University of Victoria

Department of Electrical and Computer Engineering

Infrared Image Recognition and Iterative Supervision (IIRIS)

Final Report

ECE 499 – Design Project II

August 2026

Group Information

S. No.	Name	V Number
1	John Soliman	V01028377
2	Yunu Choe	V01008760

Supervisor: Dr. Phalguni Mukhopadhyaya, P.Eng.
Department of Civil Engineering, University of Victoria

Acknowledgment

The authors would like to express their sincere gratitude to their supervisor, Dr. Phalguni Mukhopadhyaya of the Department of Civil Engineering, for his continuous guidance, domain expertise in building envelope science, and support throughout the duration of this project. The authors are especially grateful to Dr. Milad Mahmoodzadeh (Ph.D., Civil Engineering, University of Victoria) for contributing his expert knowledge of building envelope thermal performance to the definition and verification of the ground-truth dataset annotations.

The authors thank the Department of Civil Engineering for providing access to the FLIR A65 thermal camera and the thermographic survey imagery of campus buildings used to construct the dataset.

The authors also wish to acknowledge the technical staff and teaching assistants of the Department of Electrical and Computer Engineering for their assistance, and their families and friends for their patience and encouragement throughout the project.

Contents

Executive Summary

A substantial portion of the energy consumed by residential and commercial buildings is lost through defects in the building envelope, including degraded insulation, thermal bridges, and unintended air leakage. Infrared thermography (IRT) is the established non-destructive method for locating these defects; however, manual interpretation of thermograms is slow, subjective, and dependent on the availability of experienced inspectors. As jurisdictions such as British Columbia pursue aggressive building retrofit and emissions-reduction targets, there is a clear need to standardize and accelerate thermographic inspection.

This project, Infrared Image Recognition and Iterative Supervision (IIRIS), developed an automated framework that detects and segments thermal anomalies in building envelope thermograms using a deep learning approach. A custom dataset of infrared images of University of Victoria campus buildings was captured with a research-grade thermal camera and annotated at the pixel level by a domain expert across six structural and anomaly classes. Because expert-annotated thermal data is scarce, the dataset was expanded tenfold through a geometric and radiometric augmentation pipeline. A lightweight instance segmentation network (YOLOv11n-seg) was then trained to delineate the exact boundaries of each defect. To make the system usable by non-programmers, a companion web application was built that performs inference on uploaded thermograms, provides interactive confidence controls, and includes an annotation canvas through which an expert can correct the model's predictions; these corrections feed back into the training set, forming an iterative human-in-the-loop supervision cycle that gives the project its name.

The trained model achieved a mean average precision (mAP@50) of 89.84% and a recall of 95.79% on the validation set, with a 5-fold cross-validation study confirming the stability of these results and an ablation study confirming that the augmentation pipeline was responsible for a 21-point improvement in mAP@50. The high recall is particularly significant for non-destructive testing, where a missed defect corresponds directly to unaddressed energy loss. The results demonstrate that a lightweight, edge-deployable network, paired with an iterative supervision workflow, can standardize and accelerate thermal anomaly detection in building envelopes, providing energy auditors, municipalities, and building owners with a practical, quantitative diagnostic tool.

I Introduction

Buildings account for a significant share of national energy consumption, and a substantial portion of that energy is lost through the building envelope: the walls, windows, roofs, and joints that separate the conditioned interior from the exterior environment. Latent energy loss is primarily driven by compromised structural integrity, degrading insulation, thermal bridges, and unintended air infiltration. Detecting these failure modes is a prerequisite for accurate energy loss calculation and targeted retrofitting, which in turn supports climate policy goals and reduces heating costs for occupants.

Infrared thermography (IRT) is an established non-destructive testing (NDT) methodology for assessing thermal performance. By capturing emitted infrared radiation, IRT visualizes heat transfer anomalies that are generally imperceptible to the naked eye [1], [2]. While radiometrically effective, the manual interpretation of thermograms is subjective, dependent on inspector expertise, and difficult to scale across large infrastructure portfolios. Two inspectors examining the same thermogram may reach different conclusions, and a municipality with hundreds of buildings cannot practically rely on frame-by-frame expert review.

Problem definition: there is no accessible, automated tool that can localize and classify building envelope thermal anomalies in infrared imagery with expert-level consistency. **Goal:** to design, train, and validate a deep learning pipeline that performs instance segmentation of thermal anomalies, and to deliver it through a web application that supports an iterative, human-in-the-loop supervision cycle. **Scope:** the project covers dataset curation and expert annotation, data augmentation, model training and statistical validation, and the development of a working web-based prototype. Radiometric quantification (e.g., U-value computation) and generalization beyond the local building stock are identified as future work.

This project chose the problem deliberately for its social impact. Improving the speed and consistency of building energy audits directly serves the public interest: it reduces wasted energy, lowers greenhouse gas emissions, and helps building owners prioritize retrofit spending. This aligns with the Engineers and Geoscientists BC (EGBC) Code of Ethics (reproduced in the Appendix), most directly with the duty to *hold paramount the safety, health, and welfare of the public and the protection of the environment* (Principle 1). The project also engages Principle 2 (undertake work only within one's training and competence) — pixel-level ground truth was therefore defined by a building science domain expert rather than by the student team — and Principle 3 (provide accurate and honest reports), which motivated the transparent reporting of dataset limitations and potential data leakage discussed in Sections VI and VIII.

The potential user base includes professional energy auditors and thermographers seeking a consistent second opinion, municipal facility and asset managers responsible for large building portfolios, building envelope consultants, utility demand-side management programs, and researchers who require a labelled thermal dataset and an annotation tool for further model development.

II Objectives

The objectives of the project, each contributing a concrete step toward the overall goal, were:

- To curate a domain-specific dataset of building envelope infrared thermograms, annotated at the pixel level by a domain expert across six structural and anomaly classes.
- To design a data augmentation pipeline that mitigates the scarcity of expert-annotated thermal imagery and improves model generalization.
- To train a lightweight instance segmentation network (YOLOv11n-seg) capable of delineating the geometric boundaries of thermal anomalies.
- To demonstrate detection performance suitable for non-destructive testing, targeting a mean average precision (mAP@50) of at least 85% and a recall of at least 90% on validation data.
- To verify the statistical robustness of the results through an augmentation ablation study and 5-fold cross-validation.
- To design and prototype a web application providing real-time inference, adjustable confidence thresholds, and class visibility controls for non-programmer end users.
- To demonstrate an iterative, human-in-the-loop supervision workflow in which expert corrections to model predictions are exported as new training annotations.

III Design Specifications

The system comprises three major subsystems: the data acquisition and annotation subsystem, the deep learning subsystem, and the web application subsystem. The technical ratings and boundary conditions of each are specified below, with the reasoning behind each choice.

Data Acquisition Subsystem

Thermal imagery was captured with a FLIR A65 research thermal camera equipped with a 640 × 512 focal plane array (FPA) microbolometer with a stated accuracy of ±5 °C. This camera was selected because it was available through the Department of Civil Engineering, has an established record in prior UVic building envelope studies [8]–[11], and its resolution is representative of professional-grade survey equipment, meaning the trained model's input domain matches imagery a practicing auditor would produce. Surveys were conducted on University of Victoria campus building facades under varying thermal loading conditions to capture a range of interior–exterior temperature differentials. The base dataset comprises 24 high-resolution thermograms carrying 357 expert-defined polygon annotations across six classes: (1) Thermal Bridge, (2) Missing Insulation, (3) Window (a non-anomaly structural reference class), (4) Air Leakage, (5) Spandrel, and (6) Vent. The Window class was included deliberately so the network learns to distinguish expected thermal signatures of structural elements from genuine defects.

Deep Learning Subsystem

The segmentation model is required to (a) produce instance-level polygon masks rather than only bounding boxes, since defect area and shape carry diagnostic meaning; (b) run inference at interactive speeds on commodity hardware to support the web application; and (c) remain small enough for future edge deployment on embedded survey equipment. The YOLOv11n-seg (nano) architecture satisfies all three constraints. Input imagery is standardized to 640 × 640 pixels, matching the native sensor width and the architecture's standard input resolution. Training was performed on an NVIDIA RTX 5090 GPU. The full training configuration is specified in Table 1; these hyperparameters are reported exactly to ensure the training run is reproducible.

Table 1. Training configuration of the YOLOv11n-seg model.

Hyperparameter	Value	Description
Input resolution	640 × 640	Standardized uniform training dimensions
Batch size	32	Images processed per gradient update
Epochs	100	Maximum training epochs
Patience	20	Early-stopping patience (best model retained)
lr0 / lrf	0.01 / 0.01	Initial and final learning rate scalars
Momentum	0.937	Gradient descent momentum
Weight decay	0.0005	L2 regularization penalty
Warmup epochs	3.0	Warmup phase for momentum/bias learning rates
Box / Cls / DFL loss weights	7.5 / 0.5 / 1.5	Loss coefficient weights
Optimizer	Auto (AdamW)	Selected automatically for task scale
overlap_mask	False	Permits physically overlapping instances

The overlap_mask setting is a deliberate boundary condition: thermal defects physically overlap structural elements (for example, an air leak at a window frame), so masks must be allowed to coexist rather than being merged.

Web Application Subsystem

The delivery platform is a Flask (Python) web application. It is specified to: accept uploaded thermal images in standard formats; execute model inference and render predicted polygons over the source image; provide a confidence threshold control continuously adjustable between 0 and 1, with per-class visibility toggles; and provide an annotation canvas on which an expert can create or correct polygon boundaries. Corrected annotations must export losslessly to the YOLO segmentation label format so they can be merged directly into the training set. This closes the iterative supervision loop that is central to the design. Flask was chosen over heavier frameworks because the interface requirements are modest and the deployment target is a single containerized service.

Where citations appear in this report they use bracketed reference numbers [1], [2], and all referenced material has been paraphrased. Regarding standards and regulations relevant to EGBC professional practice: thermographic building surveys are governed by recognized procedures (e.g., ASTM C1060, Standard Practice for Thermographic Inspection of Insulation Installations in Envelope Cavities of Frame Buildings), and the acquisition protocol followed the survey practices established in prior peer-reviewed UVic campus studies [8]–[11]. The system is a diagnostic decision aid; final assessments remain the responsibility of a qualified professional, consistent with the EGBC principle of practicing only within one's area of competence.

IV Literature Survey

The application of computer vision to infrared thermography for building diagnostics is an active research domain. Early foundational work established IRT as a building diagnostic methodology and emphasized the challenges that still motivate automation today: environmental radiative noise, reflective surfaces, and the need for standardized inspection protocols. Balaras and Argiriou [1] surveyed infrared thermography for building diagnostics, while Fox et al. [2] catalogued thermography methodologies for detecting energy-related building defects and highlighted the dependence of manual approaches on inspector skill.

A first generation of automation efforts applied deterministic image processing. Asdrubali et al. [3] detected thermal bridges by enhancing thermograms with sampling Kantorovich operators prior to histogram-based thresholding. De Filippo et al. [4] proposed a computer vision algorithm that identifies spatial temperature gradients in reinforced concrete, demonstrating efficacy for water leakage and debonding detection. These methods are computationally cheap and interpretable, but they are sensitive to varying environmental conditions, reflective surfaces, and complex thermal noise, and they generally cannot classify the type of defect they detect.

The transition to deep learning has improved robustness. Mirzabeigi et al. [5] used drone-based data collection to demonstrate deep learning semantic segmentation for in-situ anomaly detection, and Dehghani et al. [6] reported pixel-level semantic segmentation of thermal anomalies in building facades. Semantic segmentation, however, labels pixels by class without separating individual defect instances, which limits per-defect quantification (count, area, and location of each anomaly). Instance segmentation architectures such as YOLOv11 [7] combine detection and mask prediction in a single efficient network, enabling per-instance diagnostics at real-time speeds.

Three solution families were therefore compared when selecting the approach. Classical thresholding pipelines [3], [4] offer low cost and no training data requirement, but their brittleness under environmental variation makes them unsuitable for a general-purpose tool. Two-stage instance segmentation networks (e.g., Mask R-CNN) offer strong accuracy but carry a parameter count and inference latency that conflict with the project's edge deployment goal and interactive web interface. Single-stage instance segmentation with a nano-scale YOLOv11 variant [7] offers near-real-time inference on commodity hardware, native polygon mask output, and a mature open-source training ecosystem — at zero licensing cost. Given the project's constraints (limited expert-annotated data, interactive latency requirements, and a small computational budget), YOLOv11n-seg was selected.

The dataset annotation methodology is grounded in the building envelope literature produced at UVic under the project supervisor. Mahmoodzadeh et al. evaluated air leakage patterns with infrared thermography [8], assessed the thermal performance of campus building envelopes [9], and established external IRT methods for determining envelope thermal performance in wood-framed construction [10], [11]. These studies define the defect taxonomy and survey practices that the six-class annotation ontology of this project encodes. Consistent with the EGBC Code of Ethics, the survey distinguishes established findings from the assumptions of this project, and applicable inspection standards (e.g., ASTM C1060) are cited where relevant in Section III.

V Team Duties & Project Planning

The project was executed by a two-member team under the guidance of Dr. Phalguni Mukhopadhyaya, with domain expert support from Dr. Milad Mahmoodzadeh for annotation ground truth. Duties were divided so that each member owned complete, testable deliverables while remaining jointly responsible for integration and the final report.

Table 2. Division of team duties and deliverables.

Member	Deliverables	Significance
John Soliman	Dataset curation and augmentation pipeline; YOLOv11n-seg training and hyperparameter configuration; Flask web application (inference interface, annotation canvas, YOLO label export); deployment and containerization	Core detection capability and the delivery platform through which end users access it
Yunu Choe	Literature survey; annotation verification and dataset quality assurance; validation experiments (ablation study, 5-fold cross-validation); results analysis and report preparation	Statistical credibility of the reported metrics and the scholarly grounding of the design choices

Project planning followed a milestone-based schedule: (1) literature review and data acquisition; (2) expert annotation and dataset construction; (3) augmentation pipeline and baseline training; (4) hyperparameter refinement and full training; (5) validation studies (ablation and cross-validation); (6) web application development and integration; and (7) final reporting. Milestones (3) through (6) were deliberately overlapped, since the web application's annotation canvas was itself used to refine the dataset — the iterative supervision loop meant that tooling and data quality improved together.

Several challenges were encountered and resolved. First, expert-annotated thermal data proved far scarcer than natural-image data, which redirected effort into the augmentation pipeline and its validation; the ablation study in Section VIII quantifies why this was necessary. Second, physically overlapping defect instances (e.g., an air leak bordering a window) initially produced corrupted training masks; this was traced to mask merging behaviour and resolved with the overlap_mask=False configuration. Third, the random split of augmented data was identified as a potential source of data leakage between training and validation sets; rather than concealing this, the team documented it and added 5-fold cross-validation to bound its effect. Finally, coordinating with domain experts required translating between civil engineering terminology and machine learning label ontologies, which the team addressed through joint annotation review sessions.

VI Design Methodology & Analysis

Dataset Acquisition and Expert Annotation

A custom dataset of infrared thermograms was curated to train the segmentation model. The original dataset comprises 24 high-resolution thermal images captured with the FLIR A65 camera specified in Section III, covering various building facades under different thermal loading conditions. Every image was manually annotated with polygon boundaries to ensure exact spatial delineation of defects. Annotations were rigorously defined by Dr. Milad Mahmoodzadeh, a domain expert with extensive publications in building envelope thermal performance evaluation [8]–[11]. A total of 357 annotations were categorized across the six classes defined in Section III. Grounding the labels in a single expert's published diagnostic framework distinguishes documented fact from assumption in the training signal — a methodological safeguard aligned with the EGBC Code of Ethics requirement to distinguish facts from assumptions and opinions.

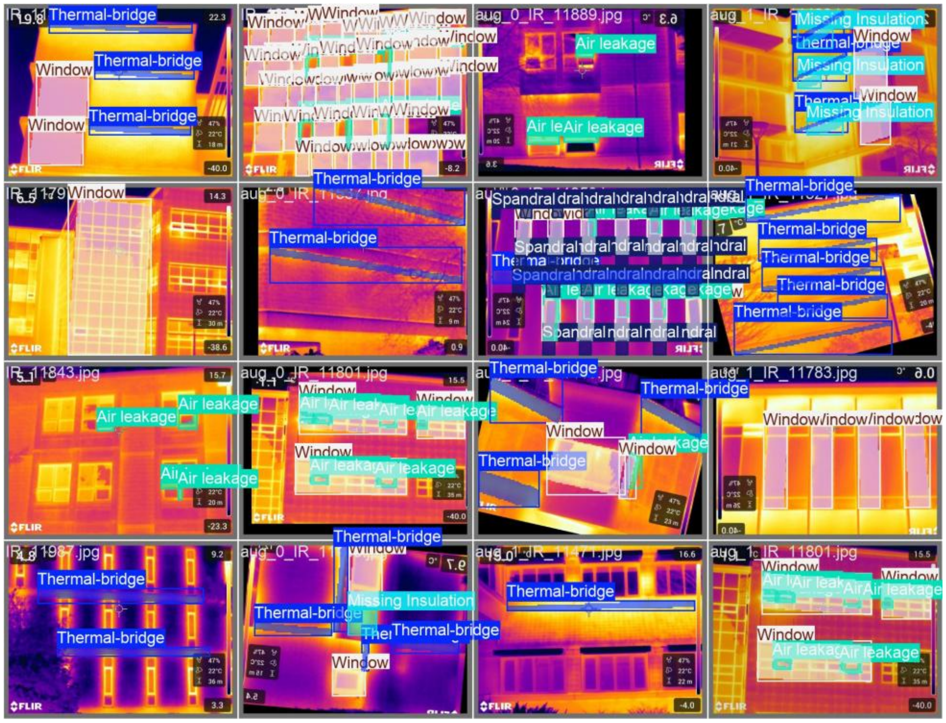

Fig. 1. A batch of expert-annotated training thermograms with class labels and polygon masks overlaid.

Data Augmentation

To mitigate overfitting and improve generalization across varying thermal contrasts, an augmentation pipeline was implemented using the Albumentations library, expanding the 24 base images by a factor of 10 to 240 images. The transformation pipeline applied the following geometric and radiometric distortions:

- Horizontal flipping (p = 0.5), exploiting the lateral symmetry of facade imagery.
- Random brightness and contrast (p = 0.2), simulating varying sensor gain and thermal scaling.
- Rotation (limit = 15°, p = 0.5), accounting for varying camera pitch during handheld surveys.
- Affine transforms (translation 5%, scaling 0.9–1.1×, p = 0.3), simulating variation in standoff distance and framing.

The dataset was partitioned 85% / 15% into training and validation sets via a pseudo-random image-level shuffle (seed = 42) applied after augmentation. This maximizes training diversity but introduces a potential for data leakage, since augmented variants of one source image may appear on both sides of the split. This is stated explicitly as an assumption-versus-fact boundary of the main experiment, and Section VIII presents the cross-validation protocol used to bound its effect.

Network Architecture and Alternatives Considered

The YOLOv11n-seg model was selected for inference. YOLOv11 utilizes Cross Stage Partial (CSP) networks and a Path Aggregation Network (PANet) for multiscale feature fusion — critical here because thermal defects range in scale from thin linear thermal bridges to wall-sized regions of missing insulation. The segmentation head is trained with a composite loss combining box regression loss, classification (focal) loss, and a binary cross-entropy mask loss, with the coefficient weights listed in Table 1.

Two design alternatives were explored before converging on this approach. A deterministic thresholding pipeline in the style of [3] was prototyped conceptually but rejected: campus thermograms exhibit reflective glazing and sky backgrounds that defeat global histogram methods, and thresholding cannot assign defect classes. A semantic segmentation approach following [5], [6] was also considered; it was rejected because merging all instances of a class into one pixel mask discards the per-defect count and geometry needed for audit reporting. Instance segmentation preserves both, and the analysis of the loss composition above shows it can be trained end-to-end without a separate region proposal stage, keeping inference latency compatible with the interactive web interface.

All figures in this report follow the labelling conventions requested in the course template: axes are labelled, legends identify multiple series, and each figure carries an explanatory caption.

VII Design & Prototype

The final working prototype consists of the trained YOLOv11n-seg model embedded in a Flask web application, together forming the complete IIRIS system. The application permits inference on raw thermal images with dynamic controls for confidence thresholding and per-class visibility, and an integrated annotation canvas through which experts correct inaccurate polygon predictions. Corrections are parsed and exported in the YOLO annotation format, establishing the iterative feedback loop for future retraining that gives the project its name.

Fig. 2. The IIRIS web-based detection interface performing inference on a building thermogram, with confidence threshold and per-class visibility controls.

Dynamic confidence thresholding mitigates the risk of overlooking subtle anomalies, a limitation often encountered with static thresholds: an auditor can lower the threshold to sweep for faint signatures, then raise it to isolate high-confidence defects.

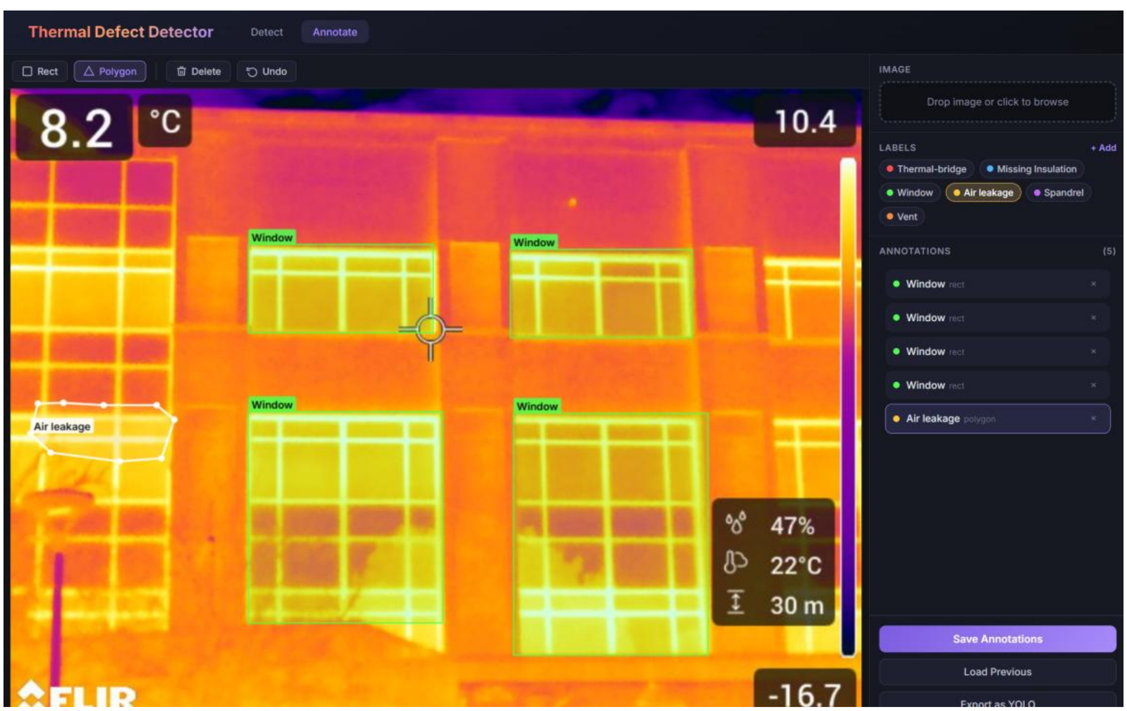

Fig. 3. The active learning annotation canvas used for expert polygon correction; corrected polygons export directly to YOLO training format.

Each project objective is addressed by the final design as follows. The dataset objective was met with 24 expert-annotated base images carrying 357 annotations, expanded to 240 images by the augmentation objective's pipeline. The training objective was met with the configuration in Table 1. The performance objective (mAP@50 ≥ 85%, recall ≥ 90%) was exceeded, with 89.84% and 95.79% respectively (Section VIII). The robustness objective was met through the ablation and cross-validation studies. The web application and iterative supervision objectives were met by the prototype shown in Figs. 2 and 3.

The boundary conditions of the prototype are stated explicitly. The model expects single-frame thermal imagery visually comparable to FLIR palette output at approximately 640 × 512 resolution; radiometric raw data is not required for detection. The learned thermal signatures are localized to the construction materials and Pacific Northwest climate of the UVic campus dataset; generalization to arid or severe freezing environments is unverified. The class ontology contains no exterior door class, so doors are frequently misclassified or bisected as windows. Inference latency on the development GPU supports interactive use; CPU-only deployment increases latency but remains functional for single-image auditing. Within these bounds the prototype meets all stated objectives; no objective was descoped.

VIII Testing & Validation

Test Plan

The system was tested at the subsystem level according to the plans in Table 3. Each plan specifies its operating conditions so results are repeatable from any geography, provided the same dataset, software versions, and random seeds are used. All model evaluations were executed on an NVIDIA RTX 5090 GPU with the training configuration of Table 1; the web application was tested against the same trained weights.

Table 3. Test plans, operating conditions, and pass criteria.

Test	Operating conditions	Expected outcome
T1: Validation metrics	Held-out 15% validation split; IoU threshold 0.50; seed 42	mAP@50 ≥ 85%, recall ≥ 90%
T2: Augmentation ablation	Retrain on raw 24-image dataset, identical hyperparameters	Quantify augmentation benefit; augmented model superior
T3: 5-fold cross-validation	5 folds over all 240 images, 100 epochs per fold (500 total)	Low variance across folds; mean consistent with T1
T4: Qualitative comparison	Side-by-side expert annotation vs. model prediction on validation images	Predicted polygons consistent with expert ground truth

T5: Web application	Trained weights served via Flask; threshold sweep 0–1; polygon edit and export	Correct rendering, threshold response, lossless YOLO export

Evaluation Metrics

Model performance was evaluated using standard detection and segmentation metrics, where TP, FP, and FN denote true positives, false positives, and false negatives respectively:

$$P = TP / (TP + FP) \tag{1}$$

$$R = TP / (TP + FN) \tag{2}$$

Mean average precision (mAP@50) averages the per-class area under the precision–recall curve at an intersection-over-union (IoU) threshold of 0.50:

$$mAP@50 = (1/N)\ \Sigma\ AP_i\ ,\ \ IoU \geq 0.50 \tag{3}$$

T1: Quantitative Metrics and Training Dynamics

Validation metrics on the segmentation masks were: mAP@50 = 89.84%, mAP@50-95 = 66.74%, precision = 86.95%, and recall = 95.79%. Both pass criteria of test T1 were exceeded. The high recall indicates a low false negative rate — the most important property for non-destructive testing, where a missed defect corresponds directly to unaddressed energy loss. Training converged stably: box loss decreased from 1.97 to 0.56 over 100 epochs, and validation curves showed no divergent behaviour, indicating the augmentation strategy prevented overfitting at this network capacity (Fig. 4).

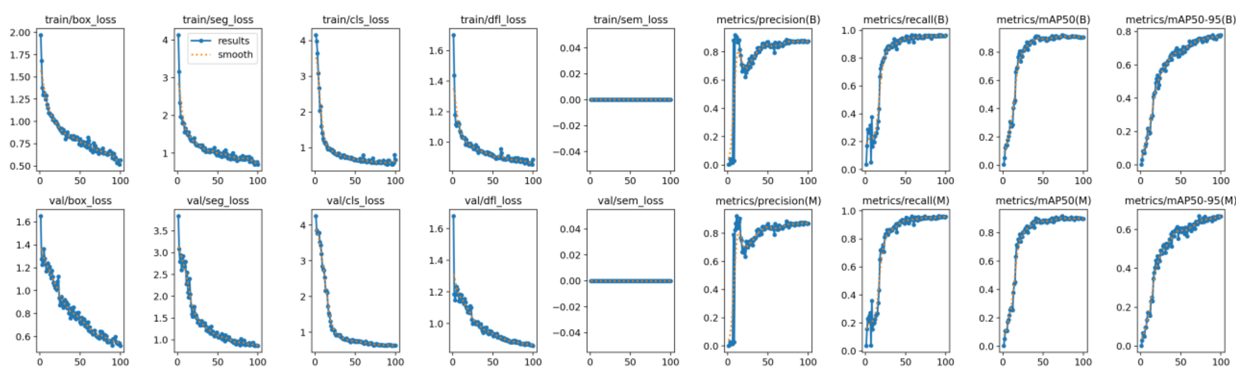

Fig. 4. Training and validation loss dynamics alongside precision and recall convergence over 100 epochs. Axes are labelled per curve; each panel shows one loss or metric component.

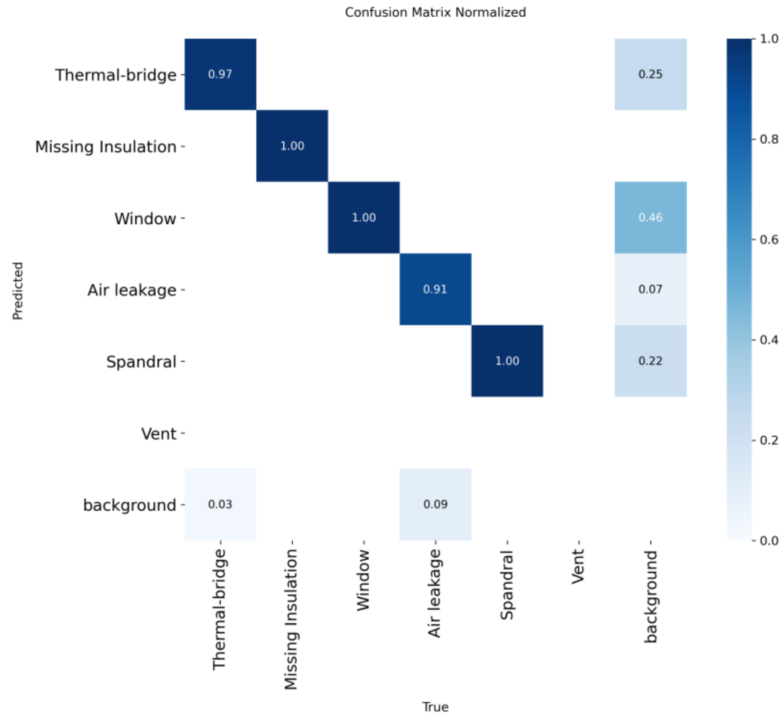

Fig. 5. Normalized confusion matrix detailing class-wise prediction accuracy across the six classes.

The confusion matrix (Fig. 5) shows high true positive rates for prevalent classes (Missing Insulation, Window). Minor cross-class confusion occurs primarily between structurally adjacent features — e.g., thermal bridges at panel boundaries — reflecting genuine physical ambiguity at thermodynamic borders rather than model failure.

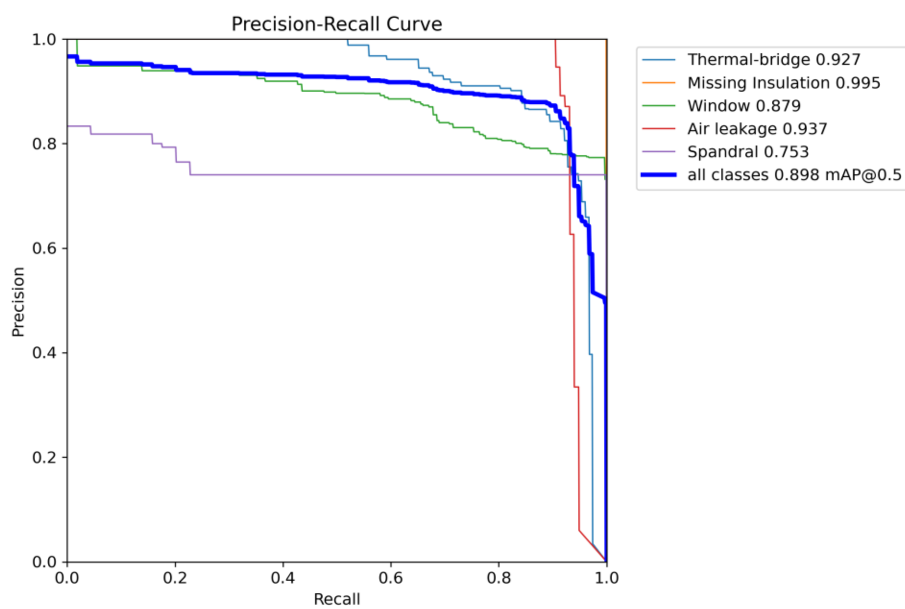

Fig. 6. Mask precision–recall curves per class and combined (0.898 mAP@0.5), with legend identifying each class.

T2: Augmentation Ablation Study

A baseline model was trained strictly on the raw 24-image unaugmented dataset with otherwise identical configuration, to quantify the benefit of the augmentation pipeline. The baseline achieved mAP@50 = 68.8% (a 21.0-point drop) and recall = 84.1% (an 11.7-point drop). The unaugmented network exhibited severe overfitting and domain locality bias, confirming that the variance induced by geometric and radiometric augmentation is necessary for generalizable instance segmentation on small NDT datasets (Fig. 7). Test T2 passed: the augmented model is superior on every metric.

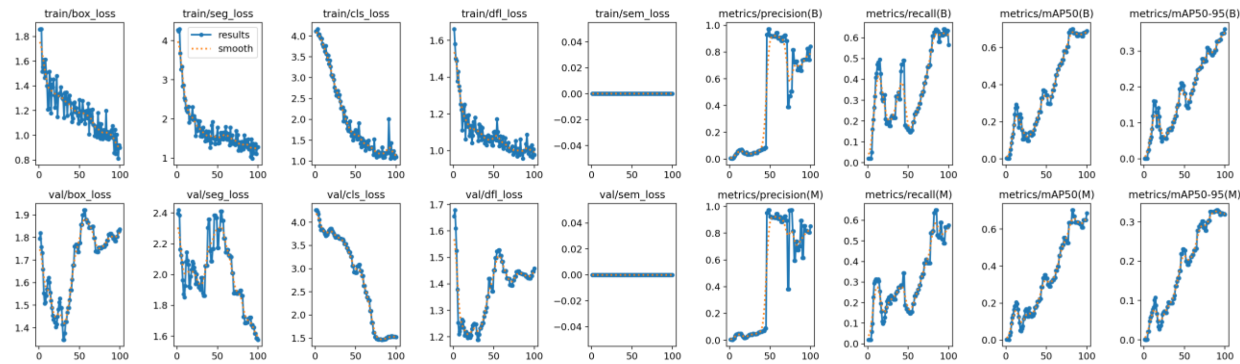

Fig. 7. Training dynamics of the unaugmented 24-image baseline, exhibiting unstable, divergent convergence relative to the augmented model of Fig. 4.

T3: 5-Fold Cross-Validation

To dispel the possibility that the T1 metrics relied on a fortunate 15% validation shuffle — and to bound the data leakage risk noted in Section VI — a programmatic 5-fold cross-validation was run over all 240 augmented images, training 100 epochs per fold (500 epochs total). The protocol yielded a mean mAP@50 of 97.99% ± 0.52% and mean recall of 95.68% ± 1.39%. The small standard deviation across folds demonstrates that performance is stable and not an artifact of a single split. Test T3 passed. It is noted transparently that the fold means exceed the single-split T1 result, which is consistent with augmented variants of source images appearing across folds; the two protocols therefore bracket the model's true generalization performance, and the more conservative T1 figures are quoted as the headline results throughout this report.

T4: Qualitative Validation

A qualitative comparison was conducted between expert ground-truth annotations and model predictions on validation images (Figs. 8–11). The model successfully segments complex anomaly geometries. In some instances, the prediction produces more tightly constrained boundaries around diffuse thermal gradients than the human annotation — a consequence of cross-entropy optimization rather than human perceptual grouping. The overlap_mask=False configuration enabled successful separation of overlapping classes, such as localized thermal bridges adjacent to windows. Test T4 passed.

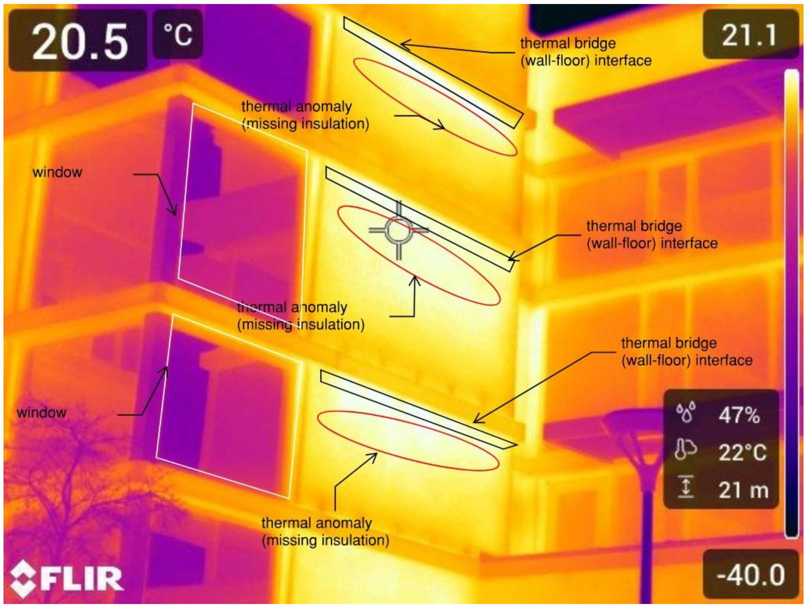

Fig. 8. Expert ground-truth annotation (image IR_11489).

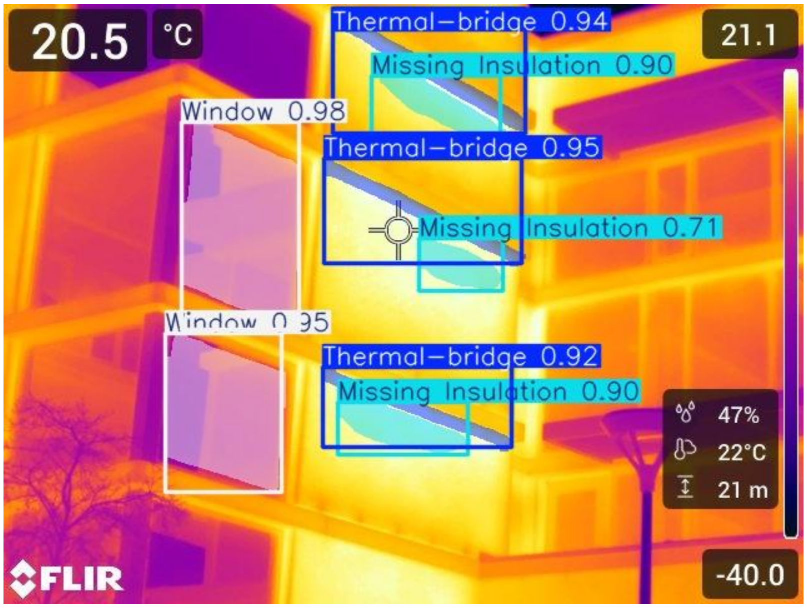

Fig. 9. YOLOv11n-seg prediction (image IR_11489).

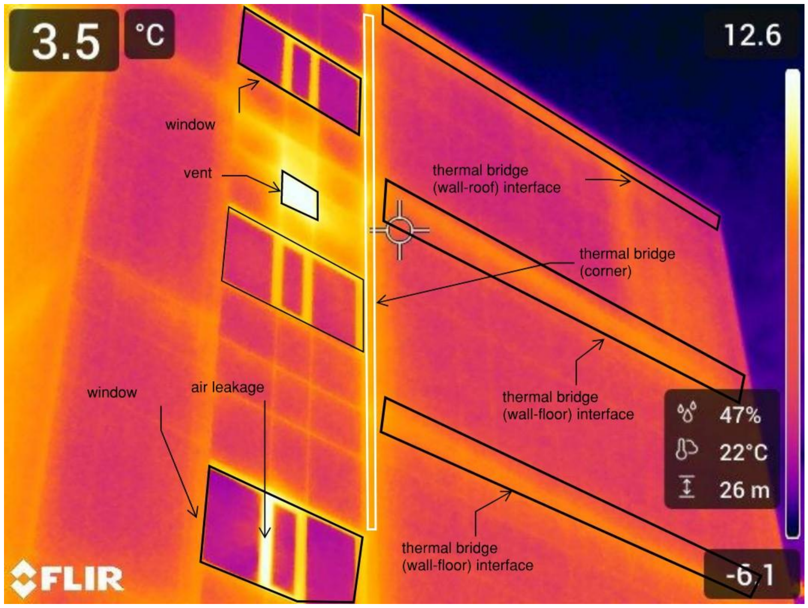

Fig. 10. Expert ground-truth annotation (image IR_12079).

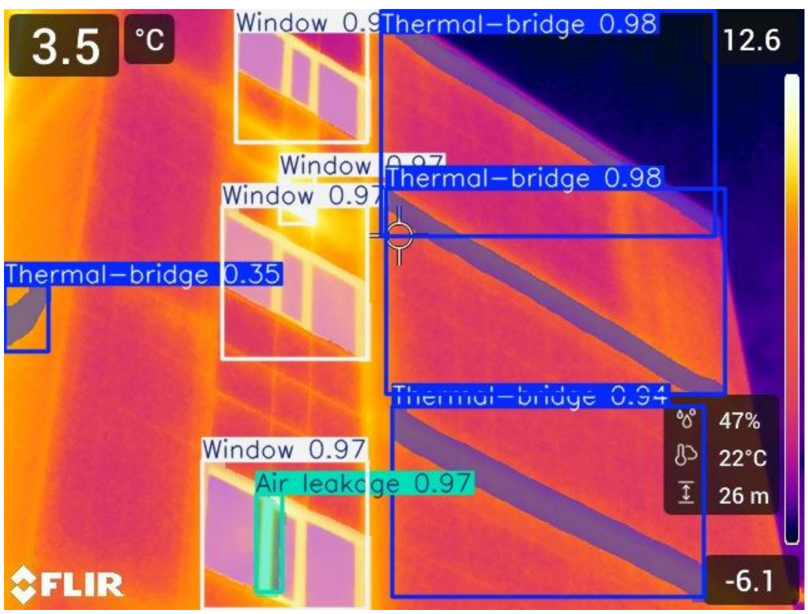

Fig. 11. YOLOv11n-seg prediction (image IR_12079).

T5: Web Application Validation

The prototype application was exercised with the trained weights: uploaded thermograms rendered predicted polygons correctly; sweeping the confidence threshold from 0 to 1 continuously added and removed detections as expected; per-class visibility toggles behaved correctly; and polygons corrected on the annotation canvas exported losslessly to the YOLO segmentation format and were successfully merged back into the training set. Test T5 passed. All five test plans passed; no test was marked as failed.

Consistent with the EGBC Code of Ethics, results are reported with their limitations: the leakage-affected split, the geographic locality of the dataset, and the door misclassification behaviour are disclosed rather than omitted, and the conservative metric set is quoted as the headline performance.

IX Cost Analysis

Table 4 summarizes direct project costs. Direct outlays were minimal because the major capital items were provided in-kind: the thermal camera and survey imagery through the Department of Civil Engineering, and the GPU workstation from existing team hardware. In-kind items are listed at estimated market value to represent the true replication cost of the project. All figures are in Canadian dollars.

Table 4. Project cost summary (CAD).

Item	Cost	Notes
FLIR A65 thermal camera	$0 ($10,000 in-kind)	Provided by Dept. of Civil Engineering
GPU workstation (RTX 5090)	$0 ($3,500 in-kind)	Existing team hardware
Software licences	$0	Entire stack open-source (Python, PyTorch, Ultralytics, Flask, Albumentations)
Hosting and domain	$60	Web application deployment
Total direct cost	$60	Excluding in-kind equipment

Tentative pricing and return on investment: a professional thermographic audit of a mid-size commercial building typically requires several hours of on-site capture plus a comparable period of manual thermogram interpretation. IIRIS automates the interpretation stage. Offered as a software service to auditing firms at an illustrative $200 per building portfolio scan, and conservatively assuming the tool halves interpretation labour (saving roughly two hours of a thermographer's time, valued at approximately $150–200), the tool pays for itself on the first building. With a direct outlay of only about $60 — and roughly $13,500 in equipment if the in-kind camera and workstation had to be purchased outright — even a small auditing practice would recover full replication cost within its first portfolio of scans, and for a municipal portfolio of several hundred buildings the return is achievable within a single heating season.

Cost reduction opportunities include: replacing the research-grade FLIR A65 with lower-cost handheld or smartphone-attached thermal cameras once the model is fine-tuned on their imagery; using the nano model's small footprint to run on inexpensive edge hardware instead of a GPU workstation; expanding the dataset through the built-in annotation canvas rather than commissioning new expert annotation campaigns (the iterative supervision loop converts routine use into free training data); and training future model iterations via transfer learning from the current weights, which reduces GPU hours substantially.

X Conclusion & Recommendations

This project demonstrated the feasibility of using a lightweight instance segmentation network for quantitative diagnostic assessment of building thermograms. All project objectives were accomplished: a six-class expert-annotated thermal dataset was curated and expanded tenfold through augmentation; a YOLOv11n-seg model was trained to a validation mAP@50 of 89.84% and recall of 95.79%, exceeding the 85% and 90% targets; the augmentation ablation and 5-fold cross-validation studies confirmed that these results are statistically stable rather than artifacts of a single split; and a working web application prototype delivers inference, adjustable thresholds, and a human-in-the-loop annotation canvas that exports corrections directly into the training format — the iterative supervision cycle at the core of IIRIS.

The limitations of the prototype are equally clear. The learned thermal signatures are localized to UVic campus construction and the Pacific Northwest climate, and generalization to other climates and building archetypes is unverified. The class ontology omits common structural elements — most visibly exterior doors, which the model tends to misclassify as windows. The augmented-then-split methodology admits data leakage between training and validation sets, which is why the conservative single-split figures are quoted as headline results.

The following recommendations are made for future work:

1. Expand the dataset to diverse building archetypes, construction materials, and seasonal weather conditions, using the annotation canvas to capture expert corrections during routine use.
2. Broaden the class ontology to include doors, pillars, and roof seams, eliminating the dominant current false positive mode.
3. Adopt a source-image-level (grouped) train/validation split before augmentation to eliminate data leakage in future evaluations.
4. Investigate real-time video inference from uncrewed aerial vehicle (UAV) thermal cameras for rapid scanning of large commercial envelopes.
5. Deploy the trained weights on embedded hardware using TensorRT optimization for low-latency, offline field inference.
6. Integrate the detector's polygon output with radiometric analysis tools so that detected regions feed directly into quantitative U-value and energy loss reporting.

References

[1] C. A. Balaras and A. A. Argiriou, “Infrared thermography for building diagnostics,” Energy and Buildings, vol. 34, no. 2, pp. 171–183, 2002.

[2] M. Fox, D. Coley, S. Goodhew, and P. De Wilde, “Thermography methodologies for detecting energy related building defects,” Renewable and Sustainable Energy Reviews, vol. 40, pp. 296–310, 2014.

[3] F. Asdrubali, G. Baldinelli, F. Bianchi, D. Costarelli, A. Rotili, M. Seracini, and G. Vinti, “Detection of thermal bridges from thermographic images by means of image processing approximation algorithms,” Applied Mathematics and Computation, vol. 317, pp. 160–171, 2018.

[4] M. De Filippo, S. Asadiabadi, N. Ko, and H. Sun, “Concept of computer vision based algorithm for detecting thermal anomalies in reinforced concrete structures,” in Proc. 15th Int. Workshop on Advanced Infrared Technology and Applications (AITA), Florence, Italy, 2019.

[5] S. Mirzabeigi, M. Razkenari, and P. Crovella, “Automated thermal anomaly detection through deep learning-based semantic segmentation of building envelope images,” in Computing in Civil Engineering 2023, ASCE, 2024. doi: 10.1061/9780784486115.004.

[6] M. Dehghani et al., “Deep learning based semantic segmentation for thermal anomaly detection in building facades using infrared thermography,” Energy and Buildings, vol. 316, p. 114389, 2024. doi: 10.1016/j.enbuild.2024.114389.

[7] G. Jocher et al., “YOLO by Ultralytics,” GitHub repository, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics. [Accessed: Mar. 2026].

[8] M. Mahmoodzadeh, V. Gretka, S. Wong, T. Froese, and P. Mukhopadhyaya, “Evaluating patterns of building envelope air leakage with infrared thermography,” Energies, vol. 13, no. 14, p. 3545, 2020.

[9] M. Mahmoodzadeh, V. Gretka, A. Blue, D. Adams, B. Dallimore, and P. Mukhopadhyaya, “Evaluating thermal performance of vertical building envelopes: Case studies in a Canadian university campus,” Journal of Building Engineering, vol. 40, p. 102712, 2021.

[10] M. Mahmoodzadeh, V. Gretka, K. Hay, C. Steele, and P. Mukhopadhyaya, “Determining overall heat transfer coefficient (U-value) of wood-framed wall assemblies in Canada using external infrared thermography,” Building and Environment, vol. 199, p. 107897, 2021.

[11] M. Mahmoodzadeh, V. Gretka, I. Lee, and P. Mukhopadhyaya, “Utilizing external infrared thermography to assess thermal performance of wood-framed building envelopes in Canada,” Energy and Buildings, p. 111807, 2021.

Appendix

A. Code of Ethics — Engineers and Geoscientists BC (EGBC)

Engineers and Geoscientists BC registrants must uphold the following ethical principles (paraphrased from the EGBC Code of Ethics; the authoritative text is available at https://www.egbc.ca):

1. Hold paramount the safety, health, and welfare of the public, including the protection of the environment and the promotion of health and safety in the workplace.
2. Undertake and accept responsibility for professional assignments only when qualified by training or experience.
3. Provide an opinion on a professional subject only when it is founded upon adequate knowledge and honest conviction.
4. Act with honesty and integrity toward clients and employers, maintain confidentiality, and avoid or disclose conflicts of interest.
5. Uphold the principle of appropriate and adequate compensation for professional work.
6. Keep informed to maintain competence, strive to advance the body of knowledge of the profession, and provide opportunities for the professional development of associates.
7. Conduct themselves with fairness, courtesy, and good faith toward clients, colleagues, and others; give credit where it is due; and accept, as well as give, honest and fair professional comment.
8. Clearly present to employers and clients the possible consequences if professional decisions or judgments are overruled or disregarded.
9. Report to the appropriate authority any hazardous, illegal, or unethical professional decisions or practices by registrants or others.
10. Extend public knowledge and appreciation of engineering and geoscience, and protect the professions from misrepresentation and misunderstanding.

Application to this project is discussed in Sections I, III, IV, VI, and VIII: the project serves public welfare and the environment (Principle 1); annotation ground truth was delegated to a qualified domain expert (Principle 2); and results are reported with their limitations explicitly disclosed (Principle 3).

B. Project Data and Artifacts

- Dataset: 24 base FLIR A65 thermograms (640 × 512), 357 expert polygon annotations across 6 classes, expanded to 240 images via the Albumentations pipeline of Section VI; YOLO segmentation label format.

- Model artifact: YOLOv11n-seg weights trained per Table 1; validation metrics mAP@50 = 89.84%, mAP@50-95 = 66.74%, precision = 86.95%, recall = 95.79%.
- Validation artifacts: ablation baseline weights and metrics; 5-fold cross-validation logs (mean mAP@50 = 97.99% ± 0.52%).
- Software: Flask web application source (inference interface, annotation canvas, YOLO label export), training pipeline scripts, and augmentation configuration, maintained in the project repository.