Accelerating Symptom Detection in Canine Ocular Diseases: A Comparative Analysis and Heterogeneous Ensemble of YOLOv8 and RT-DETR

1st Venzhower M. Manlangit
Dept. of Computer, Information Sciences and Mathematics
University of San Carlos
Cebu City, Philippines
venzhowermanlangit@gmail.com

2nd Kristian Angelo Ray C. Rosillosa
Dept. of Computer, Information Sciences and Mathematics
University of San Carlos
Cebu City, Philippines
kristianrosillosa@gmail.com

3rd Christine D. Bandalan, M.Eng.
Dept. of Computer, Information Sciences and Mathematics
University of San Carlos
Cebu City, Philippines
cdbandalan@usc.edu.ph

***Abstract*—Canine ocular diseases such as cataracts, conjunctivitis, and cherry eye present significant challenges in veterinary practice due to overlapping symptoms and the patients' inability to verbalize discomfort. Delayed diagnosis often leads to irreversible vision impairment. While modern object detectors like the CNN-based YOLOv8 and the Transformer-based RT-DETR offer potential for automated symptom detection, their comparative effectiveness for this specific, fine-grained veterinary task has remained unexplored. This study conducted a direct comparative analysis of these two architecturally distinct models to determine their efficacy in detecting and localizing visual indicators of canine ocular pathologies. To establish a reliable foundation, a publicly available corpus was subjected to a rigorous, expert-guided curation and re-annotation protocol governed by a novel "shrink-wrap" bounding box rule. Both models were trained for 100 epochs to ensure methodological parity. The empirical results indicate that while the CNN-based YOLOv8n achieved a robust Mean Average Precision (mAP@50) of 0.952, the Transformer-based RT-DETR-l outperformed it with an overall mAP@50 of 0.962. Notably, RT-DETR demonstrated superior sensitivity in detecting Conjunctivitis, achieving a 3.1% performance gain over the baseline (0.912 vs. 0.881) and significantly reducing background false positives from 10% to 3%. Furthermore, a heterogeneous ensemble model was developed using the Weighted Boxes Fusion (WBF) algorithm to synergize the localization speed of the CNN with the classification sensitivity of the Transformer, serving as a robust clinical safety net.**

***Index Terms*—computer vision, veterinary ophthalmology, YOLOv8, RT-DETR, ensemble learning, fine-grained visual categorization**

I. Introduction

The health and welfare of companion animals are significantly impacted by ocular conditions. Diseases such as cataracts, cherry eye, and conjunctivitis are highly prevalent in canine populations and can lead to pain, blindness, or systemic infection if left untreated. Diagnosis in general veterinary practice faces three primary obstacles: (1) overlapping symptoms, where a generic "red eye" may signal benign irritation or a vision-threatening emergency; (2) the patient's inability to verbalize symptoms; and (3) limited access to specialized equipment like slit lamps or tonometers in resource-limited settings.

Recent advancements in Deep Learning (DL) have revolutionized medical imaging, moving from manual feature extraction to automated pattern recognition. However, a critical gap remains in the specific domain of veterinary ophthalmology. Existing studies often rely on older Convolutional Neural Network (CNN) architectures or classification-only models that lack localization capabilities. Furthermore, the architectural debate between the established CNN paradigm (known for local feature extraction) and the emerging Vision Transformer paradigm (known for global context awareness) has not been empirically tested in this specific domain.

This study addresses this gap by conducting a head-to-head comparison of **YOLOv8** (You Only Look Once), representing the state-of-the-art in CNN efficiency, and **RT-DETR** (Real-Time Detection Transformer), representing the hybrid Transformer approach. The primary objective is to determine their relative effectiveness in detecting and localizing visual indicators of common ocular diseases from standard clinical photographs.

II. Related Literature

A. Evolution of Veterinary Diagnostics

Before the widespread adoption of deep learning, automated diagnosis relied on classical machine learning paradigms defined by manual, "hand-crafted" feature engineering. Early

studies utilized Support Vector Machines (SVMs) to classify cataracts based on texture histograms [1]. However, these methods were brittle, struggling with variations in lighting and camera angles inherent in clinical photography.

B. CNNs in Medical Imaging

The advent of Convolutional Neural Networks (CNNs) introduced automated representation learning. Models like ResNet and earlier YOLO iterations have proven effective in detecting distinct structural anomalies. A landmark study by Lee et al. (2019) utilized InceptionV3 to classify ulcerative keratitis with high accuracy [2]. However, traditional CNNs inherently struggle with long-range dependencies, often failing to capture the global context required to distinguish diffuse pathologies from normal anatomical variations.

C. The Rise of Vision Transformers

The introduction of the Transformer architecture has shifted the landscape of computer vision. Unlike CNNs, Transformers utilize self-attention mechanisms to process the entire image simultaneously, weighing the importance of different regions relative to one another [6]. In human medical imaging (e.g., diabetic retinopathy detection), Transformers have shown superior performance in identifying subtle, texture-based lesions. This study investigates whether these advantages translate to the veterinary domain using the RT-DETR architecture [4], specifically for Fine-Grained Visual Categorization (FGVC) tasks where inter-class variance is low.

III. Methodology

The research followed a phased development model consisting of data curation, model training, and ensemble evaluation.

A. Data Acquisition and Curation

The initial dataset was sourced from the "dog-diseases-9class-augm" public repository. Recognizing the prevalence of label noise in public datasets, a rigorous expert-guided curation protocol was implemented under the supervision of a licensed veterinarian.

1) Programmatic Filtration: Images belonging to dermatological (skin) classes were removed. Glaucoma was also excluded as its diagnosis requires tonometry (intraocular pressure measurement) rather than visual inspection alone.

2) The "Shrink-Wrap" Protocol: A strict re-annotation rule was applied to the remaining 1,557 images. Bounding boxes were redrawn to tightly enclose *only* the visible pathology, explicitly excluding healthy anatomical features:

- **Cataracts:** Boxes enclosed only the lens opacity, excluding the iris/pupil boundary to prevent the model from learning "circular shapes" as a feature.
- **Cherry Eye:** Boxes enclosed the prolapsed nictitating membrane mass, excluding the healthy canthus and surrounding fur.
- **Conjunctivitis:** Boxes enclosed contiguous areas of abnormal scleral redness, distinguishing pathological inflammation from normal vascularization.

As shown in Table I, this rigorous process reduced the dataset size but significantly increased clinical label quality.

TABLE I
Data Curation Impact

Curation Stage	Image Count
Raw Data Acquisition	4,900
Programmatic Filtering	$\sim$2,500
Quality Control	$\sim$1,800
Re-annotation (Shrink-Wrap)	**1,557**
Augmentation (Train set only)	**3,725**

B. Data Augmentation

To improve robustness against varying clinical environments, a 3x augmentation pipeline was applied exclusively to the training set. This included geometric transformations (Horizontal Flip, Rotation $\pm 15^{\circ}$) and photometric adjustments (Brightness $\pm 10\%$) to simulate poor lighting conditions common in rural clinics.

C. Model Architectures

1) YOLOv8 (Baseline): The YOLOv8n (Nano) model was selected as the baseline. It utilizes a CSPDarknet53 backbone and a Path Aggregation Network (PAN) neck. It represents the "Edge Deployment" scenario, optimized for low-latency inference on mobile devices with limited computational resources.

2) RT-DETR (Challenger): The RT-DETR-l (Large) model was selected as the challenger. It employs a hybrid encoder that processes multi-scale features using self-attention. It represents the "Clinical Server" scenario, prioritizing maximum diagnostic sensitivity and global context awareness over raw inference speed.

D. Experimental Setup

To ensure methodological parity, both models were trained using the Google Colab platform with an NVIDIA Tesla T4 GPU. The training configuration was standardized:

- **Epochs:** 100 (ensuring full convergence).
- **Image Size:** 640x640 pixels.
- **Optimizer:** AdamW with a learning rate of $1e^{-3}$.
- **Batch Size:** 16 (YOLO) and 8 (RT-DETR, adjusted for VRAM).
- **Split:** 80% Training, 10% Validation, 10% Testing (Stratified).

E. Heterogeneous Ensemble Strategy

A heterogeneous ensemble was constructed using the **Weighted Boxes Fusion (WBF)** algorithm [5]. Unlike Non-Maximum Suppression (NMS), which discards redundant boxes, WBF averages the coordinates of overlapping predictions from different models, weighted by their confidence scores. This allows the system to leverage the consensus between the CNN and Transformer.

IV. Results and Analysis

A. Baseline Model Performance (YOLOv8n)

The YOLOv8n model demonstrated exceptional performance for a lightweight architecture, achieving an overall Mean Average Precision (mAP@50) of **0.952**. As shown in Table II, the model achieved near-perfect detection for **Cherry Eye (0.992)** and **Cataracts (0.984)**. This success is attributed to the distinct morphological features of these conditions (structural deformity and high-contrast opacity), which are effectively captured by the local convolution filters of the CNN.

However, the model exhibited a relative weakness in detecting **Conjunctivitis**, with a lower mAP@50 of **0.881**. The confusion matrix (Fig. 1) reveals a higher rate of background confusion for this class (0.10 false positives), suggesting the CNN struggles to distinguish diffuse inflammatory redness from normal vascularization.

TABLE II
YOLOv8n Performance Metrics (Baseline)

Class	Precision	Recall	mAP@50	mAP@50-95
All	0.930	0.937	**0.952**	0.522
Cataracts	0.975	0.980	0.984	0.493
Cherry Eye	0.961	0.994	0.992	0.697
Conjunctivitis	0.855	0.837	0.881	0.376

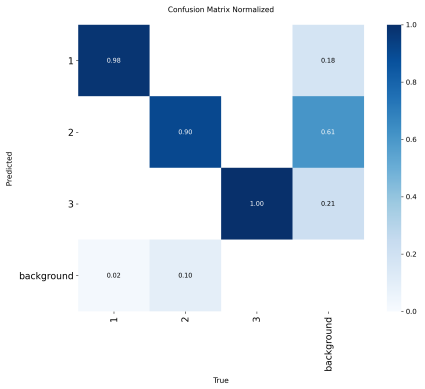

Fig. 1. Normalized Confusion Matrix for YOLOv8n Baseline.

B. Challenger Model Performance (RT-DETR-l)

The RT-DETR model achieved a superior overall mAP@50 of **0.962**. While performance on the structural classes remained statistically similar to the baseline, the critical differentiator was observed in the **Conjunctivitis** class.

As detailed in Table III, RT-DETR achieved a mAP@50 of **0.912** for Conjunctivitis, a **3.1% improvement** over the baseline. This finding validates the hypothesis that the global attention mechanism of the Transformer is superior at interpreting diffuse, texture-based pathologies.

TABLE III
RT-DETR-l Performance Metrics (Challenger)

Class	Precision	Recall	mAP@50	mAP@50-95
All	0.961	0.963	**0.962**	0.505
Cataracts	0.976	0.986	0.983	0.479
Cherry Eye	0.990	0.994	0.991	0.657
Conjunctivitis	0.918	0.909	**0.912**	0.378

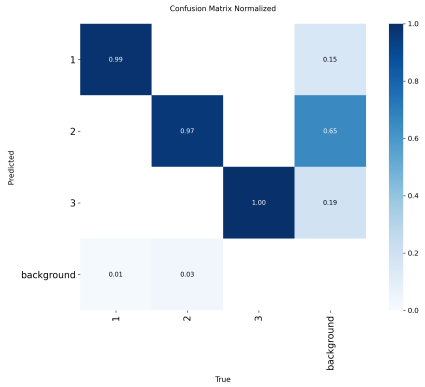

Fig. 2. Normalized Confusion Matrix for RT-DETR-l Challenger.

C. Analysis of Confidence Disparity

Beyond raw detection rates, the RT-DETR model exhibited higher confidence scores when identifying inflammatory conditions. Qualitative analysis revealed that for identical correct detections of Conjunctivitis, the RT-DETR model consistently assigned higher confidence scores (e.g., 0.83) compared to the YOLOv8 baseline (0.67).

This 16% "Confidence Gap" suggests that the Transformer is more decisive. In a clinical decision support system, the confidence score is often used as a threshold for alerting the user. A detection with 67% confidence might be filtered out to prevent spam, while an 83% confidence detection triggers an alert. Therefore, the RT-DETR model is not just more accurate; it is operationally more robust for automated triage.

D. Comparative Analysis & Clinical Translation

Table IV presents the definitive resource-to-performance trade-off. The RT-DETR-l model required significantly more computational resources (32M params vs 3.2M) and training time (7.09h vs 1.49h).

However, in a medical context, this cost is justified. While the 3.1% gain in Conjunctivitis detection may appear marginal statistically, in a high-volume veterinary practice evaluating 1,000 ocular cases annually, this differential translates to approximately **31 additional patients** receiving a correct diagnosis who might otherwise have been missed by the baseline model. Furthermore, RT-DETR reduced background false positives from 10% (YOLOv8) to 3%, significantly reducing potential "alert fatigue" for clinicians.

TABLE IV
COMPARATIVE SUMMARY OF ARCHITECTURAL PERFORMANCE

Model	Params	Time	mAP@50 (Conj.)
YOLOv8n	3.2M	1.49h	0.881
RT-DETR-l	32M	7.09h	**0.912**

E. Heterogeneous Ensemble Results

The ensemble model successfully fused predictions from both architectures. Qualitative inspection (Fig. 3) demonstrates that the ensemble serves as a "clinical safety net," refining bounding boxes in ambiguous cases by requiring consensus between the local features detected by the CNN and the global context analyzed by the Transformer.

Fig. 3. Qualitative Results of the Heterogeneous Ensemble.

V. CONCLUSION

This study successfully demonstrated that while lightweight CNNs (YOLOv8) are highly effective for distinct structural pathologies, Transformer-based models (RT-DETR) provide superior sensitivity for complex, inflammatory conditions. Specifically, the global attention mechanism of RT-DETR proved essential for distinguishing the diffuse redness of Conjunctivitis, achieving a 3.1% performance gain and reducing false positives.

A primary limitation of this study is the reliance on a single expert for the dataset re-annotation process. Future work will focus on expanding the dataset to include ocular manifestations of Canine Distemper, which were excluded due to data scarcity. Based on the findings, a **dual-deployment strategy** is recommended: utilizing YOLOv8n for offline mobile triage applications to ensure accessibility in rural areas, and deploying RT-DETR on cloud servers for second-opinion screening where diagnostic sensitivity is paramount.

REFERENCES

[1] N. Tawfik et al., "Cataract detection using support vector machines and artificial neural networks," *International Journal of Computer Applications*, vol. 180, no. 20, pp. 1-6, 2018.

[2] H. Lee et al., "CNN-based diagnosis models for canine ulcerative keratitis," *Scientific Reports*, vol. 9, no. 1, p. 14209, 2019.

[3] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680-1716, 2023.

[4] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 102, p. 104117, 2021.

[6] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] G. Jouliani et al., "Health and welfare of Brachycephalic (Flat-faced) Companion Animals," *Veterinary Sciences*, 2024.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[9] N. Carion et al., "End-to-end object detection with transformers," in *European Conference on Computer Vision*, Springer, 2020, pp. 213-229.

[10] C. Li et al., "An improved YOLOv8-based lightweight attention mechanism for cross-scale feature fusion," *Remote Sensing*, vol. 17, no. 6, p. 1044, 2024.

[11] M. Buric, S. Grozdanic, and M. Ivasic-Kos, "Diagnosis of ophthalmologic diseases in canines based on images using neural networks for image segmentation," *Heliyon*, vol. 10, no. 19, p. e38287, 2024.