Frequency and Geometry Guided Graph Clustering for Weakly Supervised Skin Lesion Segmentation

Zhaoxin Deng[1*], Dongang Wang[2*], and Hualei Shen[1(✉)]

[1] College of Computer and Information Engineering, College of Artificial Intelligence, Henan Normal University, Xinxiang, China
shenhualei@htu.edu.cn
[2] Brain and Mind Centre, The University of Sydney, New South Wales, Australia

Abstract. Pixel-level annotation for skin lesion segmentation is costly and subjective. This limits large-scale clinical deployment. Weakly supervised methods based on image-level labels alleviate this burden. However, most existing methods rely on class activation maps (CAMs), which often fail to cover the entire lesion. To address this, we introduce a Frequency and Geometry Guided Graph Clustering framework, reformulating weakly supervised skin lesion segmentation as a differentiable node-clustering problem. First, a Frequency-aware Feature Modulation module injects local spectral priors into deep features. This enhances the model's sensitivity to ambiguous boundaries. Second, we construct a semantic-spatial graph from these features. A graph neural network (GNN) then captures long-range dependencies to ensure structural completeness. Finally, we propose a tri-source fusion strategy that integrates three cues (GNN cluster, CAM, and the input image) via an improved side-window mean filtering. Experiments on ISIC 2017, ISIC 2018, and PH^2 show that our method achieves state-of-the-art performance. It significantly narrows the gap with fully supervised models, providing a reliable solution for label-efficient skin lesion analysis.

Keywords: Weakly Supervised Semantic Segmentation · Skin Lesion Segmentation · Graph Neural Network · Frequency Modulation

1 Introduction

Malignant melanoma is among the fastest-growing malignancies worldwide. According to the American Cancer Society, an estimated 212,000 new cases with approximately 8,400 deaths are attributed to the disease in the United States in 2025 [2]. Although fully supervised methods have achieved substantial progress in skin lesion analysis, their clinical deployment remains challenging. It is time-consuming and expensive to label pixel-level skin lesion annotations. To alleviate this limitation, weakly supervised semantic segmentation (WSSS) employing

* Z. Deng and D. Wang contributed equally to this paper and are considered co-first authors.

image-level labels has been investigated in recent years [5], which aims to achieve competitive segmentation results with reduced annotation cost.

Existing WSSS methods primarily rely on class activation maps (CAMs) [32] to generate initial pseudo-masks. However, original CAMs often activate only the most discriminative regions [1,21,14], such as the lesion core, and may miss non-salient regions. This limitation stems from the discriminative objective of classification tasks and the local inductive bias of convolutional neural networks (CNNs) [5], leading to pseudo-labels that are fragmented or structurally incomplete. Recent studies mitigate this issue using multi-scale feature fusion [31], intermediate contrastive supervision [22,19,23], or boundary constraints [1,8,29,9]. However, these methods rely primarily on local receptive fields of CNNs or independent pixel-level constraints, and they do not explicitly model global topology.

Graph neural networks (GNNs) offer a complementary perspective by modeling irregular topological structures and capturing long-range dependencies among pixels. For instance, Yang et al. proposed MoRe [24], which models ViT class-patch attention as a directed graph and regularizes it via graph-based category representation to capture long-range token dependencies. Zhang et al. introduced A^2GNN [28] to build a weighted graph using an affinity CNN and aggregate features from confident seeds. Furthermore, Zhang et al. [30] incorporated external commonsense knowledge via dual graph reasoning across spatial and interaction spaces.

However, existing GNN-based WSSS methods still largely rely on smoothing or diffusion driven by low-level cues, without an explicit semantic grouping mechanism that can dynamically aggregate pixels with similar semantic meaning. In addition, most methods operate purely in the spatial domain and may underutilize implicit spectral information in the frequency domain [13,25] that is informative for distinguishing complex textures such as hair occlusions. Finally, graph-based feature propagation is commonly performed on downsampled feature maps [1], which can blur spatial detail and reduce geometric consistency with sharp lesion boundaries in the original image [18].

To address these challenges, we reformulate weakly supervised skin lesion segmentation as a graph node-clustering problem and propose a Frequency and Geometry Guided Graph Clustering (**FG3-Cluster**) framework. By constructing a graph with semantic and spatial constraints and integrating frequency-domain boundary priors, our method explicitly aggregates global context. This generates structurally complete pseudo-masks while preserving fine-grained boundaries. Our code is publicly available at https://github.com/hualei213/FG3-Cluster.

Our main contributions are summarized as follows:

- By leveraging GNNs to capture long-range dependencies, we propose a Differentiable Graph Clustering (DGC) framework to model the intrinsic lesion structure and thus generate better candidate clusters.
- Frequency-aware Feature Modulation (FFM) injects high-frequency priors into graph nodes to enhance discriminability along ambiguous boundaries.

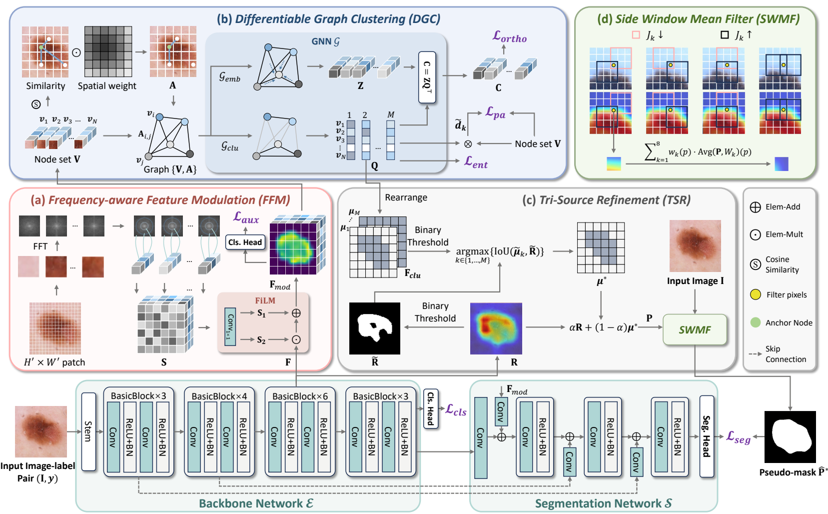

Fig. 1. Overview of the proposed FG3-Cluster. Its key components include (a) Frequency-aware Feature Modulation (FFM) module, (b) Differentiable Graph Clustering (DGC) module, and (c) Tri-Source Refinement (TSR) module.

- Tri-Source Refinement (TSR) integrates graph predictions, CAM responses, and input image guided geometric filtering via improved Side Window Mean Filter (SWMF) [26], aligning pseudo masks with lesion edges.
- Extensive experiments on ISIC 2017, ISIC 2018, and PH^2 demonstrate state-of-the-art performance and strong robustness.

2 Method

As shown in Fig. 1, FG3-Cluster is an end-to-end weakly supervised framework that integrates frequency-domain and geometry priors with graph-based clustering. Built upon a backbone network $\mathcal{E}$, FG3-Cluster comprises three core modules: (1) Frequency-aware Feature Modulation (FFM) to modulate backbone features using frequency cues, (2) Differentiable Graph Clustering (DGC) to generate discriminative semantic clusters, and (3) Tri-Source Refinement (TSR) to produce high-quality pseudo-masks.

2.1 Frequency-aware Feature Modulation

Patch-wise Spectral Modeling. As shown in Fig. 1 (a), to capture spatially varying frequency statistics, we partition the input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ along the spatial dimensions into N non-overlapping patches of size (H', W'). For the n-th patch, we compute its 2D Fast Fourier Transform (FFT) in the luminance space to obtain the spectral energy $E_n(u, v)$, where (u, v) denotes discrete frequency indices. For each patch, we then divide the centered frequency plane into K concentric radial bands, and define the energy proportion of the k-th band as:

$$s_{n,k} = \frac{\sum_{(u,v)\in\Omega_k} E_n(u,v)}{\sum_{u,v} E_n(u,v) + \epsilon}, \tag{1}$$

where Ω_k is a predefined band mask and ϵ ensures numerical stability. Aggregating $s_{n,k}$ across all patch locations yields the spectral feature map $\mathbf{S} \in \mathbb{R}^{K\times H/H'\times W/W'}$ of image $\mathbf{I}$.

Feature Modulation. We follow FiLM [17] to modulate the spatial feature $\mathbf{F}$ extracted by the backbone network $\mathcal{E}$ using $\mathbf{S}$:

$$\begin{aligned} &\mathbf{F} = \mathcal{E}(\mathbf{I}), \quad [\mathbf{S}_1, \mathbf{S}_2] = \mathrm{Split}\left(\mathrm{Conv}_{1\times1}(\mathbf{S})\right), \\ &\mathbf{F}_{mod} = (1 + \sigma(\mathbf{S}_1)) \odot \mathbf{F} + \tanh(\mathbf{S}_2), \end{aligned} \tag{2}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ denotes a 1×1 convolution, $\mathrm{Split}(\cdot)$ partitions features equally along the channel dimension, $\odot$ denotes element-wise multiplication, $\tanh(\cdot)$ is the hyperbolic tangent function, and $\sigma(\cdot)$ is the sigmoid function. Here, $\mathbf{F}, \mathbf{F}_{mod} \in \mathbb{R}^{D\times\frac{H}{16}\times\frac{W}{16}}$, where D is the channel dimension.

Auxiliary Supervision. As illustrated in Fig. 1 (a), we employ an auxiliary classification head with cross-entropy loss $\mathcal{L}_{aux}$, supervised by the image-level label y. $\mathcal{L}_{aux}$ regularizes the learning of $\mathbf{F}_{mod}$ by reinforcing class-discriminative frequency components and suppressing background interference, thereby improving representation separability.

2.2 Differentiable Graph Clustering

Graph Construction. As shown in Fig. 1 (b), we reformulate the modulated feature map $\mathbf{F}_{mod}$ as graph representation and feed it into a GNN $\mathcal{G}$ to capture long-range dependencies. Specifically, we reshape $\mathbf{F}_{mod}$ into a node set $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^N$ and build a spatial-semantic affinity matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ to jointly capture feature similarity and spatial proximity:

$$\mathbf{A}_{i,j} = \mathrm{ReLU}\left(\frac{\boldsymbol{v}_i^\top \boldsymbol{v}_j}{\|\boldsymbol{v}_i\|_2\|\boldsymbol{v}_j\|_2}\right) \cdot \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\tau^2}\right) \tag{3}$$

where $\boldsymbol{x}_i$ denotes the spatial coordinate of node $\boldsymbol{v}_i$, τ controls the Gaussian kernel bandwidth, and $\|\cdot\|_2$ is the Euclidean norm.

Graph-based Clustering. We leverage GraphSAGE [10] for feature propagation. As shown in Fig. 1 (b), the GNN $\mathcal{G}$ comprises two branches: an embedding branch $\mathcal{G}_{emb}$ and a clustering branch $\mathcal{G}_{clu}$. $\mathcal{G}_{emb}$ generates enhanced node representations $\mathbf{Z}$, while $\mathcal{G}_{clu}$ predicts a soft assignment matrix $\mathbf{Q}$:

$$\mathbf{Z} = \mathrm{ReLU}\left(\mathcal{G}_{emb}(\mathbf{V}, \mathbf{A})\right), \quad \mathbf{Q} = \mathrm{Softmax}\left(\mathcal{G}_{clu}(\mathbf{V}, \mathbf{A})\right), \tag{4}$$

where $\mathbf{Z} \in \mathbb{R}^{D\times N}, \mathbf{Q} \in \mathbb{R}^{M\times N}$, and M is the predefined number of clusters.

Each element $\mathbf{Q}(i,k)$ represents the probability that node $\boldsymbol{v}_i$ belongs to cluster k. To discourage overly smooth assignments characterized by nearly uniform probabilities across clusters, we introduce a clustering entropy loss [27]:

$$\mathcal{L}_{ent} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}\mathbf{Q}_{i,k}\log\mathbf{Q}_{i,k}. \tag{5}$$

This encourages the network to make more deterministic cluster assignments, producing sharper skin lesion boundaries in the generated pseudo-masks.

Following DiffPool [27], we construct coarse cluster centers $\mathbf{C} = \mathbf{Z}\mathbf{Q}^\top$, where $\mathbf{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_M\}, \boldsymbol{c}_i \in \mathbb{R}^{D\times 1}$. To mitigate collapse and encourage distinct cluster centers, we introduce an orthogonality loss $\mathcal{L}_{ortho}$ [3]:

$$\mathcal{L}_{ortho} = \frac{1}{M^2}\sum_{i\neq j}\left(\tilde{\mathbf{c}}_i^\top\tilde{\mathbf{c}}_j\right)^2, \quad \tilde{\mathbf{c}}_i = \frac{\boldsymbol{c}_i}{\|\boldsymbol{c}_i\|_2}. \tag{6}$$

Hypersphere Prototype Alignment. We further propose a hypersphere prototype alignment loss $\mathcal{L}_{pa}$ to enforce geometric compactness in the latent space. By projecting features onto the unit hypersphere and optimizing a cosine-distance objective, the loss enhances intra-cluster cohesion and stabilizes clustering:

$$\mathcal{L}_{pa} = \frac{2}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}\mathbf{Q}_{i,k}\left(1 - \tilde{\boldsymbol{v}}_i^\top\tilde{\boldsymbol{d}}_k\right), \quad \tilde{\boldsymbol{v}}_i = \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|_2}. \tag{7}$$

The cluster prototype is defined as:

$$\boldsymbol{d}_k = \frac{\sum_{i=1}^{N}\mathbf{Q}_{i,k}\tilde{\boldsymbol{v}}_i}{\sum_{i=1}^{N}\mathbf{Q}_{i,k}}, \quad \tilde{\boldsymbol{d}}_k = \frac{\boldsymbol{d}_k}{\|\boldsymbol{d}_k\|_2}. \tag{8}$$

Finally, the soft assignment matrix $\mathbf{Q}$ is reshaped into $\mathbf{F}_{clu} \in \mathbb{R}^{M\times\frac{H}{16}\times\frac{W}{16}}$, yielding M cluster maps.

2.3 Tri-Source Refinement

As shown in Fig. 1 (c), our proposed TSR module refines pseudo-masks using three sources: the DGC output $\mathbf{F}_{clu}$, the LayerCAM [12] response $\mathbf{R}$, and the input image $\mathbf{I}$. Specifically, both $\mathbf{F}_{clu}$ and $\mathbf{R}$ are bilinearly upsampled to $H\times W$. From the upsampled cluster maps $\boldsymbol{\mu}_k \in \mathbf{F}_{clu}$, we select the optimal cluster $\boldsymbol{\mu}^*$ whose binarized mask $\tilde{\boldsymbol{\mu}}_k$ achieves the highest IoU with the binarized response map $\tilde{\mathbf{R}}$. The initial pseudo-mask is obtained via weighted fusion, $\mathbf{P} = \alpha\mathbf{R} + (1-\alpha)\boldsymbol{\mu}^*$, where the weighting factor α is set to 0.5.

As shown in Fig. 1 (d), to sharpen skin lesion boundaries, we apply an improved Side Window Mean Filter (SWMF) [26] guided by input image $\mathbf{I}$. In detail, for each pixel p, eight directional side windows $\{W_k\}_{k=1}^{8}$ are scored by:

$$J_k = \mathrm{Var}(\mathbf{I}, W_k) + \lambda_c\left\|\mathrm{Avg}(\mathbf{I}, W_k) - \mathbf{I}(p)\right\|_2^2 + \lambda_s\mathrm{Var}(\mathbf{P}, W_k), \tag{9}$$

where $\mathrm{Avg}(\mathbf{X}, W)$ and $\mathrm{Var}(\mathbf{X}, W)$ denote the mean and variance of $\mathbf{X}$ over window W (channel-wise for $\mathbf{X}$). Finally, we obtain the refined pseudo-mask $\mathbf{P}^*$ by:

$$\mathbf{P}^*(p) = \sum_{k=1}^{8} w_k(p) \cdot \mathrm{Avg}(\mathbf{P}, W_k)(p), \qquad w_k = \frac{\exp(-J_k(p)/T)}{\sum_{s=1}^{8} \exp(-J_s(p)/T)}. \tag{10}$$

This training-free strategy improves boundary alignment with image edges. The binary mask $\hat{\mathbf{P}}^*$ is used as the target of the segmentation task.

2.4 Training Objective

We employ a progressive training schedule to mitigate noise introduced by initial CAMs and graph clusters. For **warm-up** (epochs $[0, 20)$), we optimize the network with image-level label supervision by minimizing $\mathcal{L}_1 = \mathcal{L}_{cls} + \lambda_{aux}\mathcal{L}_{aux}$. Here, $\mathcal{L}_{cls}$ denotes the cross-entropy loss supervised by the image-level label y to optimize the backbone network $\mathcal{E}$. For **structure construction** (epochs $[20, 40)$), we introduce graph-clustering regularization and optimize $\mathcal{L}_2 = \mathcal{L}_1 + \lambda_{pa}\mathcal{L}_{pa} + \lambda_{ent}\mathcal{L}_{ent} + \lambda_{ortho}\mathcal{L}_{ortho}$ to enhance the learning of discriminative semantic clusters. Finally, during **joint fine-tuning** (epochs $[40, e_{total}]$), we incorporate pixel-level supervision using the generated pseudo-mask $\hat{\mathbf{P}}^*$ and optimize both mask generation and segmentation tasks simultaneously using $\mathcal{L}_3 = \mathcal{L}_2 + \lambda_{seg}\mathcal{L}_{seg}$, where $\mathcal{L}_{seg} = \mathcal{L}_{BCE}(\mathbf{L}_{seg}, \hat{\mathbf{P}}^*) + \mathcal{L}_{Dice}(\mathbf{L}_{seg}, \hat{\mathbf{P}}^*)$ and $\mathbf{L}_{seg}$ denotes the segmentation logits predicted by the segmentation network $\mathcal{S}$.

3 Evaluation

3.1 Experimental Setting

Datasets. We evaluate all methods on three public benchmarks: **ISIC 2017** [7], **ISIC 2018** [6,20], and **PH^2** [16], containing 2,750, 3,694, and 200 images, respectively. For ISIC 2017, annotations are reorganized into two lesion classes and an "other" class. For ISIC 2018, image-level labels are extracted from diagnostic metadata. Each dataset is randomly split into training and testing sets with a 7:3 ratio, and experiments are repeated 4 times with different splits. Data augmentation includes random rotation, affine transformation, and random erasing. All images are resized and normalized before being fed into the network.

Baselines and Evaluation Metrics. We compare FG3-Cluster with seven recent baseline approaches, including (1) two-stage methods, which train a separate segmentation model using generated pseudo-masks (HAMIL [31], ARML [8], Gland [9], and ESFAN [29]); and (2) single-stage methods, which derive segmentation results in an end-to-end framework without offline mask generation (ToCo [19], CoSA [23] and MoRe [24]). We also take the fully supervised DLV3+ [4] as the performance upper bound. We report the Dice similarity coefficient (DSC) and mean intersection over union (mIoU).

Table 1. Evaluation of pseudo-mask quality on the **training sets** of ISIC 2017, ISIC 2018, and PH^2 (best results in boldface; second-best underlined).

Methods	ISIC 2017		ISIC 2018		PH^2	
	mIoU(%) ↑	DSC(%) ↑	mIoU(%) ↑	DSC(%) ↑	mIoU(%) ↑	DSC(%) ↑
HAMIL [31]	47.16	59.13	49.50	62.12	52.40	65.43
ARML [8]	60.97	71.83	<u>63.25</u>	<u>73.80</u>	69.14	79.09
Gland [9]	59.49	70.16	58.40	70.27	71.80	82.11
ESFAN [29]	60.25	70.81	59.48	70.80	67.83	78.30
ToCo [19]	61.96	72.11	53.23	66.37	<u>78.58</u>	<u>86.18</u>
CoSA [23]	<u>63.80</u>	<u>76.46</u>	56.53	70.47	73.37	83.86
MoRe [24]	63.50	75.25	57.17	70.23	69.53	80.87
Ours	**68.56**	**79.06**	**72.26**	**82.27**	**79.92**	**88.20**

Table 2. Segmentation performance on the **test sets** of ISIC 2017, ISIC 2018, and PH^2 (best results in bold; second-best underlined).

Methods	ISIC 2017		ISIC 2018		PH^2	
	mIoU(%) ↑	DSC(%) ↑	mIoU(%) ↑	DSC(%) ↑	mIoU(%) ↑	DSC(%) ↑
DLV3+ [4]	82.62	89.66	80.54	87.63	89.88	94.66
HAMIL [31]	53.21	64.81	58.35	69.63	56.88	69.07
ARML [8]	65.74	75.92	<u>69.17</u>	<u>78.56</u>	<u>79.12</u>	<u>87.75</u>
Gland [9]	62.58	73.07	62.93	73.62	75.93	85.57
ESFAN [29]	<u>67.83</u>	<u>77.04</u>	68.93	78.17	77.93	87.37
ToCo [19]	62.33	73.67	55.91	66.56	78.58	87.36
CoSA [23]	62.89	75.15	59.37	72.09	70.45	81.95
MoRe [24]	62.99	74.13	59.17	71.54	66.24	77.83
Ours	**68.78**	**79.37**	**71.72**	**81.90**	**80.70**	**88.91**

Implementation Details. FG3-Cluster is implemented in PyTorch and trained on a single NVIDIA 4090 GPU. We adopt ResNet34 [11] as the feature extraction backbone with a Feature Pyramid Network (FPN) [15] segmentation head, and resize all inputs to 224×224. We optimize the network using AdamW (weight decay $1e^{-4}$) with a cosine annealing learning-rate scheduler. For ISIC 2017/ISIC 2018, the batch size, number of training epochs, and initial learning rate are set to 256, 200, and $1e^{-4}$, respectively. For PH^2, we use a batch size of 64, train for 200 epochs, and set the learning rate to $1e^{-5}$ considering the smaller data size. The loss weights (λ_{ent}, λ_{ortho}, λ_{seg}) in Sec. 2.4 are set to (0.01, 0.10, 1.00). We set λ_{pa} to 0.05 for ISIC 2017, and 0.03 for ISIC 2018 and PH^2. Finally, λ_{aux} is set to 0.50 during warm-up and reduced to 0.20 afterwards.

Table 3. Ablation study of key components on **ISIC 2018**. Values are in mIoU (%).

	baseline	S1	S2	S3	Ours
Pseudo-Mask	48.69	58.70	68.09	69.31	**72.26**
Segmentation	48.46	58.63	67.92	69.75	**71.72**

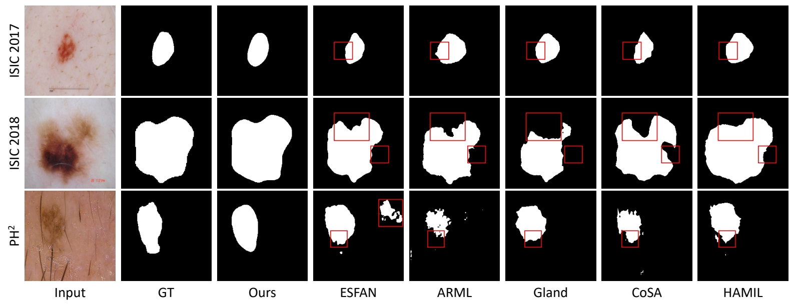

Fig. 2. Visual comparisons of FG3-Cluster and several leading baselines on testing cases. Red boxes indicate regions with challenging lesion boundaries.

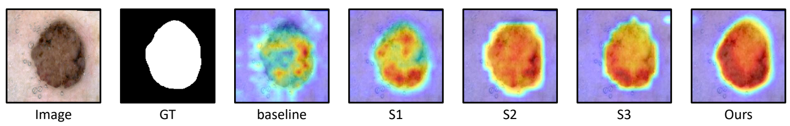

Fig. 3. CAM: DGC (S1). Clustering: DGC+$\mathbf{F}_{clu}$ (S2), FFM (S3), and TSR (Ours).

3.2 Comparison with SOTA Methods

As reported in Tables 1 and 2, our FG3-Cluster consistently outperforms single-stage and two-stage WSSS baselines across the three datasets. It achieves an mIoU of 68.56%, 72.26%, and 79.92% for pseudo-mask quality of ISIC 2017, ISIC 2018, and PH^2, outperforming the second-best methods by 7.46%, 14.25%, and 1.71%. The larger margin on ISIC 2018 suggests that FG3-Cluster is more robust to high intra-class variability. For the segmentation performance, FG3-Cluster surpasses the second-best mIoU by 1.40%, 3.69%, and 2.00%, demonstrating that improved pseudo-masks consistently benefits downstream segmentation. Fig. 2 presents visual comparisons with several leading baselines, where FG3-Cluster produces segmentations with improved connectivity and clearer boundaries.

3.3 Ablation Study

We ablate the key components of FG3-Cluster to assess their contributions. In Table 3, 'baseline' refers to generating pseudo-masks using only CAM responses. Adding DGC module (S1) enhances the CAM responses, which is further improved by using the cluster map $\mathbf{F}_{clu}$ (S2), showing that explicit semantic grouping provides more coherent region proposals. Introducing the FFM module (S3) and the TSR module (Ours) brings additional gains, suggesting that frequency-domain priors and geometry-guided refinement help disambiguate complex boundaries. The pseudo-mask responses are shown in Fig. 3, showing that the pseudo-mask quality gradually improved with modules introduced.

4 Conclusions

In this work, we propose a frequency-incorporated and geometry-aware graph clustering framework FG3-Cluster for WSSS skin lesion segmentation. We reformulate WSSS as a differentiable graph node-clustering problem. The DGC module enables the model to capture intrinsic lesion topology for improved cluster generation. By leveraging FFM, our method incorporates frequency-domain priors, rectifying incomplete clusters. Furthermore, we propose the TSR module which acts as a geometric constraint to improve boundary adherence. Experiments on three benchmark datasets validate the robustness and SOTA performance of our method, highlighting its potential for label-efficient clinical skin lesion analysis.

Acknowledgments. This work was supported by the Basic Research Program of the Key Scientific Research Project Plan of Higher Education Institutions in Henan (Project No. 26ZX010). The authors thank Prof. Yue Zhang for her insightful discussions.

Disclosure of Interests. The authors declare that they have no competing interests relevant to the content of this article.

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4981–4990 (2018)
2. American Cancer Society: Cancer facts & figures 2025. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2025/2025-cancer-facts-and-figures-acs.pdf, last accessed: 2026/02/26
3. Bianchi, F.M., Grattarola, D., Alippi, C.: Spectral clustering with graph neural networks for graph pooling. In: Proceedings of the 37th International Conference on Machine Learning. vol. 119, pp. 874–883 (2020)
4. Chen, L.C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., et al. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018)
5. Chen, Z., Sun, Q.: Weakly-supervised semantic segmentation with image-level labels: From traditional models to foundation models. ACM Comput. Surv. **57**(5) (2025)
6. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019)
7. Codella, N.C.F., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: IEEE 15th International Symposium on Biomedical Imaging. pp. 168–172 (2018)

8. Feng, S., et al.: Mining gold from the sand: Weakly supervised histological tissue segmentation with activation relocalization and mutual learning. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15008, pp. 414–423. Springer, Cham (2024)
9. Feng, S., et al.: Weakly supervised gland segmentation with class semantic consistency and purified labels filtration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2987–2995 (2025)
10. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems. vol. 30, pp. 1024–1034. Curran Associates, Inc. (2017)
11. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. Jiang, P.T., et al.: LayerCAM: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing **30**, 5875–5888 (2021)
13. Kwon, H., et al.: Phase concentration and shortcut suppression for weakly supervised semantic segmentation. In: Leonardis, A., et al. (eds.) ECCV 2024. LNCS, vol. 15090, pp. 293–312. Springer, Cham (2024)
14. Lee, J., Kim, E., Yoon, S.: Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4071–4080 (2021)
15. Lin, T.Y., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)
16. Mendonça, T., et al.: PH^2 - A dermoscopic image database for research and benchmarking. In: 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5437–5440 (2013)
17. Perez, E., et al.: FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32, pp. 3942–3951 (2018)
18. Ru, L., et al.: Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 16846–16855 (2022)
19. Ru, L., et al.: Token contrast for weakly-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3093–3102 (2023)
20. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 180161 (2018)
21. Wang, Y., et al.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12275–12284 (2020)
22. Wei, Y., et al.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1568–1576 (2017)
23. Yang, X., et al.: Weakly supervised co-training with swapping assignments for semantic segmentation. In: Leonardis, A., et al. (eds.) ECCV 2024. LNCS, vol. 15114, pp. 459–478. Springer, Cham (2024)
24. Yang, Z., et al.: MoRe: Class patch attention needs regularization for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9400–9408 (2025)

25. Yang, Z., et al.: FFR: Frequency feature rectification for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 30261–30270 (2025)
26. Yin, H., Gong, Y., Qiu, G.: Side window filtering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8750–8758 (2019)
27. Ying, Z., et al.: Hierarchical graph representation learning with differentiable pooling. Advances in Neural Information Processing Systems **31** (2018)
28. Zhang, B., et al.: Affinity attention graph neural network for weakly supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8082–8096 (2022)
29. Zhang, H., et al.: Edge-semantic synergy fusion and adaptive noise-aware for weakly supervised pathological tissue segmentation. In: Gee, J.C., et al. (eds.) MICCAI 2025. LNCS, vol. 15967, pp. 160–169. Springer, Cham (2025)
30. Zhang, J., Peng, B., Wu, X.: Dual graph inference network for weakly supervised semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology **35**(8), 8104–8118 (2025)
31. Zhong, L., et al.: HAMIL: High-resolution activation maps and interleaved learning for weakly supervised segmentation of histopathological images. IEEE Transactions on Medical Imaging **42**(10), 2912–2923 (2023)
32. Zhou, B., et al.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)