Retinopathy of Prematurity Staging via Prototype-Evidential Learning

Donghan Wu[1,5†], Wenyue Shen[2 †], Lu Yuan[3], Heng Li[4], Huaying Hao[1], Caiping Shi[3 ✉], Juan Ye[2 ✉], and Yitian Zhao[1,5✉]

[1] Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China yitian.zhao@nimte.ac.cn

[2] Zhejiang University, Eye Center of the Second Affiliated Hospital, School of Medicine; Zhejiang Provincial Key Laboratory of Ophthalmology; Zhejiang Provincial Clinical Research Center for Eye Diseases; Zhejiang Provincial Engineering Institute on Eye Diseases, Hangzhou, China yejuan@zju.edu.cn

[3] Department of Ophthalmology, Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Center for Child Health, Hangzhou, China hzshicp@zju.edu.cn

[4] Faculty of Biomedical Engineering, Shenzhen University of Advanced Technology, Shenzhen, China

[5] University of Chinese Academy of Sciences, Beijing, China

Abstract. Retinopathy of prematurity (ROP) staging via fundus imaging is critical for timely screening and treatment planning in premature infants. However, automated diagnosis is frequently impeded by low image quality and the high visual similarity between adjacent disease stages, both of which introduce significant diagnostic ambiguity. To address this, we propose ProtoEvi-ROP, an interpretable and uncertainty-aware framework designed to enhance the reliability of early-stage ROP staging. Our approach synergizes a prototype-based reasoning mechanism for feature-level interpretability with evidential uncertainty modeling to explicitly quantify decision confidence. Additionally, we introduce a region-focused regularization strategy to stabilize prototype geometry, preventing feature drift. Extensive experiments on two ROP fundus datasets demonstrate that ProtoEvi-ROP consistently outperforms state-of-the-art methods. Crucially, the estimated uncertainty serves as a robust reliability indicator, effectively filtering incorrect predictions. Selective prediction experiments reveal a clear accuracy-coverage trade-off, confirming that high uncertainty correlates with increased diagnostic risk, while decision space visualizations elucidate the synergy between interpretable prototype reasoning and uncertainty quantification.

Keywords: Retinopathy of Prematurity · Prototype · Evidential

1 Introduction

Retinopathy of prematurity (ROP) is a vision-threatening retinal vascular disorder affecting premature infants and remains a leading cause of preventable

† Donghan Wu and Wenyue Shen contributed equally to this work.

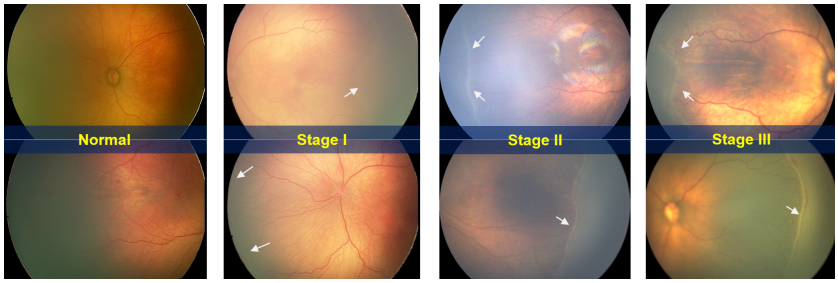

Fig. 1. Intrinsic ambiguity in early-stage ROP staging. Despite distinct clinical labels, these examples demonstrate adjacent stages lack clear visual boundaries.

childhood blindness worldwide [11,25,7]. Clinically, ROP is categorized into five stages based on the severity of morphological changes at the vascular-avascular junction. Accurate staging is vital for determining the interventions necessary to prevent blindness [21,1].

While the pathological features of advanced disease (Stages 4 and 5) are visually conspicuous, distinguishing early-stage ROP (Stages 1-3) from normal subjects, and accurately grading these stages, presents a significant challenge due to the subtlety of abnormalities [3]. From an imaging perspective, early-stage ROP exhibits highly similar visual appearances and ambiguous stage boundaries, posing substantial difficulties for both clinical interpretation and automated analysis [12]. As illustrated in Fig. 1, adjacent stages often share nearly indistinguishable global fundus appearances, while discriminative cues are confined to subtle, localized morphological changes that evolve gradually rather than discretely. This visual ambiguity creates two primary hurdles for algorithmic diagnosis. First, strong inter-stage similarity leads to significant class overlap in the feature space [4], rendering rigid decision boundaries unstable. Second, because pathological cues are often weak, spatially sparse, and incomplete, existing deep learning models typically output deterministic predictions in a "black-box" manner. Consequently, these models offer limited insight into decision reliability.

To enhance trust in clinical decision-making, prior work prioritized either interpretability through prototype-based representations[15,6] or reliability through uncertainty modeling[23,22]. However, prototype learning suffers a critical flaw in ROP: relying on feature similarity without assessing evidence quality leads to over-confident grading on samples with weak signs. This limitation is critical in early-stage ROP, where diagnostic cues are sparse and boundaries ambiguous. Evidential Deep Learning (EDL) offers a principled solution for reliability. Unlike classifiers forcing decisions based on similarity, EDL explicitly models evidence accumulation to construct a second-order probability distribution. This distinguishes between conflicting evidence (resembling multiple stages) and lack of evidence (weak signs). Yet, while EDL quantifies uncertainty, it lacks the visual interpretability required to explain morphological similarity to clinicians.

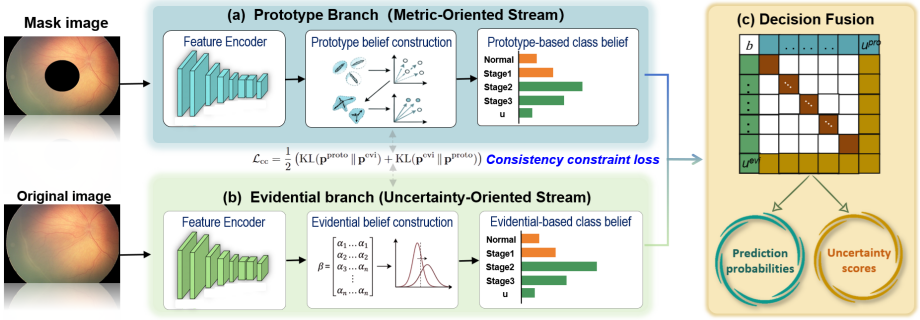

Fig. 2. Architecture of the proposed ProtoEvi-ROP framework, which synergizes prototype-based geometric reasoning with evidential uncertainty modeling for trustworthy ROP staging.

To address interpretability and reliability in early-stage ROP screening, we propose ProtoEvi-ROP. This framework introduce dual pathways to decompose diagnostic uncertainty: a prototype branch establishes geometric reference points for visual similarity (handling conflicting evidence), while an evidential branch acts as a reliability gate for signal sufficiency. By fusing these insights, our method transcends rigid "black-box" predictions to offer transparent, risk-aware staging (handling lack of evidence). Main contributions are as follows:

1) We present a decision-centric architecture that models geometric similarity (via prototypes) and evidential sufficiency (via uncertainty modeling), resolving the ambiguity inherent in subtle disease grades.

2) We develop a joint inference mechanism utilizing reduced Dempster-Shafer fusion to integrate prototype and evidential beliefs. To ensure reliability, a coordination consistency loss aligns visual similarity and predictive confidence.

3) We introduce a region-focused regularization strategy. By utilizing spatial masking as a weak inductive bias, this method stabilizes prototype learning and prevents geometric drift without relying on costly anatomical priors.

2 Methodology

As illustrated in Fig. 2, ProtoEvi-ROP integrates geometric interpretability and reliability through two complementary pathways. The prototype branch employs targeted masking to learn stage-specific geometric references, while the evidential branch quantifies the sufficiency of visual evidence. We enforce alignment via a consistency loss and integrate outputs using reduced Dempster–Shafer fusion [9], ensuring robust predictions and uncertainty estimation in high-conflict scenarios.

2.1 Prototype and Evidential Dual-stream Network

Let $x \in \mathbb{R}^{H\times W\times 3}$ denote the input fundus image and $x' \in \mathbb{R}^{H\times W\times 3}$ its masked counterpart, where H and W represent spatial dimensions. The evidential and

prototype branches process them in parallel using ResNet-50 [10] backbones. This yields embeddings $\mathbf{z}_e \in \mathbb{R}^d$ (for uncertainty modeling) and $\mathbf{z}_p \in \mathbb{R}^d$ (for stage geometry modeling), respectively, where d is the embedding dimension.

Prototype Branch: To capture stage geometry, this branch utilizes distance-based prototype learning. However, to prevent the model from overfitting to central anatomical features irrelevant to ROP, we apply a targeted spatial mask. This mechanism acts as a hard inductive bias, suppressing the central $1/3$ of the input to direct attention toward the peripheral vascular–avascular junction. The $1/3$ ratio provides a necessary safety margin, accommodating variations in camera centering while effectively removing distractors. This allows the prototype branch to specialize in lesion morphology, while the parallel evidential branch maintains access to the complete global information.

Given the feature embedding $\mathbf{z}_p$, we introduce a set of learnable stage-specific prototypes $\mathbf{p}^{proto} = \{\mathbf{p}_k \in \mathbb{R}^d \mid k = 1, \ldots, K\}$, where each prototype represents a reference position associated with one ROP stage. The model determines class membership by calculating the squared Euclidean distance from the sample embedding $\mathbf{z}_p$ to these anchors, converting the geometric proximity into a normalized probability distribution, and formulate as:

$$d_k = \|\mathbf{z}_p - \mathbf{p}_k\|_2^2, \quad p_k^{\text{proto}} = \frac{\exp(-d_k/\tau)}{\sum_{j=1}^{K} \exp(-d_j/\tau)}, \tag{1}$$

where τ is a temperature parameter controlling the sharpness of the distribution.

To capture geometry-induced ambiguity, we leverage the dispersion of the prototype similarity distribution. Samples lying in the transition zones between stages produce flattened (diffuse) distributions, which correspond to high entropy. We formalize this by decomposing the prediction into a stage-specific belief b^{proto} and a global uncertainty score u^{proto} via normalized entropy:

$$u^{\text{proto}} = \frac{-\sum_{k=1}^{K} p_k^{\text{proto}} \log p_k^{\text{proto}}}{\log K}, b_k^{\text{proto}} \quad = (1 - u^{\text{proto}})\, p_k^{\text{proto}}. \tag{2}$$

Here, the entropy is normalized by $\log K$ to strictly bound $u^{\text{proto}} \in [0, 1]$, ensuring that the total mass satisfies $\sum_{k=1}^{K} b_k^{\text{proto}} + u^{\text{proto}} = 1$. A higher entropy value signifies increased ambiguity in the feature space, indicating that the available geometric evidence is dispersed across multiple stages rather than supporting a single confident prediction.

Evidential Branch: To ensure trustworthy decision-making under conditions of limited or ambiguous visual evidence, especially in real-world ROP screening, this branch explicitly models predictive uncertainty using evidential theory [17,8]. Given the feature embedding $\mathbf{z}_e \in \mathbb{R}^d$, the network predicts non-negative evidence vector $\mathbf{e} = \{e_k\}_{k=1}^{K}$, where e_k represents the support for the k-th stage.

This evidence parametrizes a Dirichlet distribution $\mathrm{Dir}(\mathbf{p}|\boldsymbol{\alpha})$ over the class probabilities, where $\alpha_k = e_k + 1$ and $S = \sum \alpha_k$ represents the total Dirichlet strength:

$$\mathrm{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \quad B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(S)}, \; S = \sum_{k=1}^{K} \alpha_k. \tag{3}$$

Following Subjective Logic, we decompose these Dirichlet parameters into class-wise belief masses b_k^{evi} and a global epistemic uncertainty score u^{evi}:

$$b_k^{\mathrm{evi}} = \frac{\alpha_k - 1}{S}, \qquad u^{\mathrm{evi}} = \frac{K}{S}. \tag{4}$$

When visual evidence is scarce (low S), the model naturally outputs lower belief scores and higher uncertainty, enabling a risk-aware assessment of decision confidence. The class probability is definded as $p_k^{\mathrm{evi}} = \frac{\alpha_k}{S}$.

2.2 Reduced Dempster-Shafer Combination Rule

To obtain a unified diagnostic prediction, we integrate the outputs of the prototype and evidential branches during the inference phase. We employ a reduced Dempster-Shafer combination rule [9] rather than standard Dempster's rule of combination. Standard normalization often yields counter-intuitive results in high-conflict scenarios (e.g., when the prototype branch suggests one stage but the evidential branch suggests another) by completely suppressing the conflicting mass. In contrast, the reduced strategy explicitly preserves this conflict information, redistributing it into the uncertainty term or adjusting the joint belief mass. This ensures that the final prediction reflects both the agreement between branches and the reliability of their consensus, providing a robust measure of uncertainty when branch predictions diverge.

Given the belief mass distributions from the prototype branch, $\mathcal{M}^{(p)} = \{\{b_k^{\mathrm{proto}}\}_{k=1}^{K}, u^{\mathrm{proto}}\}$, and the evidential branch, $\mathcal{M}^{(e)} = \{\{b_k^{\mathrm{evi}}\}_{k=1}^{K}, u^{\mathrm{evi}}\}$, we compute the joint belief mass $\mathcal{M} = \{\{b_k\}_{k=1}^{K}, u\}$ using the reduced Dempster–Shafer combination rule. This rule fuses the independent sources of evidence while explicitly accounting for conflict:

$$b_k = \frac{b_k^{\mathrm{proto}}\, b_k^{\mathrm{evi}} + b_k^{\mathrm{proto}}\, u^{\mathrm{evi}} + b_k^{\mathrm{evi}}\, u^{\mathrm{proto}}}{1 - C}, u \quad = \frac{u^{\mathrm{proto}}\, u^{\mathrm{evi}}}{1 - C}, C = \sum_{i \neq j} b_i^{\mathrm{proto}}\, b_j^{\mathrm{evi}}. \tag{5}$$

Here, C measures the conflict between the two belief mass sets arising from mutually exclusive singleton hypotheses, and the normalization factor $1 - C$ ensures $\sum_{k=1}^{K} b_k + u = 1$. The fused belief assignment is converted into a stage prediction by conditioning on non-ignorance, yielding a probability-like output $p_k = \frac{b_k}{1-u}$, $k = 1, \ldots, K$, where p_k denotes the final predicted likelihood and u is retained as an explicit indicator of insufficient or ambiguous evidence.

2.3 Loss Function

To ensure that the geometric and evidential branches develop compatible representations, we introduce a coordination consistency loss $\mathcal{L}_{\text{cc}}$. This term employs a symmetric Kullback-Leibler (KL) divergence to penalize contradictory probability distributions between the prototype ($\mathbf{p}^{\text{proto}}$) and evidential ($\mathbf{p}^{\text{evi}}$) outputs:

$$\mathcal{L}_{\text{cc}} = \frac{1}{2}\left(\text{KL}(\mathbf{p}^{\text{proto}} \,\|\, \mathbf{p}^{\text{evi}}) + \text{KL}(\mathbf{p}^{\text{evi}} \,\|\, \mathbf{p}^{\text{proto}})\right), \tag{6}$$

The total training objective $\mathcal{L}$ combines this consistency constraint with branch-specific supervision. It aggregates the standard cross-entropy loss for the prototype branch and the evidential loss for the evidential branch:

$$\begin{aligned}\mathcal{L} = &-\sum_{k=1}^{K} y_k \log p_k^{\text{proto}} + \sum_{k=1}^{K} y_k \left[\psi\left(\sum_{j=1}^{K}\alpha_j\right) - \psi(\alpha_k)\right] \\ &+ \lambda_{\text{evi}}\,\text{KL}\big[\text{Dir}(\boldsymbol{\alpha}) \,\|\, \text{Dir}(\mathbf{1})\big] + \lambda_{\text{cc}}\mathcal{L}_{\text{cc}}.\end{aligned} \tag{7}$$

Here, y_k denotes the one-hot ground truth label, $\psi(\cdot)$ is the digamma function. The hyperparameters are fixed at $\lambda_{\text{evi}} = 0.1$ and $\lambda_{\text{cc}} = 0.1$ to balance evidence regularization and cross-branch alignment, respectively.

3 Experimental Results

Two datasets were utilized in this study: a public benchmark (SZ-ROP) [26] and a private in-house collection (PEFI-ROP). SZ-ROP contains 756 images (Normal: 236, Stage 1: 94, Stage 2: 165, Stage 3: 261). PEFI-ROP comprises 858 images (Normal: 405, Stage 1: 195, Stage 2: 133, Stage 3: 125), with diagnostic labels verified by consensus between two senior ophthalmologists. The data were split at patient-level into 80% for development (using 5-fold cross-validation) and 20% for independent testing. The training process utilized the AdamW optimizer with an initial learning rate of 1×10^{-4} and a batch size of 16, running for a total of 60 epochs. Five-fold cross-validation was applied to strictly evaluate model generalization.

3.1 ROP Staging Performance and Ablation Study

We evaluated our method against leading classification models, including ConvNeXt V2 [5], SwiftFormer [20], CLAHE-MobileNet [19], ViT-tiny [5], SABPI-Net [2], and DFOCNet [16]. As shown in Table 1, our approach outperforms the strongest baseline across all key metrics. On SZ-ROP and PEFI-ROP, our framework yields substantial gains over the next-best competitor: Accuracy (+2.90%/4.54%), Precision (+2.92%/5.77%), Recall (+2.88%/6.24%), and F1-score (+3.42%/5.98%). These results confirm the advantage of our dual-branch prototype-evidential strategy in distinguishing subtle disease stages.

Table 1. Quantitative comparison with state-of-the-art methods on different datasets

	Model	Accuracy	Precision	Recall	F1-score	AUC
SZ-ROP	[5]	80.43±2.69	79.03±4.19	73.90±3.33	73.75±4.59	94.31±0.97
	[18]	83.33±2.20	80.66±2.70	78.53±3.05	78.98±2.82	93.82±1.30
	[2]	83.55±0.90	75.83±6.73	75.51±2.56	74.82±4.80	94.96±0.65
	[24]	83.87±2.68	81.79±3.24	80.41±2.76	80.33±2.83	96.33±0.76
	[19]	84.52±2.64	81.64±2.60	81.73±3.20	81.39±2.76	96.71±0.51
	[16]	85.38±2.05	82.88±2.76	81.13±2.06	81.58±2.09	95.95±0.77
	Ours	**88.28±1.04**	**85.80±1.80**	**84.61±1.85**	**85.00±1.78**	**96.83±0.35**
PEFI-ROP	[24]	77.64±5.23	74.24±10.01	73.47±5.81	77.24±4.97	94.06±3.31
	[5]	78.36±1.06	74.06±10.88	74.32±1.98	67.11±2.69	93.91±1.05
	[19]	80.00±7.63	77.06±16.55	74.95±10.31	70.66±12.18	96.31±2.72
	[16]	80.00±6.30	77.63±8.38	77.16±7.18	73.77±8.10	94.75±2.65
	[18]	80.18±3.70	79.13±4.70	78.35±4.50	77.24±4.97	94.06±3.31
	[2]	82.91±1.49	80.26±1.50	78.04±2.14	76.76±2.72	90.32±2.03
	Ours	**87.45±1.67**	**86.03±2.27**	**84.59±2.44**	**83.22±2.33**	**98.49±0.67**

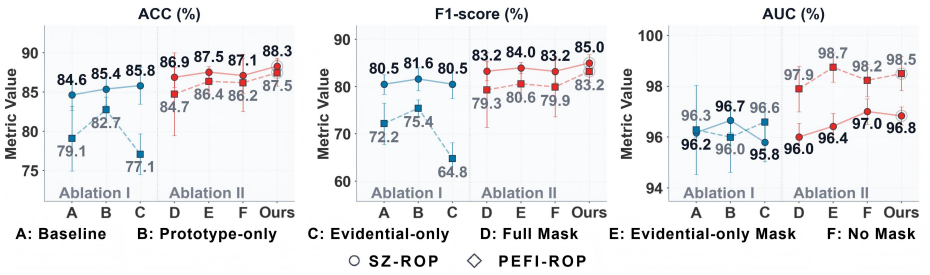

Fig. 3. Quantitative ablation study results on SZ-ROP and PEFI-ROP. The evaluation is divided into two parts: "Ablation I" compares different key components , while "Ablation II" assesses the impact of varying masking strategies.

Impact of dual-branch complementarity To validate the architectural design of ProtoEvi-ROP, we first assessed the individual contributions of the prototype and evidential branches (see 'Ablation I' in Fig. 3). Models relying exclusively on a single branch exhibit clear performance limitations. The Prototype-only variant improves Accuracy and F1-score by leveraging geometric discrimination, while the Evidential-only variant enhances AUC and Recall through explicit uncertainty modeling. However, both fail to match the full framework, as they lack the complementary cues provided by the other stream. ntegrating both branches yields consistent performance gains, boosting Accuracy by 2.5% on SZ-ROP and 4.8% on PEFI-ROP. This confirms that combining geometric similarity with evidential sufficiency creates a synergistic effect, effectively resolving ambiguous cases that single-stream models misclassify.

Analysis of masking strategy We further investigated the optimal configuration for our targeted masking mechanism (see 'Ablation I' in Fig. 3). As it can be seen, applying the mask exclusively to the prototype branch proves to be the

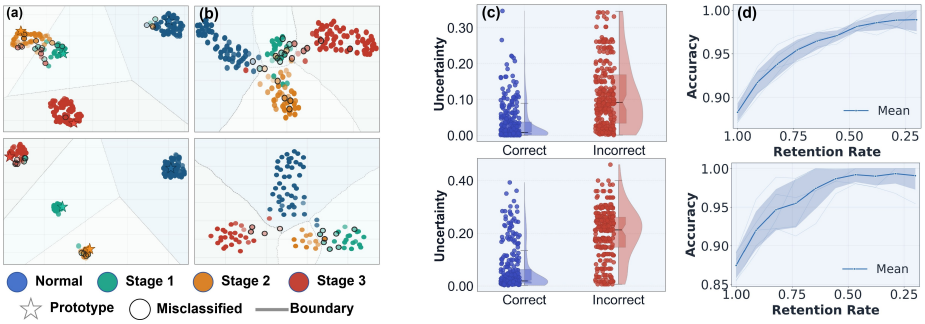

Fig. 4. Trustworthiness and interpretability analysis on SZ-ROP (top) and PEFI-ROP (bottom). (a–b) 2D visualizations of prototype and evidential representations. (c) Uncertainty box plots, (d) selective prediction curves.

superior strategy, outperforming alternative configurations by 0.8% (SZ-ROP) and 1.1% (PEFI-ROP) in Accuracy. This result validates our asymmetric design: the mask acts as a beneficial inductive bias for the prototype branch (forcing focus on the peripheral vascular ridge), while the evidential branch requires the full, unmasked image to accurately estimate the total evidence count without artificial occlusion.

3.2 Interpretability and trustworthy analysis

To assess interpretability and trustworthiness, we visualize the prototype and evidential feature manifolds using independent UMAP projections [13], as shown in Fig. 4 (a-b). The prototype space forms compact ordinal clusters, with errors concentrated near class boundaries, whereas the evidential space is more dispersed for ambiguous samples, reflecting insufficient evidence rather than forced geometric alignment. The consistency constraint preserves partial geometric structure in the evidential space, linking interpretable prototype reasoning with uncertainty-aware prediction.

Across both datasets, the estimated uncertainty exhibits a clear separation between correct and erroneous predictions. As shown in Fig. 4(c), correct predictions consistently present low uncertainty (SZ-ROP: $[0.0027, 0.0376]$; PEFI-ROP: $[0.0110, 0.0634]$), whereas erroneous predictions shift toward substantially higher ranges (SZ-ROP: $[0.0341, 0.1693]$; PEFI-ROP: $[0.1474, 0.2615]$). This difference is confirmed by the Mann–Whitney U test [14], yielding $p = 6.97 \times 10^{-28}$ for SZ-ROP and $p = 1.70 \times 10^{-23}$ for PEFI-ROP, indicating a strong association between uncertainty and prediction errors. Consistently, the accuracy–coverage curves in Fig. 4(d) show monotonically increasing accuracy as high-uncertainty samples are excluded, demonstrating the utility of uncertainty for risk-aware screening. Based on this separation, we define a distribution-driven threshold u as the midpoint between the upper quartile of correct predictions and the lower quartile of erroneous predictions, yielding $u = 0.036$ for SZ-ROP and $u = 0.105$

for PEFI-ROP. These thresholds enable automatic identification of high-risk cases for further review.

4 Conclusion

In this work, we introduced ProtoEvi-ROP, a dual-stream framework designed to bridge the gap between interpretability and reliability in ROP staging. By synergizing prototype-based geometric reasoning with evidential uncertainty modeling via a reduced Dempster–Shafer fusion strategy, our approach effectively resolves diagnostic ambiguity. Extensive experiments on two datasets demonstrate that ProtoEvi-ROP not only achieves state-of-the-art performance but also provides transparent decision support through visualized prototype spaces and actionable uncertainty quantification. These attributes establish our method as a robust tool for risk-aware clinical screening. Future efforts will focus on adapting this evidential-prototype paradigm to multi-modal data and other complex medical imaging pathologies.

Acknowledgments. This work is supported by the National Science Foundation Program of China (62422122, 62272444, U20A20386), the Key Research and Development Program of Zhejiang Province (2024C03204), the Ningbo 'KeChuang Yongjiang 2035' Key R&D Program (2024Z230).

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

References

1. Chan-Ling, T., Gole, G.A., Quinn, G.E., Adamson, S.J., Darlow, B.A.: Pathophysiology, screening and treatment of rop: a multi-disciplinary perspective. Progress in retinal and eye research **62**, 77–119 (2018)
2. Chen, S., Zhao, X., Fu, H., Tan, T., Huang, J., Xiong, X., et al.: Sabpi-net: A novel structure-aware network for accurate and domain-invariant retinopathy of prematurity diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 456–466. Springer (2025)
3. Chiang, M.F., Quinn, G.E., Fielder, A.R., Ostmo, S.R., Chan, R.P., Berrocal, A., et al.: International classification of retinopathy of prematurity. Ophthalmology **128**(10), e51–e68 (2021)
4. Ding, A., Chen, Q., Cao, Y., Liu, B.: Retinopathy of prematurity stage diagnosis using object segmentation and convolutional neural networks. In: 2020 International joint conference on neural networks (IJCNN). pp. 1–6. IEEE (2020)
5. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Gallée, L., Beer, M., Götz, M.: Interpretable medical image classification using prototype learning and privileged information. In: BVM Workshop. pp. 25–25. Springer (2024)
7. Gilbert, C., Rahi, J., Eckstein, M., O'Sullivan, J., Foster, A.: Retinopathy of prematurity in middle-income countries. The Lancet **350**(9070), 12–14 (1997)

8. Han, Z., Yang, F., Huang, J., Zhang, C., Yao, J.: Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20707–20717 (June 2022)
9. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification with dynamic evidential fusion. IEEE transactions on pattern analysis and machine intelligence **45**(2), 2551–2566 (2022)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
11. Hellström, A., Smith, L.E., Dammann, O.: Retinopathy of prematurity. The lancet **382**(9902), 1445–1457 (2013)
12. Jafarizadeh, A., Maleki, S.F., Pouya, P., Sobhi, N., Abdollahi, M., Pedrammehr, S., et al.: Current and future roles of artificial intelligence in retinopathy of prematurity. Artificial Intelligence Review **58**(6), 1–55 (2025)
13. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
14. McKnight, P.E., Najab, J.: Mann-whitney u test. The Corsini encyclopedia of psychology pp. 1–1 (2010)
15. Nauta, M., Hegeman, J.H., Geerdink, J., Schlötterer, J., Keulen, M.v., Seifert, C.: Interpreting and correcting medical image classification with pip-net. In: European Conference on Artificial Intelligence. pp. 198–215. Springer (2023)
16. Peng, Y., Zhu, W., Chen, Z., Wang, M., Geng, L., Yu, K., et al.: Automatic staging for retinopathy of prematurity with deep feature fusion and ordinal classification strategy. IEEE Transactions on Medical Imaging **40**(7), 1750–1762 (2021)
17. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
18. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 17425–17436 (2023)
19. Shoeibi, N., Ameri, N., Hoseinkhani, M.R., Ansari Astaneh, M.R., Motamed Shariati, M., Hosseini, S., et al.: Deep learning algorithms for timely diagnosis of retinopathy of prematurity requiring treatment. Eye (2025)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Stahl, A., Lepore, D., Fielder, A., Fleck, B., Reynolds, J.D., Chiang, M.F., et al.: Ranibizumab versus laser therapy for the treatment of very low birthweight infants with retinopathy of prematurity (rainbow): an open-label randomised controlled trial. The Lancet **394**(10208), 1551–1559 (2019)
22. Wang, M., Lin, T., Lin, A., Yu, K., Peng, Y., Wang, L., et al.: Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. Nature Communications **16**(1), 5528 (2025)
23. Wang, M., Lin, T., Wang, L., Lin, A., Zou, K., Xu, X., et al.: Uncertainty-inspired open set learning for retinal anomaly identification. Nature Communications **14**(1), 6757 (2023)
24. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., et al.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16133–16142 (2023)

25. Wu, D., Shen, W., Yuan, L., Li, H., Hao, H., Ye, J., Zhao, Y.: Fundus image quality assessment in retinopathy of prematurity via multi-label graph evidential network. Medical Image Analysis p. 103959 (2026)
26. Zhao, X., Chen, S., Zhang, S., Liu, Y., Hu, Y., Yuan, D., et al.: A fundus image dataset for intelligent retinopathy of prematurity system. Scientific Data **11**(1), 543 (2024)