity between patches. ResGAT takes the graph as input, applies a stack of residual graph attention blocks, then pools the updated node features into a bag-level representation, and passes it through an MLP classifier to obtain the final prediction $\hat{y}_s$. Fig. 1 shows the overall architecture of ResGAT.

2.2. Graph Construction

For each slide s, we construct edges using a hybrid k-NN procedure. Each node v_i is associated with a spatial coordinate $\mathbf{p}_i \in \mathbb{R}^2$ derived from the patch location on the WSI. We first identify the *d_neighbors* nearest spatial neighbors of v_i in Euclidean coordinate space, and the *f_neighbors* nearest feature neighbors based on cosine distance. We then take the intersection of these two neighbor sets and rank the intersected candidates by feature similarity. The top k nodes from this ranked list form the final adjacency of v_i; when the intersection is empty or insufficient, we fall back to up to three additional feature-nearest neighbors (excluding those already in the intersection) to ensure connectivity.

In all experiments, we fix *f_neighbors* $= 50$ and $k = 6$, and treat the resulting patch graph as undirected. The parameter *d_neighbors* is tuned as a hyperparameter based on validation performance; Section 3.4.2 reports results for all the choices. Empirically, increasing *d_neighbors* enlarges the size of the intersection, which in turn increases the average node degree and yields denser patch graphs.

2.3. ResGAT Architecture and Training Objective

Node Updates. Given a patch graph $\mathcal{G}_s$ with node features $\{\mathbf{h}_i^{(0)}\}_{i=1}^N$ initialized from $\mathbf{x}_i$. Let $\mathbf{h}_i^{(\ell)}$ denote the feature of node i at layer ℓ. ResGAT applies a stack of $L = 3$ residual blocks to obtain updated node representations $\mathbf{h}_i^{(L)}$. Each residual block updates node features through a linear projection in parallel with a GATv2Conv-based(Brody et al., 2021) multi-head graph attention convolution. The combined update takes the form

$$\begin{aligned}
e_{ij}^{(k)} &= \mathbf{a}^{(k)\top}\sigma\big(\mathbf{W}_s^{(k)}\mathbf{h}_i^{(\ell)} + \mathbf{W}_t^{(k)}\mathbf{h}_j^{(\ell)}\big), \qquad j \in \mathcal{N}(i),\\
\alpha_{ij}^{(k)} &= \frac{\exp(e_{ij}^{(k)})}{\sum_{u\in\mathcal{N}(i)}\exp(e_{iu}^{(k)})},\\
\mathbf{m}_i^{(\ell)} &= \Big\Vert_{k=1}^{K} \sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(k)}\mathbf{W}^{(k)}\mathbf{h}_j^{(\ell)},\\
\mathbf{h}_i^{(\ell+1)} &= \phi\Big(\mathrm{GN}\big(\mathbf{m}_i^{(\ell)}\big) + \mathrm{GN}\big(\mathbf{W}_{\text{res}}^{(\ell)}\mathbf{h}_i^{(\ell)}\big)\Big),
\end{aligned} \tag{1}$$

where $\mathbf{W}_s^{(k)}, \mathbf{W}_t^{(k)}, \mathbf{W}^{(k)} \in \mathbb{R}^{d_{\ell+1}\times d_\ell}$ are learnable projections for head k, $\mathbf{a}^{(k)} \in \mathbb{R}^{d_{\ell+1}}$ is the attention vector, σ is the LeakyReLU activation, $\Vert$ denotes concatenation over K heads. $\mathbf{W}_{\text{res}}^{(\ell)} \in \mathbb{R}^{d_{\ell+1}\times d_\ell}$ is the linear projection on the residual path, $\mathrm{GN}(\cdot)$ denotes GraphNorm, and ϕ is the ELU non-linearity. $\mathrm{GN}(\cdot)$ is applied separately to two branches. This update is applied to all nodes for $\ell = 0, \ldots, L-1$, and naturally supports progressively decreasing dimensions (e.g., $1024 \to 512 \to 256$).

Graph Normalization. To stabilize training across slides with different graph sizes and node statistics, we adopt GraphNorm (Cai et al., 2021) within each residual block. Given node features $\{\mathbf{h}_i^{(\ell)}\}_{i=1}^N$ in a graph at layer ℓ, GraphNorm normalizes each node as

$$\mathbf{u}_i^{(\ell)} = \boldsymbol{\gamma} \odot \frac{\mathbf{h}_i^{(\ell)} - \boldsymbol{\alpha} \odot \boldsymbol{\mu}^{(\ell)}}{\sqrt{\left(\boldsymbol{\sigma}^{(\ell)}\right)^2 + \epsilon}} + \boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\mu}^{(\ell)}$ and $\left(\boldsymbol{\sigma}^{(\ell)}\right)^2$ are the mean and variance of $\{\mathbf{h}_i^{(\ell)}\}_{i=1}^N$ over nodes in the graph, and $\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}$ are learnable parameters shared across nodes. The operator $\odot$ denotes element-wise multiplication. Intuitively, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ provide a channel-wise affine re-parametrization of the normalized features, while $\boldsymbol{\alpha}$ modulates the strength of graph-level centering on each feature dimension.

Pooling and Loss. After the residual blocks, we apply global mean pooling over nodes to obtain a bag-level representation $\mathbf{z}_s \in \mathbb{R}^{d_z}$. This vector is fed into an MLP classifier to produce logit vector $[\hat{y}_{s,0}, \hat{y}_{s,1}]$. The predicted probabilities are obtained via a Softmax function. We train the model using the standard cross-entropy loss. Given a slide-level label $y_s \in \{0,1\}$, we encode it as a one-hot vector $\mathbf{y}_s \in \{0,1\}^2$, the loss is

$$\mathcal{L} = -\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \Big[y_{s,1} \log \frac{\exp(\hat{y}_{s,1})}{\exp(\hat{y}_{s,0}) + \exp(\hat{y}_{s,1})} + y_{s,0} \log \frac{\exp(\hat{y}_{s,0})}{\exp(\hat{y}_{s,0}) + \exp(\hat{y}_{s,1})} \Big]. \tag{3}$$

The impact of this two-branch residual block design on overall performance is further evaluated in Section 3.4.

2.4. Heatmap Visualization

To visualize which slide regions most strongly influence the model's subtype prediction, we generate patch-level heatmaps using Grad-CAM++. Given a target class c and feature maps $\{A^k\}$ from the final residual block, Grad-CAM++ computes node-wise importance weights w_k^c from the gradients of the class logit with respect to A^k and forms a class-specific localization map

$$L^c(i,j) = \mathrm{ReLU}\Big(\sum_k w_k^c A^k(i,j) \Big), \tag{4}$$

which highlights locations with positive contribution to the model's score for class c.

3. Experiments

3.1. Dataset and Experimental Setup

3.1.1. Dataset

Appendiceal cancer cohort. The appendiceal cancer cohort consists of diagnostic WSIs from 92 patients with low-grade appendiceal mucinous neoplasm (LAMN) and mucinous adenocarcinoma (MAC). After quality control, the dataset exhibits significant class imbalance (LAMN:MAC = 32:15), with 114 slides from Wake Forest Baptist Health (WF) and 27 from Stanford Health (SF), presenting additional domain shift challenges. Clinically, MAC is regarded as the more aggressive subtype with worse prognosis than LAMN, so in our experiments MAC is treated as the positive class when computing AUC and the reported F1-score corresponds to the positive label.

TCGA cohorts. Two public cohorts were curated from The Cancer Genome Atlas (TCGA) program (Tomczak et al.). The TCGA-NSCLC cohort contains diagnostic WSIs from lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), while the TCGA-ESCA cohort consists of esophageal adenocarcinoma (EAC) and esophageal squamous cell carcinoma (ESCC). For evaluation, the clinically more aggressive subtype—LUSC in NSCLC and ESCC in ESCA—was treated as the positive class when computing AUC.

Slide counts per subtype for all three cohorts are summarized in Table A in the Appendix. All tissue segmentation and patch extraction were performed at 20× magnification.

3.1.2. Evaluation protocols

For each cohort, we perform slide-level 5-fold cross-validation with patient-wise splits. In each split, three folds are used for training, one for validation and one for testing, and we report the mean and standard deviation of metrics over the five test folds. For the appendiceal cancer cohort, which is highly imbalanced, we focus on balanced accuracy and also report AUC and F1-score with MAC as the positive class. For the TCGA-NSCLC and TCGA-ESCA cohorts, overall accuracy and AUC are reported.

For the domain adaptation analysis on the appendiceal cancer cohort, WF slides form the source domain and SF slides the target domain. The WF data are partitioned into training, validation and test subsets in a 70/15/15 ratio for pre-training each model. For the target domain, we define a fixed SF test set of 12 slides (10 LAMN and 2 MAC); this SF test set is used for all zero-shot and few-shot evaluations. Zero-shot performance is obtained by applying the WF-pretrained model directly to the SF test set. For few-shot adaptation, we fine-tune the pretrained model on small labeled SF subsets with 3, 6 and 9 training slides per class and separate validation sets of 3, 3 and 5 slides, respectively. After adaptation, we report overall accuracy on the SF test set. We also compute backward transfer (BWT), defined as the change in WF test accuracy before and after fine-tuning; large negative BWT values indicate catastrophic forgetting. Forward transfer (FWT) is computed as the improvement of SF test accuracy over the zero-shot baseline, where positive values indicate successful adaptation.

See Appendix B for implementation details.

3.2. Comparison with state-of-the-art methods

We compare our method with eight strong MIL baselines that cover diverse design paradigms: attention-based pooling MIL (CLAM-SB and CLAM-MB (Lu et al., 2021)), transformer-based MIL (TransMIL (Shao et al., 2021)), dual-stream MIL (DSMIL (Li et al., 2021)), distillation-based MIL (DTFD-MIL (Zhang et al., 2022)), graph-based MIL (WiKG (Li et al., 2024)), and hard-instance-mining MIL (MHIM-DSMIL and MHIM-TransMIL (Tang et al., 2023, 2025)). All methods are evaluated with a shared UNI-based feature extractor (Chen et al., 2024), which uses a ViT-L/16 backbone pretrained with DINOv2 on a large histopathology corpus to produce 1024-dimensional patch embeddings. Table 1 reports mean and standard deviation over five folds for all metrics on the three cohorts.

Table 1: Subtype classification performance (mean$_{\text{std}}$, %) on three datasets: appendiceal cancer, TCGA-NSCLC, and TCGA-ESCA, reported as balanced accuracy (BAcc), AUC, F1-score, and accuracy.

Method	**Appendiceal Cancer**			**TCGA-NSCLC**		**TCGA-ESCA**	
	BAcc	**AUC**	**F1**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
CLAM-SB	$90.09_{6.47}$	$94.96_{8.79}$	$86.25_{8.15}$	$\mathbf{93.72_{1.72}}$	$97.55_{1.44}$	$\mathbf{98.04_{1.60}}$	$99.83_{0.34}$
CLAM-MB	$88.62_{10.68}$	$\mathbf{96.82_{4.13}}$	$85.36_{14.95}$	$92.70_{1.53}$	$97.39_{1.57}$	$96.11_{3.16}$	$100.00_{0.00}$
DSMIL	$78.92_{13.86}$	$90.58_{9.89}$	$68.44_{24.77}$	$92.29_{1.40}$	$97.08_{1.53}$	$95.42_{4.45}$	$97.72_{2.65}$
TransMIL	$84.07_{10.71}$	$92.47_{7.64}$	$76.87_{14.33}$	$92.29_{2.13}$	$97.15_{0.82}$	$93.51_{4.08}$	$99.39_{0.52}$
WiKG	$84.31_{7.39}$	$94.37_{6.51}$	$79.16_{11.19}$	$92.09_{1.94}$	$96.35_{1.45}$	$93.48_{3.59}$	$99.63_{0.74}$
DTFD-MIL	$86.22_{9.56}$	$93.27_{11.35}$	$80.08_{13.02}$	$93.61_{1.75}$	$97.41_{1.38}$	$96.11_{3.16}$	$99.39_{0.65}$
MHIM-DSMIL	$86.42_{12.74}$	$97.03_{2.72}$	$81.15_{19.45}$	$92.70_{1.23}$	$97.48_{1.23}$	$94.82_{4.37}$	$98.88_{1.80}$
MHIM-TransMIL	$87.49_{9.45}$	$91.59_{11.69}$	$84.94_{13.47}$	$92.40_{1.31}$	$97.30_{1.51}$	$94.82_{4.37}$	$99.73_{0.22}$
ResGAT (ours)	$\mathbf{92.56_{6.36}}$	$96.41_{1.94}$	$\mathbf{90.98_{7.98}}$	$93.51_{0.75}$	$97.15_{1.47}$	$98.02_{1.62}$	$99.91_{0.17}$

Table 2: Domain adaptation performance comparison. Source refers to pre-trained test accuracy from WF dataset. Zero-shot refers to SF test performance on two classes data separately without adaptation. FWT measures forward transfer (target improvement), BWT measures backward transfer (source performance retention). Class 0 and Class 1 represent LAMN and MAC respectively.

Method	**Source(WF)**	**Zero-shot (SF)**		**3-shot (SF)**			**6-shot (SF)**			**9-shot (SF)**		
	Accuracy	**class 0**	**class 1**	**Acc**	**FWT**	**BWT**	**Acc**	**FWT**	**BWT**	**Acc**	**FWT**	**BWT**
WiKG	89.47	100	0	83.33	0	10.5	83.33	0	5.26	83.33	0	5.26
TransMIL	84.21	100	0	83.33	0	0	83.33	0	0	83.33	0	0
DSMIL	73.68	70	50	75.0	8.33	0	75.0	8.33	0	75.0	8.33	0
MHIM-DSMIL	84.21	90	0	75.0	0	5.26	75.0	0	5.26	83.33	8.33	0
MHIM-TransMIL	89.47	100	0	83.33	0	0	91.67	8.33	5.26	100	16.67	5.26
CLAM-MB	89.47	100	0	83.33	0	0	83.33	0	0	83.33	0	5.26
CLAM-SB	**94.74**	90	0	75.0	0	0	75.0	0	0	75.0	0	0
DTFT-MIL	89.47	**90**	**100**	91.67	0	0	100	8.33	0	100	8.33	5.26
ResGAT	**92.86**	100	50	**100**	**8.33**	**0**	**100**	**8.33**	**0**	**100**	**8.33**	**0**

Table 1 summarizes balanced accuracy, AUC, F1-score and accuracy for binary subtype classification on the appendiceal cancer, TCGA-NSCLC and TCGA-ESCA cohorts. On the appendiceal cancer cohort, ResGAT achieves the highest balanced accuracy at 92.56±6.36%, outperforming the best baseline CLAM-SB by roughly 2.5% and yielding the lowest standard deviation across folds. It also attains the highest F1-score and a high AUC, indicating good detection of the clinically more aggressive MAC subtype while preserving good overall discrimination between subtypes. On TCGA-NSCLC and TCGA-ESCA, CLAM-SB attains the highest mean accuracy, while ResGAT remains competitive: its accuracy is only 0.21% and 0.02% below CLAM-SB on TCGA-NSCLC and TCGA-ESCA, respectively. Notably, ResGAT's low standard deviations on TCGA cohorts shows stable performance across folds. Overall, these results indicate that ResGAT performs well on the small, class-imbalanced and label-noisy appendiceal cancer cohort, while remaining comparable to competitive MIL baselines on the larger public datasets.

The results also highlight complementary strengths of other MIL approaches. On the two TCGA cohorts, DTFD-MIL obtain the second highest accuracies and AUCs, with CLAM-

MB generally close behind. The MHIM variants (MHIM-DSMIL and MHIM-TransMIL) consistently improve over their backbones, and show the effectiveness of the hard-instance mining strategy. UNI features provide higher quality embeddings than ResNet50 used in prior WSI classification studies, which is reflected in the higher general accuracy and AUC across methods.

3.3. Domain Adaptation Analysis

In this experiment, we evaluate cross-site robustness on the appendiceal cancer cohort, where WF and SF correspond to different acquisition sites (see Section 3.1.1 for details). Such cross-site settings often introduce substantial distribution shift due to differences in scanners, staining protocols and local practice, and models trained on a single site can experience a marked performance drop when deployed elsewhere(Liu et al., 2025; PoceviVCiute et al., 2024). We therefore use this scenario to assess generalization ability of methods, which is an important consideration for realistic clinical deployment. We first evaluate zero-shot performance, where a model trained on the source site is directly applied to the target site. Then we evaluate few-shot adaptation, where only a small number of labeled SF slides are available for finetuning the source-trained model (see Section 3.1.2 for details).

3.3.1. Cross-domain Generalization

Table 2 compares our method with the same eight MIL baselines as in the previous experiment. Most MIL baselines achieve reasonably high accuracy on the WF source test set, but their zero-shot performance on the SF target set is highly variable and often subtype-imbalanced. Several baselines, including WiKG, TransMIL and the CLAM variants, rarely predict MAC samples correctly, indicating a strong bias towards the majority subtype when crossing sites. In the meanwhile, DTFD-MIL achieves the strongest zero-shot performance on the SF test set, with per-class accuracies of 90% and 100%, suggesting good cross-site generalization. ResGAT attains the second-highest source-domain accuracy on the WF test set and provides competitive zero-shot accuracy on the SF test set, indicating good performance on both sites.

3.3.2. Few-shot Adaptation

In this experiment, we analyze how pre-trained models adapt target data when fine-tuned on a small number of labeled SF slides. ResGAT reaches 100% accuracy on the SF test set at the 3-shot setting and maintains this performance at 6-shot and 9-shot. Its already high source test performance remains unchanged across all settings (BWT = 0), showing that adaptation does not induce forgetting on the source domain. This result suggests that ResGAT can be effectively adapted to a new site using only a small number of labeled slides, which is especially valuable in rare-disease scenarios where annotation is costly and limited.

DTFD-MIL attains strong zero-shot accuracy on SF and reaches 100% SF test accuracy at 6-shot and 9-shot with positive BWT at 9-shot seeting. This pattern is consistent with effective adaptation to the target domain without compromising the source domain. MHIM-TransMIL also shows increased SF accuracy as more target slides are used, together with positive BWT, indicating stable improvement under additional target supervision.

Table 3: Ablation on normalization layers for ResGAT. Values are mean_{std} over 5-fold cross-validation (%).

Normalization	Appendiceal Cancer		TCGA-NSCLC		TCGA-ESCA	
	BAcc	AUC	Accuracy	AUC	Accuracy	AUC
InstanceNorm	$89.23_{7.70}$	$95.84_{2.19}$	$\mathbf{93.51_{0.66}}$	$\mathbf{97.18_{1.41}}$	$98.02_{1.62}$	$99.91_{0.17}$
LayerNorm	$81.32_{11.15}$	$91.31_{7.19}$	$91.48_{1.98}$	$96.63_{1.77}$	$93.46_{4.08}$	$99.30_{0.57}$
GraphNorm	$\mathbf{92.56_{6.36}}$	$\mathbf{96.41_{1.94}}$	$93.51_{0.75}$	$97.15_{1.47}$	$\mathbf{98.02_{1.62}}$	$\mathbf{99.91_{0.17}}$

Table 4: Ablation on edge construction for ResGAT. Values are mean_{std} over 5-fold cross-validation (%).

Graph Variant	Appendiceal Cancer		TCGA-NSCLC		TCGA-ESCA	
	BAcc	AUC	Accuracy	AUC	Accuracy	AUC
Feature kNN	$90.97_{4.82}$	$96.45_{3.86}$	$92.70_{1.71}$	$97.21_{1.17}$	$98.02_{1.62}$	$99.91_{0.17}$
Spatial kNN	$91.79_{6.22}$	$96.06_{4.22}$	$93.10_{1.24}$	$97.35_{1.37}$	$97.35_{2.49}$	$99.45_{0.68}$
Hybrid (d=24)	$\mathbf{92.56}_{6.36}$	$96.41_{1.94}$	$92.60_{1.56}$	$97.65_{0.90}$	$\mathbf{98.02}_{1.62}$	$\mathbf{99.91}_{0.17}$
Hybrid (d=15)	$90.78_{6.05}$	$94.23_{3.69}$	$\mathbf{93.51}_{0.75}$	$97.15_{1.47}$	$98.02_{1.62}$	$99.54_{0.93}$
Disconnect	$88.95_{6.71}$	$\mathbf{96.52}_{3.08}$	$93.11_{0.59}$	$\mathbf{97.94}_{0.74}$	$97.38_{2.43}$	$99.82_{2.20}$
Node-permuted	$92.46_{6.14}$	$96.08_{3.53}$	$92.80_{0.99}$	$97.70_{1.00}$	$98.02_{1.62}$	$99.83_{0.21}$

By contrast, CLAM-SB and CLAM-MB, despite their strong performance on general subtype classification benchmarks, show little change in SF accuracy across the 3, 6, and 9-shot settings, suggesting that their architectures are less responsive to small amounts of target supervision in our cross-site few-shot protocol.

3.4. Ablation Study

3.4.1. Effectiveness of GraphNorm

Table 3 shows that GraphNorm provides the most favorable performance pattern within ResGAT. On the appendiceal cancer cohort, it yields the best balanced accuracy and AUC—improving BAcc by 3.3% over InstanceNorm and 11.2% over LayerNorm—with lower variance. GraphNorm normalizes node representations per graph and includes learnable affine parameters that control the scaling and shifting of normalized features; this design is consistent with the improved stability observed on this small and noisy cohort. On TCGA-NSCLC and TCGA-ESCA, GraphNorm's performance is comparable to InstanceNorm and above LayerNorm, supporting its use as the default normalization layer across datasets.

3.4.2. Effectiveness of Proposed Edge Construction

We evaluate a set of graph variants: Feature kNN (edges based on feature similarity), Spatial kNN (edges based on spatial proximity), Hybrid (edges combining spatial and feature criteria), Disconnected (only self-loop), and Node-permuted (hybrid adjacency with features randomly reassigned to nodes). For all connected variants we use $k = 6$ neighbors per node; in the hybrid case, we vary only the $d\_neighbors$ hyperparameter, while all other settings are kept the same (see Section 2.2 for details).

As shown in Table 4, on the appendiceal cancer cohort, all connected graph variants outperform the disconnected (MLP-only) backbone; among them, the hybrid graph ($d\_neighbors = 24$) achieves the highest balanced accuracy and AUC, and exhibits the lowest variance across folds. Feature-kNN and Spatial-kNN graphs perform worse than the hybrid variant, indicating that edges relying only on feature similarity or spatial proximity is insufficient under small-sample, noisy, and class-imbalanced conditions. This observation suggests that combining both spatial and feature criteria improves the modeling for this cohort. Moreover, the node-permuted variant — which retains the adjacency structure but disrupts the alignment between node features and spatial positions — performs similarly to other connected graphs and substantially better than the disconnected baseline. This result supports the interpretation that the benefit of the graph-based branch in this setting stems not strictly from preserving exact spatial-feature alignment, but from the graph inductive bias imposed by graph connectivity and regularized neighborhood aggregation, which helps stabilize learning and mitigate overfitting.

On the TCGA-NSCLC and TCGA-ESCA cohorts, performance differences across all graph variants are small, but the hybrid graph consistently yields the best results among them. These results indicate that, although specific graph construction has only a limited effect on overall performance in these cohorts, our method still offers a modest advantage.

3.5. Qualitative Results

Using the Grad-CAM++ procedure described in Section 2.4, we compute patch-level contribution maps and visualize them as WSI-level heatmaps. We show an example from a MAC case ($S36$) in the appendiceal cancer cohort (Appendix C), including the slide-level heatmap and twelve selected patches from those with the highest contribution scores. These visualizations highlight regions that ResGAT associates with the predicted subtype. Although the highlighted areas do not fully satisfy pathologists' diagnostic requirements and tumour localization can be inaccurate, they provide a stable, prediction-related reference that may assist slide review.

4. Conclusion

In this work, we propose ResGAT, a residual graph attention framework designed for weakly supervised WSI subtype classification under challenging clinical settings. Comprehensive evaluations demonstrate that ResGAT outperforms strong state-of-the-art MIL baselines on a rare appendiceal cancer cohort and remains competitive on two public TCGA benchmarks. Furthermore, ResGAT shows promising cross-site generalization in few-shot adaptation experiments, maintaining strong performance when limited labeled data are available in target domains. Our ablation study shows that both the hybrid graph connectivity strategy and the use of graph normalization contribute to improved performance and more stable training. These findings highlight ResGAT as a robust and generalizable approach for WSI classification tasks, especially in noise-prone and data-scarce medical scenarios.

References

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

Tianle Cai, Shengjie Luo, Keyulu Xu, Di He, Tie-yan Liu, and Liwei Wang. Graphnorm: A principled approach to accelerating graph neural network training. In *International Conference on Machine Learning*, pages 1204–1215. PMLR, 2021.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15661–15670, 2023.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International conference on machine learning*, pages 1725–1735. PMLR, 2020.

Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3): 850–862, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Matthew G Hanna, Victor E Reuter, Jennifer Samboy, Christine England, Lorraine Corsale, Samson W Fine, Narasimhan P Agaram, Evangelos Stamelos, Yukako Yagi, Meera Hameed, et al. Implementation of digital pathology offers clinical and operational increase in efficiency and cost savings. *Archives of pathology & laboratory medicine*, 143 (12):1545–1555, 2019.

Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. Hˆ 2-mil: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 933–941, 2022.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Neeta Kumar, Ruchika Gupta, and Sanjay Gupta. Whole slide imaging (wsi) in pathology: current perspectives and future directions. *Journal of digital imaging*, 33(4):1034–1040, 2020.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11323–11332, 2024.

Jingsong Liu, Han Li, Chen Yang, Michael Deutges, Ario Sadafi, Xin You, Katharina Breininger, Nassir Navab, and Peter J Schüffler. Hasd: hierarchical adaption for pathology slide-level domain-shift. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 332–342. Springer, 2025.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

M PoceviVCiute, G Eilertsen, S Garvin, and C Lundstrom. Detecting domain shift in multiple instance learning for digital pathology using fréchet domain distance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 157–167, 2024.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4078–4087, 2023.

Wenhao Tang, Sheng Huang, Heng Fang, Fengtao Zhou, Bo Liu, and Qingshan Liu. Multiple instance learning framework with masked hard instance mining for gigapixel histopathology image analysis, 2025. URL https://arxiv.org/abs/2509.11526.

Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics*, 9(1):38, 2018.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77.

Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International conference on medical image computing and computer-assisted intervention*, pages 296–306. Springer, 2024.

Fazilet Yilmaz, Arlen Brickman, Fedaa Najdawi, Evgeny Yakirevich, Robert Egger, and Murray B Resnick. Advancing artificial intelligence integration into the pathology workflow: Exploring opportunities in gastrointestinal tract biopsies. *Laboratory Investigation*, 104(5):102043, 2024.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18802–18812, 2022.

Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *European conference on computer vision*, pages 125–143. Springer, 2024.

Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022.

Appendix A. Dataset Table

Table 5: Dataset statistics for Appendiceal Cancer, TCGA-NSCLC, and TCGA-ESCA.

Dataset	**Label**	**Diagnosis**	**Number of WSIs**	
			Site 1	**Site 2**
Appendiceal Cancer	0	LAMN	74	22
	1	MAC	40	5
TCGA-NSCLC	0	LUAD	496	
	1	LUSC	490	
TCGA-ESCA	0	EAC	63	
	1	ESCC	90	

Appendix B. Implementation Details

All experiments were conducted on an NVIDIA RTX A6000 GPU with 48GB memory. For feature extraction, we adopted the CLAM (Lu et al., 2021) preprocessing pipeline with HSV-based tissue segmentation and contour-based spatial sampling to identify tissue regions. Features were extracted using UNI (Chen et al., 2024)(ViT-L/16 via DINOv2) pretrained on the Mass-100K histopathology corpus, processing 224×224 patches to produce 1024-dimensional feature vectors with standard ImageNet normalization (Deng et al., 2009).

For ResGAT model, we trained for 30 epochs using Adam optimizer with learning rate 3×10^{-4} and weight decay 1×10^{-4}. To account for randomness, each experiment was repeated with two random seeds 3 and 3407; the best-performing run is reported. Following standard MIL practice, we applied batch size of 1. For baseline methods, we used their recommended hyperparameters from official implementations to ensure fair comparison.

Appendix C. Additional Results

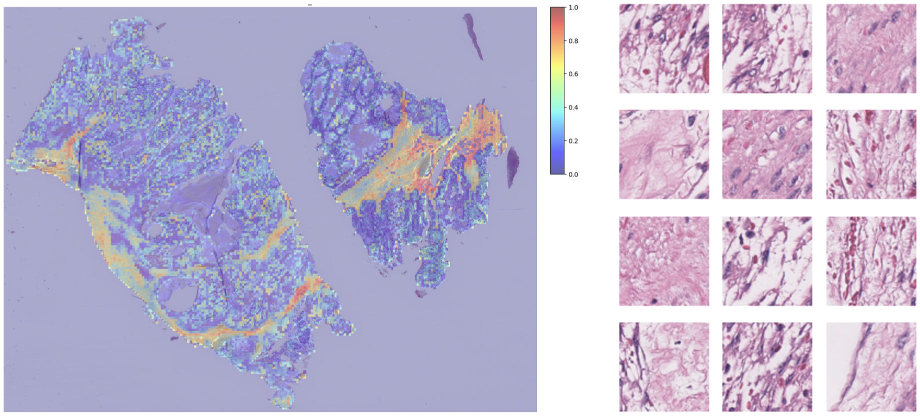

Figure 2: Heatmap visualization on a MAC case (sample S36) from the appendiceal cancer cohort. **Left:** Whole-slide image overlaid with patch-level contribution scores from ResGAT, where the colour bar encodes normalized contribution values from 0 (blue, low) to 1 (red, high). **Right:** Example image patches selected from those with the highest contribution scores, illustrating regions that the model associates with the predicted subtype.

This figure presents the qualitative result discussed in Section 3.5.