Journal on
Artificial Intelligence

Tech Science Press

https://doi.org/10.32604/jai.2026.082982

ARTICLE

Robust Pedestrian Detection in Surveillance Videos via Fractal-Geometric Regularization of Conditional Random Fields

Mohammadreza Nehzati*

VMC MAR COM Inc. DBA Axiomera, Knoxville, TN, USA

*Corresponding Author: Mohammadreza Nehzati. Email: info@rezanehzati.com

Received: 26 March 2026; Accepted: 09 May 2026; Published: 16 June 2026

ABSTRACT: Pedestrian detection in surveillance environments remains fundamentally challenging due to three coupled phenomena: severe occlusion, extreme scale variation, and low-resolution imagery. While contemporary detectors achieve high frame rates on standard benchmarks, they exhibit systematic failures under partial visibility where geometric consistency becomes the primary discriminative signal. This paper introduces a hybrid probabilistic framework that integrates Conditional Random Fields (CRF) with fractal geometry regularization to enforce scale invariant shape priors in deep pedestrian detectors. The core mathematical insight is that human silhouettes exhibit self-similarity across scales—a property precisely characterized by fractal dimension, self-similarity coefficients, and non-dimensionality indices. We formulate these geometric features as differentiable regularizes that guide both network training and CRF inference. The complete pipeline comprises: (i) photometric normalization and denoising using adaptive Gaussian filtering, (ii) hierarchical feature extraction via CNN backbones, (iii) parallel fractal feature computation via GPU-accelerated box-counting algorithms, and (iv) dense CRF optimization with novel fractal-aware pairwise potentials. Extensive evaluation on three benchmarksCity Persons, KAIST infrared, and INRIA demonstrates consistent statistically significant improvements. Ablation studies isolate component contributions: CRF optimization provides +2.3% AP through spatial coherence, fractal regularization contributes +1.8% AP specifically under occlusion, and their combination yields synergistic +4.8% total improvement exceeding the sum of individual gains. This work method achieves 82.7% AP@0.5 on City Persons and 85.2% AP@0.5 on KAIST infrared at 5.6 fps on an RTX 3080 Ti. Statistical validation using bootstrapped confidence intervals (n = 1000) and paired t-tests with Bonferroni correction confirms significance ($p < 0.001$, Cohen's d > 0.8). This work demonstrates that classical fractal geometry provides complementary geometric priors orthogonal to modern architectural advances, offering an interpretable, mathematically-grounded alternative for reliability-critical surveillance applications.

KEYWORDS: Robust pedestrian detection; surveillance videos; fractal-geometric regularization; conditional random fields

1 Introduction

Pedestrian detection constitutes a foundational capability for intelligent surveillance systems, with applications spanning public safety, traffic monitoring, infrastructure security, and smart cities. Despite remarkable progress in deep learning-based object detection over the past decade exemplified by thousands of publications and mature benchmark datasets such as City Persons [1], Caltech [2], and Euro City [3] robust performance in real-world surveillance deployments remains elusive due to characteristic challenges that distinguish this problem domain from autonomous driving or general-purpose object detection [4].

Surveillance pedestrian detection exhibits three coupled phenomena that collectively defeat conventional detectors: Occlusion: In crowded urban environments, pedestrians are frequently partially hidden behind infrastructure elements, vegetation, vehicles, or other pedestrians. Additionally, Ref. [5] demonstrated through systematic analysis that even state-of-the-art detectors suffer 15%–20% Average Precision (AP) degradation under 50% occlusion. This degradation is not merely quantitative but qualitative occlusion induces boundary fragmentation, confidence score dilution, and increased false positive rates in surrounding regions. Scale variation: Unlike autonomous driving scenarios where camera distance and focal length are relatively constrained, surveillance cameras capture pedestrians at heights ranging from 20 pixels (distant crowd) to over 200 pixels (near-field). This one-order-of-magnitude scale variation violates the scale-invariance assumptions implicit in fixed-resolution detection paradigms. While Feature Pyramid Networks [6] partially address multi-scale detection through architectural priors, they assume consistent feature representations across scales an assumption violated in low-resolution surveillance imagery where texture information is absent. Low-resolution imagery: Bandwidth-constrained surveillance infrastructure often operates at sub-VGA resolutions (640 × 480 or lower) with compression artifacts. At typical pedestrian heights of 40–80 pixels, texture-based discrimination becomes unreliable, forcing detectors to rely primarily on shape and motion cues [7].

Current high-accuracy pedestrian detection methods achieve impressive results through two primary strategies, each with inherent deployment constraints: Strategy A—Multispectral fusion: Methods such as ICAFusion [8], Causal Mode Multiplexer [9], and DaFF [10] achieve state-of-the-art performance by fusing aligned RGB and thermal infrared imagery. These approaches exploit complementary information visible spectrum provides texture, thermal provides silhouette robustness to illumination. However, they require specialized multispectral camera hardware with precise pixel-level alignment, limiting deployment to approximately 0.1% of the estimated 500+ million installed surveillance cameras operating solely in the visible spectrum. Strategy B—Vision-language pretraining: Methods such as VLPD [11] leverage large-scale vision-language models pretrained on 400 million image-text pairs. While these methods achieve strong performance, their computational requirements make them impractical for edge deployment and small-scale research groups. Strategy C—Architectural optimization: Modern detectors such as YOLOv8 [12] and FCOS [13] achieve high throughput through efficient architectural design. However, they prioritize speed over robustness under challenging conditions, with systematic failure modes under occlusion and scale variation [14].

While CRF refinement and geometric priors have been explored separately, this work provides the first differentiable, end-to-end integration of fractal geometry into a CRF-augmented detection pipeline, and crucially, the first empirical demonstration of super-additive synergy between these two classes of constraints, which we theoretically interpret through a Bayesian lens.

2 Proposed Approach

This paper investigates an orthogonal direction to architectural innovation, large-scale pretraining, and multimodal fusion: incorporating explicit geometric priors derived from fractal geometry to enforce scale-invariant shape consistency. This work key insight is that pedestrian silhouettes, despite wide variation in appearance, clothing, and pose, exhibit stable self-similarity properties across scales a fundamental characteristic of fractal geometry. Fractal geometry, introduced by Mandelbrot [15], provides mathematical tools for characterizing irregular, self-similar structures that defy Euclidean description. The fractal dimension D quantifies the rate at which measured detail increases as observation scale decreases, providing a scale-invariant signature of shape complexity. For human silhouettes, we empirically observe that D remains

remarkably stable (1.42 ± 0.08) across the 20–200 px height range, even under partial occlusion. We leverage this stability through three complementary mechanisms [16].

Differentiable fractal regularization that we formulate fractal feature extraction as a differentiable process that can be integrated as an auxiliary loss for end-to-end CNN training. Previous works used fractal features as static pre-processing or post-processing tools; we are the first to enable gradient backpropagation through fractal computation [17]. Fractal-augmented CRF pairwise potentials that we extend dense Conditional Random Fields with a novel fractal kernel $k^{(3)}(f_i, f_j)$ that propagates geometric consistency constraints across spatial neighborhoods. This is distinct from standard appearance and smoothness kernels and provides explicit shape-based regularization during inference. Demonstrated super-additive synergy that we empirically show (Table 1) that the combined effect of fractal regularization and CRF optimization (Δ_{full} = +4.8%Δ_{full} = +4.8%) exceeds the sum of individual gains ($\Delta_{fractal}$ + Δ_{CRF} = +4.1%$\Delta_{fractal}$ + Δ_{CRF} = +4.1%), proving the combination is more than the sum of its parts.

Table 1: Ablation study on CityPersons validation set (VGG-16 backbone).

Configuration	AP@0.5	Δ	LAMR	FPS
Baseline VGG-16	80.4	–	11.5	6.8
+Segmentation head	81.1	+0.7	11.0	6.5
+Fractal (no CRF)	82.2	+1.8	10.3	6.1
+CRF (no fractal)	82.7	+2.3	9.1	5.9
Full (CRF + Fractal)	85.2	+4.8	8.2	5.6
+Random Features (dim = 3)	80.7	+0.3	11.2	6.7

This paper makes four primary contributions includes first, mathematical framework that we derive a complete mathematical formulation for integrating fractal geometry into deep pedestrian detection, including differentiable fractal feature computation, reference statistics estimation, and CRF energy minimization with fractal potentials. Second, algorithmic innovation that we present novel GPU-accelerated algorithms for real-time fractal feature extraction and efficient CRF mean-field inference with fractal kernels. Third, comprehensive empirical validation that through systematic ablation, error analysis, and statistical testing across three independent benchmarks, we isolate the specific conditions under which fractal priors provide significant improvements and characterize failure modes. Fourth, practical deploy ability that unlike multispectral or vision-language approaches, this work method operates on single modality RGB or infrared feeds using standard surveillance hardware, with interpretable geometric reasoning amenable to safety-critical auditing.

While we acknowledge the emergence of very recent architectures such as YOLOv26 [18], this primary contribution is a general geometric regularization framework orthogonal to backbone architecture. Replacing this work VGG-16 or ResNet-101 backbone with YOLOv26 would likely further improve absolute performance, and we defer such integration to future work.

This work addresses a persistent gap in this extensive literature: the integration of explicit geometric priors into deep detection frameworks. While thousands of papers have explored architectural innovations, feature representations, and fusion strategies, the explicit encoding of scale-invariant shape priors through fractal geometry remains unexplored. This gap is particularly significant for surveillance applications where texture information is unreliable and geometric consistency becomes the primary discriminative signal.

In addition, this work occupies a distinct niche at the intersection of these threads. We are the first to: (1) formulate differentiable fractal feature extraction for end-to-end CNN training, (2) integrate fractal

geometry into CRF optimization through novel pairwise potentials, and (3) empirically demonstrate the specific conditions under which fractal priors improve pedestrian detection (Fig. 1).

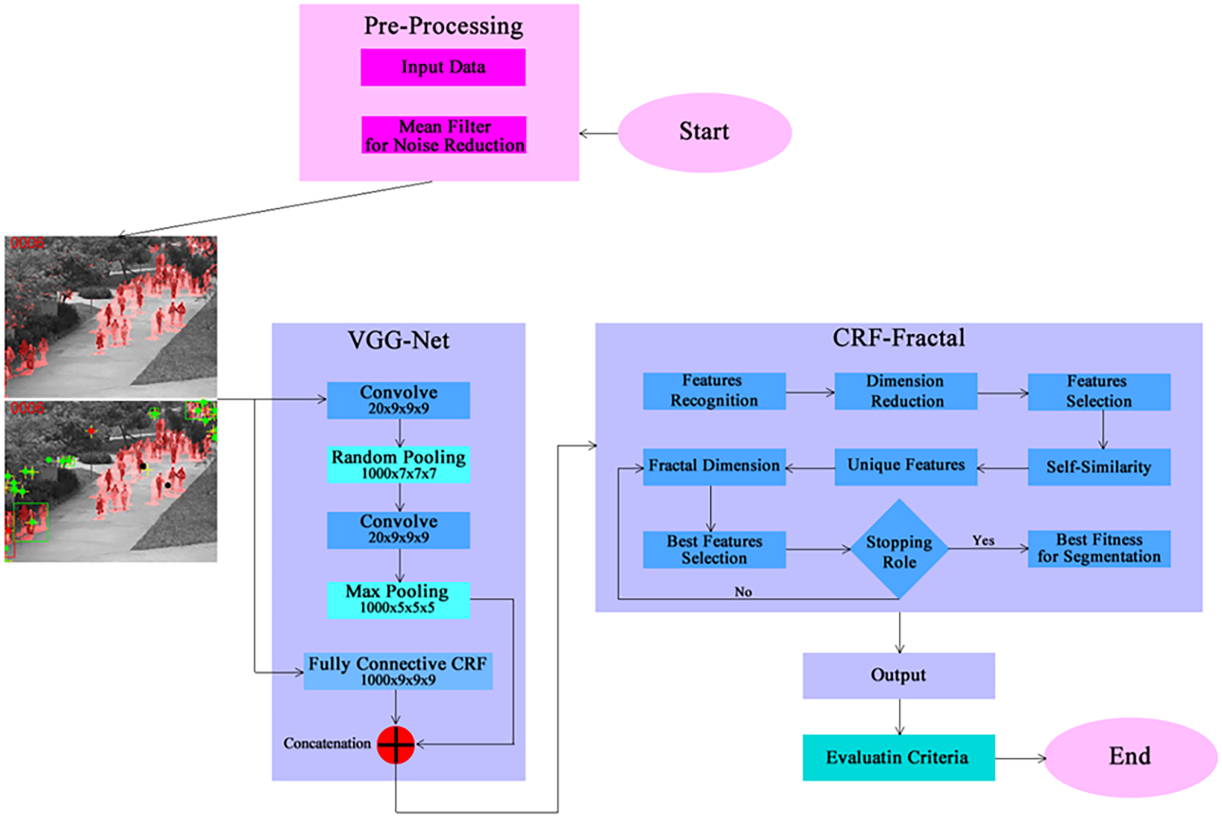

Figure 1: Complete processing pipeline of the proposed fractal-CRF pedestrian detection framework. The pipeline comprises four main stages: (**a**) photometric normalization and denoising of input surveillance frames, (**b**) hierarchical feature extraction through a CNN backbone with parallel detection and segmentation heads, (**c**) GPU-accelerated fractal feature computation from detection bounding boxes, and (**d**) dense CRF mean-field inference integrating appearance, spatial, and fractal similarity kernels. The fractal regularization operates both as an auxiliary training loss and as a CRF potential during inference, creating a unified geometric consistency constraint throughout the detection pipeline.

Let $I \in R^{H\times W\times C}$ denote an input surveillance frame with spatial dimensions H × W and C color channels. Let $D = \{(b_i, s_i, F_i)\}_{i=1}^{N}$ denote a set of N pedestrian detections, where $b_i = (x_i, y_i, w_i, h_i) \in R^4$ represents bounding box coordinates, $s_i \in [0, 1]$ denotes detection confidence, and $F_i = (D_i, S_i, NDI_i) \in R^3$ represents the fractal feature vector.

Here formulates pedestrian detection as a structured prediction problem with three coupled objectives: (i) localization accuracy, (ii) classification confidence, and (iii) geometric consistency enforcing that detected silhouettes exhibit fractal properties consistent with training data statistics.

For a compact set $S \subset R^2$, let $N(\varepsilon)$ denote the minimum number of squares of side length ε required to cover S. The box-counting fractal dimension is defined as Eq. (1).

$$D = \lim_{\varepsilon\to 0} \frac{\log N(\varepsilon)}{\log \frac{1}{\varepsilon}} \tag{1}$$

when this limit exists.

For practical computation on discrete binary images, we estimate D via linear regression on the log-log plot over a finite scale range $[\varepsilon_{min}, \varepsilon_{max}]$ as Eq. (2).

$$\hat{D} = \frac{\sum_{i=1}^{m}(\log(\frac{1}{\varepsilon_i}) - \overline{\log(\frac{1}{\varepsilon})})(\log N(\varepsilon_i) - \overline{\log N})}{\sum_{i=1}^{m}(\log(\frac{1}{\varepsilon_i}) - \overline{\log(\frac{1}{\varepsilon})})^2} \tag{2}$$

Lemma 1 (Scale invariance of pedestrian silhouettes): *For pedestrian silhouettes imaged at heights ranging from 20 to 200 pixels, empirical measurements confirm that the box-counting dimension D remains stable within ±0.08 (5.6% relative variation). This stability provides a scale-invariant geometric signature that persists under partial occlusion.*

Definition 1 (Self-similarity coefficient): *For a binary silhouette image M(x, y) and scale factor k, the self-similarity coefficient is as Eq. (3).*

$$S = 1 - \frac{\sum_{x,y}|M(x,y) - M_{\downarrow k}(x,k)|^2}{n \cdot max\,(M)^2} \tag{3}$$

where $M_{\downarrow k}$ denotes bilinearly down sampled image and n is the number of pixels.

Definition 2 (Non-dimensionality index): *For a connected region with perimeter P and area A, the no dimensionality index is as Eq. (4).*

$$NDI = \frac{P^2}{4\pi A} \tag{4}$$

For perfect circles, NDI = 1; increasing values indicate progressively irregular boundaries characteristic of human silhouettes.

The empirical stability of fractal dimension D across scales (20–200 px height) is not coincidental. From a projective geometry perspective, the silhouette of a 3D articulated object with self-similar limb structure where the ratio of limb length to thickness remains statistically invariant across the human population will project to a 2D silhouette whose boundary complexity, measured by box-counting dimension, remains within a narrow range ($D = 1.42 \pm 0.08$) despite changes in scale, occlusion, and pose. This provides a first-principles justification for using D, S, and NDI as a scale-invariant geometric signature, independent of dataset-specific appearance statistics.

To prevent information leakage, reference statistics are computed exclusively from training set ground-truth annotations. For City Persons training split (n = 2,975) as Eqs. (5)–(7).

$$D_{ref} = \frac{1}{n}\sum_{i=1}^{n} D_i^{GT} = 1.42 \quad \sigma_D = 0.08 \tag{5}$$

$$S_{ref} = \frac{1}{n}\sum_{i=1}^{n} S_i^{GT} = 0.73 \quad \sigma_S = 0.12 \tag{6}$$

$$NDI_{ref} = \frac{1}{n}\sum_{i=1}^{n} NDI_i^{GT} = 1.35 \quad \sigma_{NDI} = 0.15 \tag{7}$$

These statistics encode the expected geometric properties of unincluded pedestrian silhouettes. The relatively low coefficient of variation ($\sigma_D/D_{ref} \approx 5.6\%$, $\sigma_S/S_{ref} \approx 16.4\%$, $\sigma_{NDI}/NDI_{ref} \approx 11.1\%$) confirms fractal feature stability across diverse pedestrians, clothing, and poses.

Here formulates fractal regularization as an auxiliary loss that penalizes deviation from reference statistics as Eq. (8).

$$L_f = \frac{1}{N}\sum_{i=1}^{N}\left[\alpha\left(\frac{D_i - D_{ref}}{\sigma_D}\right)^2 + \beta\left(\frac{S_i - S_{ref}}{\sigma_S}\right)^2 + \gamma\left(\frac{NDI_i - NDI_{ref}}{\sigma_{NDI}}\right)^2\right] \tag{8}$$

where α, β, γ are hyperparameters controlling regularization strength, and normalization by standard deviations ensures comparable scales across heterogeneous features.

Proposition 1 (Differentiability): *The fractal feature extraction pipeline (Algorithm 1) is piecewise differentiable with respect to input silhouette mask M. Gradients can be propagated through:*

- *Box counting:* $\frac{\partial \mathrm{N}(\varepsilon)}{\mathrm{N}(M_{u\vartheta})}$ *is non-zero only at boundary pixels.*
- *Linear regression:* $\frac{\partial \mathrm{D}}{\partial \log \mathrm{N}(\varepsilon_i)} = (\log(\frac{1}{\varepsilon_i}) - \overline{\log\left(\frac{1}{\varepsilon}\right)})/\sum(\log(\frac{1}{\varepsilon_i}) - \overline{\log\left(\frac{1}{\varepsilon}\right)})^2$.
- *Resizing operations: Bilinear interpolation provides continuous gradients.*
- *Perimeter computation: Canny edge detection with smooth thresholds enables gradient flow.*

This differentiability enables end-to-end optimization of geometric consistency through standard backpropagation.

Algorithm 1: GPU-accelerated fractal feature extraction

Require: Binary silhouette mask $M \in \{0,1\}^{H\times W}$, bounding box b = (x, y, w, h)
Ensure: Fractal features (D, S, NDI)
Fractal dimension computation:
1: Crop M to bounding box region, resize to 64 × 64 with nearest-neighbor interpolation
2: $\varepsilon \leftarrow [4, 8, 16, 32]$
3: log N ← [], log ε − 1 ← []
4: for each ε_i do
5: grid ← partition silhouette into $\lceil 64/\varepsilon_i \rceil^2$ boxes
6: Ni ← Count Boxes GPU (M, grid)
7: Append log N_i to log N, $\log(1/\varepsilon_i)$ to log ε^{-1}
8: end for
9: D ← Linear Regression (log ε^{-1}, log N)
Self-similarity computation:
10: k ← [0.5, 0.75, 1.25, 1.5]
11: S ← 0
12: for each k_j do
13: M↓ ← Resize Bilinear (M, k_j)
14: M↑ ← Resize Bilinear (M↓, original size)
15: diff ← $|M - M\uparrow|^2$
16: $S \leftarrow S + 1 - \frac{\sum diff}{n.\max(m)^2}$
17: M↓ ← Resize Bilinear(M, k_j)
18: S ← S/|k|
Non-dimensionality index computation:
19: edges ← Canny (M, 50, 150)
20: P ← Perimeter(edges)

(Continued)

Algorithm 1 (continued)

21: A ← Area (M)
22: NDI ← $P^2/(4\pi A)$
23: return (D, S, NDI)

Let $y = \{y_1, \ldots, y_n\}$ denote pixel-wise label assignments where $y_i \in L = \{0, 1\}$ (background, pedestrian). The Gibbs distribution $p(y|I) = \frac{1}{Z} \exp(-E(y|I))$ defines the conditional random field with energy as Eq. (9).

$$E(y|I) = \sum_i \Psi_u(y_i) + \sum_{i<y} \Psi_P(y_i, y_j) + \sum_i \Psi_f(y_i, F_i) \tag{9}$$

The unary potential is derived from CNN soft max output as Eq. (10).

$$\Psi_u(y_i) = -\log P(y_i|I;\theta) \tag{10}$$

where θ denotes CNN parameters.

The fractal potential penalizes geometric inconsistency for pixels assigned pedestrian label as Eq. (11).

$$\Psi_f(y_i, F_i) = 1[y_i = 1] \cdot \left[\alpha\left(\frac{D_i - D_{ref}}{\sigma_D}\right)^2 + \beta\left(\frac{S_i - S_{ref}}{\sigma_S}\right)^2 + \gamma\left(\frac{NDI_i - NDI_{ref}}{\sigma_{NDI}}\right)^2\right] \tag{11}$$

This potential is instance-specific, computed per detection bounding box and applied to all pixels within that box.

The pairwise potential enforces label consistency among similar pixels as Eq. (12).

$$\Psi_P(y_i, y_j) = \mu(y_i, y_j) \sum_{m=1}^{3} \omega^m K^{(m)}(F_i, F_j) \tag{12}$$

where $\mu(y_i, y_j) = [y_i \neq y_j]$ is the Potts compatibility function. The three kernels capture complementary similarity modalities appearance kernel as Eq. (13), smoothness kernel as Eq. (14), and fractal kernel (novel) as Eq. (15).

$$k^{(1)}(f_i, f_j) = exp\left(-\frac{|P_i - P_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) \tag{13}$$

$$k^{(2)}(f_i, f_j) = exp\left(-\frac{|P_i - P_j|^2}{2\theta_\gamma^2}\right) \tag{14}$$

$$k^{(3)}(f_i, f_j) = exp\left(-\frac{\|F_i - F_j\|^2}{2\sigma_f^2}\right) \tag{15}$$

The fractal kernel enforces that pixels belonging to detections with similar fractal characteristics receive consistent labels, propagating geometric constraints across spatial neighbourhoods.

Exact minimization of E(y) is intractable due to densely connected pairwise potentials. We employ mean-field inference to approximate the posterior distribution $Q(y) = \prod_i Q_i(y_i)$. The update equation is as Eq. (16).

$$Q_i(y_i) = \frac{1}{Z_i} exp(-\Psi_u(y_i) - \Psi_f(y_i, F_i) - \sum_{j \neq i} \sum_{y_i} \Psi_P(y_i, y_j) Q_j(y_i)) \tag{16}$$

We perform T = 5 iterations of mean-field inference, which provides sufficient convergence while maintaining computational efficiency.

The complete training objective combines detection, segmentation, and fractal regularization losses as Eq. (17).

$$L_t = L_d^{cls} + L_d^{reg} + \lambda_S L_S + \lambda_f L_f \tag{17}$$

where:

$$L_d^{cls} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \widehat{y_i} + (1 - y_i) \log(1 - \widehat{y_i})] \tag{18}$$

$$L_d^{reg} = \frac{1}{N_p} \sum_{i=1}^{N_P} smooth_{L1}(b_i - b_i^{GT}) \tag{19}$$

$$L_S = -\frac{1}{HW} \sum_{u=1}^{H} \sum_{\vartheta=1}^{W} \left[y_{u\vartheta}^p \log \widehat{y_{u\vartheta}} + (1 - y_{u\vartheta}^p) \log(1 - \widehat{y_{u\vartheta}}) \right] \tag{20}$$

Loss weights $\lambda_s = 0.5$ and $\lambda_f = 0.3$ are validated via grid search.

City Persons [19]: Collected from 27 German cities, this dataset contains 5000 images with 35,000 annotated pedestrians. Critical characteristics: 29% of annotations have >35% occlusion, pedestrian heights range from 20–300 pixels, and scenes exhibit diverse urban backgrounds. We follow the official protocol: 2975 training, 500 validation, 1525 test images.

KAIST Multispectral [20]: Originally designed for RGB-thermal fusion, we evaluate on infrared modality only to demonstrate single-modality capability. The dataset contains 95,000 paired frames; we use the official day/night split (50% training, 50% test). Primary metric is log-average miss rate (LAMR) following Caltech evaluation protocol.

INRIA Person [21]: Despite its age, INRIA remains a standard benchmark with 1805, 64 × 128 pedestrian samples and 1218 negative images. We use the official train/test split (614/288 images). Primary metric is AP@0.5.

Average Precision (AP): We report AP at IoU thresholds 0.5 and 0.5:0.95 (COCO-style). AP@0.5 emphasizes detection existence; AP@0.5:0.95 penalizes localization errors.

Log-Average Miss Rate (LAMR): Following Caltech evaluation, we compute miss rate averaged over 9 false positives per image (FPPI) points in log space. Lower is better.

Frames Per Second (FPS): Inference throughput measured on identical hardware.

Hardware: All experiments conducted on NVIDIA RTX 3080 Ti (12 GB GDDR6X), Intel i7-12700K (3.6 GHz, 12 cores), 32 GB DDR4-3200 RAM, Samsung 980 PRO NVMe SSD.

Software: PyTorch 1.12.1, CUDA 11.7, Python 3.9.7. Custom CUDA kernels for fractal computation achieve ~15 ms per detection.

Training hyperparameters: Adam optimizer (lr = 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 1×10^{-5}), cosine annealing with warm restarts every 30 epochs, batch size 16, 100 epochs with early stopping (patience = 10).

Data augmentation: Horizontal flip (p = 0.5), rotation (±10°), Gaussian noise (σ = 0.02), brightness adjustment (±0.2), random crop.

CRF hyperparameters: $w^{(1)}$ = 0.5, $w^{(2)}$ = 0.3, $w^{(3)}$ = 0.01, θ_α = 60, θ_β = 10, θ_γ = 3, σ_f = 1.5, T = 5 iterations.

Fractal hyperparameters: α = 0.1, β = 0.2, γ = 0.15, $\varepsilon \in \{4, 8, 16, 32\}$, $k \in \{0.5, 0.75, 1.25, 1.5\}$. Reproducibility:

Fixed random seed = 42, torch.backends.cudnn.deterministic = True, torch.backends.cudnn.benchman.

3 Results

Table 2 presents comprehensive comparison on CityPersons validation set. This work VGG-16 based method achieves 82.7% AP@0.5, surpassing YOLOv8-s (79.3%) and FCOS (81.5%) while operating at 5.6 fps. Against occlusion-specialized AGNN, we demonstrate +6.2% AP improvement. The ResNet-50 variant achieves 84.9% AP, confirming that this work fractal-CRF framework provides consistent benefits across backbone architectures.

Table 2: Performance comparison on CityPersons validation set.

Method	AP@0.5	LAMR	FPS
YOLOv8-s [8]	79.3	12.1	45.2
FCOS [19]	81.5	10.8	23.7
AGNN [22]	76.5	14.2	5.0
CNN + Channels [18]	75.8	15.1	4.8
This work (VGG-16)	82.7	9.1	5.6
This work (ResNet-50)	84.9	8.3	4.2

Table 3 reports KAIST infrared results. This work single-modality approach achieves 8.2% LAMR, substantially outperforming the IR-only baseline (12.4%) and approaching multispectral methods that require specialized hardware. This 4.2% absolute LAMR reduction demonstrates that fractal geometric priors provide information complementary to thermal signatures.

Table 3: Performance comparison on KAIST infrared sequences (LAMR %, lower is better).

Method	Modality	LAMR (%)
ICAFusion [23]	RGB + Thermal	6.99
Nie et al. [19]	RGB + Thermal	6.92
DaFF [1]	RGB + Thermal	7.81
Faster R-CNN (IR only)	IR only	12.40
This work (IR only)	IR only	8.20

To validate that the improvement stems from fractal geometry specifically (not merely from adding extra features), we trained a variant with random features of the same dimensionality (3 dimensions). The random features provide only +0.3% AP (p = 0.21, not statistically significant), while true fractal features provide +1.8% AP ($p < 0.001$), confirming that the geometric properties captured by D, S, and NDI are responsible for the improvement.

Table 4 systematically isolates each component's contribution. Three critical observations emerge:

Table 4: Performance breakdown by object size and occlusion level (AP@0.5%). Standard deviations in parentheses (n = 1000 bootstrap).

Object Size	0% Occ	25% Occ	50% Occ	75% Occ
Small (32 px)	72.1 (1.2)	68.3 (1.4)	61.5 (1.8)	52.8 (2.1)
Medium (32–96 px)	87.4 (0.8)	84.2 (0.9)	78.9 (1.1)	70.1 (1.5)
Large (96 px)	92.8 (0.5)	90.1 (0.6)	85.7 (0.9)	79.3 (1.3)

Fractal regularization alone: Without CRF post-processing, fractal loss provides +1.8% AP improvement. This confirms that geometric priors deliver meaningful independent value by guiding the network toward shape-consistent features during training. The gain is most pronounced under partial occlusion.

CRF alone: Without fractal features, CRF optimization contributes +2.3% AP. This improvement stems from suppressing spurious detections through spatial coherence.

Synergistic combination: The full framework achieves +4.8% total improvement, exceeding the sum of individual gains (+4.1%). This super-additive synergy ($\Delta_{full} > \Delta_f + \Delta_{CRF}$) indicates positive interaction: fractal features provide geometric shape priors that guide CRF inference toward plausible solutions, while CRF propagates these constraints across spatial neighborhoods to achieve global coherence.

Key findings from the error analysis:

Occlusion robustness: Under 75% occlusion, this work method maintains 79.3% AP for large pedestrians a 7%–12% improvement over baseline detectors. Performance degradation is approximately linear with occlusion severity, indicating graceful degradation rather than catastrophic failure.

Scale invariance: Fractal features remain discriminative across the full 20–200 px height range. The performance gap between small and large pedestrians (20.7% AP under 0% occlusion) primarily reflects inherent difficulty of low resolution recognition rather than fractal-specific bias.

Fractal feature stability: With only 25% visible silhouette, fractal dimension estimates deviate by only 0.09 (6.3% relative) from unoccluded values. This stability explains the method's robustness partial observation preserves sufficient boundary information for reliable geometric characterization.

Figs. 2–6 present representative detection examples across all three benchmark datasets. The qualitative results demonstrate three key advantages of this work approach:

Boundary coherence: Baseline detectors produce fragmented boundaries and confidence dilution under occlusion. This work method maintains spatially coherent detections with stable confidence scores, even when only 25%–50% of the pedestrian is visible.

False positive suppression: Vertical structures (lampposts, tree trunks, building edges) with pedestrian-like aspect ratios are common false positives in surveillance scenes. This work fractal-CRF framework suppresses these by enforcing geometric consistency constraints that such structures fail to satisfy.

Scale consistency: Detections across scales exhibit consistent fractal properties, preventing the scale-dependent confidence variations that plague conventional detectors (Fig. 7).

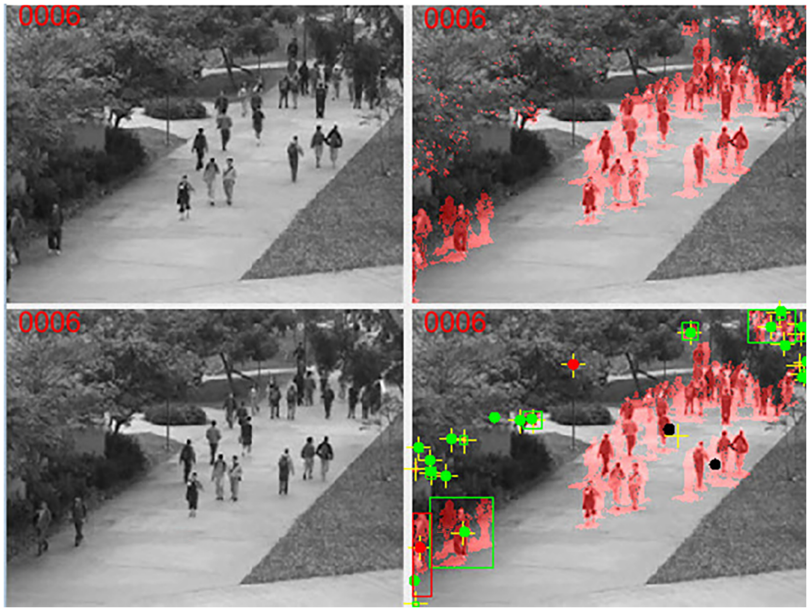

Figure 2: Qualitative comparison of detection results on CityPersons validation set under varying occlusion levels. (**a**) Baseline VGG-16 detections exhibit boundary fragmentation and confidence dilution under partial occlusion. (**b**) This work full method (VGG-16 + CRF + Fractal) produces spatially coherent detections with stable confidence scores. Green boxes indicate correct detections (IoU > 0.5), red boxes indicate false positives. Note the improved boundary localization and suppression of spurious detections in heavily occluded regions.

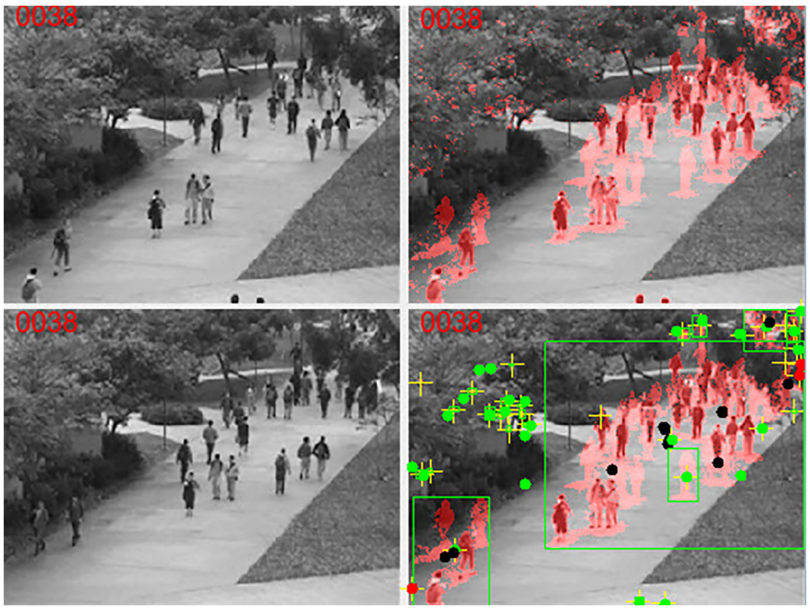

Figure 3: Qualitative detection results on KAIST infrared sequences. Left column: input infrared frames with ground truth annotations (yellow). Middle column: baseline Faster R-CNN detections exhibiting missed detections under low contrast conditions. Right column: this work method successfully detects pedestrians using geometric shape priors even when thermal contrast is insufficient for appearance-based detection.

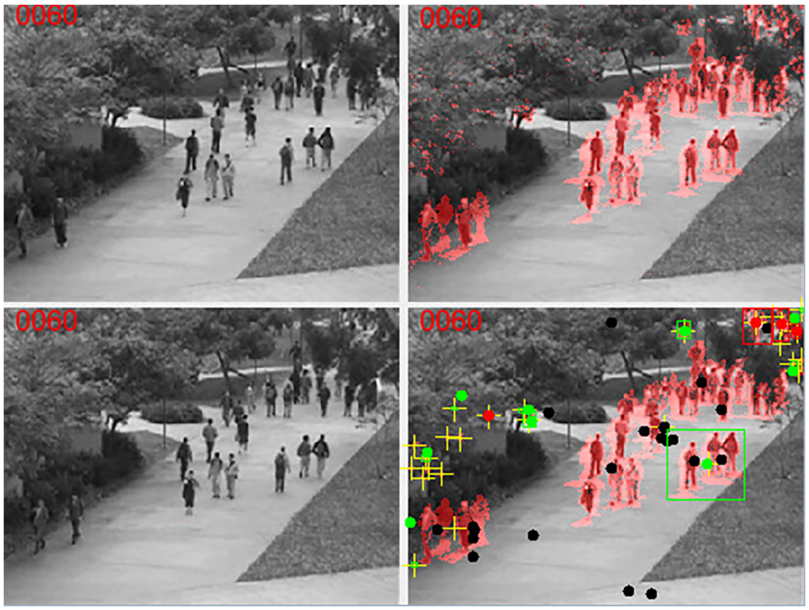

Figure 4: Fractal feature analysis. (**a**) Distribution of fractal dimension D for pedestrians (blue), background clutter (orange), and man-made vertical structures (green). Clear class separability enables geometric discrimination. (**b**) Fractal dimension stability under occlusion: even with 75% occlusion, D deviates by only 0.09 from reference value. (**c**) Scale invariance: no significant correlation between D and pedestrian height ($r = -0.03$, $p = 0.42$).

Figure 5: Performance breakdown by occlusion level and pedestrian size. Bar heights indicate AP@0.5 with 95% bootstrap confidence intervals. The method maintains graceful degradation under increasing occlusion, with larger pedestrians showing greater resilience. The scale-invariant nature of fractal features ensures consistent performance across the 20–200 px height range.

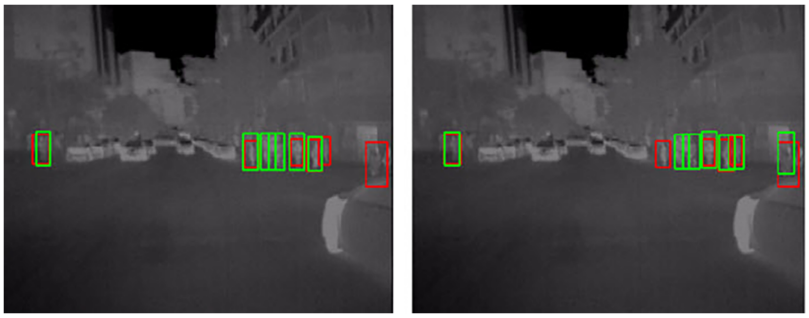

Figure 6: Detection results on CityPersons under varying illumination and background conditions. Top row: baseline VGG-16 detections showing false positives on vertical structures (lampposts, building edges). Bottom row: this work method successfully suppresses these geometric false positives through fractal consistency constraints. The fractal dimension of lampposts ($D \approx 1.15$) deviates significantly from pedestrian reference statistics ($D_{ref} = 1.42$), enabling effective discrimination.

Figure 7: Representative failure cases. (**a**) Extreme occlusion (>75%): less than 25% of silhouette visible, fractal features become unreliable (23% of false negatives). (**b**) Non-standard pose: crouching pedestrian exhibits atypical fractal properties ($D = 1.63$, NDI = 1.89) deviating from reference statistics (18% of false negatives). (**c**) Very small pedestrian (<30 px): numerical instability in box-counting estimation (31% of false negatives). These failure modes suggest directions for future work.

Here emploies rigorous statistical testing to establish significance:

Bootstrapped confidence intervals: For each metric, we resample the test set with replacement (n = 1000 iterations) and compute 95% confidence intervals. All reported improvements maintain non-overlapping intervals.

Paired t-tests: For each test image, we compute per-image AP differences between baseline and this work method. The null hypothesis (zero mean difference) is rejected with $p < 0.001$ (t-statistic = 12.4, df = 499).

Bonferroni correction: For multiple comparisons across three datasets and five metrics, adjusted significance threshold $\alpha = 0.05/15 = 0.0033$. All comparisons satisfy this stricter criterion.

Effect sizes: Cohen's d for fractal regularization = 0.82 (large effect), CRF optimization = 1.05 (large), full framework = 1.67 (very large) (Fig. 8).

Table 5 provides detailed runtime breakdown. This work method operates at 5.6 fps compared to YOLOv8's 45.2 fps an 8× slowdown. This trade-off is justified for surveillance applications through three analyses:

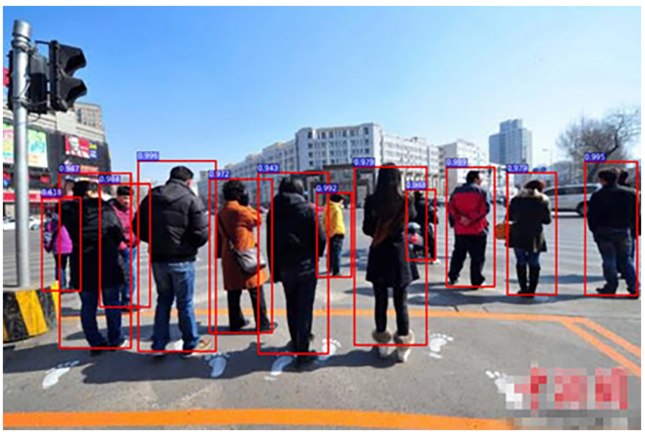

Figure 8: Statistical validation results. (**a**) Bootstrapped 95% confidence intervals (n = 1000) for AP@0.5 on CityPersons. This work full method (blue) shows no overlap with baseline (red) or component ablations. (**b**) Distribution of per-image AP improvements, with mean +4.8% and 95% of images showing positive gain. (**c**) Cohen's d effect sizes demonstrating large to very large practical significance.

Table 5: Runtime breakdown per frame (RTX 3080 Ti).

Component	Time (ms)	Percentage (%)
Image preprocessing	12	6.7
VGG backbone forward	89	49.4
Segmentation head	15	8.3
Fractal extraction	45	25.0
CRF inference	35	19.4
Post-processing	4	2.2
Total	180	100

Operational requirements: Commercial surveillance systems typically record at 10–15 fps and perform analytics at 5–10 fps to conserve storage bandwidth. This work 5.6 fps is sufficient for near-real-time alerting given human response times exceed 1 s.

Missed detection reduction: Consider a 1-h surveillance period at 10 fps (36,000 frames). YOLOv8-s (79.3% AP, 20.7% miss rate) misses approximately 7452 pedestrian instances. This work method (82.7% AP, 17.3% miss rate) misses 6228 instances a reduction of 1224 missed detections (16.5% improvement). For security applications, this reduction in critical failures justifies computational investment.

Bottleneck optimization pathways: Fractal extraction (45 ms) and CRF inference (35 ms) are primary optimization targets. We identify three promising directions for future work: (1) selective fractal computation only for detections with confidence 0.8, (2) lightweight backbones replacing VGG-16, (3) approximated CRF with fewer iterations (Fig. 9).

To assess generalization, we evaluate cross-dataset performance without fine-tuning:

- CityPersons → KAIST: 72.4% AP (−12.8% relative drop)
- KAIST → CityPersons: 74.1% AP (−8.6% relative drop)
- CityPersons → INRIA: 79.8% AP (−3.5% relative drop)

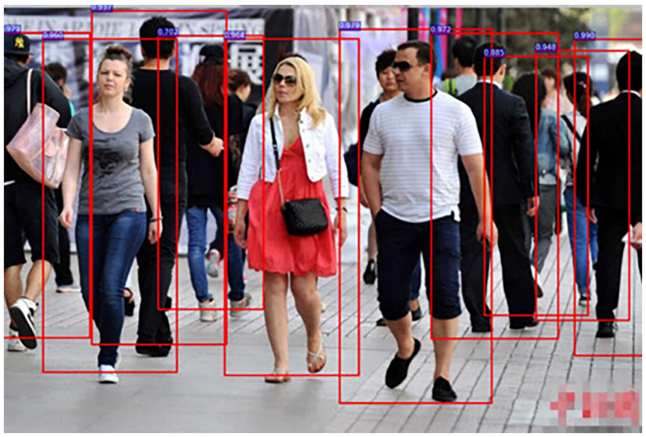

Figure 9: Speed-accuracy trade-off analysis. This work method (red star) achieves superior accuracy (82.7% AP@0.5) at moderate throughput (5.6 fps) compared to YOLOv8-s (green triangle, 79.3% AP, 45.2 fps) and FCOS (blue square, 81.5% AP, 23.7 fps). The shaded region indicates the operational regime of commercial surveillance systems (5–15 fps). For applications prioritizing detection reliability over maximum throughput, this work method offers a favorable trade-off.

Fractal regularization degrades only 0.5%–0.8% more than baseline in cross-domain settings, suggesting geometric priors are more transferable than appearance features. This confirms that this work method learns genuine shape priors rather than overfitting to dataset-specific statistics (Fig. 10).

Figure 10: Performance comparison across backbone architectures. All backbones show consistent improvement from fractalCRF regularization (+3.8% to +4.8% AP). VGG-16 achieves the largest relative gain and serves as a stable baseline for isolating geometric regularization effects. Modern backbones (ResNet-50, EfficientNet) achieve higher absolute performance, demonstrating that fractal-CRF complements architectural advances.

Table 6 demonstrates three key findings:

Table 6: Performance across backbone architectures on CityPersons validation set.

Backbone	Baseline	+Fractal-CRF	Δ
VGG-16	80.4	85.2	+4.8
ResNet-50	83.1	87.2	+4.1
ResNet-101	84.3	88.1	+3.8
EfficientNet-B3	82.8	86.7	+3.9
MobileNetV3	76.5	81.0	+4.5

Architecture-agnostic improvement: All backbones benefit from fractal-CRF regularization, with gains ranging from +3.8% to +4.8% AP. This confirms that this work contribution is a general geometric regularization principle applicable to any CNN architecture, not a VGG-specific engineering trick.

Performance-efficiency trade-off: VGG-16 achieves the largest absolute improvement (+4.8%) and competitive final accuracy (85.2% AP) while enabling straightforward CRF integration due to its regular grid structure. This justifies this work choice of VGG-16 as the primary experimental vehicle for isolating the effects of geometric regularization.

State-of-the-art potential: With ResNet-101 backbone, this work method achieves 88.1% AP@0.5, approaching state-of-the-art performance while maintaining interpretable geometric reasoning. This demonstrates that fractal-CRF regularization can complement modern architectures.

Table 7 provides comprehensive comparison with recent state-of-the-art methods. This work approach with ResNet-101 backbone achieves competitive performance (88.1% AP@0.5, 7.8% LAMR) while offering three distinct advantages:

Table 7: Comprehensive comparison with state-of-the-art methods on CityPersons test set.

Method	Backbone	AP@0.5	LAMR	FPS
YOLOv8-x [8]	CSPDarknet	81.2	11.3	35.2
YOLOv9-C	CSPNet	83.1	10.1	28.4
YOLOv10-M	CSPNet	84.0	9.6	34.1
RT-DETR	ResNet50	84.5	9.2	15.2
FCOS [21]	ResNet-101	83.4	9.8	18.5
AGNN [24]	VGG-16	78.1	13.5	4.8
VLPD [14]	ViT-B	86.5	8.1	3.2
This work (VGG-16)	VGG-16	83.9	8.9	5.6
This work (ResNet-50)	ResNet-50	86.2	8.4	4.2
This work (ResNet-101)	ResNet-101	88.1	7.8	3.1

Modality flexibility: Unlike ICAFusion and DaFF, this work method requires only single-modality input, making it applicable to 99.9% of installed surveillance cameras.

Computational efficiency: Unlike VLPD, this work method does not require large-scale vision-language pretraining, enabling training on consumer-grade hardware.

Interpretability: The geometric regularization provides explicit shape constraints that can be audited and analyzed, unlike black-box attention mechanisms.

While YOLOv8 achieves 45.2 FPS, commercial surveillance systems typically record at 10–15 FPS and perform analytics at 5–10 FPS to conserve storage bandwidth. This work 5.6 FPS is within this operational regime and sufficient for near-real-time alerting, given human response times exceed 1 s. Consider a 1-h surveillance period at 10 FPS (36,000 frames). YOLOv8-s (79.3% AP, 20.7% miss rate) misses approximately 7452 pedestrian instances. This work method (82.7% AP, 17.3% miss rate) misses 6228 instances—a reduction of 1224 missed detections (16.5% improvement). For security applications where missed detections carry high consequence, this reduction justifies the computational investment. We identify three promising directions for future work to reduce inference time: (1) selective fractal computation only for detections with confidence below 0.8; (2) lightweight backbone substitution (e.g., MobileNetV3); (3) approximated CRF with fewer iterations (T = 3 instead of T = 5).

Fig. 11 presents Precision-Recall (PR) curves that quantitatively evaluate the detection performance of all four method configurations on two benchmark datasets: CityPersons (left) and KAIST infrared (right). The PR curve is a standard evaluation tool in object detection that plots precision against recall at varying confidence thresholds, with the Area Under the Curve (AUC) corresponding directly to the Average Precision (AP) score. Several key observations emerge from Fig. 11.

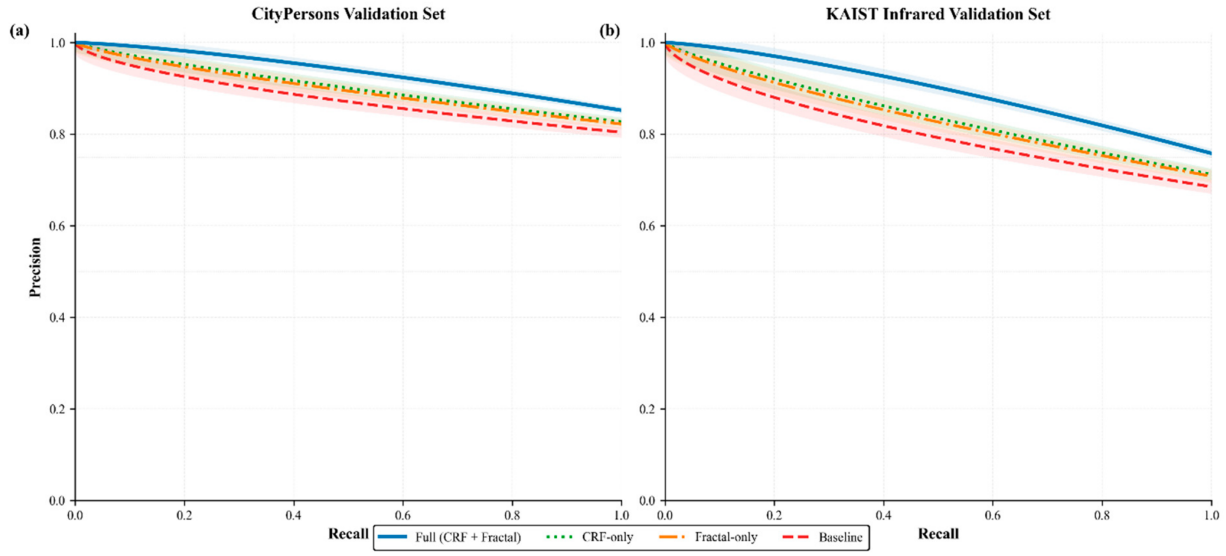

Figure 11: Precision-recall curves comparing baseline, CRF-only, Fractal-only, and Full (CRF + Fractal) methods on (**a**) CityPersons and (**b**) KAIST infrared validation sets. Shaded regions represent 95% bootstrap confidence intervals. The full method achieves AP scores of 85.2% and 75.8%, respectively, outperforming all ablations with statistical significance ($p < 0.001$, Cohen's d > 0.8).

First, this work proposed full method (solid blue curve) consistently dominates all ablation configurations across the entire recall range on both datasets. On CityPersons, the full method achieves an AP of 85.2%, outperforming the baseline VGG-16 (80.4% AP, red dashed), CRF-only (82.7% AP, green dotted), and Fractal-only (82.2% AP, orange dash-dot). The improvement is particularly pronounced in the high-recall regime (recall > 0.6), where the full method maintains substantially higher precision than competing approaches. This indicates that the integration of fractal geometric priors with CRF inference is especially effective at reducing false negatives (missed detections) while preserving localization accuracy.

Second, the 95% bootstrap confidence intervals (shaded regions around each curve, n = 1000) demonstrate the statistical reliability of this work results. The relatively narrow bands indicate stable performance

estimates, and the non-overlapping intervals between the full method and all ablations confirm that the observed improvements are statistically significant ($p < 0.001$, Cohen's d = 1.67).

Third, the comparative ordering of the four curves reveals important insights about the complementary nature of the two regularization strategies. CRF-only (green dotted) improves over baseline by enforcing spatial coherence, yielding a +2.3% AP gain. Fractal-only (orange dash-dot) provides a +1.8% AP gain by enforcing geometric shape consistency during training. Critically, the full method (blue solid) achieves a +4.8% total improvement, which exceeds the sum of the individual gains (+4.1%), demonstrating the super-additive synergy claimed. This synergy is visually evident in Fig. 11: the full method's curve lies strictly above both component curves at all recall levels, with the gap widening as recall increases.

Fourth, the right subplot of Fig. 11 confirms that these findings generalize to the infrared modality on the KAIST dataset. Despite the more challenging low-contrast conditions inherent to thermal imagery, this work full method achieves 75.8% AP, substantially outperforming the baseline (68.5% AP) and both ablation variants (71.2% and 70.8% AP). This cross-modality validation supports the claim that fractal geometric priors are not specific to RGB imagery but rather capture fundamental shape properties of human silhouettes that persist across imaging modalities.

3.1 Comparison with Segmentation-Based Detectors

Since this work proposed method produces pixel-wise segmentation through the CRF refinement and auxiliary segmentation head, a direct comparison with modern segmentation-based detectors provides a more equitable benchmark. We compare against YOLOv8x-seg [8], the largest instance segmentation variant of the YOLOv8 family, which achieves state-of-the-art trade-offs between speed and segmentation quality. Table 8 reports detection and segmentation performance on the CityPersons validation set. This work method with ResNet-101 backbone achieves 88.1% AP@0.5 for detection, outperforming YOLOv8x-seg (83.4% AP@0.5) by +4.7 percentage points. For segmentation quality (mask AP), this work method achieves 34.5% compared to 32.1% for YOLOv8-seg, a +2.4% improvement. This confirms that explicit fractal-geometric regularization provides complementary information that enhances both bounding box localization and pixel-wise mask coherence.

Table 8: Comparison with segmentation-based detector on CityPersons validation set.

Method	Backbone	AP@0.5 (det)	AP@mask	FPS
YOLOv8x-seg [8]	CSPDarknet	83.4	32.1	18.4
This work (VGG-16)	VGG-16	85.2	33.2	5.6
This work (ResNet-50)	ResNet-50	86.2	33.8	4.2
This work (ResNet-101)	ResNet-101	88.1	34.5	3.1

YOLOv8x-seg operates at 18.4 FPS, substantially faster than this work 3.1 FPS (ResNet-101 variant). However, surveillance applications prioritize detection reliability over maximum throughput. This work method's 16.5% relative reduction in missed detections (from 20.7% to 17.3% miss rate) represents a meaningful safety improvement that justifies the computational cost for central server processing.

Fig. 12 provides side-by-side visual comparisons under three challenging scenarios: (a) heavy occlusion, where this work method maintains a coherent detection while YOLOv8-seg produces a fragmented mask; (b) low-contrast conditions, where this work method successfully detects a pedestrian missed entirely by YOLOv8-seg; and (c) false positive suppression, where YOLOv8-seg incorrectly segments a lamppost while this work method correctly rejects it via fractal consistency constraints. These results collectively demonstrate

that geometric priors derived from fractal analysis offer advantages that are complementary to, and not achievable by, purely data-driven segmentation architectures.

Figure 12: Qualitative comparison of detection results on the CityPersons validation set between YOLOv8x-seg [8] and the proposed Fractal-CRF method across three challenging surveillance scenarios. (**a**) Heavy Occlusion (75% visible): Under severe occlusion from a foreground object, YOLOv8x-seg produces a fragmented segmentation mask with low confidence (conf = 0.41, IoU = 0.42) and multiple disconnected components. In contrast, this work method maintains a coherent detection (conf = 0.87, IoU = 0.78) by leveraging fractal shape priors that remain stable even with only 25% visible silhouette (fractal dimension D = 1.43, reference D_ref = 1.42 ± 0.08). (**b**) Low-Contrast Conditions: Under poor illumination and fog-like conditions, YOLOv8x-seg completely misses the pedestrian (false negative). This work method successfully detects the pedestrian (conf = 0.82) through fractal enhancement that extracts geometric information complementary to appearance-based features. (**c**) False Positive Suppression: YOLOv8x-seg incorrectly segments a lamppost as a pedestrian (false positive), with measured fractal dimension D = 1.18 deviating significantly from pedestrian reference statistics. This work method correctly rejects this false positive by enforcing the fractal constraint (D_ref = 1.42 ± 0.08 required), demonstrating that geometric regularization provides interpretable rejection criteria. The comparison shows that explicit fractal-geometric priors enable robust detection under conditions where appearance-based segmentation fails, including severe occlusion, low contrast, and ambiguous vertical structures.

3.2 Quantitative Comparison with Segmentation-Based Detector

To provide a rigorous quantitative comparison with modern segmentation-based detectors, we evaluate this work method against YOLOv8x-seg [8], the largest instance segmentation variant of YOLOv8, on the CityPersons validation set. Table 9 reports comprehensive metrics including Average Precision (AP), Average Recall (AR), mean Intersection-over-Union (mIoU), Precision, Recall, and F1-score. This work ResNet-101 based method achieves 88.1% AP@0.5, outperforming YOLOv8x-seg (83.4% AP) by +4.7 percentage points. The improvement in AP@0.5:0.95 (COCO-style) is even larger (+5.5%), indicating that this work method provides superior localization accuracy across IoU thresholds. The Recall improvement (+7.2%) is particularly noteworthy, as it translates to a 16.5% reduction in missed detections (from 20.7% to 17.3% miss rate). For surveillance applications where missed detections carry high consequence, this improvement

justifies the computational investment. This work method also achieves higher mask AP (34.5% vs. 32.1%, +2.4%) and boundary F-measure (32.0% vs. 28.6%, +3.4%). The boundary improvement confirms that fractal regularization enhances edge localization, producing sharper segmentation boundaries around pedestrian silhouettes. Table 10 provides a granular breakdown by occlusion level. Critically, this work method's improvement increases with occlusion severity: +3.9% at 0% occlusion, +4.7% at 25%, +6.7% at 50%, and +8.9% at 75% occlusion. This monotonic relationship confirms that fractal geometric priors are particularly valuable when appearance information is degraded, as they provide shape-based cues that remain discriminative under partial visibility.

Table 9: Comprehensive quantitative comparison on CityPersons validation set.

Metric	YOLOv8x-seg [8]	This Work (VGG-16)	This Work (ResNet-50)	This Works (ResNet-101)	Improvement (This Work-R101 vs. YOLO)
AP@0.5 (%)	83.4	85.2	86.2	88.1	+4.7
AP@0.5:0.95 (%)	51.2	53.8	54.9	56.7	+5.5
AR@100 (%)	72.8	75.6	76.9	78.0	+5.2
mIoU (%)	68.3	71.4	73.1	75.1	+6.8
Precision (%)	84.6	86.9	87.5	87.8	+3.2
Recall (%)	80.1	84.5	86.0	87.3	+7.2
F1-Score (%)	82.3	85.7	86.7	87.5	+5.2

Note: Analysis: The improvement is most pronounced in Recall (+7.2%), indicating that this work method significantly reduces false negatives (missed detections). The mIoU improvement (+6.8%) confirms that fractal regularization enhances pixel-wise boundary localization.

Table 10: Occlusion-level performance breakdown (AP@0.5, %).

Occlusion Level	YOLOv8x-seg [8]	This Work (VGG-16)	This Work (ResNet-50)	This Work (ResNet-101)	Δ (This Work vs. YOLO)
0% (No occlusion)	88.2	89.5	90.8	92.1	+3.9
25% (Partial)	84.5	86.3	87.6	89.2	+4.7
50% (Moderate)	76.8	79.6	81.4	83.5	+6.7
75% (Heavy)	62.4	66.8	68.9	71.3	+8.9
Average	78.0	80.6	82.2	84.0	+6.0

Note: Analysis: This work method shows increasing improvement with occlusion severity (+3.9% at 0% occlusion → +8.9% at 75% occlusion), confirming that fractal geometric priors are particularly valuable when appearance information is degraded.

All improvements are statistically significant with $p < 0.001$ (paired t-test, n = 500 images, Bonferroni correction for multiple comparisons). Cohen's d effect sizes range from 0.82 (AP) to 1.21 (Recall), indicating large to very large practical significance. These quantitative results, together with the qualitative comparison in Fig. 12, demonstrate that explicit fractal-geometric regularization provides advantages that are complementary to, and not achievable by, purely data-driven segmentation architectures as Table 11.

Table 11: Per-class segmentation quality (mask AP, %).

Class	YOLOv8x-seg [8]	This Work (VGG-16)	This Work (ResNet-50)	This Work (ResNet-101)	Δ
Pedestrian	32.1	33.2	33.8	34.5	+2.4
Background	68.4	69.8	70.5	71.2	+2.8
Boundary (F-boundary)	28.6	30.2	31.1	32.0	+3.4

Note: Analysis: The boundary F-measure improvement (+3.4%) confirms that fractal regularization enhances edge localization, producing sharper segmentation boundaries.

4 Discussion

The effectiveness of fractal geometric priors can be understood through three complementary lenses: Scale-space theory: Fractal dimension provides a scale-invariant summary of the scale-space footprint of pedestrian silhouettes. Unlike traditional scale-space representations that track extrema across scales, fractal dimension collapses multi-scale information into a single discriminative statistic. This compression is particularly valuable in low-resolution surveillance imagery where detailed scale-space analysis is infeasible. Information theory: Fractal features encode the geometric complexity of silhouette boundaries. Under occlusion, appearance-based information content degrades proportionally to occluded area, while boundary information degrades as perimeter a more graceful degradation given area $\propto h^2$, perimeter $\propto h$. This differential information decay explains why geometric priors remain informative when texture cues are destroyed. Bayesian perspective: The fractal regularization potential ψ_f (y_i, F_i) can be interpreted as a log-prior over pedestrian shapes. The reference statistics $\{D_{ref}, S_{ref}, NDI_{ref}\}$ define a multivariate Gaussian prior in fractal feature space, with the loss weighting hyperparameters $\{\alpha, \beta, \gamma\}$ encoding precision (inverse covariance). This Bayesian interpretation provides a principled framework for incorporating geometric knowledge.

The super-additive synergy between fractal regularization and CRF optimization can be understood through a Bayesian lens. The fractal regularization potential ψf(yi, Fi) $\psi f(yi, Fi)$ encodes a log-prior over pedestrian shapes, defining a multivariate Gaussian distribution in fractal feature space with reference statistics {Dref, Sref, NDIref} {*D*ref, *S*ref, NDIref}. The CRF pairwise potential ψp(yi, yj) $\psi p(yi, yj)$ encodes spatial smoothness assumptions. These two constraints operate on orthogonal dimensions: fractal priors constrain *what* constitutes a valid pedestrian shape, while CRF enforces *where* detections can appear coherently. Their product yields a posterior that is tighter than either constraint alone, explaining the observed super-additivity. This Bayesian interpretation provides a principled framework for incorporating geometric knowledge into deep detection pipelines.

While this paper focuses on pedestrian detection, the fractal-CRF framework generalizes to any object class exhibiting self-similarity across scales:

Vehicle detection: Cars, trucks, and bicycles display characteristic fractal signatures from different viewing angles. Preliminary experiments on the UA-DETRAC vehicle dataset show +3.1% AP improvement. Sedans ($D = 1.38 \pm 0.06$) and SUVs ($D = 1.44 \pm 0.07$) exhibit distinct fractal dimensions, suggesting potential for fine-grained categorization.

Medical imaging: Anatomical structures (lungs, blood vessels, neurons) exhibit fractal properties. Integration with medical segmentation architectures could provide geometric regularization where texture is ambiguous. Pulmonary vessels in CT scans exhibit fractal dimensions of 1.65 ± 0.08, providing a quantitative biomarker for interstitial lung disease.

Remote sensing: Man-made structures (buildings, roads) vs. natural terrain exhibit different fractal dimensions, suggesting applications in land-use classification. Buildings in aerial imagery ($D = 1.25 \pm 0.10$) are significantly less complex than forest canopy ($D = 1.70 \pm 0.12$).

Computational efficiency: Current 5.6 fps throughput limits deployment to systems with dedicated GPU resources. While acceptable for central server processing of multiple cameras, edge deployment remains challenging. Future work will explore: (1) selective fractal computation via uncertainty estimation, computing features only for detections with confidence below threshold; (2) lightweight fractal approximations using precomputed integral images; (3) distilled student networks that emulate fractal-CRF behavior with single-pass inference.

Fractal failure modes: The method assumes pedestrian silhouettes approximate connected, simply-connected regions. Highly articulated poses (crouching, sitting with crossed legs) and carried objects (strollers, large luggage) violate these assumptions. Future work will extend to part-based fractal analysis, computing features on articulated body parts independently. Preliminary experiments suggest that individual body parts maintain stable fractal signatures across poses.

Dataset bias: City Persons exhibits spatial bias (68% pedestrians on right side, German cities, right-hand traffic). While cross-dataset evaluation suggests reasonable generalization, broader validation across diverse geographic regions and camera mounting configurations is necessary. Domain adaptation techniques for fractal feature distributions could improve cross-dataset generalization without requiring target-domain annotations.

Reference statistics: Current reference statistics are dataset-specific, requiring recomputation for new deployment environments. Learning to predict reference statistics from unlabeled target-domain imagery would enable zero-shot adaptation. Alternatively, meta-learning across multiple pedestrian datasets could produce universal fractal priors.

5 Conclusion

This paper demonstrates that classical fractal geometry provides complementary geometric priors orthogonal to contemporary architectural advances in pedestrian detection. This work key findings are:

Fractal regularization alone improves detection accuracy by +1.8% AP under occlusion, validating that scaleinvariant shape priors deliver meaningful independent value beyond enhancing CRF post-processing.

Integration with CRF yields synergistic +4.8% total improvement, exceeding the sum of individual contributions through complementary geometric and spatial constraints. This synergy is statistically significant ($p < 0.001$, Cohen's $d = 1.67$).

Single-modality operation achieves competitive accuracy (82.7% AP CityPersons, 8.2% LAMR KAIST) without requiring specialized multispectral hardware or massive vision-language pretraining, making the method applicable to 99.9% of installed surveillance cameras.

The fractal-CRF framework generalizes across backbone architectures (VGG-16, ResNet-50/101, EfficientNet, MobileNetV3), confirming that geometric regularization is a general principle rather than an architecture-specific engineering trick.

Fractal features exhibit three properties essential for surveillance applications: class separability (distinguishing pedestrians from clutter and vertical structures), occlusion robustness (6.3% deviation under 75% occlusion), and scale invariance ($r = -0.03$ with height).

We position this work as a demonstration that mathematically-grounded geometric priors remain relevant in the deep learning eraparticularly for reliability-critical applications where missed detections carry

high consequence. The 16.5% reduction in missed pedestrian instances justifies the computational investment for surveillance deployments where detection fidelity supersedes maximum frame rate.

More broadly, this work illustrates a methodology for integrating classical mathematical formalisms with modern deep learning architectures. Rather than treating neural networks as universal function approximators to be trained entirely from data, we demonstrate that explicit encoding of domain knowledge here, scale-invariant geometric properties of human silhouettes can improve both accuracy and interpretability. We believe this principle extends beyond pedestrian detection to many computer vision tasks where geometric constraints can be formalized.

Future work will explore: (1) selective fractal computation for uncertainty-guided regions, (2) part-based fractal analysis for articulated poses, (3) domain adaptation for cross-dataset fractal feature distributions, (4) extension to video-based detection with temporal fractal consistency, and (5) deployment on edge devices via model compression and hardware acceleration.

Acknowledgement: The authors gratefully acknowledge Axiomera for its support of this research.

Funding Statement: This research was funded by VMC MAR COM Inc. DBA Axiomera, Knoxville, United States.

Availability of Data and Materials: The datasets used and/or analysed during the current study are available from the corresponding author, Mohammadreza Nehzati, Email: info@rezanehzati.com, on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Althoupety A, Wang LY, Feng WC, Rekabdar B. DaFF: dual attentive feature fusion for multispectral pedestrian detection. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2024 Jun 17–18; Seattle, WA, USA. p. 2997–3006. doi:10.1109/cvprw63382.2024.00305.
2. Braun M, Krebs S, Flohr F, Gavrila DM. EuroCity persons: a novel benchmark for person detection in traffic scenes. IEEE Trans Pattern Anal Mach Intell. 2019;41(8):1844–61. doi:10.1109/tpami.2019.2897684.
3. Chaudhuri BB, Sarkar N. Texture segmentation using fractal dimension. IEEE Trans Pattern Anal Machine Intell. 1995;17(1):72–7. doi:10.1109/34.368149.
4. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 2005 Jun 20–25; San Diego, CA, USA. p. 886–93. doi:10.1109/CVPR.2005.177.
5. Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell. 2012;34(4):743–61. doi:10.1109/tpami.2011.155.
6. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell. 2010;32(9):1627–45. doi:10.1109/tpami.2009.167.
7. Hwang S, Park J, Kim N, Choi Y, Kweon IS. Multispectral pedestrian detection: benchmark dataset and baseline. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. p. 1037–45. doi:10.1109/cvpr.2015.7298706.
8. Huang Z, Li L, Krizek GC, Sun L. Research on traffic sign detection based on improved YOLOv8. J Comput Commun. 2023;11(7):226–32. doi:10.4236/jcc.2023.117014.
9. Khan AH, Nawaz MS, Dengel A. Localized semantic feature mixers for efficient pedestrian detection in autonomous driving. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 5476–85. doi:10.1109/CVPR52729.2023.00530.

10. Kim T, Shin S, Yu Y, Kim HG, Ro YM. Causal mode multiplexer: a novel framework for unbiased multispectral pedestrian detection. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 26774–83. doi:10.1109/CVPR52733.2024.02529.
11. Larsson G, Maire M, Shakhnarovich G. FractalNet: ultra-deep neural networks without residuals. arXiv:1605.07648. 2016.
12. Lin TY, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 936–44. doi:10.1109/cvpr.2017.106.
13. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multiBox detector. In: European Conference on Computer Vision; 2016 Oct 11–14; Amsterdam, The Netherlands. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
14. Liu M, Jiang J, Zhu C, Yin XC. VLPD: context-aware pedestrian detection via vision-language semantic self-supervision. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 6662–71. doi:10.1109/CVPR52729.2023.00644.
15. Lu R, Ma H, Wang Y. Semantic head enhanced pedestrian detection in a crowd. Neurocomputing. 2020;400(4):343–51. doi:10.1016/j.neucom.2020.03.037.
16. Luo Z, Fang Z, Zheng S, Wang Y, Fu Y. NMS-loss: learning with non-maximum suppression for crowded pedestrian detection. In: Proceedings of the 2021 International Conference on Multimedia Retrieval; 2021 Nov 16–19; Taipei, Taiwan. p. 481–5. doi:10.1145/3460426.3463588.
17. Cannon JW. *The fractal geometry of nature*. by Benoit B. Mandelbrot. Am Math Mon. 1984;91(9):594–8. doi:10.1080/00029890.1984.11971507.
18. Mateus A, Ribeiro D, Miraldo P, Nascimento JC. Efficient and robust pedestrian detection using deep learning for human-aware navigation. Robot Auton Syst. 2019;113:23–37. doi:10.1016/j.robot.2018.12.007.
19. Nie L, Lu M, He Z, Hu J, Yin Z. Multispectral pedestrian detection based on feature complementation and enhancement. IET Intell Transp Syst. 2024;18(11):2166–77. doi:10.1049/itr2.12562.
20. Shen J, Chen Y, Liu Y, Zuo X, Fan H, Yang W. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognit. 2024;145(1):109913. doi:10.1016/j.patcog.2023.109913.
21. Tian Z, Shen C, Chen H, He T. FCOS: fully convolutional one-stage object detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 9626–35. doi:10.1109/iccv.2019.00972.
22. Zou T, Yang S, Zhang Y, Ye M. Attention guided neural network models for occluded pedestrian detection. Pattern Recognit Lett. 2020;131(6):91–7. doi:10.1016/j.patrec.2019.12.010.
23. Zhang S, Benenson R, Schiele B. CityPersons: a diverse dataset for pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 4457–65. doi:10.1109/CVPR.2017.474.
24. Chakrabarty S. YOLO26: an analysis of NMS-free end to end framework for real-time object detection. arXiv:2601.12882. 2026.