Two-Stage Dataset Curation via Class Pruning for Improved AI-Based Thermal Image Object Detection

Boyang Chen, Tianyang Fang, and Jafar Saniie
Embedded Computing and Signal Processing (ECASP) Research Laboratory (http://ecasp.ece.iit.edu)
Department of Electrical and Computer Engineering
Illinois Institute of Technology, Chicago, IL, U.S.A.

Abstract - This study examines how dataset curation and contrast enhancement affect thermal image object detection using YOLOv11, focusing on pedestrians, vehicles, and other road objects. Experiments were conducted on the FLIR ADAS Thermal Dataset v2 through a two-stage class refinement process. Six severely underrepresented classes were first removed from the original 16-class dataset, producing a 10-class subset. Subsequently, three additional classes exhibiting unstable precision–recall behavior were eliminated based on quantitative per-class PR and F1–confidence analysis, yielding a refined 7-class dataset. Across Nano, Small, Medium, and Large YOLOv11 variants, dataset curation produced substantial and consistent performance improvements. For example, the Nano model's precision increased from 61.7% to 78.6%, while mAP@0.5 (mean Average Precision at 50% Intersection over Union (IoU) for object detection) improved from 49.3% to 63%. For larger model variants, recall improved by approximately 10–11%, indicating enhanced object coverage and reduced missed detections. Notably, precision values converged near 80% under the refined 7-class dataset, suggesting that dataset quality establishes a performance ceiling largely independent of architectural scaling. Contrast Limited Adaptive Histogram Equalization (CLAHE) preprocessing was also evaluated; while it provided modest gains for the medium YOLOv11 model in both 7-class and 10-class datasets, systematic dataset-level refinement via class pruning was the primary factor driving performance improvements, outperforming both contrast enhancement and model scaling.

Keywords—YOLOv11, thermal image detection, dataset curation, class pruning

I. INTRODUCTION

Thermal infrared (TIR) imaging enables object detection under challenging lighting conditions where visible-spectrum cameras fail, such as nighttime driving, low visibility, and adverse weather. The FLIR ADAS Thermal Dataset v2 [1], captured using a FLIR Tau 2 long-wave infrared (LWIR) camera, provides thermal imagery suitable for detecting pedestrians, vehicles, and other road objects under these conditions [2]. Accurate detection in such scenarios is critical for road safety, as machine learning–based perception systems can reduce collisions and improve situational awareness.[3] Thermal imaging helps fill perceptual gaps left by conventional sensors, such as RGB cameras and LiDAR, which may fail in darkness, fog, or smoke. Prior work has shown that integrating LWIR thermal cameras with learning-based detection and tracking algorithms can improve vehicle perception and reliability in real-world conditions. [4]

Recent advances in deep learning-based object detection, particularly the YOLO (You Only Look Once) family of architectures, have significantly improved real-time detection performance. Modern variants such as YOLOv11 [5] provide an effective trade-off between inference speed and mean Average Precision (mAP), making them suitable for both high-performance and edge deployment scenarios. While architectural scaling and network optimization have been widely explored, comparatively less attention has been given to systematic dataset-level refinement in thermal detection tasks.

In many detection studies, performance improvements are pursued through deeper architecture, data augmentation strategies, or image preprocessing techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE). However, the influence of dataset composition, particularly the presence of underrepresented or semantically ambiguous classes, remains underexamined. Poorly represented classes can introduce label imbalance, unstable gradients, and inter-class confusion, potentially degrading overall detection reliability.

This work investigates whether dataset curation can provide greater performance gains than preprocessing or model scaling in thermal object detection. Using the FLIR ADAS Thermal Dataset v2 [4], we implement a two-stage refinement strategy. First, severely underrepresented classes are removed to ensure sufficient training representation. Second, per-class precision–recall (PR) and F1-confidence analysis is conducted to identify classes exhibiting unstable detection behavior. This process yields a refined 7-class dataset derived from the original 16 classes.

Experiments are conducted across Nano, Small, Medium, and Large YOLOv11 variants, with and without CLAHE preprocessing, to isolate the effects of dataset composition, contrast enhancement, and architectural scaling. Results demonstrate that dataset curation produces consistent improvements in precision, recall, and mAP scores across all model sizes, while CLAHE yields marginal and inconsistent gains. Notably, precision values converge across model scales under the refined dataset, indicating that dataset quality establishes a practical upper-bound performance level largely independent of network capacity.

The primary contributions of this work are summarized as follows:

A two-stage dataset curation framework for thermal object detection, separating data insufficiency from metric-driven class instability.

A quantitative diagnostic methodology using PR-curve and F1-confidence analysis to identify poorly separable classes.

A controlled evaluation demonstrates that dataset refinement yields larger and more consistent performance gains than CLAHE preprocessing or model scaling.

II. RELATED WORK

Contrast enhancement has been widely used to improve deep learning performance in both visible and thermal images. Techniques such as gamma encoding and histogram equalization have been shown to improve feature extraction and model accuracy for image recognition tasks [7, 8]. For thermal and infrared imagery, adaptive methods including neighborhood histogram-based contrast enhancement and CLAHE have demonstrated better local contrast preservation while mitigating noise amplification, helping convolutional neural networks detect objects under low-contrast or challenging conditions [9, 10].

In addition to preprocessing, multi-modal approaches have been explored to improve thermal object detection, such as static early fusion of visible and thermal images, which enhances CNN performance in low-light and occluded scenarios [11]. However, most prior work focuses on preprocessing, fusion, or network scaling, with limited investigation of dataset-level factors such as class pruning. Our method addresses this gap by systematically evaluating the effects of dataset curation, CLAHE preprocessing, and YOLOv11 model capacity, demonstrating that careful dataset design can provide more consistent performance gains than contrast enhancement alone.

III. PROPOSED METHOD

A. System Overview

The proposed thermal object detection framework evaluates the effects of class pruning, contrast enhancement, and model scaling within a unified experimental pipeline. Specifically, the experiment examines the impact of removing underperforming classes from the original 16-class dataset and the optional application of Contrast Limited Adaptive Histogram Equalization (CLAHE) on YOLOv11 model variants (Nano, Small, Medium, and Large). The objective is to isolate how dataset composition, preprocessing, and model capacity influence detection performance.

Fig. 1 presents the overall system design and process flow. The framework begins with the original thermal dataset, followed by per-class performance analysis to identify unstable classes. After class pruning, the refined dataset optionally undergoes CLAHE preprocessing before being used for training and validation across multiple YOLOv11 model scales. Detection performance is evaluated using precision, recall, mAP@0.5, and mAP@0.5–0.95 metrics to enable controlled comparison across experimental conditions.

All training and evaluation experiments were conducted on a desktop workstation equipped with an Intel Core Ultra 265K processor (CPU), 96 GB DDR5 RAM, and an NVIDIA RTX 5070 Ti GPU. This configuration ensures sufficient computational capacity for systematic experimentation across dataset variations and model scales.

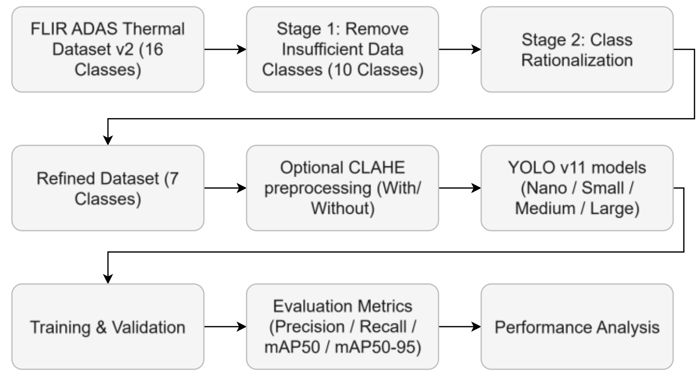

Fig. 1. Block diagram of system design flow.

B. Dataset Description

The YOLOv11 object detection model was employed for thermal image detection due to its favorable trade-off between inference speed and accuracy. Table I summarizes the specifications of different YOLOv11 model sizes used in this experiment, including mAP50–95, number of parameters, and FLOPs. The choice of model size (Nano (n), Small (s), Medium (m), Large (l)) allows evaluation of performance scaling relative to dataset quality and preprocessing. [12]

TABLE I. YOLOV11 PERFORMANCE TABLE

Model	mAP50-95	params (millions)	FLOPs (billions)
YOLO11n	39.5	2.6	6.5
YOLO11s	47	9.4	21.5
YOLO11m	51.5	20.1	68
YOLO11l	53.4	25.3	86.9

The dataset used in this study is the Teledyne FLIR Free ADAS Thermal Dataset v2, captured with a FLIR Tau 2 13 mm LWIR thermal camera in 8-bit grayscale format. The dataset contains 15,635 thermal images divided into 10742 training images, 1144 validation images, and 3749 testing images. The dataset includes 16 object classes, including “person”, “bike”, “car”, “motor”, “bus”, “train”, “truck”, “light”, “hydrant”, “sign”, “dog”, “deer”, “skateboard”, “stroller”, “scooter”, and “other vehicle” classes. The thermal images were taken under diverse environmental conditions, such as day, night, and low visibility scenarios.

To improve detection reliability, in Table II, an initial statistical inspection revealed that six classes (train, dog, deer, skateboard, stroller, and scooter) contained extremely limited training and validation annotations, with some classes exhibiting near-zero validation samples. Due to insufficient representation for stable model learning, these classes were excluded before experimentation, resulting in a reduced 10-class dataset. To further improve detection stability, a second-stage quantitative evaluation was performed on the 10-class dataset using per-class precision–recall (PR) curves and F1-confidence analysis. After evaluating per-class performance using precision-recall (PR) and F1-confidence curves, three classes (“truck”, “hydrant”, and “other vehicle”) with low separability and unstable detection behavior were systematically removed, yielding a refined 7-class dataset. This class pruning reduced inter-class confusion and provided a more stable foundation for evaluating the impact of preprocessing and model scaling.

Training and evaluation were performed across the four experimental conditions: 7 classes without preprocessing, 7 classes with CLAHE, and, for reference, the original 10-class dataset with and without CLAHE. This setup allows clear attribution of performance differences to dataset curation, contrast enhancement, and model capacity.

TABLE II. THERMAL IMAGE ANNOTATIONS

Label	**Train**	**Validation**
person	50,478	4,470
bike	7,237	170
car	73,623	7,133
motor	1,116	55
bus	2,245	179
train	5	0
truck	829	46
light	16,198	2,005
hydrant	1,095	94
sign	20,770	2,472
dog	4	0
deer	8	0
skateboard	29	3
stroller	15	6
scooter	15	0
other vehicle	1,373	63
Total	175,040	16,696

1) Contrast Limited Adaptive Histogram Equalizer (CLAHE):

To evaluate the effect of contrast enhancement on thermal object detection, Contrast Limited Adaptive Histogram Equalization (CLAHE) was optionally applied as a preprocessing step before model training. Thermal infrared imagery often exhibits limited dynamic range and reduced edge contrast, which may affect feature extraction in convolutional neural networks. CLAHE was investigated to determine whether localized contrast normalization improves detection performance.

Histogram equalization enhances image contrast by redistributing pixel intensities according to the cumulative distribution function (CDF). For a grayscale image with intensity levels r_k. The transformation is defined as:

$$s_k = (L-1)\sum_{j=0}^{k}\frac{n_j}{N}$$

Where L represents the number of gray levels (typically 256), n_j is the number of pixels with intensity r_j, and $N$is the total number of pixels in the image. This transformation maps the input intensities to a wider dynamic range, improving global contrast.

However, global histogram equalization may over-amplify noise or distort local structures in images with heterogeneous intensity distributions. To address this limitation, CLAHE performs histogram equalization independently within small spatial regions (tiles) of the image. Let $h_i$denote the histogram bin value within a tile. Contrast limiting is applied such that:

$$h_i' = \min\,(h_i, C)$$

Where $C$is the clip limit. Histogram bins exceeding $C$are are truncated, and the excess counts are redistributed uniformly across all bins before computing the CDF. This mechanism constrains local contrast amplification and mitigates noise enhancement in low-texture regions. After equalization, bilinear interpolation is performed between neighboring tiles to eliminate boundary artifacts.

In this study, CLAHE was implemented using the OpenCV library with a clip limit set to 2.0, and a tile grid size set to 8 × 8. The 8 × 8 tile configuration provides localized contrast enhancement while preserving structural continuity across object boundaries. A clip limit of 2.0 was selected to limit amplification of thermal noise and avoid artificial gradient formation.

CLAHE was applied as an optional preprocessing step enabling controlled comparison between raw thermal images and contrast-enhanced inputs to quantify its impact on precision, recall, and mean Average Precision (mAP).

2) Model Training:

YOLOv11 models were trained on both 7-class and 10-class datasets, with and without CLAHE, to systematically evaluate the impact of dataset quality, preprocessing, and model capacity. All models were trained for a maximum of 100 epochs with early-stopping patience set to 10 epochs on validation loss to prevent early overfitting. The batch size was configured in auto mode to utilize approximately 60% of available GPU memory, preventing resource overflow during training. [11]

3) Metrics Evaluation:

To quantitatively evaluate detection performance, we employ precision, recall, Intersection over Union (IoU), and mean Average Precision (mAP).

a) Intersection over Union (IoU)

Object detection correctness is determined using the Intersection over Union (IoU) metric, defined as:

$$\text{IoU} = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}}$$

where B_p denotes the predicted bounding box and B_{gt} represents the ground-truth bounding box. A detection is considered correct if its IoU exceeds a predefined threshold. In this work, we report results using IoU thresholds of 0.5 and 0.5–0.95.

b) Precision and Recall

Precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Precision reflects detection reliability (low false positives). Recall reflects the completeness of detection (low missed detections).

c) Precision–Recall (PR) Curve

Precision and recall are computed across varying confidence thresholds, generating a Precision–Recall (PR) curve. A model with strong detection capability produces a curve that approaches the top-right region of the plot, indicating both high precision and high recall.

The Average Precision (AP) corresponds to the area under the PR curve:

$$\text{AP} = \int_0^1 P(R)\, dR$$

The mean Average Precision (mAP) is computed as the mean AP across all classes:

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

mAP@0.5 evaluates performance at IoU ≥ 0.5.

mAP@0.5–0.95 averages AP over IoU thresholds from 0.5 to 0.95 in increments of 0.05, providing a stricter localization assessment.

d) F1-Confidence Analysis

To further analyze class stability, we evaluate the F1 score, defined as the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

The F1-confidence curve plots F1 as a function of the detection confidence threshold. The peak of this curve indicates the optimal operating threshold that balances false positives and false negatives.

Classes exhibiting unstable PR behavior or low peak F1 scores are indicative of ambiguous or poorly separable classes. These diagnostic signals are used in Stage 2 of the proposed class pruning framework to identify and remove performance-degrading classes.

IV. Results and Discussion

Fig. 2 displays the detection results of different objects along with their confidence scores. The top-left image shows a nearly completely dark scene, with a person walking across a parking lot. The top-right and bottom-left images also capture nighttime scenarios, with multiple cars and pedestrians moving in a parking area and a neighborhood street, respectively; the bottom-left image additionally shows parked motorcycles on the right, moving vehicles in the center, and pedestrians on the sidewalk. The bottom-right image shows an intersection with traffic lights, signs, cars, motorcycles, and pedestrians.

Training was first conducted using the original 10-class dataset to establish baseline detection performance. Although overall mAP@0.5 values ranged from 0.493 to 0.595 across model sizes, detailed per-class analysis revealed a significant performance imbalance. As shown in Fig. 3, several classes,

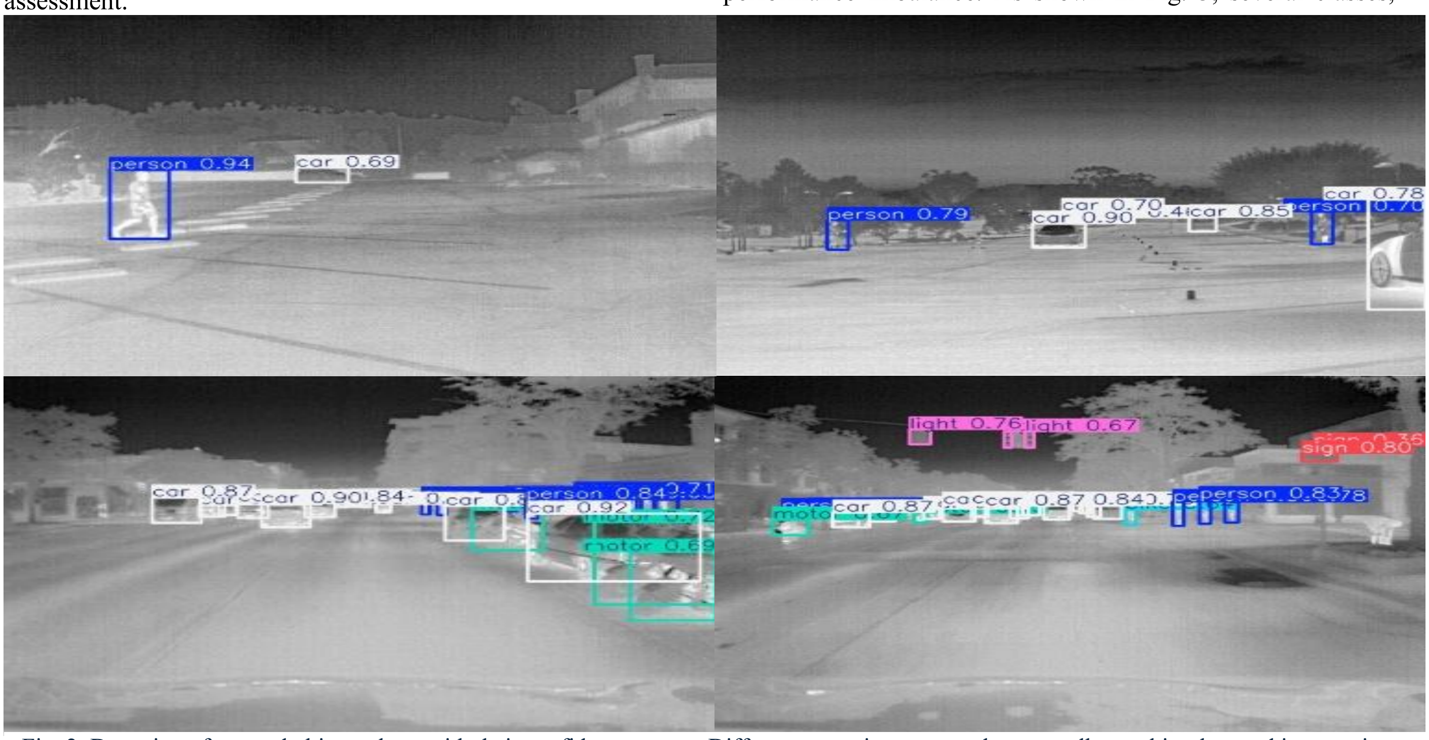

Fig. 2. Detection of on road objects along with their confidence scores. Different scenarios are tested: crosswalks, parking lots and intersections.

such as car and person, exhibit strong precision–recall trade-offs with high average precision values. In contrast, other classes demonstrate rapid precision degradation at low recall levels. In particular, three classes (truck, hydrant, and other vehicle) show shallow PR curves and low area under the curve, indicating poor class separability and unstable detection behavior. To further evaluate class stability, in Fig. 4, F1-confidence curves were analyzed. The F1-confidence analysis confirms that the same three classes produce low peak F1 scores and lack a clear optimal confidence threshold. This instability suggests that these classes introduce ambiguity during training and negatively affect overall detection reliability.

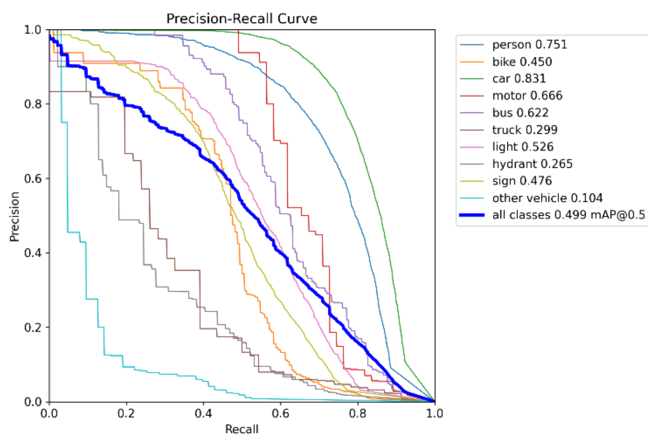

Fig. 3. Per-class precision–recall (PR) curves for the original 10-class dataset.

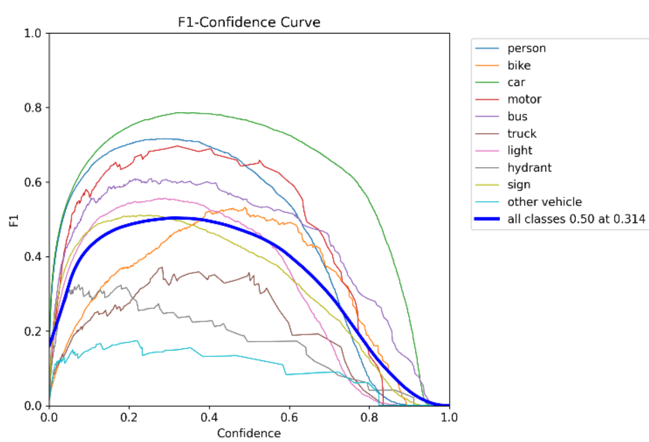

Fig. 4. F1-Confidence curves for the original 10-class dataset.

Based on this quantitative evidence, three poorly performing classes were removed, resulting in a refined 7-class dataset. This decision was driven by systematic performance analysis rather than heuristic simplification. Fig. 5 shows the nano-sized model achieves approximately a 13% increase in mAP@0.5, and Fig. 6 shows approximately a 13% of improvement under various thresholds.

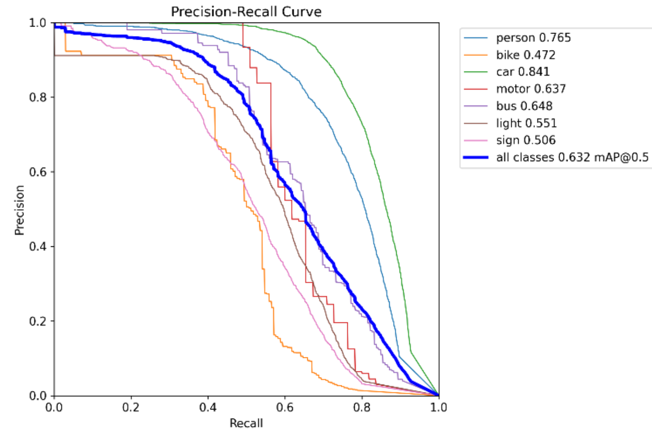

Fig. 5. Per-class precision–recall (PR) curves for the original 7-class dataset.

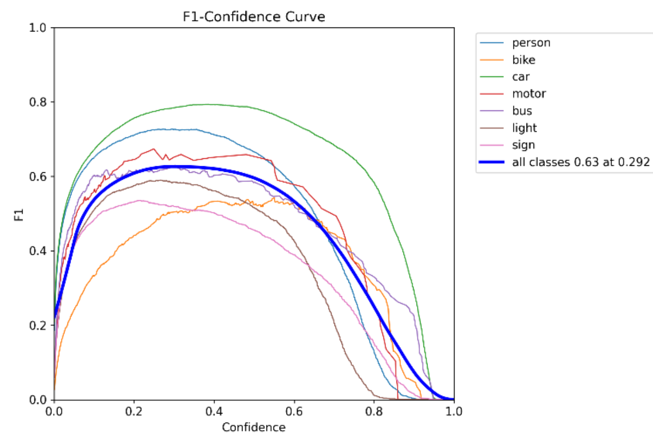

Fig. 6. F1-Confidence curves for the original 7-class dataset.

To assess the impact of preprocessing and model scaling on detection performance, we trained YOLOv11 models of nano, small, medium, and large sizes on the 7-class dataset with and without CLAHE preprocessing. Table III summarizes the precision, recall, mAP50, and mAP50–95 results for the four experimental conditions:

1. 7 classes (no CLAHE)
2. 7 classes + CLAHE
3. 10 classes (no CLAHE)
4. 10 classes + CLAHE

Model Comparison: The refined 7-class dataset was trained under identical experimental settings to isolate the effect of class-level refinement. In Table III, dataset curation leads to substantial performance improvements across all model sizes. For example, the nano model's precision increased from 61.7% to 78.6%, while mAP@0.5 improved from 49.3% to 63%. Similar gains were observed for small, medium, and large variants.

TABLE III. MODEL COMPARISONS TABLE

model type	size	Precision	Recall	mAP50	mAP50-95
7 classes	nano	78.583%	53.504%	63.023%	38.156%
	small	79.569%	58.763%	68.174%	42.611%
	medium	76.768%	65.365%	72.498%	47.06%
	large	79.69%	64.372%	72.754%	47.536%
7 classes + CLAHE	nano	76.208%	52.469%	62.206%	37.302%
	small	75.366%	58.056%	66.724%	41.633%
	medium	78.94%	63.396%	72.116%	46.592%
	large	78.918%	65.728%	73.776%	47.86%
10 classes	nano	61.743%	45.572%	49.263%	29.504%
	small	71.089%	47.391%	54.212%	33.898%
	medium	64.59%	54.72%	59.061%	38.182%
	large	68.332%	53.375%	59.53%	33.82%
10 classes + CLAHE	nano	58.461%	47.059%	48.051%	28.412%
	small	62.055%	46.061%	52.402%	32.27%
	medium	67.704%	55.153%	59.912%	38.415%
	large	66.007%	54.864%	59.346%	38.276%

CLAHE Preprocessing: Across nano to large model sizes, applying CLAHE did not consistently improve performance. For example, nano and small models showed slightly lower precision when CLAHE was applied (nano: 78.5% → 76.2%; small: 79.6% → 75.4%). Medium and large models showed minor improvements in mAP50–95, but gains were marginal compared to the effect of dataset curation. This suggests that contrast enhancement via CLAHE may not be beneficial for thermal grayscale images, possibly because the original images already have sufficient thermal contrast for the detector to learn discriminative features.

Model Size Effects: Precision remained similar across nano, small, medium, and large models, indicating that dataset quality influences precision more than model capacity. In contrast, recall and mAP metrics increased with model size (medium ≈ large> small > nano), consistent with the notion that larger models have higher feature extraction capacity to detect more objects, but do not necessarily improve precision if the dataset is already clean and well-curated.

Overall, dataset curation produces more substantial and consistent improvements than CLAHE preprocessing. Removing the three poorly performing classes improved precision by roughly 10%, stabilized class-wise F1-confidence curves, and reduced inter-class confusion, while CLAHE effects were minor or inconsistent. Consequently, the combination of a refined 7-class dataset and moderate model sizes (medium or large) achieves the best balance of accuracy and efficiency for edge deployment.

V. CONCLUSION

This study demonstrates that dataset curation significantly enhances thermal image detection performance. By reducing the original 16-class dataset to a refined 7-class subset, we achieved improved precision, higher mAP, and more stable F1-confidence behavior across all YOLOv11 model sizes. CLAHE preprocessing provided only marginal benefits, suggesting that thermal grayscale images generally contain sufficient contrast for effective object detection.

Future work will investigate expanding the dataset and incorporating data fusion techniques to support inclusion of previously removed classes while maintaining detection stability. Additional thermal images captured using the same sensor will be integrated to improve class representation and reduce variability caused by limited samples. This approach aims to increase data coverage for underrepresented objects without introducing inconsistencies in thermal characteristics. In addition, data fusion methods, such as combining thermal images with visible-spectrum data, will be explored to improve feature separability for challenging classes. These strategies are expected to support more stable learning and extend the proposed framework toward full multi-class detection.

REFERENCES

[1] "Download Free ADAS Dataset v2 - Teledyne FLIR," Flirconservator.com, 2025. https://adas-dataset-v2.flirconservator.com/#fulldownloadsection (accessed Mar. 04, 2025).

[2] "Teledyne FLIR Tau 2 640 x 512 Lensless – LWIR Thermal Imaging Camera Core," *flircameras*. https://www.flircameras.com/product/46640001x-htm/ (accessed Mar. 07, 2026).

[3] E. Dilek and M. Dener, "Computer Vision Applications in Intelligent Transportation Systems: A Survey," Sensors, vol. 23, no. 6, p. 2938, Mar. 2023, doi: https://doi.org/10.3390/s23062938.

[4] A. S. Bhadoriya, V. Vegamoor, and S. Rathinam, "Vehicle Detection and Tracking Using Thermal Cameras in Adverse Visibility Conditions," Sensors, vol. 22, no. 12, p. 4567, Jun. 2022, doi: https://doi.org/10.3390/s22124567.

[5] Ultralytics, "Home," docs.ultralytics.com. https://docs.ultralytics.com/#yolo-a-brief-history

[6] P. Kaur, B. S. Khehra, and A. P. S. Pharwaha, "Color Image Enhancement based on Gamma Encoding and Histogram Equalization," Materials Today: Proceedings, vol. 46, pp. 4025–4030, 2021, doi: https://doi.org/10.1016/j.matpr.2021.02.543.

[7] J. H. Chowdhury, Q. Liu, and Sheela Ramanna, "Simple Histogram Equalization Technique Improves Performance of VGG Models on Facial Emotion Recognition Datasets," Algorithms, vol. 17, no. 6, pp. 238–238, Jun. 2024, doi: https://doi.org/10.3390/a17060238.

[8] C. Liu, X. Sui, X. Kuang, Y. Liu, G. Gu, and Q. Chen, "Adaptive Contrast Enhancement for Infrared Images Based on the Neighborhood Conditional Histogram," Remote Sensing, vol. 11, no. 11, p. 1381, Jun. 2019, doi: https://doi.org/10.3390/rs11111381.

[9] T. Wang, G. T. Kim, M. Kim, and J. Jang, "Contrast Enhancement-Based Preprocessing Process to Improve Deep Learning Object Task Performance and Results," Applied Sciences, vol. 13, no. 19, p. 10760, Jan. 2023, doi: https://doi.org/10.3390/app131910760.

[10] E. Heredia-Aguado, J. J. Cabrera, L. M. Jiménez, D. Valiente, and A. Gil, "Static Early Fusion Techniques for Visible and Thermal Images to Enhance Convolutional Neural Network Detection: A Performance Analysis," Remote Sensing, vol. 17, no. 6, p. 1060, Mar. 2025, doi: https://doi.org/10.3390/rs17061060.

[11] A. A. Murat and M. S. Kiran, "A comprehensive review on YOLO versions for object detection," Engineering Science and Technology, an International Journal, vol. 70, p. 102161, Aug. 2025, doi: https://doi.org/10.1016/j.jestch.2025.102161.

[12] B. Chen, T. Fang and J. Saniie, "On-Road Thermal Image Recognition by Efficient Edge Computing," 2025 IEEE International Conference on Electro Information Technology (eIT), Valparaiso, IN, USA, 2025, pp. 338-343, doi: 10.1109/eIT64391.2025.11103682.