Supplementary Material for: Robust Few-Shot Medical Anomaly Detection via Feature Manifold Densification

I. Detailed experimental data for hyperparameter selection

This supplementary document provides comprehensive numerical results supporting the sensitivity analysis presented in Section IV-C of the main manuscript (referencing Fig. 3). The experiments were conducted on six public medical datasets under a 4-shot setting to determine the optimal hyperparameters for the proposed SGGP framework: the mixing probability (mixProb) and the ensemble size (maxN).

Impact of Mixing Probability (mixProb): Table S.I details the performance variations across different mixProb values. The first row (mixProb=0) serves as the baseline. In the table, red bold values denote performance improvements over the baseline, whereas blue values indicate a decline. The results indicate that setting mixProb to 0.25 yields the most robust generalization. Specifically, as shown in the last column of Table S.I, mixProb=0.25 achieves improvements in 8 out of 9 metrics, the highest count among all configurations. Although the configuration of mixProb=0.5 results in a marginally higher cumulative percentage gain (2.78% vs. 1.98%), it exhibits lower consistency (improving only 6 metrics). To ensure broader applicability and stability across diverse datasets, we selected mixProb=0.25 for all subsequent experiments.

Impact of Ensemble Size (maxN): Table S.II presents the ablation results for the ensemble size maxN, where maxN=0 represents the baseline. Consistent with Table S.I, red bold and blue texts represent improvements and declines, respectively. The data demonstrate that the model's performance peaks at maxN=5. At this setting, the method achieves the highest cumulative percentage improvement (+3.36%) across all metrics. Furthermore, both maxN=3 and maxN=5 demonstrate superior stability, improving 8 out of 9 metrics compared to the baseline. Consequently, to maximize both the magnitude of improvement and the consistency of the results, we fixed maxN=5 as the default configuration for the SGGP framework.

II. Detailed Experimental Data for Ablation Studies

Table S.III presents the comprehensive numerical results for the ablation experiments discussed in Section IV-D of the main manuscript (Fig. 4 and 5 of the main manuscript). Using the MVFA method as the baseline (first row), the subsequent rows detail the performance of comparative methods. In the table, red bold, blue, and orange text denote performance improvements, degradations, and consistency relative to the baseline, respectively. The final two columns summarize overall efficacy, with "Sum (%)" representing the cumulative percentage improvement and "Count" indicating the number of metrics surpassing the baseline; notably, the proposed max5ensemble strategy achieves the highest cumulative gain (+3.36%) and improves the most metrics (8 out of 9), thereby empirically validating the effectiveness of the Salience-Guided Group-PnMix framework.

III. Experiments with Identical Shot-Number but Varied Sample Settings

List of Specific Samples. As analyzed in Section IV-F of the main manuscript, sample selection bias exerts a significant impact on few-shot medical anomaly detection performance. To visualize this instability, Fig. 6 of the main manuscript contrasts the best and worst outcomes across different splits. To facilitate reproducibility, we provide the specific indices for the 10 randomly selected sample sets used in our experiments on our open-source code website. For detailed sample information, please refer to the .txt files located in the relevant folders for each dataset. All experiments were conducted under a 4-shot setting, containing four normal and four anomalous samples per set.

Detailed Comparative Data. Table S.IV details the performance of the baseline method (MVFA) across the ten random splits. Table S.V presents the corresponding results for our proposed method (SGGP). In Table S.V, red bold text indicates an improvement over the baseline result for that specific split (comparing to the corresponding entry in Table S.IV), whereas blue text indicates a performance decline. Columns with the suffix 'p' denote the pixel-level pAUC metric for segmentation tasks, whereas others represent the image-level AUC metric.

IV. Detailed Experimental Data of Group-Based PnMix

To empirically validate the impact of feature granularity on feature grouping, we conducted a comprehensive sweep of the pnmix_afterConv strategy. As detailed in Section III-A of the main manuscript, we intervene after the first convolutional layer of the CLIP encoder ($C = 1024$). Under a 4-shot setting, we partitioned these feature maps into 11 distinct granularities, ranging from 2^0 (1 group) to 2^{10} (1024 groups).

Table S.VI presents the exhaustive numerical results. The first row establishes the baseline (MVFA). Rows 2 through

TABLE S.I
EXPERIMENTS WITH DIFFERENT VALUES OF MIXPROB.

mixProb	OCT	HIS	Chest	Brain		Liver		RESC		Sum (%)	Count
	AUC	AUC	AUC	AUC	pAUC	AUC	pAUC	AUC	pAUC		
0	99.57	83.58	82.46	90.94	96.89	86.94	99.57	95.86	99.16	-	-
0.1	**99.68**	83.27	79.98	**92.50**	**97.15**	**87.95**	99.38	**96.42**	99.13	0.38	4
0.25	**99.60**	81.22	**83.61**	**93.00**	**97.32**	**87.27**	**99.62**	**96.10**	**99.24**	1.98	8
0.5	99.38	**83.65**	82.30	90.83	**97.48**	**88.04**	**99.74**	**96.96**	**99.18**	2.78	6
0.75	**99.69**	79.35	79.44	87.81	**97.15**	**88.71**	**99.79**	**96.29**	99.15	-7.71	5

TABLE S.II
EXPERIMENTS WITH DIFFERENT VALUES OF MAXN.

maxN	OCT	HIS	Chest	Brain		Liver		RESC		Sum (%)	Count
	AUC	AUC	AUC	AUC	pAUC	AUC	pAUC	AUC	pAUC		
0	99.57	83.58	82.46	90.94	96.89	86.94	99.57	95.86	99.16	-	-
1	**99.71**	82.65	81.45	**92.63**	**97.21**	**88.83**	**99.74**	95.52	**99.19**	1.82	6
3	**99.59**	**83.68**	**83.16**	**92.03**	**97.52**	**87.04**	**99.70**	**96.21**	99.08	3.02	8
5	**99.67**	**83.84**	**83.27**	**91.74**	**97.75**	**87.13**	**99.71**	**96.29**	99.03	3.36	8
7	**99.70**	**84.01**	79.81	**92.64**	**97.72**	**87.08**	99.55	95.84	99.15	0.40	5
9	**99.73**	83.06	78.70	**92.41**	**97.61**	**87.04**	**99.61**	**95.92**	99.07	-1.98	6

TABLE S.III
DETAILED EXPERIMENTAL RESULTS FOR ABLATION STUDY.

Method	OCT	HIS	Chest	Brain		Liver		RESC		Sum (%)	Count
	AUC	AUC	AUC	AUC	pAUC	AUC	pAUC	AUC	pAUC		
baseline	99.57	83.58	82.46	90.94	96.89	86.94	99.57	95.86	99.16	-	-
mixup	99.57	81.84	81.76	**92.59**	**97.48**	85.67	99.47	**97.61**	99.10	0.12	3
pnmix	**99.65**	81.68	80.61	**92.70**	**97.66**	**87.05**	99.57	95.33	98.96	-1.84	4
mixup_afterConv	**99.63**	82.91	81.17	90.18	**97.63**	86.04	99.51	**96.18**	**99.17**	-2.61	4
pnmix_afterConv	**99.60**	81.22	**83.61**	**93.00**	**97.32**	**87.27**	**99.62**	**96.10**	**99.24**	1.98	8
ensembleAll	99.36	83.51	80.42	**92.78**	**97.65**	86.37	**99.58**	95.81	99.14	-0.14	3
max5ensemble	**99.67**	**83.84**	**83.27**	**91.74**	**97.75**	**87.13**	**99.71**	**96.29**	99.03	3.36	8

TABLE S.IV
BASELINE EXPERIMENTAL DATA FOR TEN DIFFERENT SPECIFIC SAMPLES

Index	OCT	HIS	Chest	Brain		Liver		RESC	
	AUC	AUC	AUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
1	99.57	78.23	58.85	94.41	97.81	86.76	99.38	95.18	98.78
2	93.53	78.71	84.92	90.94	97.16	82.76	97.67	97.82	98.99
3	98.58	71.21	75.07	86.04	97.13	85.84	99.06	93.40	97.93
4	98.54	63.63	78.36	86.39	97.45	87.91	99.52	96.86	99.12
5	96.27	70.46	79.48	94.16	98.08	84.45	96.82	93.23	97.98
6	99.97	80.33	80.86	91.45	97.42	86.23	98.63	91.37	98.41
7	99.45	75.64	80.94	92.50	97.50	81.40	97.41	95.73	98.37
8	98.68	80.52	75.65	92.64	97.67	84.92	98.78	96.88	98.71
9	98.50	77.12	79.96	94.45	97.40	78.47	96.14	93.24	98.58
10	99.57	83.58	82.46	90.94	96.89	86.94	99.57	95.86	99.16

TABLE S.V
EXPERIMENTS ON DIFFERENT SAMPLES WITH THE SAME SHOT NUMBER (OURS)

Index	OCT	HIS	Chest	Brain		Liver		RESC	
	AUC	AUC	AUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
1	**99.68**	**78.90**	**62.77**	**95.09**	**98.03**	**87.63**	**99.64**	**95.32**	**98.93**
2	**94.68**	**80.79**	**85.04**	**93.87**	**97.70**	**84.75**	**98.25**	96.71	**99.05**
3	**99.04**	66.49	74.10	**89.69**	**97.81**	**86.60**	**99.24**	**94.26**	**98.07**
4	98.27	**64.89**	**80.61**	**88.74**	**97.80**	**89.47**	99.49	**97.59**	**99.13**
5	**96.96**	**70.69**	79.48	94.15	**98.23**	**86.70**	**97.86**	**94.87**	**98.26**
6	**99.98**	**83.93**	78.90	**91.46**	**97.62**	**86.65**	**99.02**	**92.31**	**98.71**
7	**99.79**	74.44	**81.28**	**93.59**	**97.81**	**85.39**	**97.48**	**96.67**	**98.52**
8	**99.00**	**80.67**	72.44	92.43	**98.09**	**86.86**	**99.31**	**97.11**	**98.91**
9	97.44	**78.34**	**80.42**	**95.47**	**97.57**	**79.90**	**96.52**	**94.68**	**98.65**
10	**99.67**	**83.84**	**83.27**	**91.74**	**97.75**	**87.13**	**99.71**	**96.29**	99.03

TABLE S.VI
DETAILED EXPERIMENTAL DATA FOR GROUP PNMIX_AFTERCONV WITH DIFFERENT GROUPING GRANULARITIES

Settings	OCT	HIS	Chest	Brain		Liver		RESC	
	AUC	AUC	AUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
baseline	99.57	83.58	82.46	90.94	96.89	86.94	99.57	95.86	99.16
1 group	**99.60**	81.22	**83.61**	**93.00**	**97.32**	**87.27**	**99.62**	**96.10**	**99.24**
2 groups	**99.59**	82.03	81.74	**91.69**	**97.64**	**87.90**	**99.60**	**95.90**	99.07
4 groups	**99.68**	83.13	81.20	**92.52**	**97.73**	**88.22**	**99.74**	**96.32**	98.98
8 groups	**99.65**	81.32	**83.99**	**92.36**	**97.30**	**88.24**	**99.74**	**96.55**	**99.19**
16 groups	**99.76**	82.73	80.53	**92.62**	**97.37**	**87.33**	**99.68**	95.72	98.97
32 groups	**99.64**	82.61	80.93	**92.40**	**97.33**	**87.37**	**99.58**	**95.88**	**99.20**
64 groups	**99.76**	82.23	80.50	**93.00**	**97.36**	**88.23**	**99.73**	**95.95**	99.05
128 groups	**99.87**	**84.42**	82.07	90.05	**97.47**	86.85	**99.68**	**96.68**	99.16
256 groups	**99.80**	80.93	**82.67**	**91.92**	**97.24**	86.46	99.56	**97.50**	99.11
512 groups	**99.96**	83.24	82.16	**93.10**	**97.06**	86.36	99.15	**96.56**	99.12
1024 groups	**99.63**	82.91	81.17	90.18	**97.63**	86.04	99.51	**96.18**	**99.17**
Best Result	**99.96**	**84.42**	**83.99**	**93.10**	**97.73**	**88.24**	**99.74**	**97.50**	**99.24**

12 detail the performance of each specific grouping configuration. Consistent with previous supplementary tables, red bold, blue, and orange texts denote improvement, decline, and consistency relative to the baseline, respectively. The final row, labeled "Best Result," aggregates the theoretical optimal score achievable for each metric if the best fixed grouping were manually selected for that specific dataset.

Key Observations:

1) Theoretical Potential vs. Practical Feasibility: The "Best Result" row represents the theoretical upper bound of feature-manifold augmentation. The data reveal that if computational resources were unlimited, performing an exhaustive search to manually select the optimal fixed grouping for each dataset would yield results slightly better than our proposed max5ensemble. However, training 11 distinct models for each new deployment scenario to find the optimal one is computationally prohibitive.

2) Sensitivity to Granularity and Data Dependency: The results highlight that optimal granularity is highly data-dependent, implying that no single fixed strategy is universally applicable.

- **General Trend:** Coarser granularities (e.g., 1 or 8

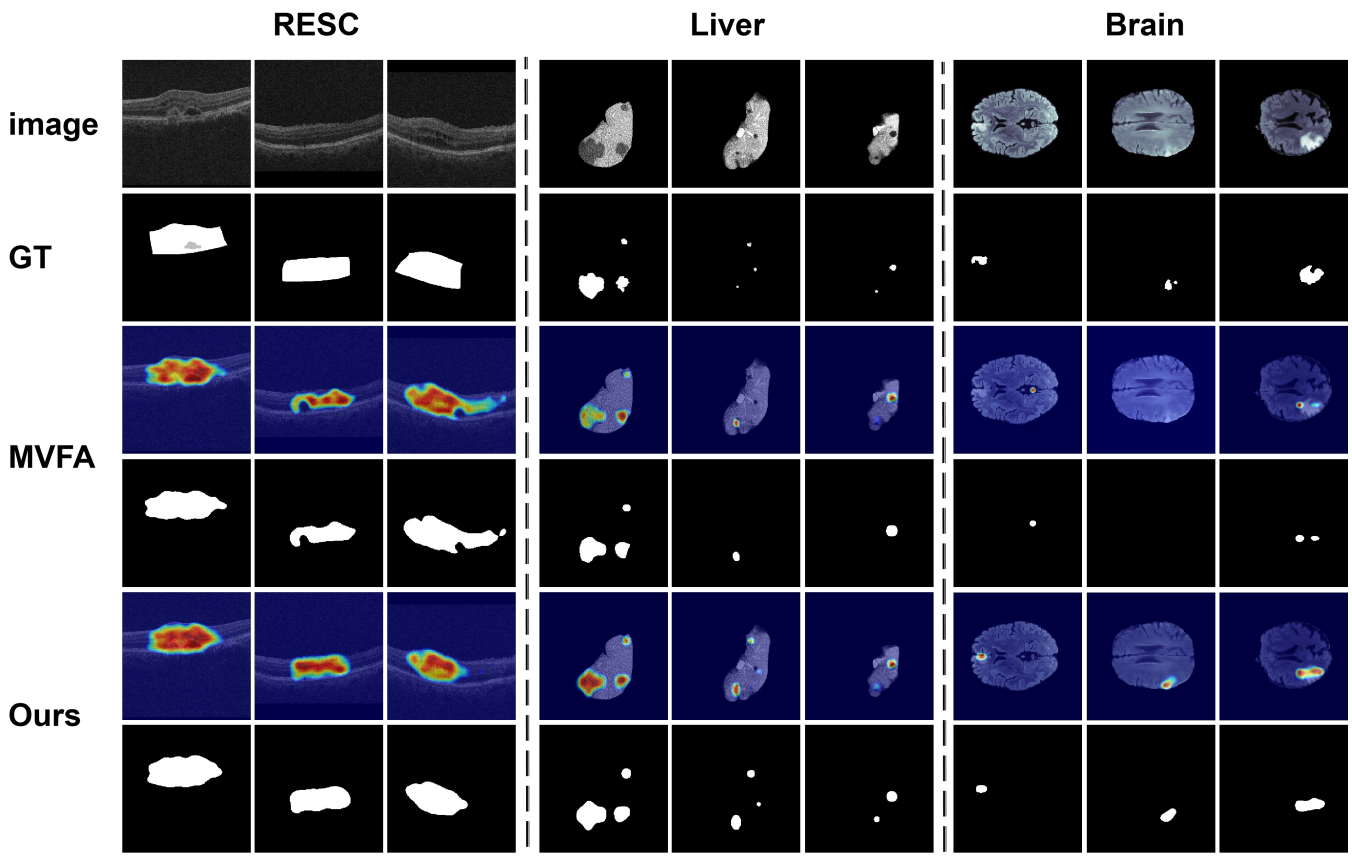

Fig. S.1. Qualitative visualization of anomaly segmentation results on the Brain, Liver, and RESC datasets. Visual inspection confirms that our method yields sharper boundaries and better alignment with the ground truth than the baseline.

groups) generally exhibit stability, whereas fine-grained settings (e.g., 512 or 1024 groups) yield higher volatility.

- **Case Study (HIS Dataset):** The necessity of adaptive selection is most distinctively illustrated by the **HIS** dataset. As shown in Table S.VI, HIS exhibits extreme sensitivity, achieving meaningful improvements only at a specific configuration (128 groups), whereas degrading under almost all others. This sharp contrast demonstrates that relying on a fixed hyperparameter for all medical domains is precarious.

3) Validation of SGGP: This trade-off justifies the design of our Salience-Guided Group-PnMix (SGGP). Although manually identifying the optimal granularity (e.g., the 128-group configuration for HIS) is challenging, our max5ensemble strategy outperforms the baseline on HIS (83.84% vs. 83.58%). This confirms that the Salience Energy criterion effectively identifies and ensembles the sparse discriminative feature groups (approximating the "Best Result") without requiring expensive exhaustive searches.

V. Visualization of Anomaly Segmentation Results

Supplementary Fig. S.1 provides a visualization of the anomaly localization results across the Brain, Liver, and RESC datasets. As illustrated, our proposed data augmentation framework effectively enhances the segmentation precision of the MVFA baseline. Visual inspection reveals that our method yields anomaly maps with sharper boundaries and better alignment with the Ground Truth masks than the original MVFA results. These visualizations empirically validate the effectiveness of our approach in accurately delineating pathological lesions and suppressing background noise.