International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Infrared Image Colorization and Enhancement for Improved Object Interpretation Using SWINIR

AZMEERA RAMESH[1], P.CHANDRASEKHAR[2]
SHAHISTHA[3], P.SINDOOJA[4]

[1]Department of CSM, CMR Technical Campus, Hyderabad, Telangana, India
rammy.sandi11@gmail.com
[2]Department of CSM, CMR Technical Campus, Hyderabad, Telangana, India
237r1a66a8@cmrtc.ac.in
[3]Department of CSM, CMR Technical Campus, Hyderabad, Telangana, India
237r1a66b4@cmrtc.ac.in
[4]Department of CSM, CMR Technical Campus, Hyderabad, Telangana, India
237r1a66b0@cmrtc.ac.in

Abstract - Thermal infrared (IR) cameras see in darkness, smoke and glare, but their images are low in resolution, noisy and monochrome, which makes both human and machine interpretation difficult. This paper presents an end-to-end pipeline that enhances infrared images with a Swin-Transformer image-restoration network (SwinIR), colorizes them for human viewing and interprets them with an object detector. The official lightweight SwinIR (0.90 M parameters) is converted exactly from three-channel to single-channel weights and fine-tuned on the FLIR ADAS v2 thermal dataset for joint ×2 super-resolution and denoising, and a YOLOv8n detector is fine-tuned on the same thermal domain. All models are evaluated on 3,749 held-out frames from eight video sequences that never appear in training. Under ×2 down-sampling with Gaussian noise ($\sigma = 10/255$), the proposed enhancer reaches 32.80 dB PSNR, 0.734 SSIM and 0.476 LPIPS, compared with 31.20 dB, 0.700 and 0.545 for the existing classical pipeline, and it is better on all 8 test videos. End to end, object-detection mAP@0.5 on the degraded frames rises from 0.021 to 0.154, recovering 76% of the accuracy obtained on clean frames. The experiments also show that pseudo-colour images, while useful for operators, degrade machine detection, so detection is performed on the enhanced IR image. Under the protocol of a 2025 benchmark study the thermal detector attains 0.523 mAP@0.5 on FLIR ADAS v2. An auxiliary satellite water-body segmentation model reaches 91.94% pixel accuracy and 0.960 ROC-AUC.

Key Words: *Infrared Imaging, Thermal Image Enhancement, SwinIR, Swin Transformer, Super-Resolution, Denoising, Colorization, Object Detection, YOLOv8, FLIR ADAS.*

1. INTRODUCTION

Thermal infrared cameras record the long-wave radiation that every object emits, so they form an image without any visible light. They keep working at night and through smoke, light fog and headlight glare [1]. This makes them valuable for video surveillance, driver assistance, search and rescue, industrial inspection and remote sensing.

The images have well-known weaknesses. Uncooled microbolometer sensors have low resolution (640×512 pixels is typical), add sensor noise and produce low-contrast, texture-poor grayscale images. People find such images hard to read, and object detectors pretrained on visible-light photographs such as COCO [2] transfer poorly to them. Higher-resolution thermal hardware is expensive, so software enhancement (super-resolution and denoising) followed by an interpretation stage is an attractive alternative.

Deep learning has advanced image super-resolution from convolutional networks such as SRCNN [3] and EDSR [4] to transformer models such as SwinIR [5] and HAT [6]. SwinIR builds on the Swin Transformer [7]: self-attention is computed inside shifted local windows, which captures long-range context at modest cost. These models are, however, usually trained on visible-light images such as DIV2K [8], and thermal images differ in noise characteristics, intensity statistics and texture.

The starting point of this work is an existing prototype of a modular infrared pipeline (upload, validation, preprocessing, enhancement, colorization, detection and reporting). Its enhancement module was a small convolutional network with no trained weights, so it always fell back to Gaussian smoothing with cubic up-sampling. Its detector was a COCO-pretrained YOLOv8n applied to the pseudo-coloured image. This paper replaces these components with a genuine SwinIR fine-tuned on thermal data and a thermal-domain detector. It then measures, on held-out data, how much each change improves image quality and object interpretation.

The rest of the paper is organised as follows. Section 2 lists the contributions and Section 3 reviews related work. Section 4 states the novelty and Section 5 describes the methodology. Section 6 gives the experimental setup and Section 7 defines the metrics.

Sections 8 and 9 present the results and the comparative analysis, and Section 10 concludes.

2. CONTRIBUTION OF THE PAPER

The contributions of this paper are:

- An exact conversion of the official three-channel SwinIR weights to a single-channel thermal model, followed by fine-tuning on FLIR ADAS v2 for joint ×2 super-resolution and denoising.
- A thermal-domain YOLOv8n detector and a task-driven evaluation that measures how each enhancement method changes object-detection accuracy on identical degraded inputs.
- A held-out evaluation on 3,749 frames from eight unseen video sequences at three noise levels, comparing the proposed model against bicubic interpolation, the existing pipeline, zero-shot SwinIR and the original SwinIRLite network, using PSNR, SSIM, LPIPS, precision, recall, F1-score and mAP.
- An analysis showing that pseudo-colour images aid human viewing but harm machine detection, which leads to a design that detects on the enhanced infrared image and only draws the results on the coloured view.
- An auxiliary satellite water-body segmentation model with full classification metrics (confusion matrix, ROC, precision, recall, F1-score).
- A working end-to-end web pipeline with the trained models, verified by an automated inference test.

3. RELATED WORK

Thermal imaging and its applications are surveyed by Gade and Moeslund [1]. For image super-resolution, SRCNN [3] introduced end-to-end convolutional learning and EDSR [4] deepened residual networks. SwinIR [5] replaced convolutional bodies with Residual Swin Transformer Blocks and reached state-of-the-art accuracy with fewer parameters, and HAT [6] activated more input pixels through hybrid attention.

For infrared super-resolution, Huang et al. [9] proposed SwinAIR, which adds an average-pooling branch to the residual Swin Transformer block, and SwinAIR-GAN for real-world degradations. Their models were trained on visible-light images and evaluated on several infrared sets including a FLIR set. Zou et al. [10] proposed a contourlet refinement gate with a thermal-spectrum fidelity loss and evaluated it on the M3FD, TNO and RoadScene infrared images.

Thermal-to-visible colorization has been studied with generative networks trained on registered image pairs [11]. Deployed systems more often use fixed pseudo-colour palettes, such as the ironbow map used in this project, because they need no paired training data.

For thermal object detection, the YOLO family [12] is widely used. Zhang et al. [13] evaluated several YOLO variants on FLIR ADAS v2 and proposed MS-YOLO, which uses a MobileNetV4 backbone and SlideLoss. Li et al. [14] proposed CRT-YOLO, which uses centralized feature regulation, for the FLIR-aligned and LLVIP datasets. Enhancement methods are mostly judged by PSNR alone. Their effect on downstream detection is rarely measured, and few studies evaluate on held-out sequences that share no video with the training data. This paper addresses both gaps.

Table -1: Related Work

Ref.	Year	Method	Data	Task / reported result
[5]	2021	SwinIR (RSTB, shifted windows)	DIV2K, Flickr2K	Visible SR and denoising
[9]	2024	SwinAIR / SwinAIR-GAN	Visible train; Flir, CVC-14, Iray test	IR ×2 SR, 34.55 dB on Flir
[10]	2024	Contourlet refinement gate	M3FD, TNO, RoadScene	IR ×2 / ×4 SR
[11]	2018	GAN thermal-to-visible	Paired TIR–RGB	Thermal colorization
[13]	2025	MS-YOLO (MobileNetV4 + SlideLoss)	FLIR ADAS v2, 9 classes	Detection, mAP@0.5 0.484
[14]	2025	CRT-YOLO	FLIR-aligned, LLVIP	Detection, mAP@0.5 79.6% (FLIR-aligned)

4. NOVELTY OF PROPOSED WORK

The novelty of this work lies in how enhancement, colorization and interpretation are combined and evaluated, not in a new network layer. First, the official SwinIR weights trained on colour images are converted to a single thermal channel exactly. Every pretrained parameter is reused and the network starts from a strong initialization rather than from scratch. Second, enhancement is judged both by image-quality metrics and by its effect on object detection. Both are measured on byte-identical, seeded degraded inputs from video sequences never seen in training, so the gains cannot come from data leakage. Third, the study measures the effect of colorization on machine interpretation and redesigns the pipeline accordingly: the detector sees the enhanced infrared image, while the operator sees the colorized view with the detections drawn on it. Finally, the complete system is delivered as a working web application with automated tests.

5. PROPOSED METHODOLOGY

The methodology has two parts: offline training of the enhancement and detection models on thermal data, and an online pipeline that applies them to new infrared images. Each step is described below.

5.1 System Overview

Fig -1 shows the pipeline. An uploaded infrared image (a CCTV or vehicle thermal frame, or a satellite IR image) is validated and normalised, then enhanced by SwinIR. The enhanced image follows two branches. One produces a pseudo-colour view for the operator; the other runs the thermal detector. The outputs are the enhanced

image, the colorized view with the detections drawn on it, and a JSON report.

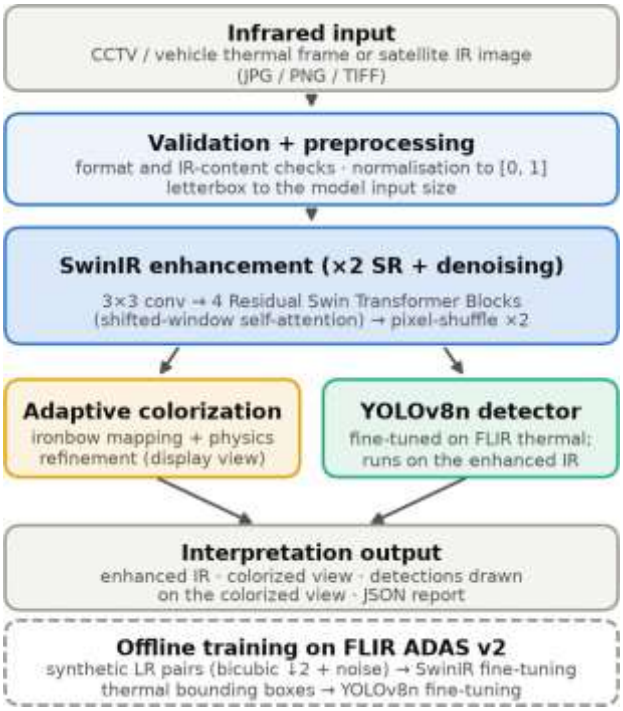


Fig -1: System Architecture of the Proposed Pipeline

5.2 Dataset Description

The main dataset is the FLIR ADAS v2 thermal dataset [15]. It provides 640×512 thermal frames from driving sequences as 8-bit images (with automatic gain control) and 16-bit radiometric TIFF files, with COCO-format bounding boxes for 16 object categories. The auxiliary satellite dataset [16] contains Sentinel-2 RGB tiles with water masks derived from a water index. It is visible-light imagery and is used only for the satellite interpretation task.

Datasets used:

- FLIR ADAS v2 – thermal frames with object boxes (enhancement and detection)
- Satellite Images of Water Bodies – RGB tiles with water masks (segmentation)

Object categories in the thermal test set:

- car, person, light, sign, motor, truck, other vehicle, hydrant, bike, dog

5.3 Data Validation and Splitting

Every file was decoded and checked before training. None of the 15,635 thermal JPEG frames or their radiometric TIFF files were corrupted, and no bounding box had zero size. The official FLIR split is sequence-disjoint: the training, validation and test splits share no video identifier, so near-duplicate frames cannot leak between them. It was therefore used as-is (Table -2). The test split contains 3,749 frames from 8 videos that were used only once, after training. For the satellite data, 21 degenerate image slivers (shorter side under 32 pixels) were removed, and a seeded 70/15/15 split was made.

Table -2: Dataset Splits Used in This Work

Split	Thermal frames	Videos	Boxes	Water tiles
Train	10,742	133	175,040	1,974
Val	1,144	17	16,696	423
Test	3,749	8	62,317	423

5.4 Degradation Model

Low-quality inputs are produced from the clean frames by the degradation model in Eq. (1). The clean frame x is down-sampled ×2 with an anti-aliased bicubic kernel, Gaussian noise n is added and the result y is quantised to 8 bits:

$$y = Q((x \otimes k)_{\downarrow 2} + n), \quad n \sim N(0, \sigma^2) \tag{1}$$

During training, σ is drawn uniformly from [0, 20]/255 for every patch, so one model learns both super-resolution and denoising. For testing, σ is fixed at 0, 10 or 20 (out of 255) and the noise is seeded from each frame's name, so every method receives exactly the same input.

5.5 SwinIR Enhancement Network

The enhancement network is the lightweight SwinIR [5]. A 3×3 convolution extracts shallow features. Four Residual Swin Transformer Blocks (RSTB) then extract deep features, each containing six Swin layers (embedding dimension 60, six attention heads, window size 8) and a convolution. A pixel-shuffle layer finally produces the ×2 output. Inside each window, attention is computed with a learned relative position bias B:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right) V \tag{2}$$

Alternate layers shift the windows by half their size so that information flows between neighbouring windows. The official weights were trained on three-channel colour images. They are converted to one channel by summing the first-layer kernels over the colour channels, folding the training mean into the bias, and combining the output-layer kernels with the luminance weights (0.299, 0.587, 0.114). The converted checkpoint loads with all keys matching. The network is then fine-tuned with the L1 loss in Eq. (3), where F is the network and N the number of pixels:

$$L = \frac{1}{N}\sum |F(y) - x| \tag{3}$$

Layers used:

- Shallow 3×3 convolution
- 4 × Residual Swin Transformer Block (6 Swin layers each)
- Shifted-window multi-head self-attention with relative position bias
- Pixel-shuffle ×2 reconstruction

5.6 Adaptive Colorization

The existing rule-based colorizer is kept unchanged. When no aligned visible image is available, it maps the normalised temperature to an ironbow-like palette, uses image gradients to keep edges visible, and adjusts brightness with a physics-inspired term proportional to the fourth power of normalised temperature (a Stefan–Boltzmann proxy). When an aligned visible image is available, it keeps the visible chrominance and modulates it with the thermal luminance. The colorized output is used only for display.

5.7 Thermal Object Detection

Objects are detected with YOLOv8n [12] (3.01 M parameters, 8.2 GFLOPs), the same detector the existing pipeline used. It is initialised from COCO weights and fine-tuned on the FLIR thermal training split. At inference time it runs on the enhanced infrared image, and its boxes are drawn onto the colorized view for the operator.

5.8 Satellite Water-Body Segmentation

For the satellite data, a U-Net [17] with four resolution levels and a base width of 32 channels (7.76 M parameters) classifies every pixel as water or non-water. It is trained with the sum of binary cross-entropy and Dice losses:

$$L_{\text{seg}} = \text{BCE}(p,t) + 1 - \frac{2\sum pt + 1}{\sum p + \sum t + 1} \quad (4)$$

6. EXPERIMENTAL SETUP

6.1 Datasets Details

The thermal experiments use 10,742 training, 1,144 validation and 3,749 test frames. Super-resolution training takes 128×128 clean patches with random flips and rotations and degrades them on the fly. Validation uses every fourth validation frame at full resolution with a fixed σ = 10. The detector is trained on the 175,040 training boxes. The satellite experiments use 1,974/423/423 tiles resized to 256×256.

6.2 Hardware Configuration

All experiments ran on a workstation with an NVIDIA GeForce RTX 5060 Ti GPU (16 GB), an Intel Core i9-14900K CPU (24 cores) and 16 GB of system memory, running Windows 11.

6.3 Software Environment

The system is implemented in Python 3.11 with PyTorch 2.11 [18] and CUDA 12.8. Ultralytics 8.4 is used for YOLOv8, scikit-image for the reference metric checks and the LPIPS package for perceptual distance. The existing application uses a FastAPI back end and a React front end.

6.4 Training Procedure

Both enhancement networks are trained with the Adam optimiser [19] at a learning rate of 2×10^{-4}, with a 1,000-iteration warm-up and cosine decay, a batch of 16 patches and 800 iterations per epoch, for up to 24 epochs. Early stopping (patience 6) and best-checkpoint selection are based on validation PSNR. Table -3 summarises the training of every model.

Table -3: Training Summary of the Models

Model	Initialisation	Epochs run	Best epoch	Best validation
SwinIR (proposed)	Official DIV2K ×2	24	23	31.54 dB PSNR
SwinIRLite	Random	24	19	31.23 dB PSNR
YOLOv8n	COCO	50	40	0.424 mAP@0.5
Water U-Net	Random	60	58	0.776 IoU

6.5 Implementation Details

Training uses bfloat16 mixed precision and a fixed random seed (42). The thermal frames are cached as memory-mapped arrays so that training does not wait on image decoding. Window attention uses PyTorch's fused scaled-dot-product kernel; zero-padding the head dimension makes it numerically equivalent to the reference implementation. The PSNR and SSIM values computed on the GPU were cross-checked against scikit-image on sample frames: the largest differences were 1.7e-06 dB and 4.0e-05.

7. EVALUATION METRICS

Image enhancement is evaluated with reference-based image-quality metrics, object detection with detection metrics, and segmentation with classification metrics. Classification accuracy or F1-score is never used as a measure of image quality.

7.1 Peak Signal-to-Noise Ratio

PSNR measures pixel fidelity between the restored image and the clean reference on a [0, 1] scale, excluding a 2-pixel border:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{1}{\text{MSE}}\right) \quad (5)$$

7.2 Structural Similarity

SSIM [20] compares local luminance, contrast and structure using an 11×11 Gaussian window (σ = 1.5):

$$\text{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (6)$$

7.3 Learned Perceptual Similarity

LPIPS [21] measures the distance between deep AlexNet features of the two images. Lower values mean the images look more alike to a human observer.

7.4 Precision, Recall and F1-Score

A detection is a true positive (TP) when it overlaps an unmatched ground-truth box of the same class with an intersection-over-union (IoU) of at least 0.5. Otherwise it

is a false positive (FP), and unmatched ground-truth boxes are false negatives (FN):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

7.5 Mean Average Precision

Average precision (AP) is the area under the precision–recall curve of one class (101-point interpolation). mAP@0.5 averages AP over the C classes at IoU 0.5, and mAP@0.5:0.95 also averages over IoU thresholds from 0.5 to 0.95 (written mAP50 and mAP50-95 in the tables):

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \quad (10)$$

$$\text{mAP} = \frac{1}{C}\sum \text{AP}_c \quad (11)$$

7.6 Accuracy, IoU and ROC-AUC

For water segmentation, every pixel is a binary classification. Pixel accuracy, per-class precision, recall and F1-score, water IoU and the area under the ROC curve (ROC-AUC) are reported:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (12)$$

8. RESULTS AND ANALYSIS

All results in this section were computed once, after training, on held-out test data. For the thermal tasks, every method received identical degraded inputs built from the 3,749 test frames.

8.1 Training Behaviour

Fig -2 shows the training and validation loss of both enhancement networks. Fig -3 shows the validation PSNR per epoch. Fine-tuned SwinIR reached its best validation PSNR of 31.54 dB at epoch 23, against 28.04 dB for bicubic interpolation; SwinIRLite, trained from scratch, reached 31.23 dB. The validation loss is lower than the training loss because training uses harder, randomly sampled noise levels on small patches. Fig -4 shows the detector's validation scores. Its best checkpoint (epoch 40) reached 0.424 mAP@0.5.

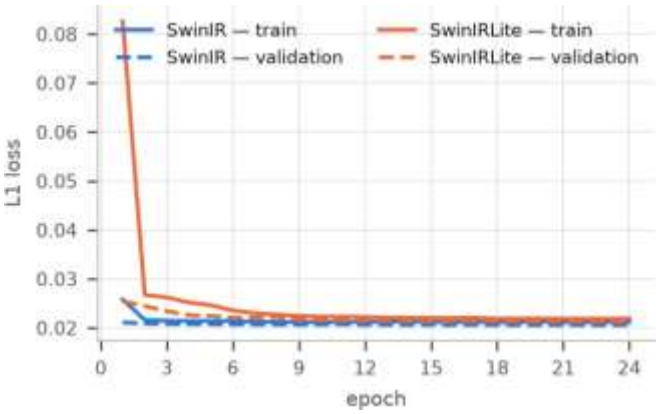


Fig -2: Training and Validation Loss of the Enhancement Networks

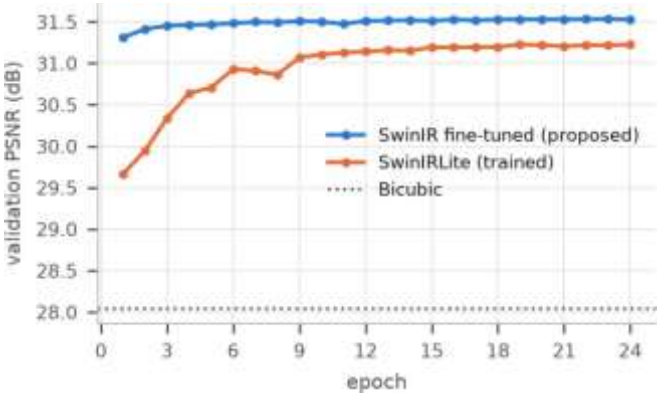


Fig -3: Validation PSNR per Epoch

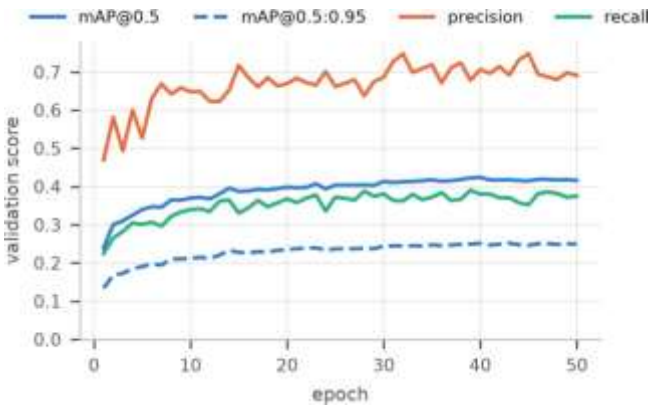


Fig -4: Validation Metrics of the Thermal Detector per Epoch

8.2 Enhancement Quality

Table -4 and Fig -5 compare the enhancement methods at the main test condition (×2, σ = 10). Fine-tuned SwinIR gives the best value on all three metrics: 32.80 dB PSNR, 0.734 SSIM and 0.476 LPIPS. This is 1.60 dB better than the existing pipeline and 4.50 dB better than bicubic interpolation. The original SwinIRLite network, once trained, comes close (32.56 dB). SwinIR without fine-tuning fails (24.75 dB) because its visible-light training makes it sharpen the noise instead of removing it.

Table -4: Enhancement Results on the Test Set (×2, σ = 10)

Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
Bicubic	28.30	0.5318	0.4879
Existing pipeline	31.20	0.7000	0.5454
SwinIR zero-shot	24.75	0.3508	0.4951
SwinIRLite (trained)	32.56	0.7291	0.4884
SwinIR fine-tuned (proposed)	**32.80**	**0.7342**	**0.4756**

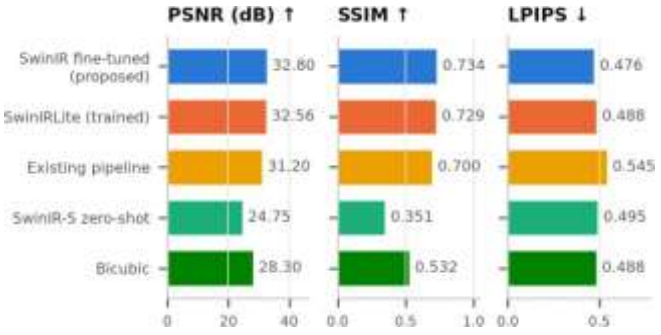


Fig -5: Comparison of Enhancement Methods (×2, σ = 10)

8.3 Robustness to Noise

Table -5 and Fig -6 report PSNR at the three noise levels. The proposed model gives the highest PSNR whenever noise is present, and the margin over the existing pipeline grows to 2.77 dB at σ = 20. On noise-free input, zero-shot SwinIR is marginally ahead (35.36 vs 35.33 dB): training for denoising costs a few hundredths of a dB on clean images. The existing classical enhancer is 2.27 dB worse than plain bicubic interpolation on clean input, because its Gaussian pre-blur removes real detail. At σ = 20, SwinIRLite has a slightly better LPIPS (0.513 vs 0.544), so the perceptual advantage of the proposed model is not uniform.

Table -5: PSNR (dB) at Different Noise Levels (×2)

Method	σ = 0	σ = 10	σ = 20
Bicubic	34.78	28.30	23.21
Existing pipeline	32.51	31.20	28.76
SwinIR zero-shot	35.36	24.75	18.62
SwinIRLite (trained)	35.09	32.56	31.08
SwinIR fine-tuned (proposed)	35.33	32.80	31.53

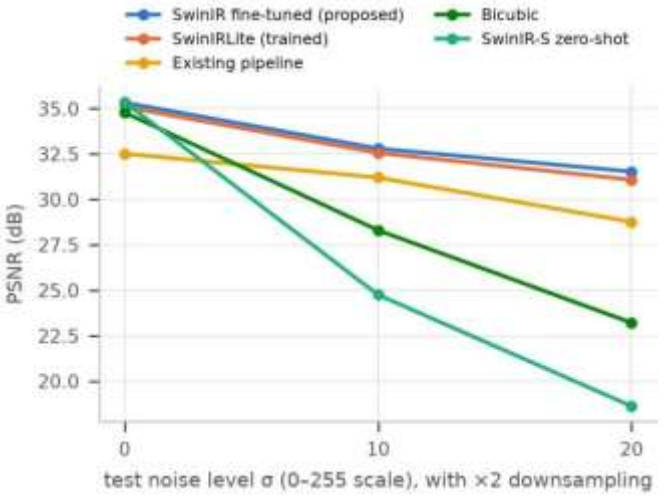


Fig -6: PSNR Versus Test Noise Level

8.4 Qualitative Results

Fig -7 shows a crop of a busy street scene from the test set. The degraded input and the bicubic result are dominated by noise. The existing pipeline removes the noise but blurs the cars and the pedestrian. Both trained networks recover clean outlines, and the proposed SwinIR gives the highest PSNR and SSIM for the frame.

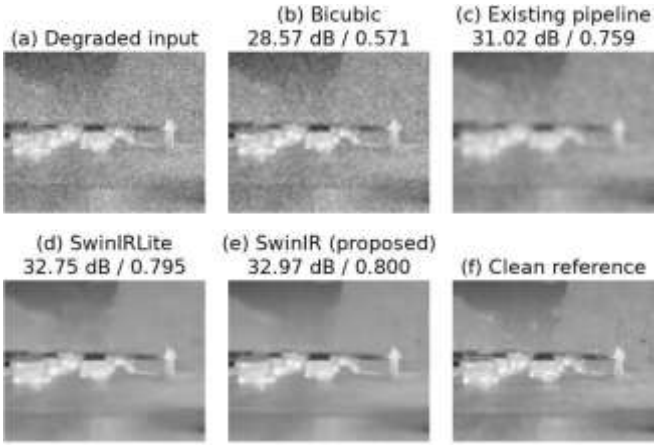


Fig -7: Visual Comparison on a Test Frame (crop; PSNR / SSIM of the full frame)

8.5 Object Detection Performance

Table -6 and Fig -8 show detection accuracy on the degraded test frames for each pipeline. The existing pipeline (classical enhancement, colorization, COCO detector) reaches only 0.021 mAP@0.5. The proposed pipeline reaches 0.154, which is 7.5 times higher and 76% of the 0.202 obtained on clean frames. Two changes contribute. Fine-tuning the detector on thermal data raises mAP@0.5 to 0.086 with the existing enhancer, and replacing that enhancer with SwinIR raises it further to 0.154. Detection on the test videos is harder than on the validation set: even on clean frames the detector reaches 0.202 mAP@0.5, against 0.424 on validation.

Table -6: Detection Results on the Test Set (×2, σ = 10)

Pipeline	mAP50	mAP50-95	P	R	F1
Existing pipeline (COCO detector)	0.021	0.008	0.076	0.033	0.036
Bicubic + thermal detector	0.055	0.031	0.089	0.098	0.082
Existing enhancer + thermal detector	0.086	0.045	0.225	0.088	0.116
SwinIRLite + thermal detector	0.149	0.077	0.364	0.159	0.180
Proposed: SwinIR + thermal detector	0.154	0.079	0.353	0.163	0.182
Clean frames (upper bound)	0.202	0.110	0.326	0.201	0.227

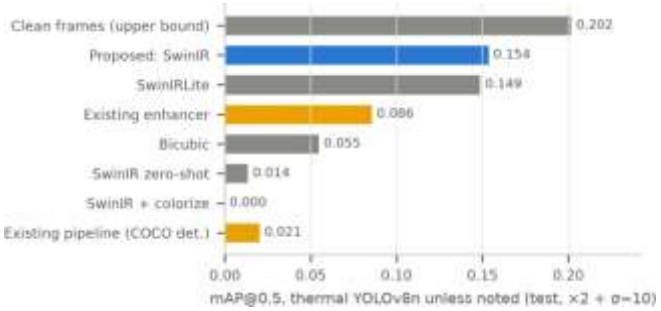


Fig -8: Detection Accuracy of Different Pipelines (×2, σ = 10)

8.5.1 Class-wise Detection Performance

Table -7 lists the per-class results of the proposed pipeline. Fig -9 shows its confusion matrix (confidence ≥ 0.25, IoU ≥ 0.45), Fig -10 compares per-class F1-scores with the existing pipeline, and Fig -11 shows the precision–recall curves of the two most frequent classes. Cars and persons, which make up most of the test boxes, are detected well. Rare classes such as bike, hydrant and dog are almost never detected; most of their errors are missed objects (the background column) rather than confusions with other classes. The most common confusion is truck predicted as car.

Table -7: Class-wise Detection Performance of the Proposed System

Class	Boxes	P	R	F1	AP50
person	12,323	0.566	0.605	0.585	0.610
bike	113	0.010	0.018	0.013	0.001
car	30,517	0.626	0.589	0.607	0.606
motor	3,314	0.519	0.117	0.191	0.140
truck	2,634	0.207	0.040	0.067	0.025
light	6,758	0.260	0.061	0.099	0.034
hydrant	277	0.000	0.000	0.000	0.000
sign	5,660	0.197	0.113	0.143	0.078
dog	25	1.000	0.000	0.000	0.000
other vehicle	696	0.144	0.092	0.112	0.046
Macro avg	62,317	0.353	0.163	0.182	0.154
Weighted avg		0.503	0.434	0.451	–

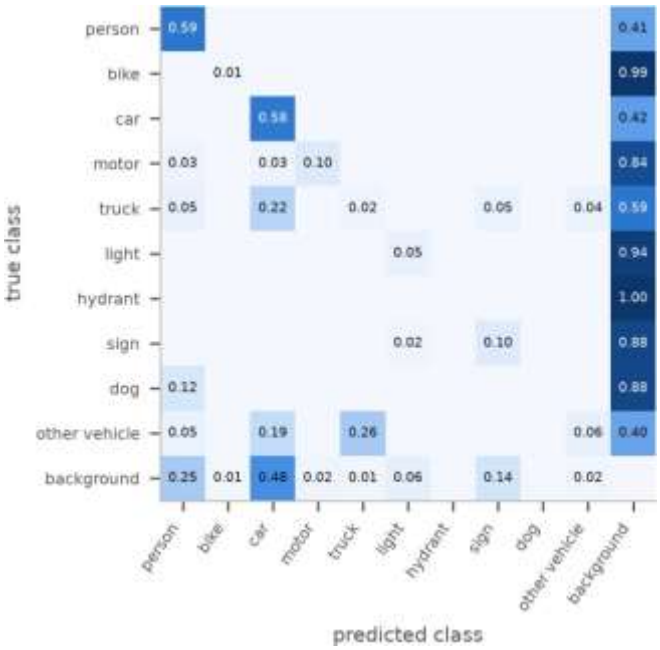


Fig -9: Confusion Matrix of the Proposed Detection Pipeline (row-normalised)

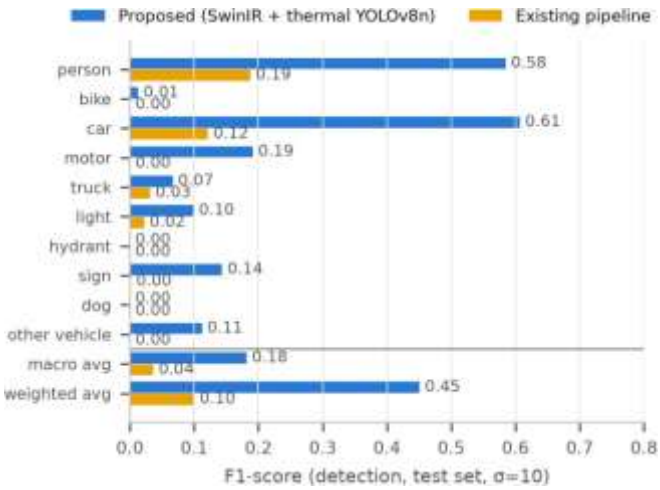


Fig -10: Per-class F1-Score: Proposed vs Existing Pipeline

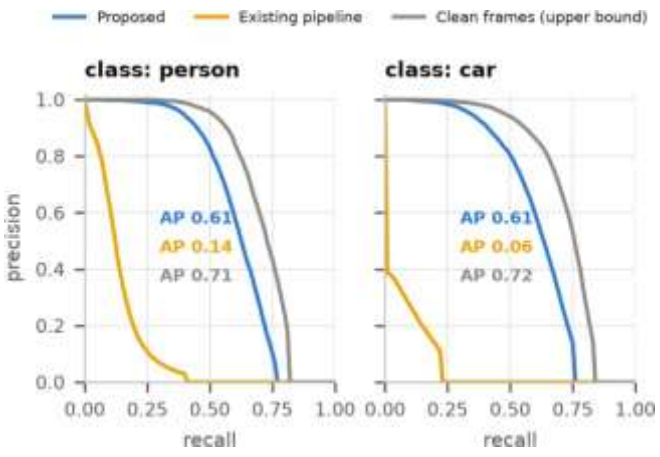


Fig -11: Precision–Recall Curves for the Person and Car Classes

8.6 Effect of Colorization on Interpretation

Colorization helps people read thermal images, but it does not help the detector. When the colorized SwinIR

output is given to the thermal detector, mAP@0.5 falls to 0.0002, because the detector learned grey thermal intensities. The COCO detector does better on the colorized SwinIR output (0.042) than on the existing pipeline's output (0.021), but it is still far below the thermal detector on the grey enhanced image (0.154). The proposed system therefore keeps the two roles separate: detection runs on the enhanced infrared image, and the colorized image is used only to display the result. Colorization quality itself was not scored, because the dataset contains no pixel-aligned colour ground truth (its RGB frames come from a separate, unregistered camera).

8.7 Satellite Water-Body Segmentation

Table -8 gives the pixel-level results on the 423 test tiles. The model reaches 91.94% accuracy, a macro F1-score of 0.9090, a water IoU of 0.7829 and a ROC-AUC of 0.9603. Fig -12 shows that training and validation accuracy rise together without a gap, Fig -13 shows the ROC curve and Fig -14 the confusion matrix. Most errors are water pixels labelled as land, often along thin rivers and shorelines where the masks themselves are uncertain.

Table -8: Water Segmentation Results on the Test Set

Class	Precision	Recall	F1-score	Pixels
non-water	0.9322	0.9475	0.9398	18,398,657
water	0.8929	0.8640	0.8782	9,323,071
Macro avg	0.9125	0.9057	0.9090	
Weighted avg	0.9190	0.9194	0.9191	
Accuracy			0.9194	27,721,728

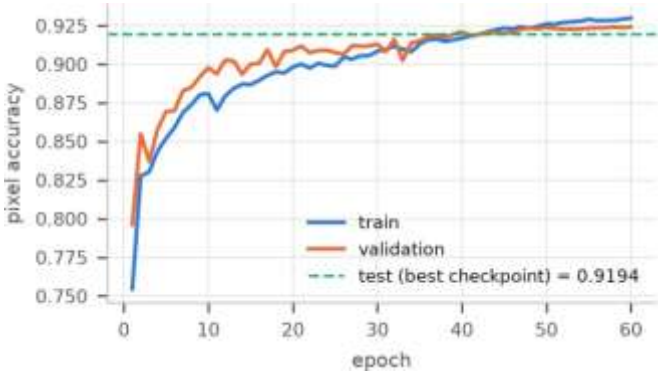


Fig -12: Training, Validation and Test Accuracy of the Water Segmentation Model

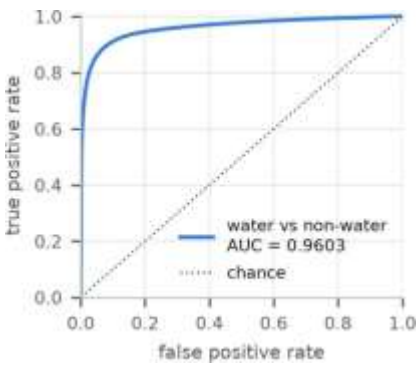


Fig -13: ROC Curve of the Water Segmentation Model

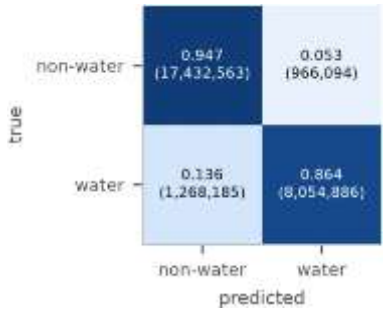


Fig -14: Confusion Matrix of the Water Segmentation Model

8.8 End-to-End System Output

Fig -15 shows the output of the deployed web pipeline for a test frame: the input, the SwinIR-enhanced image, the colorized view and the detections. An automated inference test sent clean and degraded test frames through the back end and checked model loading, output files, image sizes and detections. All 6 of 6 cases passed.

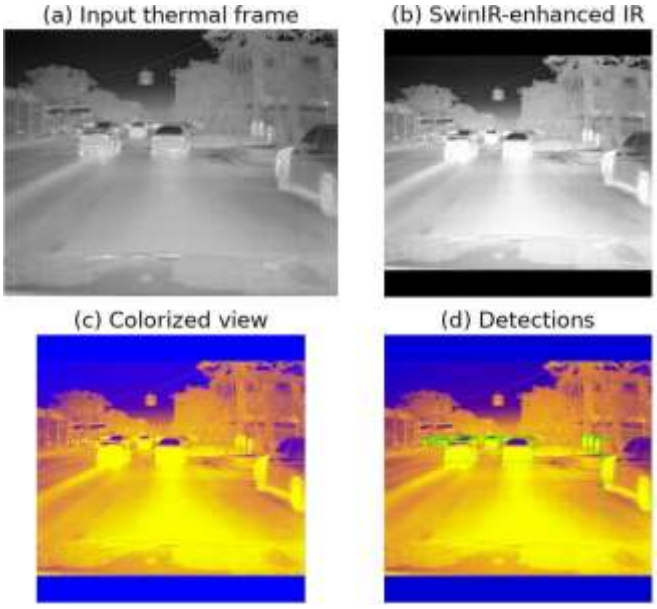


Fig -15: Output of the Deployed Pipeline on a Test Frame

9. COMPARATIVE AND STATISTICAL ANALYSIS

9.1 Comparison with Existing Algorithms (2024–2026)

9.1.1 Thermal Object Detection

The detector was evaluated under the exact protocol of the 2025 MS-YOLO study [13]: the FLIR ADAS v2 thermal validation split (1,144 frames) with nine classes. The published results are in Table -9 and Fig -16. To check that the protocols agree, the COCO-pretrained YOLOv8n was re-measured with this project's evaluation code. It scored 0.165 mAP@0.5, against 0.181 published. The fine-tuned thermal detector reaches 0.523 mAP@0.5 and 0.312 mAP@0.5:0.95. This is level with the published retrained YOLOv8n* (0.517 / 0.318) and above MS-YOLO (0.484 / 0.282), after 50 instead of 200 training epochs. As in the cited study, the validation split was also used for model selection. CRT-YOLO [14]

reports 79.6% mAP@0.5 on the older FLIR-aligned dataset with three classes, which is a different dataset and cannot be compared directly.

Table -9: Comparison with Existing Detectors on FLIR ADAS v2 (9 classes; * = retrained)

Method	Source	P	R	mAP50	mAP50-95
YOLOv8n (COCO weights, not retrained)	[13]	0.426	0.171	0.181	0.106
YOLOv11n (COCO weights, not retrained)	[13]	0.375	0.182	0.189	0.109
YOLOv5n*	[13]	0.619	0.413	0.478	0.288
YOLOv8n*	[13]	0.624	0.463	0.517	0.318
YOLOv10n*	[13]	0.593	0.422	0.477	0.281
YOLOv11n*	[13]	0.628	0.418	0.462	0.274
MS-YOLO	[13]	0.649	0.430	0.484	0.282
YOLOv8n (COCO), re-measured	This work	0.414	0.163	0.165	0.097
Ours: YOLOv8n thermal fine-tuned	This work	**0.668**	**0.485**	**0.523**	**0.312**

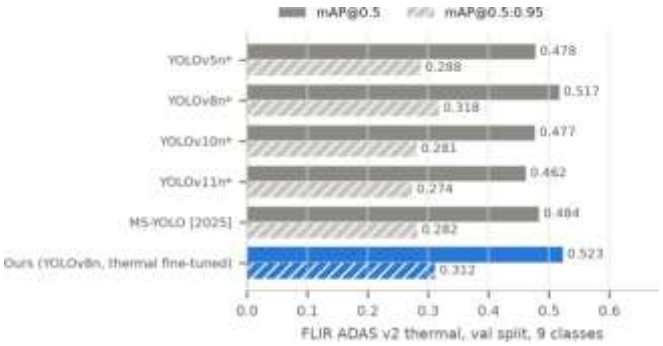


Fig -16: Detection Accuracy Compared with Published Results (2025)

9.1.2 Thermal Image Super-Resolution

Table -10 sets the noise-free (σ = 0) ×2 results beside those published for SwinAIR [9] on its FLIR test set. The two protocols differ in the test images, the down-sampling kernel and the training data, so only the gain over each table's own bicubic row is comparable. On that basis the gain obtained here (+0.54 dB) is smaller than the published gains (+2.32 dB for SwinAIR). Two factors contribute: the proposed model is trained for noisy input, and the FLIR v2 reference frames already contain sensor noise and compression artefacts, which limits the achievable PSNR. The proposed model's advantage is in noisy conditions (Section 8.3), which are closer to real sensor output.

Table -10: ×2 Super-Resolution Without Noise: Published vs This Work

Method	Source	Params	PSNR (dB)	SSIM
Bicubic	[9]	-	32.23	0.7633
SRCNN	[9]	0.1M	32.60	0.7834
EDSR	[9]	43.0M	33.17	0.7930
SRFBN	[9]	3.5M	34.53	0.8214
SwinIR (classical, M)	[9]	11.5M	34.53	0.8208
HAT	[9]	20.8M	33.10	0.7964
SwinAIR	[9]	9.2M	34.55	0.8221
Bicubic	This work	–	34.78	0.8078
SwinIR-S zero-shot	This work	0.90M	35.36	0.8253
SwinIRLite (trained)	This work	0.50M	35.09	0.8159
SwinIR fine-tuned (proposed)	This work	0.90M	35.33	0.8227

9.2 Ablation Analysis

Table -11 adds the components one at a time and reports enhancement PSNR and detection mAP@0.5 at σ = 10. Each component contributes: the thermal detector, then a trained enhancement network, then pretrained SwinIR weights. Removing the fine-tuning (the zero-shot row) or feeding the colorized image to the detector (last row) causes the largest losses.

Table -11: Ablation Analysis (×2, σ = 10)

Configuration	PSNR (dB)	mAP@0.5
Existing pipeline (classical + COCO detector on colour)	31.20	0.021
+ thermal fine-tuned detector	31.20	0.086
SwinIR without fine-tuning + thermal detector	24.75	0.014
SwinIRLite (trained) + thermal detector	32.56	0.149
SwinIR fine-tuned + thermal detector (proposed)	**32.80**	**0.154**
Proposed, detector fed the colorized image	32.80	0.0002

9.3 Statistical Consistency

Consecutive frames of one video are strongly correlated, so a per-frame significance test would overstate the evidence. The video is therefore treated as the unit of replication. The proposed enhancer has a higher mean PSNR than the existing pipeline on 8/8 test videos at σ = 0, 8/8 at σ = 10 and 8/8 at σ = 20. Across frames, the PSNR standard deviation at σ = 10 is 1.70 dB for the proposed model and 1.21 dB for the existing pipeline.

10. CONCLUSION AND FUTURE SCOPE

This paper turned a prototype infrared pipeline into an effective enhancement, colorization and interpretation system and evaluated it on held-out data. A genuine SwinIR, converted exactly from colour to thermal weights and fine-tuned on FLIR ADAS v2, improves PSNR by 1.60 dB over the existing pipeline at σ = 10 and wins on every test video at every noise level tested. Combined with a thermal-domain detector, it raises detection mAP@0.5 on degraded frames from 0.021 to

0.154. The experiments also show that colorization should serve the human operator, not the detector.

The study has limitations. The degradations are synthetic, and training was much shorter than in the cited studies. Colorization quality could not be scored without aligned colour references. In the deployed web pipeline, the existing 512-pixel letterbox step resamples small inputs before enhancement, which reduces detections on low-resolution uploads.

Future work will train on real sensor degradations, train the enhancement network and the detector jointly with a detection-aware loss, use the 16-bit radiometric data and temporal information from video, learn colorization from aligned thermal–visible pairs, and optimise the models for edge devices.

REFERENCES

1. R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," Machine Vision and Applications, vol. 25, no. 1, pp. 245–262, 2014.
2. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in Proc. European Conference on Computer Vision (ECCV), 2014, pp. 740–755.
3. C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295–307, 2016.
4. B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in Proc. IEEE CVPR Workshops, 2017, pp. 136–144.
5. J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin Transformer," in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 1833–1844.
6. X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22367–22377.
7. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.
8. E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in Proc. IEEE CVPR Workshops, 2017, pp. 126–135.
9. F. Huang, Y. Li, X. Ye, and J. Wu, "Infrared image super-resolution network utilizing the enhanced transformer and U-Net," Sensors, vol. 24, no. 14, p. 4686, 2024, doi: 10.3390/s24144686.
10. Y. Zou, Z. Chen, Z. Zhang, X. Li, L. Ma, J. Liu, P. Wang, and Y. Zhang, "Contourlet refinement gate framework for thermal spectrum distribution regularized infrared image super-resolution," arXiv:2411.12530, 2024.
11. A. Berg, J. Ahlberg, and M. Felsberg, "Generating visible spectrum images from thermal infrared," in Proc. IEEE/CVF CVPR Workshops (PBVS), 2018.
12. G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
13. J. Zhang, T. S. White, H. Zhang, W. Hu, D. C. Wunsch II, and J. Liu, "MS-YOLO: Infrared object detection for edge deployment via MobileNetV4 and SlideLoss," arXiv:2509.21696, 2025.
14. J. Li, Y. Wu, and X. Yang, "A high-performance thermal infrared object detection framework with centralized regulation," arXiv:2505.10825, 2025.
15. Teledyne FLIR, "FLIR ADAS thermal dataset, version 2 (thermal and RGB frames with COCO-format annotations)." [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/
16. "Satellite Images of Water Bodies" (Sentinel-2 RGB tiles with NDWI-derived water masks), Kaggle dataset. [Online]. Available: Satellite Images of Water Bodies" (Sentinel-2 RGB tiles with NDWI-derived water masks), Kaggle dataset. [Online]. Available: https://www.kaggle.com/datasets/franciscoescobar/satellite-images-of-water-bodies
17. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, LNCS vol. 9351, 2015, pp. 234–241.
18. A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 8024–8035.
19. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. International Conference on Learning Representations (ICLR), 2015.
20. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.
21. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proc. IEEE/CVF CVPR, 2018, pp. 586–595.