Tackling Mask Imbalance: Prototypical Mixture-of-Experts with Hardness-Aware Mining for Medical Segmentation

Junjie Jiao*, Guoqian Liu*, Yanhao Chen*, Shuhao Hu*, Wei Liu†, Yin Zhang†, Lei Wang†‡, Qingqiang Wu*‡
* Xiamen University, Xiamen, China
{jiaojj1204, lgqian, cyhao, horacehsh}@stu.xmu.edu.cn, wuqq@xmu.edu.cn
† Academy of Military Medical Sciences, Academy of Military Sciences, China
liuweilcf@163.com, aring2010@163.com, wangleienjoy@163.com
‡ Corresponding authors: Lei Wang and Qingqiang Wu

***Abstract*—Medical image segmentation often suffers from severe mask and difficulty imbalance with ambiguous boundaries, where sparse foreground regions are overwhelmed by dominant background pixels. To address these challenges, we propose a Prototypical Mixture-of-Experts framework with a Collaborative Expert Module (CEM) that encourages complementary specialization while preserving global consistency, using a harmonic-mean objective to balance expert competence and inter-expert distillation to enforce consensus. Furthermore, we introduce a Prototypical Contrastive Strategy (PCS) to enhance pixel-level discriminability. PCS leverages consensus predictions to mine hard pixels and contrast their representations against global prototypes, sharpening decision boundaries. Extensive experiments on MoNuSeg, GlaS, COVID-19, and ISIC 2018 demonstrate competitive segmentation performance and improved boundary robustness over strong baselines.**

***Index Terms*—Medical Image Segmentation, Mixture of Experts, Prototypical Contrastive Learning, Boundary Refinement**

I. INTRODUCTION

Precise medical image segmentation is fundamental for clinical diagnosis and treatment planning. Recent advances are dominated by U-Net [1]/nnU-Net [2] and Transformer-based hybrids that balance local details with global context, while foundation models such as SAM2 [3] extend segmentation toward promptable settings. However, even with stronger backbones, boundary quality and robustness often degrade in real-world cases with severe mask imbalance.

We further observe that the performance bottleneck often stems not from the backbone itself, but from the widely adopted static single-stream segmentation head. This design fails to systematically alleviate the "Triple Imbalance" inherent in medical imagery: (1) *Size Imbalance*, where foregrounds can be extremely sparse (e.g., COVID-19 [4] masks contain only 1.76% foreground pixels across 2,729 samples), and are overwhelmed by dominant background pixels [5], [6]; (2) *Difficulty Imbalance*, where gradients are dominated by "easy" background pixels, leaving critical "hard" boundary pixels consistently under-optimized [6], [7]; and (3) *Representation Imbalance*, where a single head struggles to capture heterogeneous targets across datasets, from small lesions to larger structures (e.g., ISIC2018 [8] and MoNuSeg [9] have 22.74% and 24.27% foreground ratios, respectively). Consequently, model predictions frequently exhibit fractured contours and boundary drift. Loss reweighting alone improves supervision but does not address representation heterogeneity and consensus under multi-branch learning.

To address these limitations, we propose **PME-Net** (Prototypical Mixture-of-Experts Network), a model-agnostic framework that augments strong backbones (e.g., UDTransNet [10]) with a plug-and-play Prototypical MoE head. PME-Net replaces the fixed segmentation head with a dynamic divide-and-conquer structure, promoting complementary feature learning and improving boundary robustness under mask imbalance.

Specifically, we introduce two components. First, the **Collaborative Expert Module (CEM)** replaces the traditional head with lightweight parallel experts, encouraging complementary specialization via diversity constraints while enforcing prediction consensus through distillation. Second, to mitigate difficulty imbalance, we propose the **Prototypical Contrastive Strategy (PCS)**, which maintains global prototypes and mines the top-k hardest pixels to optimize a pixel-to-prototype contrastive loss, sharpening ambiguous decision boundaries.

Our main contributions are summarized as follows:

- We propose **PME-Net**, a model-agnostic segmentation framework that enhances strong backbones via a Prototypical Mixture-of-Experts head to better handle mask imbalance.
- We design **CEM** for complementary yet consistent expert learning, and **PCS** to mine and optimize hard pixels, improving performance at fuzzy boundaries.
- Experimental results on four public multi-modal datasets demonstrate that our method achieves competitive segmentation performance.

II. RELATED WORK

A. Backbone Networks for Medical Image Segmentation

Medical image segmentation has long been dominated by U-shaped CNN architectures. U-Net [1] and its variants (e.g.,

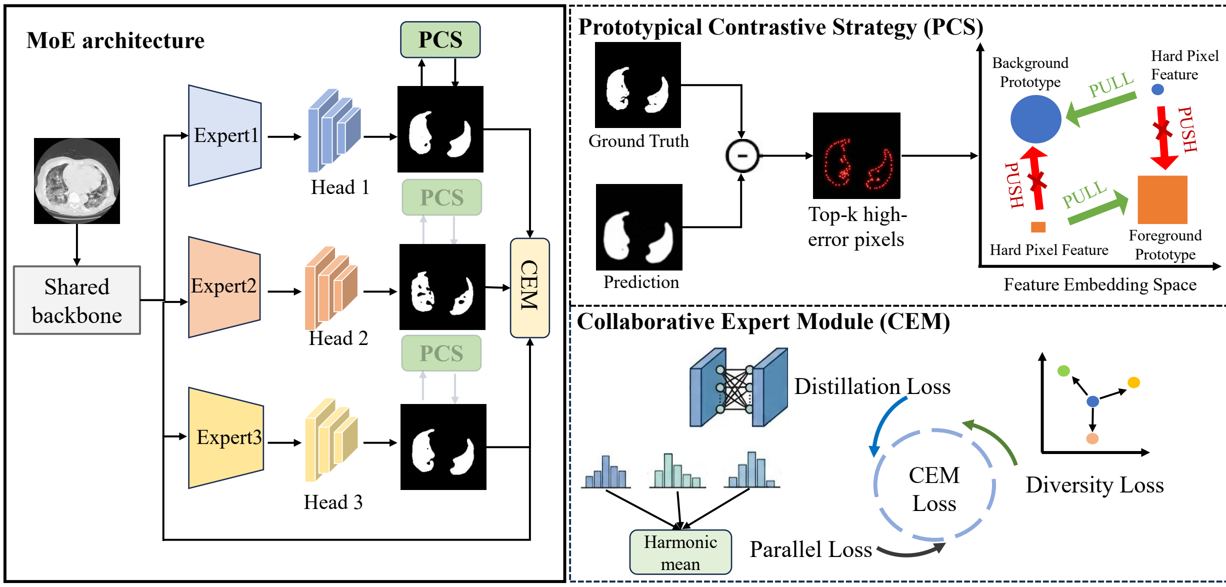

Fig. 1. **Overview of PME-Net.** (a) MoE architecture with a shared backbone and multiple experts/heads. (b) Prototypical Contrastive Strategy (PCS) that mines top-k hard pixels and contrasts them against class prototypes. (c) Collaborative Expert Module (CEM) with parallel loss, distillation loss, and diversity.

UNet++ [11], Attention U-Net [12], MultiResUNet [13], and Non-local U-Net [14]) have enhanced feature fusion and context modeling by incorporating skip connections, attention mechanisms, or global aggregation modules. Concurrently, nnU-Net [2] has established a robust baseline across tasks through its ”self-configuring” training paradigm, serving as a benchmark framework. With the advent of Transformers and hybrid architectures, methods like TransUNet [15] and MedT [16] integrate self-attention into the encoding/decoding process to capture long-range dependencies, while Swin-unet [17] uses hierarchical window-based attention to balance efficiency with global context. Furthermore, approaches such as MCTrans [18], UCTransNet [19], and UDTransNet [10] focus on multi-scale interactions and learnable skip connections to bridge the semantic gap between encoder and decoder. Recently, State Space Models (SSMs) have also been introduced into segmentation backbones; for instance, Mamba-unet [20] and VM-UNet [21] replace parts of attention with more efficient sequence modeling, exploring a new trade-off for long-range context modeling.

Beyond these backbones, foundation segmentation models are expanding the paradigm. For example, promptable segmentation frameworks (e.g., the SAM series [3]) offer pathways for cross-domain generalization, although they require targeted adaptation and training for specific medical scenarios. Overall, while backbone networks continue to evolve, performance on small targets and weak boundaries remains constrained by current prediction and optimization mechanisms.

B. Mask Imbalance in Medical Segmentation

Medical image segmentation suffers from foreground sparsity, where background pixels dominate optimization and rare boundary pixels are under-updated, degrading boundary-sensitive metrics (e.g., HD95). Remedies include Focal Loss [7], Tversky Loss [5], Unified Focal Loss [22], and Boundary Loss [6]. Beyond loss design, MoE-style architectures allocate capacity to rare foreground patterns, e.g., MoE-Polyp [23]. Long-tail MoE recognition further adopts independent and collaborative learning to avoid marginalizing experts while maintaining optimality [24]. Nonetheless, focusing on ambiguous boundaries while preserving diverse representations remains challenging.

III. METHOD

A. PME-Net Overall Framework

As illustrated in Fig. 1, PME-Net is a model-agnostic framework designed to address the challenge of mask imbalance in medical segmentation. Given an input image x, a shared backbone $\mathcal{B}(\cdot)$ extracts multi-scale feature representations $F = \mathcal{B}(x)$. Instead of a single static head, we couple a Collaborative Expert Module (CEM) with a prototype-guided hard-region refinement strategy (PCS) to jointly improve representation diversity and boundary robustness under severe mask imbalance. Specifically, CEM constructs E lightweight parallel experts $\{\mathcal{H}_e\}_{e=1}^{E}$ to generate diverse predictions, aggregating them via parameter-free averaging ($p_{\text{avg}} = \frac{1}{E}\sum_{e=1}^{E} p_e$) to achieve robust consensus. Complementary to CEM which coordinates learning at the prediction level, PCS operates on the feature level. It applies hardness-aware refinement to the last-layer decoder feature map G (prior to the individual expert heads), utilizing prototype-based constraints to sharpen decision boundaries in ambiguous regions.

B. Collaborative Expert Module

We employ multiple lightweight heads sharing decoder features. Unlike ensembles, experts are jointly trained via CEM to

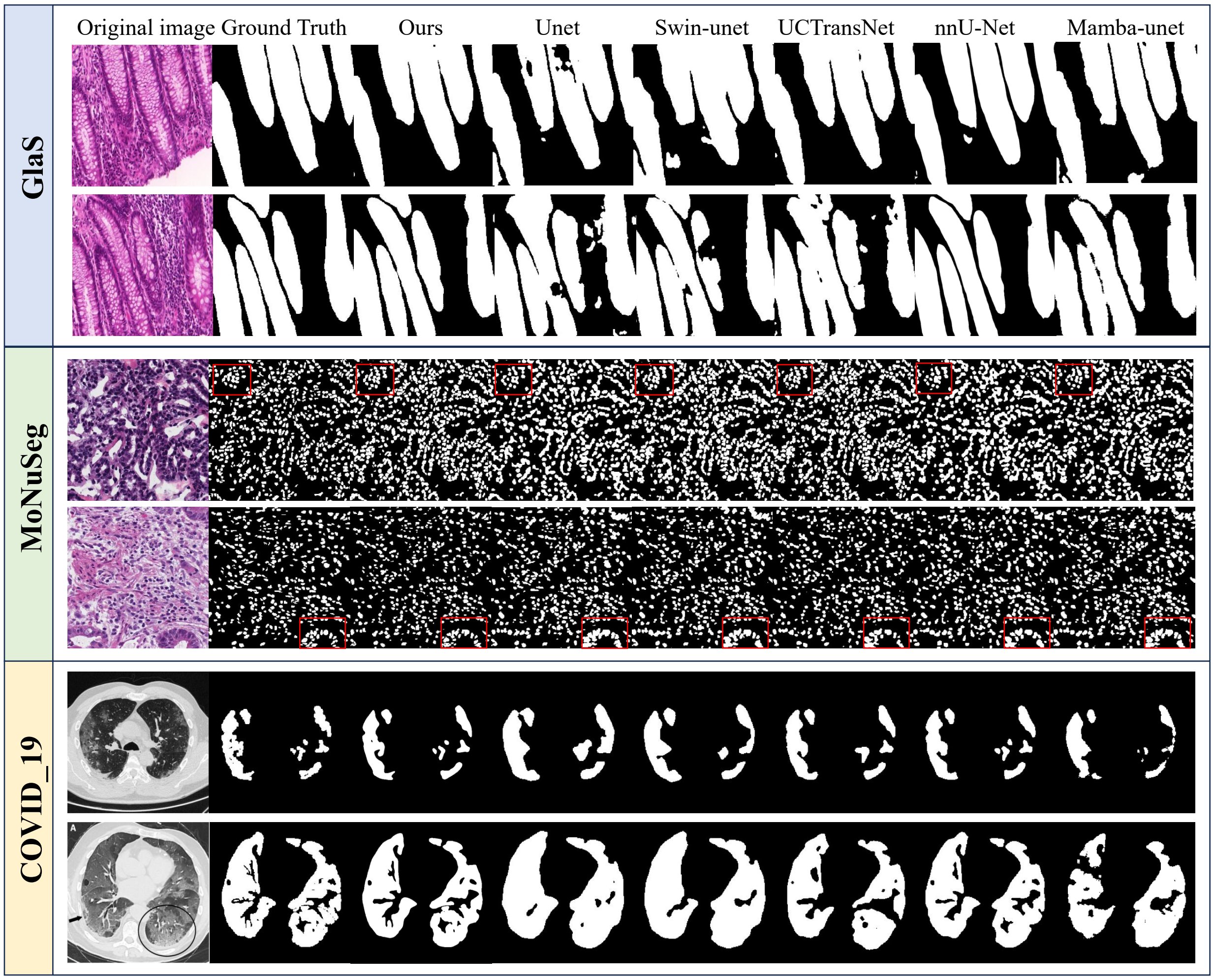

Fig. 2. **Qualitative comparisons on GlaS, MoNuSeg, and COVID-19.** From left to right: original image, ground truth, and predictions of our method and baselines (UNet, Swin-unet, UCTransNet, nnU-Net, Mamba-unet). Red boxes indicate challenging local regions (e.g., thin structures, noisy boundaries, or small target areas) for a clearer visual inspection.

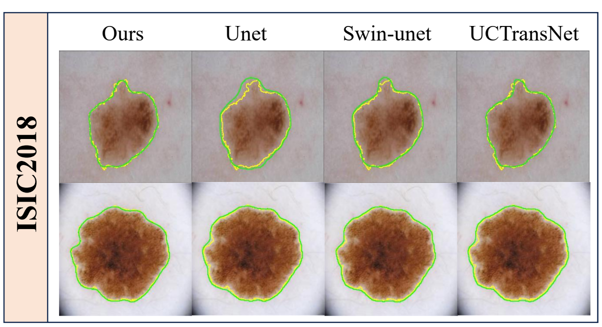

Fig. 3. **Qualitative comparisons on ISIC 2018.** Yellow contours indicate the ground-truth lesion boundaries, and green contours indicate the predicted boundaries.

encourage complementary representations while maintaining global consistency. We adopt a composite segmentation loss:

$$\mathcal{L}_{\text{seg}}(p, y) = \mathcal{L}_{\text{wBCE}}(p, y) + \mathcal{L}_{\text{wDice}}(p, y), \tag{1}$$

where p is the predicted probability map and y is the ground-truth mask. $\mathcal{L}_{\text{wBCE}}$ is a class-balanced weighted BCE that reweights foreground/background terms and normalizes by pixel counts to mitigate mask imbalance. $\mathcal{L}_{\text{wDice}}$ is a weighted squared Dice loss with L_2-style normalization for stable optimization under sparse foregrounds.

For E experts, we compute $\mathcal{L}_e = \mathcal{L}_{\text{seg}}(p_e, y)$ for each expert prediction p_e. To avoid domination by a subset of experts, we aggregate expert losses using a harmonic-mean style objective:

$$\mathcal{L}_{\text{par}} = \frac{E}{\sum_{e=1}^{E} \frac{1}{\mathcal{L}_e + \epsilon}}, \tag{2}$$

where ϵ is a small constant for numerical stability. This aggregation provides a smoother joint objective and stabilizes optimization under heterogeneous expert convergence, reducing the dominance of low-loss experts and emphasizing underperforming ones.

To enforce prediction-level consensus, we minimize pairwise mean squared error between expert probability maps:

$$\mathcal{L}_{\text{dist}} = \frac{1}{E(E-1)} \sum_{i \neq j} \|p_i - \text{sg}(p_j)\|_2^2, \tag{3}$$

where $\text{sg}(\cdot)$ is the stop-gradient operator.

To encourage complementary feature representations, we impose a diversity regularizer on expert features before the prediction heads. After global average pooling and ℓ_2 normalization to obtain z_e, we minimize the mean off-diagonal cosine similarity:

$$\mathcal{L}_{\text{div}} = \mathbb{E}\left[\frac{1}{E(E-1)} \sum_{i \neq j} z_i^\top z_j\right], \tag{4}$$

where z_i and z_j are normalized pooled features. Minimizing $\mathcal{L}_{\text{div}}$ reduces inter-expert similarity and promotes complementary cues, while $\mathcal{L}_{\text{dist}}$ and the consensus supervision below preserve global consistency. Finally, we add an additional supervision on the consensus map using the same base loss, denoted as $\mathcal{L}_{\text{avg}}$. The overall CEM objective is:

$$\mathcal{L}_{\text{CEM}} = \mathcal{L}_{\text{par}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{avg}}\mathcal{L}_{\text{avg}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}, \tag{5}$$

where λ_{dist}, λ_{avg}, and λ_{div} control the trade-off among prediction consensus, ensemble supervision, and representation diversity.

C. Prototypical Contrastive Strategy

While CEM coordinates experts at the model level, boundary-sensitive feature discrimination is further improved by PCS. Different from boundary-aware losses that reweight pixel-level supervision, PCS refines feature discriminability by contrasting mined hard pixels against class prototypes. Moreover, hard-pixel mining is driven by the stop-gradient consensus prediction, providing a more stable hardness signal than using a single expert head. PCS operates on the last-layer decoder feature map without upsampling. The ground-truth mask is downsampled to the same resolution for prototype update and hard-pixel selection. We maintain a prototype bank for background and foreground, updated via momentum:

$$\mathbf{P}_c \leftarrow m\mathbf{P}_c + (1-m) \cdot \text{Mean}\left(\{G_u \mid y'_u = c\}\right), \quad c \in \{0, 1\}, \tag{6}$$

where $\mathbf{P}_c$ is the prototype of class c, m is the momentum, G_u is the decoder feature at location u, and y'_u is the downsampled label. Prototype updates use detached features to stabilize training. To identify ambiguous regions, we leverage the consensus prediction to mine the top-k high-error locations per class. We first downsample the consensus prediction to the decoder resolution and define an error score:

$$e_u = \left|y'_u - p'_{\text{avg},u}\right|, \tag{7}$$

in which e_u is the error at location u and $p'_{\text{avg},u}$ is the downsampled consensus probability. For each sampled location, we extract its normalized embedding and compute similarities to the positive and negative prototypes with temperature:

$$s_u^+ = \frac{\hat{g}_u^\top \hat{\mathbf{P}}_c}{\tau}, \qquad s_u^- = \frac{\hat{g}_u^\top \hat{\mathbf{P}}_{1-c}}{\tau}, \tag{8}$$

where $\hat{g}_u$ is the normalized embedding at u, $\hat{\mathbf{P}}$ denotes normalized prototypes, and τ is the temperature. The PCS objective is minimized via a two-class InfoNCE loss:

$$\mathcal{L}_{\text{PCS}} = -\frac{1}{|\mathcal{S}|} \sum_{u \in \mathcal{S}} \log \frac{\exp(s_u^+)}{\exp(s_u^+) + \exp(s_u^-)}, \tag{9}$$

where $\mathcal{S}$ is the set of mined hard locations and $|\mathcal{S}|$ is its size. Hard-location mining uses the consensus prediction as a stable signal with stop-gradient, while $\mathcal{L}_{\text{PCS}}$ backpropagates through decoder features to refine boundary-sensitive representations.

D. Joint Objective

The final training objective integrates collaborative expert learning and prototypical contrastive refinement:

$$\mathcal{L} = \mathcal{L}_{\text{CEM}} + \lambda_{\text{pcs}}\mathcal{L}_{\text{PCS}}, \tag{10}$$

where λ_{pcs} balances the contribution of PCS. This joint design enables PME-Net to learn diverse yet consistent expert representations and explicitly refine ambiguous regions under severe mask imbalance.

IV. EXPERIMENT

A. Experimental Setup

1) *Datasets*: We evaluate the proposed method on four public datasets covering diverse modalities and target scales: MoNuSeg [9] (nuclei), GlaS [26] (gland), COVID-19 [4] (CT lung infection), and ISIC 2018 [8] (skin lesion). These datasets vary significantly in target morphology and boundary ambiguity, facilitating a robust assessment of the method's generalization across different clinical conditions.

2) *Evaluation Metrics*: Following previous work [10], we employ Dice Similarity Coefficient (Dice) and 95% Hausdorff Distance (HD95) to evaluate regional overlap and boundary precision, respectively. Higher Dice and lower HD95 values indicate superior segmentation performance.

3) *Implementation Details*: We implement PME-Net in PyTorch and adopt a two-stage training protocol. We use UDTransNet as the shared backbone/decoder to extract multi-scale features. In Stage 1, we train the base UDTransNet model with a single segmentation head to obtain stable representations. In Stage 2, we enable the Collaborative Expert Module (CEM) with E=3 lightweight expert heads and optimize the full CEM objective. For CEM, we set λ_{dist}=2.0, λ_{avg}=0.2, and λ_{div}=0.1. For PCS, we update the prototype bank with momentum m=0.9, use temperature τ=0.07, and mine the top-k hard locations with k=512 per class on the downsampled decoder feature map. Both stages are trained with the Adam optimizer for 1,000 epochs, using a cosine annealing learning rate schedule and an early-stopping criterion based on validation performance. We apply online data augmentations (random flips and rotations) to reduce overfitting. Unless otherwise specified, the input resolution is 224×224 with patch size 4. We report all results using 5-fold cross-validation with image-level splits.

TABLE I
SEGMENTATION PERFORMANCE (MEAN ± STD) ON **GLAS, MONUSEG, AND COVID-19**. HIGHER DICE IS BETTER; LOWER HD95 IS BETTER. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	GlaS		MoNuSeg		COVID-19	
	Avg Dice ↑	Avg HD95 ↓	Avg Dice ↑	Avg HD95 ↓	Avg Dice ↑	Avg HD95 ↓
U-Net [1]	85.45 ± 1.25	15.30 ± 1.67	76.45 ± 2.62	5.58 ± 2.04	60.29 ± 10.55	50.43 ± 19.30
R34-U-Net [1]	87.15 ± 1.08	10.27 ± 2.42	77.70 ± 2.31	4.71 ± 1.64	65.33 ± 8.19	38.19 ± 9.76
UNet++ [11]	87.56 ± 1.17	12.72 ± 1.76	77.01 ± 2.10	4.18 ± 1.29	78.58 ± 0.75	13.62 ± 1.47
Attention U-Net [12]	88.80 ± 1.07	9.10 ± 2.48	76.67 ± 1.06	5.38 ± 1.76	77.23 ± 2.45	16.21 ± 1.38
MultiResUNet [13]	88.21 ± 1.05	8.77 ± 1.89	78.22 ± 2.47	3.34 ± 0.97	77.64 ± 1.75	13.28 ± 2.10
Non-local U-Net [14]	88.83 ± 0.63	9.27 ± 1.06	77.75 ± 2.30	3.84 ± 1.43	77.31 ± 1.32	12.65 ± 1.52
MedT [16]	85.92 ± 2.93	16.02 ± 2.71	77.46 ± 2.38	6.74 ± 2.65	78.87 ± 1.78	15.28 ± 1.30
TransUNet [15]	88.40 ± 0.74	9.00 ± 1.17	78.53 ± 1.06	3.20 ± 1.25	78.46 ± 11.93	12.23 ± 1.03
Swin-unet [17]	89.58 ± 0.57	8.27 ± 1.64	77.69 ± 0.94	2.77 ± 0.84	77.93 ± 2.84	14.73 ± 1.88
MCTrans [18]	88.44 ± 0.74	9.96 ± 1.62	77.87 ± 3.04	3.74 ± 0.97	79.70 ± 0.38	10.95 ± 1.37
UCTransNet [19]	90.18 ± 0.71	7.52 ± 1.38	79.08 ± 0.67	2.72 ± 0.57	79.29 ± 0.70	10.82 ± 1.37
UDTransNet [10]	91.03 ± 0.56	<u>6.71 ± 0.99</u>	<u>79.47 ± 0.80</u>	<u>2.35 ± 0.25</u>	79.80 ± 0.73	<u>10.23 ± 0.62</u>
nnU-Net [2]	90.74 ± 0.89	9.96 ± 2.35	78.84 ± 0.46	4.35 ± 0.75	<u>79.86 ± 0.89</u>	10.96 ± 0.56
Mamba-unet [20]	<u>91.35 ± 0.29</u>	10.39 ± 4.89	71.56 ± 0.79	8.35 ± 0.72	77.59 ± 0.11	13.03 ± 3.09
Ours	**91.55 ± 0.40**	**6.48 ± 0.72**	**79.87 ± 0.69**	**2.29 ± 0.52**	**80.19 ± 0.21**	**10.04 ± 1.09**

TABLE II
ISIC 2018 PERFORMANCE (MEAN ± STD): HIGHER DICE AND LOWER HD95 ARE BETTER.

Method	Avg Dice ↑	Avg HD95 ↓
U-Net [1]	88.15 ± 0.27	14.57 ± 0.81
R34-U-Net [1]	88.47 ± 0.62	13.67 ± 0.76
UNet++ [11]	88.97 ± 0.38	12.31 ± 0.91
Attention U-Net [12]	89.05 ± 0.37	13.34 ± 0.58
MultiResUNet [13]	89.20 ± 0.23	12.70 ± 0.57
MedT [16]	87.67 ± 0.37	15.28 ± 1.30
TransUNet [15]	89.32 ± 0.19	12.02 ± 0.57
Swin-unet [17]	88.80 ± 0.52	12.75 ± 0.95
MCTrans [18]	89.50 ± 0.32	11.45 ± 0.36
UCTransNet [19]	89.18 ± 0.40	12.37 ± 0.61
UDTransNet [10]	89.91 ± 0.07	10.85 ± 0.26
nnU-Net [2]	89.29 ± 0.56	12.65 ± 0.86
Mamba-unet [20]	88.44 ± 0.15	11.35 ± 0.21
Ours	**90.35 ± 0.32**	**10.31 ± 0.76**

TABLE III
COMPARISON WITH STATE-OF-THE-ART LOSS FUNCTIONS ON THREE DATASETS IN TERMS OF **DICE COEFFICIENT (%)**.

Method	**MoNuSeg**	**GlaS**	**COVID-19**
CE Loss	77.45	89.12	76.35
Dice Loss [25]	78.30	89.85	77.10
CE + Dice	79.05	90.20	78.85
Focal Loss [7]	78.95	90.45	78.40
Tversky Loss [5]	79.15	90.68	78.92
Boundary Loss [6]	78.88	90.35	78.65
Ours (CEM + PCS)	**79.87**	**91.55**	**80.19**

B. Experimental Results

Fig. 2 visualizes segmentation outputs on GlaS, MoNuSeg, and COVID-19. The highlighted regions show that PME-Net better preserves thin structures and produces cleaner boundaries under noisy/ambiguous areas compared with representative baselines. Fig. 3 presents qualitative results on ISIC 2018, where PME-Net produces predictions that more closely follow lesion contours; this tighter boundary alignment is consistent with the quantitative improvements in Table II.

TABLE IV
ABLATION RESULTS OF PME-NET ON MONUSEG, GLAS, AND COVID-19. WE REPORT DICE (%) TO QUANTIFY EACH MODULE.

Base	**MoE**	**CEM**	**PCS**	**Average Dice Score (%)**		
				MoNuSeg	**GlaS**	**COVID-19**
✓				79.08	91.03	79.80
✓	✓			79.12	90.95	79.60
✓	✓	✓		79.67	91.35	80.03
✓	✓	✓	✓	**79.87**	**91.55**	**80.19**

All experiments follow a unified setting for fair comparison. Table I summarizes the quantitative comparison on GlaS, MoNuSeg, and COVID-19 in terms of Dice (%) and HD95. PME-Net achieves the best results on all three datasets, improving Dice by about $+0.2 \sim +0.4$ over the strongest prior baseline while simultaneously reducing HD95 by roughly $0.06 \sim 0.23$. These consistent gains on *histology* (GlaS, MoNuSeg) and *CT* (COVID-19) validate that our CEM+PCS design improves both accuracy and boundary quality under severe mask imbalance, especially around hard boundary pixels and small/fragmented targets, with limited extra computation since only lightweight heads are introduced here. We additionally validate PME-Net on ISIC 2018 (Table II), where it again delivers the strongest performance, leading both Dice and HD95 among all compared methods, suggesting consistent transferability from histology/CT to dermoscopic lesion segmentation. To disentangle architectural gains from loss engineering, we compare PME-Net (CEM+PCS) with commonly used loss designs in Table III. For fair comparison, all loss baselines are trained on the same backbone with identical settings. While stronger objectives such as CE+Dice, Focal, and Tversky improve over plain CE, PME-Net still achieves the highest Dice on all three datasets (79.87 on MoNuSeg, 91.55 on GlaS, and 80.19 on COVID-19). This suggests that the improvements are not merely due to adopting a different loss, but stem from the proposed expert collaboration and prototype-driven refinement mechanisms.

C. Ablation Studies

To investigate the effectiveness of each component, we conduct step-by-step ablation studies on MoNuSeg, GlaS, and COVID-19 datasets, as summarized in Table IV. Directly applying the MoE structure (Row 2) yields only marginal changes or even performance degradation compared to the baseline (e.g., a drop of 0.08% on GlaS and 0.20% on COVID-19). This phenomenon suggests that without explicit constraints, unguided experts may learn redundant or conflicting representations, increasing optimization difficulty. In contrast, integrating the proposed CEM strategy (Row 3) consistently improves performance across all tasks. This substantial improvement confirms that *the MoE architecture relies on CEM* to enforce expert complementarity and global consistency, thereby preventing mode collapse and unlocking the full potential of the multi-expert design. Finally, introducing PCS (Row 4) further elevates the Dice scores to 79.87%, 91.55%, and 80.19% respectively. This demonstrates that mining hard prototypes effectively refines pixel-level discriminability, particularly for ambiguous boundaries.

V. CONCLUSION

We propose PME-Net, a model-agnostic Prototypical Mixture-of-Experts framework for medical image segmentation under severe mask imbalance. PME-Net combines a Collaborative Expert Module to promote complementary experts with consistent predictions, and a Prototypical Contrastive Strategy (PCS) that mines hard pixels to refine ambiguous boundaries. Experiments on four public datasets show consistent improvements over strong CNN/Transformer baselines, and ablations confirm the effectiveness of each component. Future work will extend PME-Net to 3D volumetric tasks and more diverse clinical scenarios.

VI. ACKNOWLEDGMENT

This work is supported by the Solfeggio ear training intelligent robot and cloud platform research and development project for music education (No.2024CXY0102), the public technology service platform project of Xiamen City (No.3502Z20231043) and Fujian Provincial Science and Technology Major Project (No.2024HZ022003), the 3D visualization digital twin integrated control system (No.2023CXY0111)

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[2] F. Isensee, P. F. Jaeger, S. A. A. Kohl *et al.*, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.
[3] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, Ryali *et al.*, "SAM 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.
[4] J. Ma, Y. Wang, X. An, C. Ge, Yu *et al.*, "Toward data-efficient learning: A benchmark for COVID-19 ct lung and infection segmentation," *Med. Phys.*, vol. 48, no. 3, pp. 1197–1210, 2021.
[5] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE Int. Symp. Biomed. Imaging*, 2019, pp. 683–687.
[6] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *Proc. Int. Conf. Med. Imag. with Deep Learn.*, 2019, pp. 285–296.
[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[8] N. Codella, V. Rotemberg, P. Tschandl, Celebi *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.
[9] N. Kumar, R. Verma, S. Sharma *et al.*, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Trans. Med. Imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.
[10] H. Wang, P. Cao, J. Yang, and O. Zaiane, "Narrowing the semantic gaps in U-Net with learnable skip connections: The case of medical image segmentation," *Neural Netw.*, vol. 178, p. 106546, 2024.
[11] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.
[12] O. Oktay, J. Schlemper, L. L. Folgoc *et al.*, "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.
[13] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, 2020.
[14] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local U-Nets for biomedical image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 6315–6322.
[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
[16] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 36–46.
[17] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: U-Net-like pure transformer for medical image segmentation," in *Eur. Conf. Comput. Vis.*, 2022, pp. 205–218.
[18] Y. Ji, R. Zhang, H. Wang *et al.*, "Multi-compound transformer for accurate biomedical image segmentation," in *Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 326–336.
[19] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 2441–2449.
[20] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li, "Mamba-UNet: U-Net-like pure visual mamba for medical image segmentation," *arXiv preprint arXiv:2402.05079*, 2024.
[21] J. Ruan, J. Li, and S. Xiang, "VM-UNet: Vision mamba U-Net for medical image segmentation," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2024.
[22] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Comput. Med. Imaging Graph.*, vol. 95, p. 102026, 2022.
[23] Z. Wu, X. Xiong, Y. Chen, S. Li, and H. Chen, "MoE-Polyp: Shifting more attention to small polyp segmentation via mixture-of-experts," in *Proc. ACM Int. Conf. Multimedia Asia*, 2024, p. 1.
[24] Y. Chen, Z. Jian, N. Ke, S. Hu, J. Jiao, Q. Hong, and Q. Wu, "Enhancing mixture of experts with independent and collaborative learning for long-tail visual recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2025, pp. 828–836.
[25] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis.*, 2016, pp. 565–571.
[26] K. Sirinukunwattana, J. P. W. Pluim, H. Chen *et al.*, "Gland segmentation in colon histology images: The GlaS challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, 2017.