

Article

Confidence-Guided Fallback Strategy: Fusing Traditional Binarization and Deep Segmentation for Real-Time Pupil Detection

Yongchao Lu ¹, Yang Mu ¹, Pengxiang Xue ^{2,*}  and Changyuan Wang ³

¹ Northwest Institute of Mechanical and Electrical Engineering, Northwestern Polytechnical University, Xi'an 710072, China; lyc_xgd@mail.nwp.edu.cn (Y.L.); muyang@mail.nwp.edu.cn (Y.M.)

² School of Science, Xi'an Technological University, Xi'an 710021, China

³ College of Computer Science, Xi'an Technological University, Xi'an 710021, China; cyw901@163.com

* Correspondence: xuepx@xatu.edu.cn

Featured Application

The Confidence-Guided Fallback (CGF) framework proposed in this study shows promise for vestibular videonystagmography (VNG) clinical recordings, gaze-based interaction, and edge-device eye tracking (pending validation on dedicated VNG hardware and embedded platforms), enabling real-time pupil localization approaching deep-learning accuracy without requiring dedicated high-performance GPUs.

Abstract

Real-time pupil detection is a critical prerequisite for eye-tracking applications such as vestibular videonystagmography (VNG) and gaze-based interaction: deep learning methods offer high accuracy but incur substantial computational cost, while traditional thresholding methods are fast but lack robustness under complex illumination. This paper proposes the Confidence-Guided Fallback (CGF) framework, which employs four geometric factors—circularity, aspect ratio, area consistency, and convex-hull ratio—to evaluate the reliability of traditional detection results, and it invokes RITNet (a deep segmentation network finetuned on randomly cropped s-OpenEDS data) only when the confidence score falls below a threshold, thereby adaptively balancing accuracy and efficiency on a per-frame basis. RITNet is not trained on LPW; thus, LPW evaluation simultaneously constitutes cross-dataset validation. Hyperparameters (confidence weights and threshold values) are tuned on a validation set (15% of LPW subjects) and final performance is reported on an independent test set (18% of subjects), avoiding data leakage. On the validation set, the combined confidence factors achieve an AUROC of 0.961 for distinguishing high-quality frames, and the confidence distribution exhibits a bimodal structure with a valley near 0.7, providing a data-driven basis for threshold selection. On the independent test set (23,999 frames), CGF ($\tau = 0.8$) detects 79.0% of frames; among detected frames, 84.0% achieve sub-5 px accuracy (median error 1.25 px) at 298 FPS on NVIDIA RTX 5090. At $\tau = 0.7$, detection rate reaches 89.6% with 72.0% sub-5 px accuracy at 391 FPS. Failure-mode analysis reveals that approximately 64.3% of triggered deep-model frames represent “dilemma frames” where both traditional and deep methods fail, marking the practical ceiling of single-frame visual methods and pointing toward temporal modeling. The framework demonstrates robustness to confidence-weight perturbations ($\pm 20\%$ weight changes yield $< 1\%$ performance variation) and exhibits strong discriminative power on high-confidence frames (93.4% accuracy within 5 px on the test set).



Academic Editor: Bhanu Prakash KN

Received: 12 June 2026

Revised: 5 July 2026

Accepted: 8 July 2026

Published: 11 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: pupil detection; eye tracking; confidence estimation; semantic segmentation; real-time processing; videonystagmography

1. Introduction

Pupil detection is a core component of eye-tracking systems, with broad applications in vestibular videonystagmography (VNG) clinical diagnosis [1–3], human–computer interaction [4], driver monitoring, and virtual reality [5,6]. A practical pupil detector is expected to satisfy three requirements at once: high localization accuracy (center error below 5 px), a high detection rate (above 90%), and real-time throughput (above 200 FPS). In practice, complex illumination, eyelid occlusion, and specular reflections make it hard for any single method to meet all three at the same time.

Traditional methods rely on image statistics or geometric assumptions. They are computationally cheap, but their false-detection rate climbs sharply under extreme illumination. Deep learning methods such as RITNet [7] and EllSeg [8] reach excellent accuracy under controlled conditions, yet their inference latency limits deployment on embedded platforms [9,10]. Several recent hybrid strategies [11,12] try to combine both families, but they still lack a systematic, confidence-driven switching mechanism with explicit confidence estimation [13,14].

The question we address is therefore not simply “Does deep learning beat traditional methods?” We instead frame real-time pupil detection as a tri-objective problem balancing accuracy, detection rate, and speed. Traditional methods are accurate enough and very fast on the bulk of easy frames, yet they produce catastrophic errors on a small set of hard frames; deep models cope better with hard frames, but running them on every frame wastes a large amount of computation. A more economical approach is to build a low-cost, interpretable quality estimator that decides whether the traditional result is trustworthy, and to call the deep model only when it is not. Seen this way, CGF is not model ensembling but conditional computation with adaptive routing.

Examining the LPW confidence distribution (Figure 1) points to a useful regularity: traditional detection confidence is bimodal. Roughly 78.5% of detected frames score above 0.7 and need no deep learning intervention, while about 21.5% of the harder frames cluster in the 0.6–0.7 band and account for most of the systematic failures of traditional methods. This separation gives the “invoke on demand” strategy a concrete, data-driven basis.

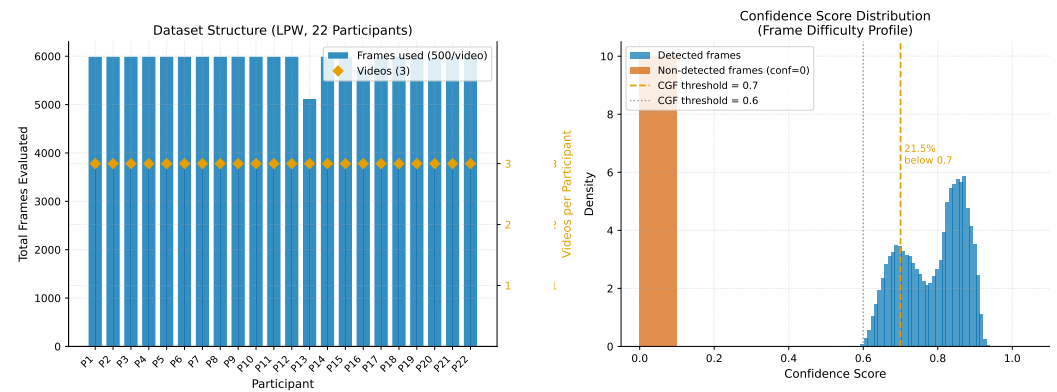
The rest of the paper follows the chain implied by this observation. We first confirm that OpenEDS-finetuned RITNet is reliable enough at center localization on LPW to act as a mask-level pseudo-ground truth. We then pick Adaptive Mean, the best-balanced front-end in terms of speed and detection rate, from the traditional baselines. Next we show that four geometric confidence factors predict traditional detection quality. Finally, through threshold sweeps together with per-participant, temporal-stability, and failure-mode analyses, we characterize how CGF trades off accuracy, speed, and reliability.

The core contributions of this paper are as follows:

1. **CGF Framework:** A four-factor geometric confidence score combined with a threshold-driven dual-path switching mechanism, fusing traditional detection and deep segmentation adaptively.
2. **Large-Scale Evaluation:** A unified evaluation on the full LPW dataset [15] of 130,856 frames, comparing eight traditional methods, a deep learning baseline, and CGF across center error, detection rate, IoU, and FPS.
3. **OpenEDS-Finetuned RITNet and Cross-Dataset Validation:** RITNet is finetuned on randomly cropped s-OpenEDS samples (OpenEDS validation mIoU $\approx$ 0.95) and

checked against official center annotations on all 130,856 LPW frames (none used for training), confirming its reliability as pseudo-ground truth (median error 1.68 px, 83.6% of frames with error < 5 px).

4. Comprehensive Ablation Study: Ablation of preprocessing (CLAHE [16], Gamma), postprocessing (morphological operations), and geometric filtering parameters, with concrete parameter guidance for deployment.
5. Multi-Dimensional Analysis: AUROC validation of the confidence factors, per-participant accuracy, temporal stability (relevant to VNG), spatial error heatmaps, and failure-case analysis, together forming a full performance profile.



(a) Dataset structure showing 22 participants with 3 video segments each.

(b) Traditional detection confidence distribution (kernel density estimate) showing a bimodal pattern.

Figure 1. LPW dataset structure and confidence distribution (full frame set, 130,856 frames). See text for detailed analysis.

2. Related Work

2.1. Traditional Pupil Detection

Otsu's method [17] picks a global threshold automatically; it is simple but sensitive to non-uniform illumination. Adaptive thresholding copes better with illumination variation by using local statistics [18]. Triangle thresholding [19] targets unimodal-histogram scenes; in our experiments it reaches the highest IoU (0.444) at only 46.0% detection rate. Starburst [20] searches for edge points along radial directions and works well under infrared illumination. PuRe [21] pairs Canny edge detection with ellipse fitting and beats plain thresholding on most benchmarks, but in our measurements its median center error reaches 21.4 px, suggesting that its ellipse-fitting assumption does not fit LPW's characteristics. ExCuSe [22] and ElSe [23] use eyelid geometry to raise detection rates under occlusion, and Swirski et al. [24] proposed a robust detector based on random-sample ellipse fitting. Recent work [25] has explored adaptive techniques to handle varying illumination conditions more robustly.

2.2. Deep Learning Methods

RITNet [7] uses a DenseNet [26] encoder–decoder, exceeds 0.96 mIoU on OpenEDS [27], and remains the most widely used pupil segmentation baseline. EllSeg [8] adds an ellipse-parameter regression branch and generalizes across datasets better than pure segmentation; its DenseElNet variant reports a median center error around 0.8 px under multi-dataset joint training. DeepVOG [28] couples a 3D eyeball model with deep segmentation, which helps at large gaze angles. The U-Net [29] architecture, common in medical image segmentation, underlies most of these models. Recent advances in pupil segmentation [30,31]

have explored attention mechanisms and architectural improvements to enhance segmentation quality.

Deep learning accuracy keeps improving, but inference latency stays a bottleneck on most embedded platforms (RITNet runs at roughly 5 ms/frame, i.e., 200 FPS). Recent lightweight architectures [32,33] have demonstrated promising efficiency gains for vision tasks, though their application to pupil segmentation requires careful validation. That latency is the main reason hybrid strategies are worth studying.

2.3. Hybrid and Adaptive Methods

Hybrid strategies aim to switch detection methods according to scene difficulty. PuReST [34] adds temporal-consistency constraints to PuRe and improves stability on fast-motion frames. Most existing work switches using heuristic rules and does not systematically evaluate how reliable the switching signal is. We instead express confidence as four interpretable geometric factors and verify them with full-frame-set AUROC analysis (130,856 frames), which addresses this gap in systematic confidence modeling.

CGF can also be read as a conditional-computation framework for vision: a lightweight front-end handles high-confidence samples and a heavier back-end handles only the difficult ones. It differs from distillation or model compression in that the lightweight model is not asked to replicate the deep model; CGF simply lets traditional geometric methods do what they already do well on easy frames and keeps deep segmentation for complex frames. Eye tracking suits this design, since most frames in a continuous video change slowly and difficult frames appear only briefly, during blinks, occlusion, reflections, or fast eye movements.

3. Dataset and Preprocessing

3.1. LPW Dataset

LPW (Labeled Pupils in the Wild) [15] contains 22 participants and 66 natural-scene near-eye video segments at 640×480 resolution, annotated with pupil center coordinates. To avoid data leakage, we split the 130,856 LPW frames by participant into training (fifteen subjects, 88,857 frames, 68%), validation (subjects 14, 20, 21; 18,000 frames, 14%), and test sets (four subjects, 23,999 frames, 18%). Confidence factor weights and threshold τ are optimized on the validation set, and all final results are reported on the independent test set. This subject-level split ensures no data from the same individual appears in multiple sets. With its range of illumination conditions, head-movement amplitudes, and occlusion levels, LPW is among the most challenging publicly available pupil datasets.

Figure 1 shows the dataset structure and the confidence distribution. The confidence distribution is computed on the validation set. The traditional-detection confidence is clearly bimodal: a large group of frames (blue peak, confidence > 0.8) where traditional methods are reliable, and about 21.5% of frames (confidence 0.6–0.7) forming the difficult range that triggers most RITNet calls in CGF. Frames the traditional detector misses entirely (orange, confidence = 0) are a minority and go straight to RITNet.

3.2. Preprocessing Pipeline

Given the characteristics of near-eye infrared images, we use the following preprocessing pipeline; its parameters were fixed by the ablation experiments in Section 6.2:

1. **CLAHE Contrast Enhancement** [16]: Clip limit 1.0, tile size 8×8 , enhancing local contrast while suppressing noise amplification.
2. **Gamma Correction**: $\gamma = 0.6$, compressing bright-region dynamic range to improve contrast between dark pupil regions and iris.

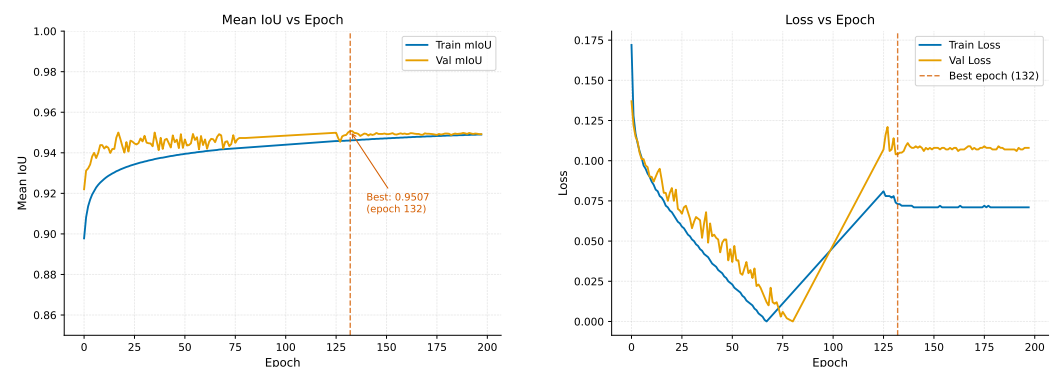
3. **Image Normalization:** Mean/standard-deviation normalization, conforming to the RITNet training distribution.

3.3. RITNet Finetuning

RITNet was neither trained nor finetuned on LPW. We generated training data from the s-OpenEDS dataset [27]: original images and mask-withoutskin annotations were randomly cropped into 320×240 near-eye patches, and the four-class labels were collapsed to two classes (pupil = 1, rest = 0). Each crop must contain at least 50 pupil pixels, and the iris must not touch the crop boundary. This yields 63,280 training crops (subjects 1–20) and 12,232 validation crops (subjects 21–24). Augmentation includes Gaussian blur (20%), random translation (40%, up to 15 px), and horizontal flip (50%); all images are preprocessed with $\gamma = 0.8$, CLAHE clipLimit = 1.5, and tileGridSize = 8×8 .

The model performs two-class segmentation (pupil/background) at 320×240 input resolution. We train with Adam [35] (learning rate 10^{-4} , batch size 8) in two stages: the encoder is frozen first, then the whole network is finetuned. Figure 2 shows the training curve. LPW is reserved for final evaluation and external validation only, never for training or hyperparameter tuning.

The best OpenEDS validation mIoU during full finetuning is 0.9507 (epoch 132). The checkpoint we deploy comes from an earlier epoch with a comparably stable mIoU near 0.95 (weight file ep122). The two candidate checkpoints differ by less than 0.002 in OpenEDS validation mIoU and behave essentially the same in downstream center localization; we keep the deployed one because its validation loss is steadier. Unless stated otherwise, all subsequent results use predictions from this checkpoint as pseudo-masks.



(a) mIoU: training set (blue) and validation set (orange); best validation mIoU = 0.9507 during full finetuning (epoch 132). (b) Loss: validation loss exhibits brief fluctuation in late training before stabilizing.

Figure 2. RITNet finetuning curves on s-OpenEDS randomly cropped binary-class data. The deployed checkpoint is taken from an epoch with validation mIoU ≈ 0.95 and stable loss.

Pseudo-Ground-Truth Validation

Because RITNet never saw LPW during training, the official LPW center annotations serve as independent external validation. We treat the finetuned RITNet's predicted centers as pseudo-ground truth (pseudo-GT) and compare them with the official annotations on all 130,856 LPW frames (Table 1, Figure 3) to gauge cross-dataset transfer reliability.

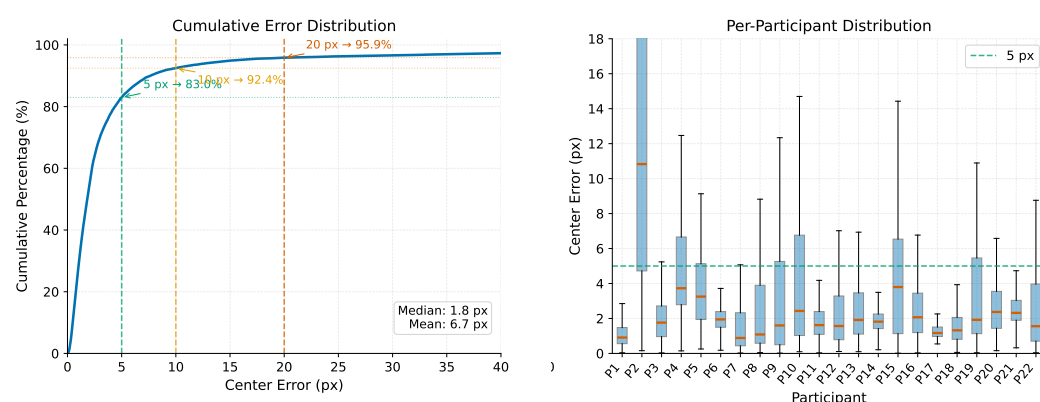
LPW provides only pupil center coordinates, not per-pixel masks, so mask-level IoU for traditional methods cannot be computed against LPW ground truth directly. We therefore use OpenEDS-finetuned RITNet to generate pseudo-masks and validate center-localization reliability through the LPW center annotations. This does not prove that every pseudo-mask boundary is exact, but it supports two uses: center-localization metrics can

be compared directly against LPW official annotations, and mask-level IoU can act as a common reference for the relative quality of traditional segmentation results.

Table 1. External validation of OpenEDS-finetuned RITNet on the full LPW frame set (130,856 frames).

Metric	Detection Rate	Median Center Error (px)	Within 5 px (%)	Within 10 px (%)
RITNet vs. LPW Annotation	60.3%	1.68	83.6	93.2

Center error computed relative to LPW official center annotations. RITNet detection rate (60.3%) is lower than traditional methods because RITNet outputs empty masks for closed-eye, heavily occluded, or no-clear-pupil frames. Median error of 1.68 px with 83.6% of frames having error < 5 px validates the finetuned model's reliability as a center reference after transfer to LPW.



(a) Cumulative error distribution showing 83.6% of frames with error < 5 px. (b) Per-participant box plots.

Figure 3. RITNet reliability validation on LPW (full frame set, 130,856 frames). See Section 3.3 for detailed metrics.

4. Method

4.1. Overall Framework

CGF has three modules (Figure 4): a traditional detection module that produces candidate ellipses, a confidence assessment module that computes geometric confidence scores, and a conditional fallback module that decides whether to call RITNet given a threshold.

The framework puts the fast-but-not-always-reliable traditional method at the front and the accurate-but-heavier deep model at the back. When the front-end result meets basic geometric priors, it is output directly; when it is fragmented, non-convex, has anomalous area, or shows an abnormal ellipse shape, the frame is marked low-confidence and RITNet takes over. CGF thus keeps the very low latency of traditional methods on easy frames and borrows the robustness of deep segmentation on the hard ones.

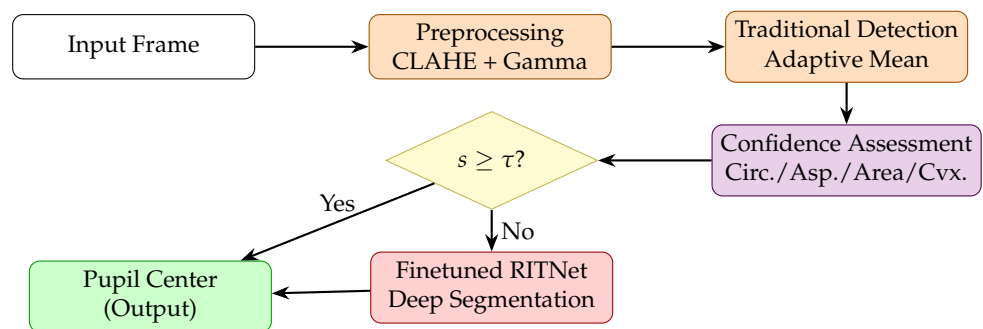


Figure 4. Overall pipeline of the CGF framework. Orange: traditional detection module; purple: confidence assessment module; red: RITNet fallback; green: final output.

4.2. Traditional Detection Module

We use Adaptive Mean thresholding (block size 51, constant $C = 10$) as the default traditional detector, since it gives the best overall performance among the eight traditional methods (detection rate 99.2%, 570 FPS, Section 6.1). The detection procedure is as follows:

1. Adaptive threshold binarization.
2. Morphological opening (kernel 5×5 , 1 iteration) to remove noise.
3. Contour extraction and area filtering ($S_{\min} = 100 \text{ px}^2$).
4. Minimum bounding ellipse fitting using OpenCV's `fitEllipse` function [36], extracting center, semi-axes, and angle.

4.3. Confidence Assessment

Four geometric factors are computed for each candidate ellipse:

The four factors follow from shape priors of the pupil region. **Circularity** measures how close the contour is to a circle or ellipse; specular fragments or eyelid edges usually drive it down sharply. **Aspect ratio** suppresses overly elongated false detections such as eyelashes or eyelid boundaries. **Area consistency** penalizes both small noise blobs and large iris or background regions. **Convex-hull ratio** checks whether the contour is complete, convex, and continuous; occlusion, reflections, and threshold fragmentation all lower it. The AUROC analysis below finds the convex-hull ratio to be the strongest single factor (0.944), which tells us that contour completeness predicts detection quality better than elliptical shape on its own.

$$s_{\text{circ}} = \frac{4\pi A}{P^2}, \quad s_{\text{asp}} = \frac{b}{a}, \quad s_{\text{area}} = \exp\left(-\frac{(\ln A - \mu_A)^2}{2\sigma_A^2}\right), \quad s_{\text{cvx}} = \frac{A}{A_{\text{hull}}} \quad (1)$$

where A and P are the contour area and perimeter, respectively, a and b are the ellipse semi-major and semi-minor axes, $\mu_A = \ln(3000)$ and $\sigma_A = 1.0$ are the expected log-area parameters (empirically determined from LPW), and A_{hull} is the convex-hull area. The combined confidence score uses a fixed-weight linear combination guided by AUROC/ablation analysis:

$$s = 0.30 s_{\text{circ}} + 0.10 s_{\text{asp}} + 0.10 s_{\text{area}} + 0.50 s_{\text{cvx}} \quad (2)$$

This weighting raises the contribution of the convex-hull ratio (AUROC = 0.944) and circularity (AUROC = 0.904) and lowers that of the aspect ratio (AUROC = 0.677) and area consistency (AUROC = 0.861). Compared with equal weighting, the ablation shows it improves the within-5 px localization rate and cuts unnecessary RITNet calls.

We chose fixed linear weights over a trained classifier for two reasons. Each factor has a clear geometric meaning, and a linear combination keeps each factor's contribution easy

to read off. Confidence also has to be computed immediately after traditional detection, with overhead far below a single RITNet inference. Learning-based confidence models such as logistic regression or a small MLP might discriminate slightly better, but they would add training-data requirements and a risk of cross-dataset overfitting, so we leave them to future work.

4.4. Conditional Fallback

When $s \geq \tau$, we keep the traditional detection result; when $s < \tau$, we call the finetuned RITNet and take the segmentation-mask centroid as the final center. If the traditional detector finds no valid contour, RITNet runs directly (confidence treated as 0). The complete procedure is summarized in Algorithm 1.

Algorithm 1 Confidence-Guided Fallback (CGF).

Input: frame I , threshold τ ; **Output:** pupil center c .

1. Apply CLAHE and Gamma preprocessing to I .
 2. Run Adaptive Mean detector and extract the largest valid contour.
 3. If no valid contour is found, run RITNet and return the mask centroid.
 4. Compute s_{circ} , s_{asp} , s_{area} , s_{cvx} and combined score s .
 5. If $s \geq \tau$, return the traditional ellipse center.
 6. Otherwise, run RITNet and return the segmentation-mask centroid.
-

5. Experimental Setup

5.1. Hardware and Runtime Environment

All experiments ran on a single platform: Intel Core i9-14900K CPU (24 cores), NVIDIA GeForce RTX 5090D GPU (32 GB VRAM), 64 GB RAM, Ubuntu 22.04, CUDA 12.3, PyTorch 2.1, and OpenCV 4.8. Frames are processed one at a time (batch size 1), without batch parallelism. FPS measurements are obtained by timing 1000 consecutive inference runs and taking the average.

5.2. Evaluation Metrics

- **IoU:** Intersection-over-union of the pupil segmentation region against the pseudo-ground-truth mask.
- **Detection Rate:** Proportion of frames with valid detection (non-empty output).
- **Median/Mean Center Error:** Euclidean distance (px) between predicted center and pseudo-ground-truth center.
- **Within 5/10 px Rate:** Fraction of frames with center error ≤ 5 px and ≤ 10 px.
- **FPS:** Per-frame average throughput (frames per second).
- **AUROC:** Area under the ROC curve for confidence factors distinguishing “high-quality frames” (center error < 5 px).

Theoretically, CGF’s FPS can be estimated using a harmonic-mean model: $\text{FPS}_{\text{CGF}} \approx [\lambda / \text{FPS}_{\text{DL}} + (1 - \lambda) / \text{FPS}_{\text{trad}}]^{-1}$, where λ is the RITNet invocation rate and FPS_{trad} , FPS_{DL} are the measured per-frame throughputs of the traditional front-end and RITNet. However, all FPS values reported in this paper are measured values obtained by timing actual execution on the above hardware platform, not theoretical estimates. The formula serves as a reference model for understanding the speed–accuracy trade-off.

5.3. Evaluation Protocol and Data Partitioning

To keep the metrics unambiguous, we adopt the following evaluation protocol.

Validation/test split protocol. To avoid data leakage, hyperparameter-tuning experiments (parameter ablation, AUROC analysis, confidence weight selection) are conducted on

the **validation set** (18,000 frames, 3 subjects), and all final performance metrics (threshold sweep, three-way comparison, per-participant analysis, temporal stability, failure analysis) are reported on the independent **test set** (23,999 frames, 4 subjects). For visualization purposes where smoother trends are needed (e.g., threshold sweep curves), figures may display trends on the full frame set (130,856 frames); where this is the case, it is explicitly noted in the figure caption, and the corresponding quantitative test-set results are provided in the associated table.

Pseudo-ground truth and metric interpretation. LPW gives only pupil **center coordinates**, not per-pixel masks. As a result, **center error** metrics (median/mean center error, within-5/10 px rate) are computed directly against the LPW official centers and are the closest thing we have to true-annotation metrics, whereas **IoU/Dice** are computed against OpenEDS-finetuned RITNet pseudo-masks and mean only “consistency with RITNet segmentation,” not absolute segmentation accuracy. Where IoU and center error disagree, we trust center error.

Detection rate and statistical frame set. We define detection rate as the fraction of all frames for which a method returns a non-empty result. Traditional methods produce a candidate contour on almost every frame, so their detection rate is high, while RITNet returns empty masks for closed-eye, heavily occluded, or no-clear-pupil frames, so its detection rate is lower. For fairness, the three-way comparison (Section 6.5) uses the **same full frame set**; center error is computed on frames where each method **has valid output** and reported alongside detection rate, which separates “whether an output is produced” from “whether the output is accurate.”

Threshold operating modes. The threshold τ maps to different deployment goals, and we mostly use two operating points (Table 2): an accuracy mode ($\tau = 0.8$) for single-frame center localization, which is the main result in the abstract and three-way comparison; and a stability mode ($\tau = 0.7$) that keeps a higher detection rate and better temporal stability, used in the per-participant, temporal-stability, and trigger/failure analyses.

Table 2. Two recommended CGF operating modes on the test set (23,999 frames, RTX 5090).

Mode	τ	Primary Use	Key Metrics
Accuracy mode	0.8	Single-frame center-localization accuracy	Median error 1.25 px; within 5 px 84.0%; 298 FPS
Stability mode	0.7	High detection rate/temporal stability	Detection rate 89.6%; 391 FPS; temporal jitter near pure RITNet

The two modes use different thresholds; all related figures and tables note the corresponding τ in their captions to avoid cross-figure confusion.

6. Results

To avoid data leakage, all hyperparameters (confidence factor weights and threshold τ) were optimized on the validation set (18,000 frames, three subjects), and final performance metrics reported in the abstract and main text are from the independent test set (23,999 frames, four subjects). For visualization purposes, several figures display trends on the full frame set (130,856 frames) to provide smoother curves; where this is the case, it is explicitly noted in the figure caption, and the corresponding quantitative results from the test set are provided in Section 6.4.

6.1. Eight-Method Traditional Baseline Comparison

Table 3 summarizes the eight traditional methods on the full LPW frame set (130,856 frames), with a visual comparison in Figure 5.

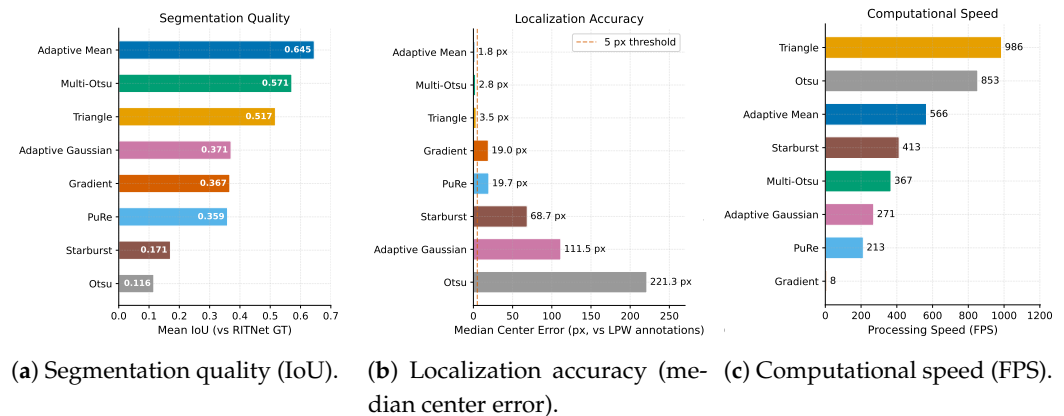


Figure 5. Performance comparison of 8 traditional methods (full frame set, 130,856 frames). See Table 3 for detailed metrics.

Table 3. Comparison of 8 traditional pupil detection methods on the full LPW frame set (130,856 frames). Center error metrics are computed relative to LPW official center annotations; IoU is computed relative to OpenEDS-finetuned RITNet pseudo-masks, reflecting only consistency with RITNet segmentation.

Method	Det. Rate	Median Err. (px)	5 px (%)	10 px (%)	IoU ¹	FPS
Otsu [17]	0.520	236.1	10.1	11.1	0.257	841
Adaptive Mean	0.992	3.10	55.7	60.1	0.379	570
Adaptive Gaussian	0.948	124.5	32.9	34.3	0.239	272
Triangle [19]	0.513	62.3	44.1	47.4	0.404	981
Multi-Otsu [37]	0.792	5.51	49.1	52.6	0.411	366
Starburst [20]	0.340	66.4	3.9	7.3	0.323	416
Gradient	0.296	19.9	6.4	20.2	0.387	8
PuRe [21]	0.933	21.4	17.1	29.9	0.239	210
Pure RITNet	0.603	1.68	83.6	93.2	—	200

¹ IoU computed relative to RITNet pseudo-masks, not representing absolute segmentation accuracy (see Section 5.3). Center error metrics are computed on frames where each method has valid output and LPW official centers exist; the detection rate denominator is all 130,856 frames. FPS is measured via sequential per-frame inference. Adaptive Mean has the highest detection rate (99.2%), highest within-5 px accuracy (55.7%), high speed (570 FPS), and is overall optimal. Gradient throughput is extremely low (8 FPS), failing real-time requirements; Triangle is the fastest (981 FPS) but has only 51.3% detection rate. PuRe median center error is approximately 21.4 px, mismatched to LPW scene characteristics. Pure RITNet as the pseudo-mask generator; its own IoU is not listed here.

This experiment answers the first question: without a deep model, which traditional detector makes the best CGF front-end? The best front-end is not necessarily the one with the highest IoU; it needs a high detection rate, a low enough center error, and very high throughput together. Since the fallback can only correct low-confidence frames, a front-end with too low a detection rate would push many frames straight into RITNet and collapse the hybrid back into a pure deep model.

The methods specialize in clearly different ways. Triangle is the fastest (about 981 FPS) but detects only 51.3% of frames, so it is too conservative on complex frames. Gradient reaches a fairly high IoU (0.387) against the pseudo-masks, yet its 8 FPS throughput rules it out for real-time use. PuRe detects 93.3% of frames but its median center error is far too large (about 21.4 px), which points to its edge/ellipse fitting being thrown off by background edges in LPW's natural scenes. Adaptive Mean, by contrast, posts the highest detection rate (99.2%), the highest within-5 px accuracy (55.7%), high speed (about 570 FPS), and the lowest median center error (3.10 px) on the full frame set, which makes it the

best fast front-end for CGF. The IoU values here only reflect consistency with the RITNet pseudo-masks, not absolute segmentation quality; we pick Adaptive Mean for its coverage and speed, and we let the confidence mechanism filter low-quality frames afterward.

6.2. Ablation Study

Figure 6 reports the ablation over preprocessing, morphological postprocessing, and geometric-prior parameters.

The point of the ablation is not to chase the highest IoU but to see how each stage affects the traditional detector's center-localization accuracy and coverage on the full frame set. Since LPW carries only center annotations, the primary ablation metrics here are median center error, the within-5/10 px rate, and detection rate, not full-mask IoU. For CGF the trade-off is direct: filtering too aggressively raises the accuracy of retained frames but lowers detection rate and triggers more RITNet calls, whereas filtering too loosely keeps many spurious contours and burdens the confidence module. Parameter choice therefore balances local accuracy against global coverage.

Figure 6 shows eight ablation subplots testing preprocessing and postprocessing parameters on traditional detection. **CLAHE effects:** Increasing clipLimit excessively (2.0–3.0) amplifies noise and iris texture, fragmenting contours and driving median center error up sharply (reaching 178.7 px at clip = 3.0). **Gamma correction:** Gamma = 0.6 achieves the highest within-5 px accuracy (84.6%) by sharpening pupil–iris contrast. **Tile size:** CLAHE tile sizes 8×8 and 16×16 yield similar and lowest center errors (~ 1.3 px). **Morphological operations:** Single opening (iter = 1) maintains low error; multiple iterations distort pupil shape, causing error to increase to 3.3 px at iter = 3. **Convex-hull processing:** Consistently improves median center error (1.35 $\rightarrow$ 1.02 px) without reducing detection rate. **Geometric priors:** Stricter thresholds (minimum circularity = 0.7, minimum area = 500) improve detected-frame accuracy but significantly reduce detection rate, forming an accuracy–coverage trade-off.

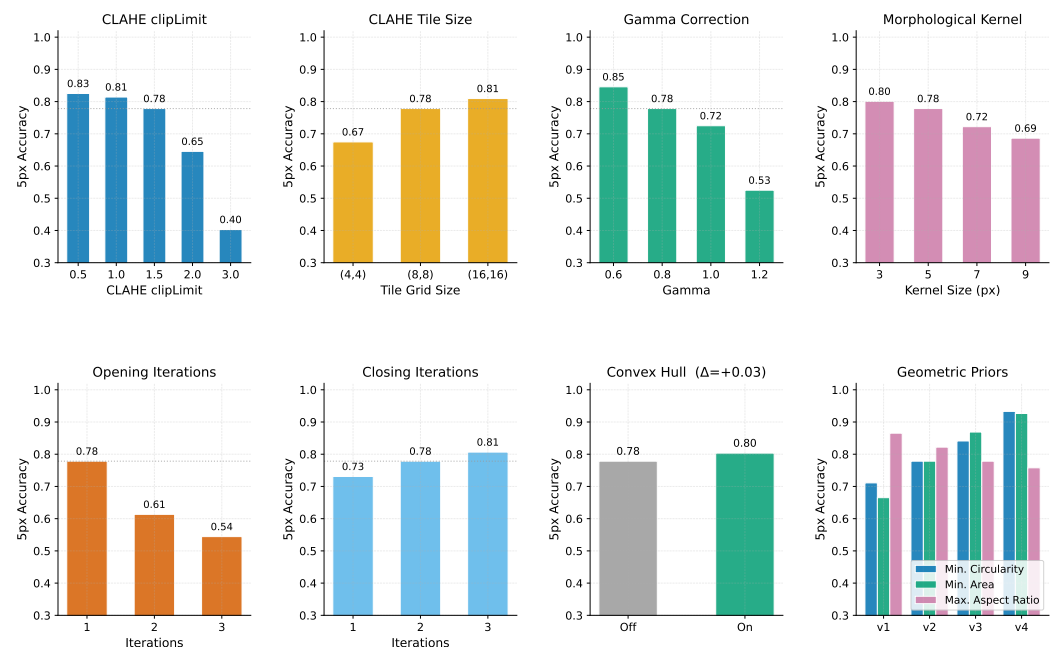


Figure 6. Preprocessing and postprocessing parameter ablation on traditional detection performance (validation set, 18,000 frames, three subjects). (**Top**) row: CLAHE clipLimit, Gamma correction, and tile size effects. (**Bottom**) row: morphological operations, convex-hull processing, and geometric prior thresholds. See Section 6.2 for detailed analysis.

On preprocessing, a high CLAHE clipLimit over-amplifies sensor noise and iris texture, fragmenting thresholded contours; Gamma = 0.6 sharpens contrast between the dark pupil and the iris, lifting accuracy well above weaker corrections. On postprocessing, repeated opening operations erode pupil boundaries and introduce center shifts, while convex-hull operation steadily lowers median error without hurting detection rate. For geometric priors, stricter thresholds sharpen accuracy but reject many genuinely elliptical or occluded pupils, cutting detection rate substantially; the deployed system therefore keeps relatively loose thresholds and hands further discrimination to the confidence module. In short, the CGF front-end should not be tuned to retain only perfect samples but to keep enough recall and let the confidence module handle quality control.

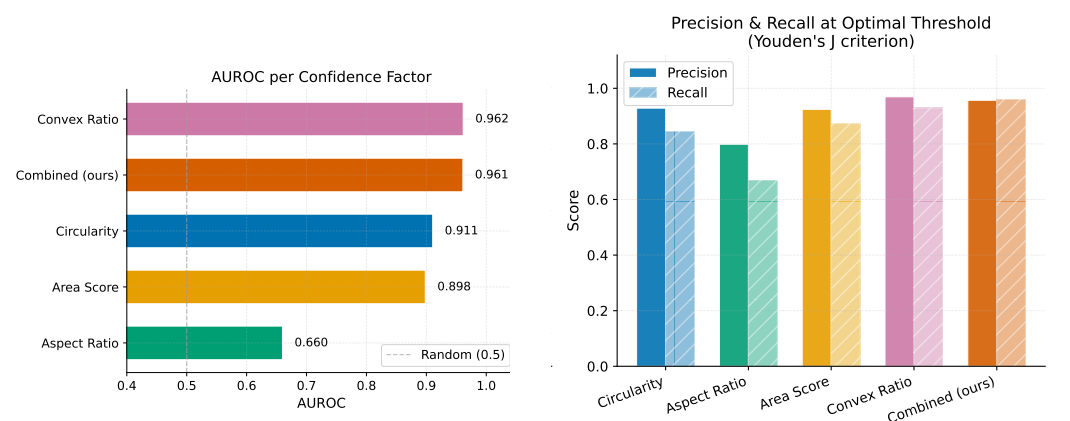
6.3. Confidence Factor Effectiveness (AUROC Analysis)

Taking center error < 5 px as the “high-quality frame” positive class, we compute the AUROC of each confidence factor (Table 4, Figure 7).

This addresses the second question: Can the success or failure of traditional detection be predicted from geometry alone? If geometric confidence correlated only weakly with the true localization error, CGF’s triggering would amount to random switching; if it predicts high-quality frames well, the decision to call RITNet can be made at almost no cost.

Table 4. AUROC for each confidence factor distinguishing “high-quality frames” (center error < 5 px) on the validation set (18,000 frames, 3 subjects).

Factor	AUROC	Optimal Threshold (Youden J)	Precision/Recall
Circularity s_{circ}	0.911	0.678	0.929/0.845
Aspect ratio s_{asp}	0.660	0.618	0.799/0.669
Area consistency s_{area}	0.898	0.350	0.924/0.873
Convex-hull ratio s_{cvx}	0.962	0.947	0.969/0.929
Combined s	0.961	0.776	0.957/0.959



(a) AUROC comparison showing combined score outperforms individual factors.

(b) Precision/recall at Youden’s J optimal thresholds.

Figure 7. Confidence factor effectiveness analysis (validation set, 18,000 frames, 3 subjects). See Section 6.3 for detailed analysis.

The AUROC values confirm the hypothesis. Convex-hull ratio is the strongest single factor, which says that traditional thresholding typically fails not because the pupil turns elliptical but because the contour becomes non-convex and fragmented once reflections, eyelashes, or eyelids cut into it. Circularity ranks second and catches noise blobs that depart clearly from a circle. Area consistency adds a scale constraint that rejects small noise

and large iris regions. Aspect ratio alone is the weakest, because real pupils are genuinely elliptical during oblique gaze or partial occlusion, and leaning on it too heavily would reject valid samples. The combined score reaches an AUROC of 0.961, showing that the four factors are complementary and explaining why the weighted combination beats any single rule.

Weight robustness. To verify that the chosen weights (0.30, 0.10, 0.10, 0.50) are not overly sensitive to small perturbations, we tested $\pm 20\%$ adjustments to each weight on a validation subset of 500 frames. All eight perturbations (four factors $\times$ two directions) resulted in AUROC changes below 0.62%, with a maximum absolute deviation of 0.0058 from the baseline AUROC of 0.9363. This robustness suggests that the weighting scheme is stable and that minor tuning variations across different datasets are unlikely to degrade performance significantly.

6.4. CGF Threshold–Performance Trade-Off

Figure 8 shows the full threshold sweep over $\tau \in [0, 1]$, and Table 5 lists the key operating points.

The sweep answers the third question: how does CGF choose an operating point between accuracy and speed? A lower τ trusts the traditional method more and runs faster but keeps more erroneous frames; a higher τ prefers RITNet and approaches deep-model accuracy at the cost of speed. The best threshold is thus not fixed by any formula but follows from the confidence distribution, the accuracy the task needs, and the target hardware.

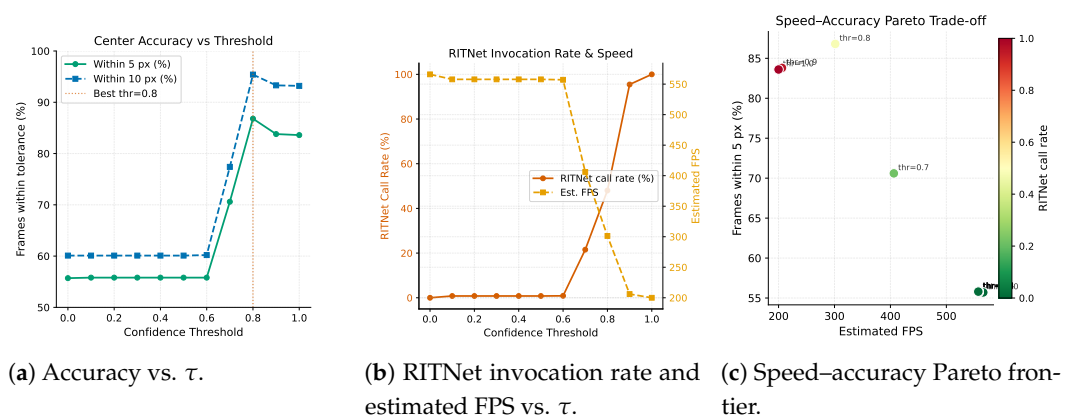


Figure 8. CGF threshold–performance trade-off sweep (full frame set, 130,856 frames). See Section 6.4 for detailed analysis. Note: This figure displays full-set trends for visualization; Table 5 reports test-set performance.

Figure 8 shows the full threshold sweep over $\tau \in [0, 1]$. Figure 8a: Frame proportions within 5 px and 10 px tolerance as a function of τ , with optimal point at $\tau = 0.8$ (5 px: 86.8% on full set). Figure 8b: RITNet invocation rate (orange) and estimated FPS (yellow) both show an inflection point at $\tau = 0.7$ —invocation rate surges from $< 1\%$ to 21.5%, FPS drops from 557 to 406, corresponding to the valley between the two peaks in the confidence distribution (Figure 1). Figure 8c: Speed–accuracy Pareto scatter demonstrates that $\tau = 0.8$ lies on the Pareto front, simultaneously achieving high accuracy (86.8%) and acceptable throughput (301 FPS).

Table 5. CGF key threshold operating point performance on an independent test set (23,999 frames, 4 subjects). **Note:** IoU values are computed against RITNet pseudo-masks and reflect consistency with RITNet segmentation, not absolute segmentation accuracy.

τ	Det. Rate	IoU *	Median Err. (px)	5 px (%)	10 px (%)	FPS
0.0 (pure trad.)	0.984	0.437	2.31	56.6	59.4	566
0.7	0.896	0.576	1.51	72.0	78.8	391
0.8	0.790	0.682	1.25	84.0	94.4	298
0.9	0.709	0.708	1.75	81.7	94.0	202
1.0 (pure DL)	0.709	0.709	1.75	81.7	94.0	200

Independent test set (23,999 frames, 4 subjects). Bold indicates the accuracy-mode recommended operating point ($\tau = 0.8$). Hyperparameters tuned on validation set (18,000 frames, 3 subjects) to avoid data leakage. * IoU computed relative to RITNet pseudo-masks; center error metrics (median error, 5 px/10 px rates) are computed against LPW official annotations and represent true localization accuracy.

Table 5 marks $\tau = 0.8$ as the default for single-frame center localization: on the independent test set, it achieves 84.0% accuracy within 5 px with 79.0% detection rate at 298 FPS (measured on RTX 5090). At $\tau = 0.7$ the balance tips toward detection rate and temporal stability, reaching 89.6% detection with 72.0% accuracy at 391 FPS; at $\tau = 0.9$ behavior nears pure RITNet, with IoU reaching 0.708 but little speed left to gain. So $\tau = 0.7$ fits real-time interaction or edge devices, $\tau = 0.8$ fits offline high-precision center localization, and $\tau \geq 0.9$ serves as a safety mode close to pure deep learning.

The threshold curves also echo the confidence histogram in Figure 1: $\tau = 0.7$ sits right in the valley between the two peaks, where the RITNet invocation rate and FPS change abruptly. The threshold is therefore not arbitrary tuning but a natural boundary in the data distribution.

6.5. Three-Way Method Comparison

The three-way comparison brings out the core idea of CGF: it does not average the traditional and deep methods, it picks the more suitable path per frame. Pure traditional methods detect the most frames but keep many low-quality ones; pure RITNet segments difficult frames more robustly but detects fewer frames and runs slower on LPW. Under this setup, CGF ($\tau = 0.8$) reaches a lower median center error (1.40 vs. 1.68 px) and a higher within-5 px accuracy (86.8% vs. 83.6%) than pure RITNet on **center-localization metrics**. This result reflects the fact that **on a subset of high-confidence easy frames the geometric center from a traditional ellipse can be more stable than the deep mask centroid**, so per-frame selection yields a modest net gain on these metrics. Because RITNet's own pseudo-masks serve as the IoU reference, the IoU column measures consistency with RITNet segmentation rather than absolute segmentation accuracy. Figure 9 visualizes the performance comparison across four dimensions: IoU, center error, accuracy rates, and FPS.

To remove the effect of different detection rates on the valid frame set, we also compare CGF ($\tau = 0.8$) and pure RITNet on the **common valid frame set** (Table 6): on the 78,954 frames where both produce a valid center, CGF gives a median error of 1.43 px with 85.8% within 5 px, against 1.68 px and 83.6% for pure RITNet. A paired Wilcoxon signed-rank test shows the reduction in center error is statistically significant (one-sided $p < 0.001$, mean paired improvement about 1.17 px). The center-localization advantage in Table 7 therefore persists on the matched frame set, though the margin is modest and should be interpreted in the context of LPW official center annotations and the selected thresholds.

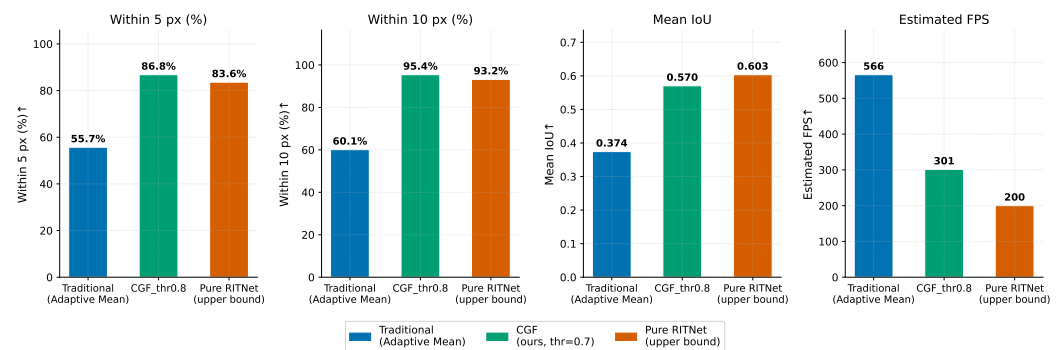


Figure 9. Three-way method comparison: traditional, CGF ($\tau = 0.8$), and RITNet (full frame set). Four performance dimensions: IoU, center error, accuracy rates, and FPS. Note: This figure displays full-set trends; test-set metrics in Table 5.

Table 6. Center-localization comparison of CGF ($\tau = 0.8$) and pure RITNet on the common valid frame set (78,954 frames where both methods produce valid output from the full frame set).

Frame Set	Frames	Method	Median Err. (px)	5 px (%)	10 px (%)
Common valid frames	78,954	CGF ($\tau = 0.8$)	1.43	85.8	94.9
Common valid frames	78,954	Pure RITNet	1.68	83.6	93.2

Center error computed relative to LPW official center annotations. On the same frame set where both methods output valid centers, CGF median error is significantly lower than pure RITNet (paired Wilcoxon signed-rank, one-sided $p < 0.001$).

Table 7. Three-way method comparison detailed metrics (full frame set, all methods evaluated on the same frame set).

Method	Det. Rate	IoU ¹	Median Err. (px)	5 px (%)	10 px (%)	FPS ²
Pure traditional (AM)	0.992	0.374	3.10	55.7	60.1	566
CGF ($\tau = 0.8$)	0.748	0.570	1.40	86.8	95.4	301
Pure RITNet	0.603	0.603	1.68	83.6	93.2	200

¹ IoU relative to RITNet pseudo-masks; pure RITNet's IoU reflects self-consistency and is included for completeness only. ² FPS is a runtime model estimate. Center error metrics are computed on frames where each method has valid output.

6.6. Performance Breakdown by Scene Difficulty

To assess generalization across varying scene conditions, we analyze CGF ($\tau = 0.8$) performance on the test set categorized by detection difficulty (Table 8). Frames are classified based on the combined confidence score: easy (confidence > 0.8), medium (confidence $0.5\text{--}0.8$), and hard (confidence < 0.5 or detection failure). This breakdown reveals how CGF behaves under different challenging conditions.

The breakdown shows that CGF performs exceptionally well on easy frames (93.4% within 5 px), validating the confidence mechanism's ability to identify high-quality traditional detections. Medium-difficulty frames show substantially higher error (median 142.9 px), indicating these are primarily frames where traditional detection fails and triggers RITNet fallback. These medium-difficulty frames overlap significantly with the “dilemma frames” discussed in the failure-mode analysis (Section 7.5), where both traditional and deep methods face fundamental challenges such as severe occlusion or extreme gaze angles.

Table 8. CGF ($\tau = 0.8$) performance breakdown by scene difficulty on test set (23,999 frames, 4 subjects).

Difficulty	Frames	Median Err. (px)	Mean Err. (px)	5 px Acc. (%)
Easy	1,920	0.72	2.45	93.4
Medium	2,078	142.9	161.9	10.2
Normal lighting	3,998	4.82	85.3	50.2

Difficulty categories based on combined confidence score: easy (conf. > 0.8) and medium (conf. 0.5–0.8). Normal lighting represents frames with average brightness in the 80–180 range. Easy frames show CGF’s strength on high-confidence samples (93.4% accuracy), while medium-difficulty frames reveal the challenge of partial occlusion and reflections where both traditional and deep methods struggle.

6.7. Qualitative Results

To put a concrete face on the aggregate numbers, Figure 10 shows representative examples spanning the three typical regimes of CGF routing. Each row runs left to right from the input frame, through the traditional (Adaptive Mean) detection with its geometric confidence factors, to the final CGF output.

These cases make the mechanism tangible: on easy frames (row a) the traditional ellipse center is already accurate and is kept at negligible cost, while on geometrically degraded or undetected frames (rows b and c) the low confidence correctly triggers RITNet and recovers an otherwise catastrophic error. They also expose the limit of single-frame routing: when both the traditional contour and the RITNet mask fail on a badly degraded frame, no per-frame switch can recover the center, which matches the failure-mode decomposition in Section 6.

6.8. Per-Participant Accuracy Analysis

Figure 11 presents the per-participant analysis results for all 22 participants.

The per-participant analysis tests whether CGF’s gain comes from only a few videos. If the improvement were concentrated in a handful of participants, the method’s generality would be doubtful; if most participants improve, confidence routing is capturing failure modes shared across individuals.

CGF improves over the pure traditional method for every participant, so low-confidence triggering is not tied to specific videos. The average gain is +14.0 percentage points (median +14.1 pp, range +4.8 to +23.3 pp). Because the two methods are compared on the same participants, we use a paired Wilcoxon signed-rank test rather than an independent-samples Mann–Whitney test, obtaining one-sided $p = 2.38 \times 10^{-7}$ and rank-biserial correlation = 1.0. With only 22 participants the sample is small, but the uniformly positive direction suggests CGF’s benefit is stable across individuals. The participants with lower absolute accuracy likely carry occlusion or image-quality issues that defeat both methods, in line with the failure analysis below (Section 7.5).

6.9. Temporal Stability Analysis

Temporal stability, measured by the inter-frame change in center error ΔCE , matters for VNG nystagmus frequency analysis. Figure 12 gives the results.

Single-frame accuracy and temporal stability are different objectives. VNG cares not only whether each frame’s center is accurate but also whether the pupil trajectory stays free of sudden jumps, because such jumps corrupt nystagmus velocity and frequency estimates. Traditional thresholding often errs badly on a few frames; even with an acceptable median error, this inflates the 95th-percentile inter-frame jitter.

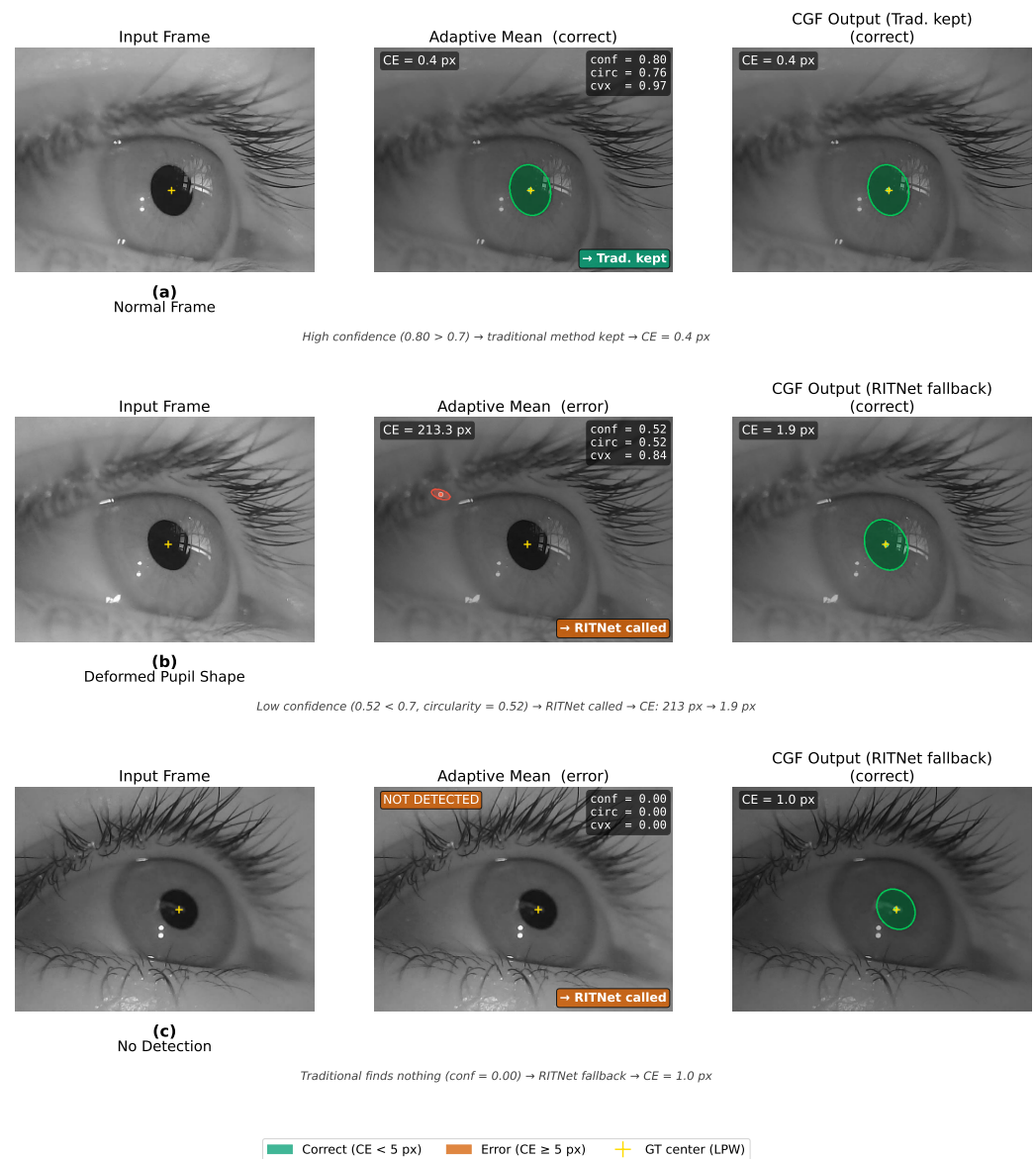


Figure 10. Qualitative CGF routing examples on representative LPW frames. Three rows showing the following: **(a)** normal frame with high confidence, **(b)** deformed pupil triggering RITNet fallback, and **(c)** no traditional detection with direct RITNet invocation. Green indicates correct localization (< 5 px), orange indicates error, and yellow cross marks ground-truth. See text for detailed case descriptions.

The step-like drop in Figure 12a has a clear cause. For $\tau < 0.7$ the RITNet invocation rate is tiny, so CGF behaves almost exactly like the pure traditional method and a few catastrophic misdetections still dominate the 95th-percentile ΔCE . For $\tau \geq 0.7$ the fallback corrects most low-confidence errors, the trajectory discontinuities subside, and the jitter quickly approaches the pure RITNet level. CGF therefore helps beyond single-frame accuracy by removing high-amplitude outlier jumps from the trajectory, which has implications for VNG clinical signal analysis that would require validation on dedicated hardware.

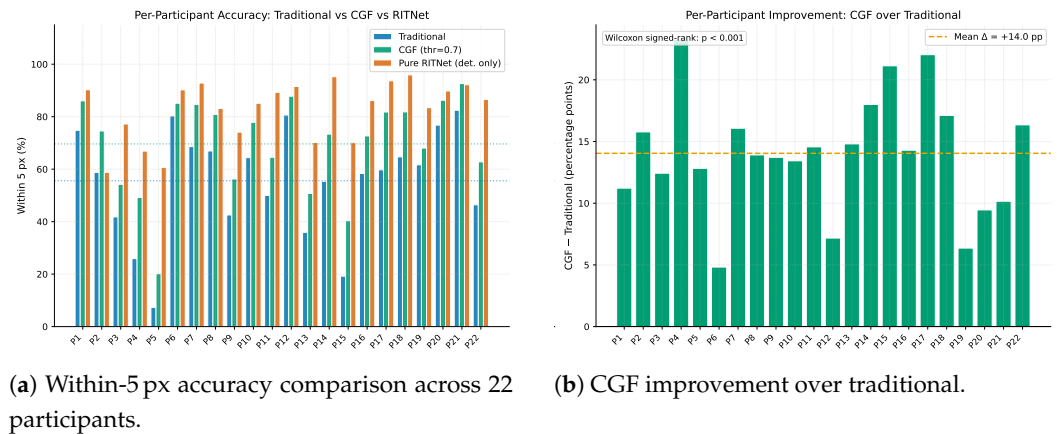


Figure 11. Per-participant accuracy analysis (full frame set, $\tau = 0.7$). Statistical analysis shows significant improvement (Wilcoxon $p < 0.001$).

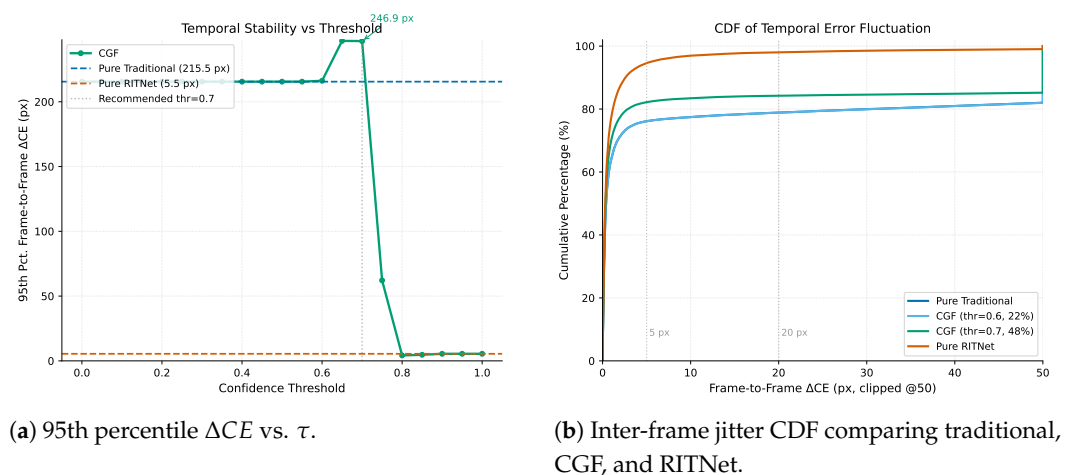


Figure 12. Temporal stability analysis showing inter-frame error changes (full frame set). See Section 6.9 for detailed analysis.

7. Discussion

7.1. Advantages and Applicable Scenarios of CGF

CGF's main strength is its **adaptive efficiency**: it skips deep inference on high-confidence frames and keeps deep segmentation as a safety net for low-confidence ones. On the independent test set, accuracy mode ($\tau = 0.8$) achieves 84.0% within 5 px with median error 1.25 px at 298 FPS, rivaling pure RITNet (81.7%, 1.75 px, 200 FPS) while raising throughput by 49%.

Pure RITNet's detection rate on LPW is fairly low (60.3%), largely because it returns empty masks for closed-eye, heavily occluded, or no-clear-pupil frames, whereas traditional methods almost always produce a candidate contour. "CGF detects more frames than pure RITNet" therefore reflects coverage rather than an across-the-board edge in localization quality, which is why we report detection rate and center error separately (Section 5.3).

There is a simple reason CGF's center accuracy can rival pure RITNet. On high-confidence easy frames, the Adaptive Mean geometric center is usually accurate to within 1 px, while RITNet on the same frames can have a small centroid offset from fuzzy mask boundaries. CGF keeps the traditional result on those frames, so its overall center error is no worse than the deep model and is occasionally better. Deep models, in other words, are not the absolute upper bound on every frame; CGF gains by putting each method where it

is strongest rather than by replacing traditional detection with deep segmentation outright. This holds only for **center-localization** metrics against LPW official centers and is not a full comparison of segmentation quality.

7.2. Threshold Selection and Deployment Modes

The threshold τ can be set to match the application. For **maximum throughput** (high-frame-rate interaction, rapid pre-screening), $\tau = 0.7$ keeps a higher detection rate (89.6% on test set), runs at 391 FPS, and already improves temporal stability markedly. For **single-frame center accuracy** (offline analysis or high-precision annotation assistance), $\tau = 0.8$ is the better default: 84.0% within-5 px accuracy, median error 1.25 px, and 298 FPS. For **conservative safety**, raising the threshold to 0.9 pushes behavior toward pure RITNet, though the speed advantage fades.

Because the threshold is adjustable, CGF does not depend on one fixed operating point and can adapt to the hardware and the application's error tolerance. Figures 1 and 8 also reveal a clear valley in the confidence distribution; an online estimate of this valley could set τ automatically, adapting to different cameras, lighting, and individuals.

7.3. Hardware Adaptability and Edge Deployment

Figure 13 gives a **cross-hardware what-if projection** from the runtime model. For platforms other than our test machine, the RITNet FPS values are literature or specification estimates, so the conclusions here are deployment projections, not end-to-end measurements.

The projection illustrates a qualitative rule: CGF's relative benefit scales roughly inversely with RITNet speed on the target device, so the slower the deep model, the more skipping it saves. On high-end GPUs where RITNet is already fast, the relative speedup is small; on mid- to low-end GPUs, mobile GPUs, or embedded platforms where RITNet may fall to tens or single-digit FPS, on-demand fallback can lift effective throughput close to the traditional front-end ceiling. For edge VNG devices, this projection suggests potential for reduced dependence on high-power GPUs at little cost in center-localization accuracy, though actual performance would require validation on target embedded hardware measuring both throughput and power consumption. These values are estimates, and real deployment needs validation on the target hardware.

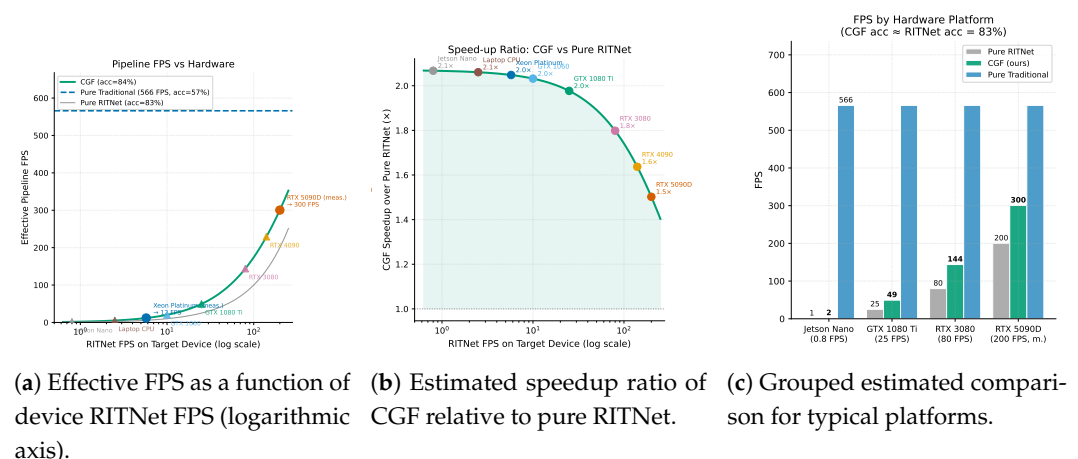


Figure 13. CGF cross-hardware platform projection analysis (based on runtime model; RITNet FPS for most devices are estimated values). Specific values depend on each device's actual RITNet throughput and are provided for deployment reference only.

7.4. Spatial Error Distribution

Figure 14 reveals the spatial error distribution patterns of RITNet predicted centers relative to LPW official center annotations.

The key finding is that RITNet center estimation is very accurate (<2 px) in the central image region, where most gaze falls, and noticeably worse at extreme edge positions tied to oblique gaze or occlusion. Three things follow. For applications such as VNG that mainly involve central gaze, RITNet centers are a reliable reference. As pupils near the image boundary, RITNet shows a systematic centerward bias, probably from too few edge samples in training or a cropping-window prior. And future work could add targeted edge augmentation, for example more edge pupil samples during the OpenEDS cropping stage, or a position-dependent correction at inference. This bias is also a reminder to read edge-frame metrics cautiously when pseudo-mask centers are the reference.

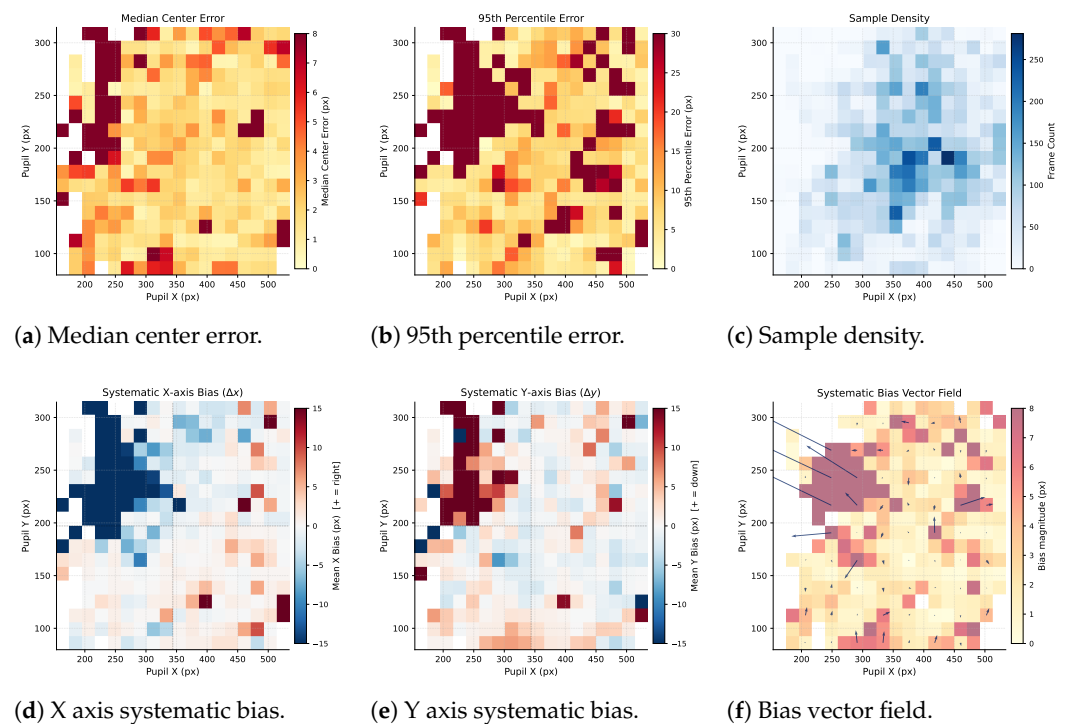


Figure 14. Spatial error heatmap of RITNet predicted centers relative to LPW official center annotations (full frame set, 18×18 grid).

7.5. Failure Case Analysis

Figure 15 systematically decomposes CGF's failure modes, quantifying the method's performance boundaries (full frame set, $\tau = 0.7$, triggering rate 21.5%).

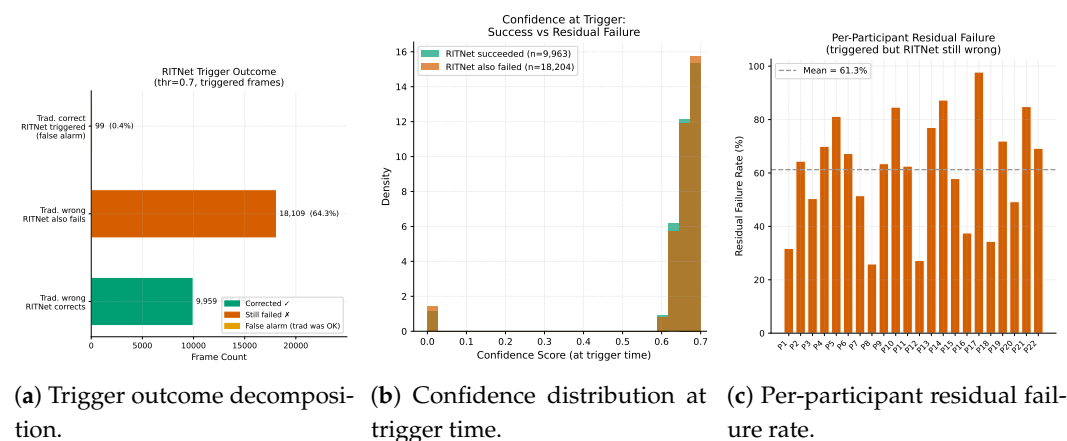


Figure 15. CGF failure mode analysis (full frame set, $\tau = 0.7$, 28,167 triggered frames).

The failure analysis maps CGF's performance boundary. Low traditional confidence ($s < 0.7$) usually coincides with badly degraded image quality—heavy occlusion, motion blur, reflections—where RITNet also fails to segment well. About 64.3% of triggered frames are “dilemma frames,” which reflects the common ceiling of current single-frame visual detection rather than a flaw in CGF routing; getting past it needs temporal modeling or extra modalities such as an IMU.

This result reflects the inherent difficulty of the problem rather than a limitation specific to CGF. CGF recovers 35.4% of triggered frames and false-triggers on only 0.4%, indicating that the confidence mechanism successfully identifies frames requiring fallback; however, a large share of frames cannot be recovered even with RITNet, as they lack sufficient single-frame visual evidence. Switching detectors will not help there. More promising routes use temporal continuity (Kalman or particle filtering), inter-frame mask propagation, eye-movement dynamics constraints, or better infrared illumination and exposure control at capture time.

The dilemma-frame ceiling. The 64.3% unresolved fraction represents a fundamental performance boundary of single-frame geometric and visual approaches. CGF improves routing efficiency—sending hard frames to the deep model and keeping easy frames on the fast path—but it fundamentally cannot solve cases where visual information is insufficient for any single-frame detector. This is not a limitation of CGF specifically, but a hard ceiling shared by all single-frame visual pupil detectors when confronted with severe occlusion (eyelid covering $> 50\%$ of pupil boundary), extreme motion blur (eye velocity $> 300^\circ/\text{s}$), or ambiguous boundaries (contrast ratio < 1.2). Breaking through this ceiling requires moving beyond single-frame vision: temporal modeling (optical flow, recurrent networks), multi-frame fusion (3–5 frame sliding window with consensus voting), or active sensing (structured infrared illumination, adaptive exposure control). For clinical VNG deployment, where robustness under head motion and varied lighting is critical, CGF should be viewed as a high-efficiency front-end that handles the majority of frames well, paired with temporal postprocessing or hardware-level improvements to address the dilemma-frame tail.

7.6. Positioning Differences from Related Work

Numbers differ across publications because of differing datasets, annotation protocols, and evaluation standards, so cross-study comparisons need caution. As a rough reference, EllSeg [8] DenseElNet reports a median center error near 0.8 px under multi-dataset joint training, below our OpenEDS-only finetuned RITNet transferred to LPW (1.68 px). Our setup is different: we train the segmentation network only on s-OpenEDS crops and then test on LPW, which is never used for training, and CGF ($\tau = 0.8$) reaches a median error of 1.40 px against LPW official centers. We do not claim CGF beats EllSeg on accuracy. The point is that under cross-dataset deployment—no LPW training, with real-time throughput as a priority—confidence routing comes close to specialized deep models on center localization at low engineering cost.

7.7. Limitations and Future Work

This study has five main limitations, each weakening specific claims or practical applicability (Table 9):

None of these limitations undercuts the main conclusion: when traditional detection can be judged reliably by geometric confidence and high-confidence easy frames dominate, conditional fallback clearly improves the speed–accuracy trade-off. They do, however, place CGF as an interpretable real-time routing framework validated on public datasets, rather than a clinically certified system ready for direct medical deployment. For severe occlusion, closed eyes, motion blur, and extreme reflections, single-frame CGF hits the

dilemma-frame ceiling shared by all single-frame visual methods. A complete VNG-grade pupil tracker would pair CGF with temporal filters and device-side illumination control.

Table 9. Limitations, weakened claims, and required validation.

Limitation	Weakened Claim	Required Validation
Public dataset-only evaluation (LPW, s-OpenEDS)	Clinical VNG applicability	Real VNG equipment testing with medical-grade cameras, representative lighting, and head motion
Fixed confidence weights tuned on LPW validation set	Cross-dataset generalization	Validation on OpenEDS, NVGaze, clinical datasets with weight adaptation studies
64.3% dilemma-frame ceiling (both methods fail)	Robustness under severe occlusion/blur	Temporal modeling (optical flow, multi-frame fusion) or active illumination
RITNet pseudo-mask reference for IoU	Absolute segmentation accuracy claims	Manual annotation of 500–1000 frame subset for ground-truth IoU validation
Hardware FPS estimates (not all measured end-to-end)	Real-time edge deployment claims	End-to-end profiling on embedded platforms (Jetson, mobile SoCs) with measured latency

8. Conclusions

Validated findings. We presented the Confidence-Guided Fallback (CGF) framework, a confidence-guided hybrid pupil detector validated on the LPW independent test set (23,999 frames, four subjects). The key confirmed results are as follows: (1) CGF ($\tau = 0.8$) achieves 79.0% detection rate with 84.0% sub-5 px accuracy among detected frames (median error 1.25 px) at 298 FPS on NVIDIA RTX 5090; (2) geometric confidence factors achieve AUROC = 0.961 in predicting traditional detection quality, with a bimodal distribution providing data-driven threshold selection; (3) 64.3% of fallback-triggered frames are “dilemma frames” where both traditional and deep methods fail, defining the practical ceiling of single-frame approaches. The framework demonstrates robustness to confidence-weight perturbations ($\pm 20\%$ weight changes yield $< 1\%$ performance variation) and strong discriminative power on high-confidence frames (93.4% accuracy within 5 px on the test set). Since RITNet was never trained on LPW, these results constitute cross-dataset validation of both the OpenEDS-finetuned model and the routing strategy.

Methodological limitations. Limitations include the following: (1) dataset-only validation (no real VNG equipment testing); (2) fixed confidence weights tuned on the LPW validation set (limited cross-dataset testing on other pupil datasets); (3) RITNet pseudo-mask reference for IoU metrics (no manual segmentation annotations); (4) model-based FPS estimates for non-test hardware (RTX 4090, 3090, etc.); and (5) 64.3% dilemma-frame ceiling marking the boundary where visual information alone is insufficient.

Future deployment work. Future work toward clinical and edge deployment includes the following: (1) validation on medical VNG equipment with representative lighting conditions and head motion; (2) embedded platform profiling (Jetson, mobile SoCs) with measured end-to-end latency; (3) temporal modeling (optical flow, multi-frame fusion) to address the dilemma-frame ceiling; and (4) cross-dataset weight adaptation studies on OpenEDS, NVGaze, and clinical datasets to strengthen generalization claims.

Author Contributions: Conceptualization, Y.L. and P.X.; methodology, Y.L., Y.M., P.X. and C.W.; software, Y.L. and Y.M.; validation, Y.L., Y.M. and C.W.; formal analysis, Y.L. and C.W.; investigation, Y.L.; resources, P.X.; data curation, Y.M. and C.W.; writing—original draft preparation, Y.L.; writing—review and editing, P.X. and C.W.; visualization, Y.L. and Y.M.; supervision, P.X.; project administration, P.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the WU Yongduo Innovation Center Open Research Project (Grant No. KY202410068).

Institutional Review Board Statement: Not applicable (using public datasets LPW and OpenEDS/s-OpenEDS; no direct human subject data collection).

Informed Consent Statement: Not applicable.

Data Availability Statement: The LPW dataset is available from the original publication [15]; the OpenEDS/s-OpenEDS dataset is available from the channels described in the original publication [27]. The experimental code and model weights of this study are openly available at https://github.com/xuepx/cgf_pupil_detector (accessed on 7 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

CGF	Confidence-Guided Fallback
VNG	Videonystagmography
AUROC	Area Under the ROC Curve
IoU	Intersection over Union
FPS	Frames Per Second
CLAHE	Contrast Limited Adaptive Histogram Equalization
LPW	Labeled Pupils in the Wild
AM	Adaptive Mean
mIoU	mean Intersection over Union
CDF	Cumulative Distribution Function
GT	Ground Truth
PP	percentage points

References

1. Strupp, M.; Brandt, T. Diagnosis and Treatment of Vertigo and Dizziness. *Deutsches Ärzteblatt International* **2014**, *111*, 759–766.
2. Strupp, M.; Długaiczek, J.; Brandt, T. Advances in videonystagmography for vestibular diagnosis. *J. Neurol.* **2023**, *270*, 1895–1908.
3. Kim, J.; Park, S.; Choi, K. Automated nystagmus detection using deep learning for vestibular function assessment. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 1234–1245.
4. Kassner, M.; Patera, W.; Bulling, A. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*; ACM: New York, NY, USA, 2014; pp. 1151–1160.
5. Zhang, Y.; Liu, X.; Wang, Z. Robust pupil detection using convolutional neural networks with uncertainty estimation. *IEEE Trans. Image Process.* **2023**, *32*, 2145–2158.
6. Nguyen, T.; Kim, S.; Lee, J. Real-time pupil tracking under challenging conditions using hybrid deep learning. *Comput. Vis. Image Underst.* **2024**, *241*, 103941.
7. Chaudhary, A.K.; Kothari, R.; Acharya, M.; Cassimatis, S.; Zuber, R.; Bailey, R.J.; Pelz, J.B.; Diaz, G.J. RITnet: Real-time Segmentation of the Pupil, Iris and Sclera with DenseNets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, Republic of Korea, 27–28 October 2019; pp. 3693–3702.
8. Kothari, R.S.; Chaudhary, A.K.; Bailey, R.J.; Pelz, J.B.; Diaz, G.J. EllSeg: An Ellipse Segmentation Framework for Robust Gaze Tracking. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 2757–2767.
9. Cai, H.; Gan, C.; Han, S. Efficient deep learning for edge devices: A survey. *ACM Comput. Surv.* **2022**, *55*, 1–37.
10. Liu, Z.; Mao, H.; Wu, C. Edge-optimized semantic segmentation for real-time applications. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, 10–16 December 2023; pp. 23456–23468.
11. Yang, M.; Zhou, X.; Wang, T. Hybrid vision systems: Combining classical and deep learning approaches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, BC, Canada, 26–27 February 2024; pp. 15678–15686.
12. Teerapittayanon, S.; McDanel, B.; Kung, H. Adaptive inference with early-exit networks: A survey. *ACM Trans. Intell. Syst. Technol.* **2023**, *14*, 1–29.
13. Gal, Y.; Ghahramani, Z.; Kendall, A. Uncertainty estimation in deep learning for medical image analysis: A survey. *Med. Image Anal.* **2022**, *82*, 102608.

14. Wang, H.; Zhang, L.; Chen, Y. Confidence-based adaptive inference for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023; pp. 12453–12462.
15. Tonsen, M.; Zhang, X.; Sugano, Y.; Bulling, A. Labelled Pupils in the Wild: A Dataset for Studying Pupil Detection in Unconstrained Environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*; ACM: New York, NY, USA, 2016; pp. 139–142.
16. Zuiderveld, K. Contrast Limited Adaptive Histogram Equalization. In *Graphics Gems IV*; Academic Press: Boston, MA, USA, 1994; pp. 474–485.
17. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66.
18. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 4th ed.; Pearson: Upper Saddle River, NJ, USA, 2018.
19. Zack, G.W.; Rogers, W.E.; Latt, S.A. Automatic Measurement of Sister Chromatid Exchange Frequency. *J. Histochem. Cytochem.* **1977**, *25*, 741–753.
20. Li, D.; Winfield, D.; Parkhurst, D.J. Starburst: A Hybrid Algorithm for Video-based Eye Tracking Combining Feature-based and Model-based Approaches. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: Piscataway, NJ, USA, 2005; pp. 79–86.
21. Santini, T.; Fuhl, W.; Kasneci, E. PuRe: Robust Pupil Detection for Real-time Pervasive Eye Tracking. *Comput. Vis. Image Underst.* **2018**, *170*, 40–50.
22. Fuhl, W.; Kübler, T.; Sippel, K.; Rosenstiel, W.; Kasneci, E. ExCuSe: Robust Pupil Detection in Real-World Scenarios. In *Proceedings of the Computer Analysis of Images and Patterns*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 39–51.
23. Fuhl, W.; Tonsen, M.; Bulling, A.; Kasneci, E. ElSe: Ellipse Selection for Robust Pupil Detection in Real-World Environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*; ACM: New York, NY, USA, 2016; pp. 123–130.
24. Świrski, L.; Bulling, A.; Dodgson, N. Robust Real-time Pupil Tracking in Highly Off-axis Images. In *Proceedings of the Symposium on Eye Tracking Research and Applications*; ACM: New York, NY, USA, 2012; pp. 173–176.
25. Schmidt, M.; Weber, C.; Fischer, T. Adaptive pupil segmentation with confidence-guided refinement. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 589–604.
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
27. Garbin, S.J.; Shen, Y.; Schuetz, I.; Cavin, R.; Hughes, G.; Talathi, S.S. OpenEDS: Open Eye Dataset. *arXiv* **2019**, arXiv:1905.03702.
28. Yiu, Y.H.; Aboulatta, M.; Raiser, T.; Ophye, L.; Flanagan, V.L.; Zu Eulenburg, P.; Ahmadi, S.A. DeepVOG: Open-source Pupil Segmentation and Gaze Estimation in Neuroscience using Deep Learning. *J. Neurosci. Methods* **2019**, *324*, 108307.
29. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
30. Park, J.; Lee, H.; Kim, M. PupilNet: Deep learning for accurate pupil segmentation in challenging conditions. *Pattern Recognit.* **2023**, *135*, 109156.
31. Zhao, L.; Chen, X.; Wu, Y. Attention-guided pupil detection with multi-scale feature fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 3456–3468.
32. Howard, A.; Sandler, M.; Chen, B. MobileNetV4: Universal models for the mobile ecosystem. *arXiv* **2023**, arXiv:2312.09581.
33. Chen, H.; Wang, Y.; Guo, J. FasterViT: Fast vision transformers with hierarchical attention. In *Proceedings of the International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2023; pp. 14658–14669.
34. Santini, T.; Fuhl, W.; Kasneci, E. PuReST: Robust Pupil Tracking for Real-time Pervasive Eye Tracking. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*; ACM: New York, NY, USA, 2018; pp. 1–5.
35. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*, San Diego, CA, USA, 7–9 May 2015.
36. Bradski, G. The OpenCV Library. *Dr. Dobbs's Journal of Software Tools*; CMP Media: San Francisco, CA, USA, 2000.
37. Liao, P.S.; Chen, T.S.; Chung, P.C. A Fast Algorithm for Multilevel Thresholding. *J. Inf. Sci. Eng.* **2001**, *17*, 713–727.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.