Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evidence-Gated Local and Ring Context Modeling for Tiny Person Detection

Peipei Song
Yantai Engineering & Technology College
Fengxiang Pan
Yantai Engineering & Technology College
Jian Li
lijian@ytetc.edu.cn

Yantai Engineering & Technology College

Research Article

Keywords: tiny person detection, contextual bias, region proposals, background intervention, evidence gating, consistency regularization

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11704377/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Evidence-Gated Local and Ring Context Modeling for Tiny Person Detection

Peipei Song[1], Fengxiang Pan[1], Jian Li[1*]

[1] Yantai Engineering & Technology College, Yantai, Shandong, China

* Corresponding author: Jian Li, lijian@ytetc.edu.cn

Abstract Tiny pedestrian detectors need context when the target contains too few pixels for reliable appearance recognition, but context can also encode background correlations that are unstable across scenes. We formulate the problem as conditional use of context at each predicted proposal. Two independently parameterized encoders receive a target-neighborhood crop and a wider ring crop sampled directly from the RGB image; the central neighborhood is removed from the ring before resizing. An evidence gate modulates a residual context contribution using proposal scale, local classification entropy, valid ring fraction, and disagreement between the two embeddings. During training, a protected intervention changes distant background while leaving annotated objects, ignored regions, and their neighborhoods intact. Detached proposals from the original image are reused at identical coordinates in both views for paired classification and localization consistency; the proposal network is supervised separately on each view. A single exploratory run on TinyPerson subsets completed 24 epochs and 1,536 optimization steps. Development/test AP50 was 0.0299/0.0420, tiny-target AP50 was 0.0266/0.0334, and proposal Recall@1000 was 0.1366/0.1124. The gate remained finite and nonsaturated, but no matched baseline, multiple-seed estimate, or external-domain test is available. The evidence supports the feasibility of explicit pixel-space context separation and identifies proposal coverage as a major bottleneck; it does not yet establish an accuracy improvement or cross-scene robustness.

Keywords tiny person detection; contextual bias; region proposals; background intervention; evidence gating; consistency regularization

1 Introduction

Tiny pedestrians are common in long-range surveillance, elevated viewpoints, and maritime search. Their bounding boxes may contain only a small number of discriminative pixels. Blur, compression, occlusion, and the downsampling inherent in convolutional detectors further weaken appearance cues. The TinyPerson benchmark and scale-matched training work make clear that object size is a central difficulty rather than a minor variant of standard detection[1]. Feature pyramids and multi-scale processing preserve finer representations[2–5], but a candidate may still be recognized largely through the scene surrounding it.

Context is not inherently spurious. For an ambiguous cluster of pixels, a nearby human, boat, or platform can provide useful evidence. Yet a model that equates a particular coastline, water texture, building, or group layout with the presence of a pedestrian may fail when the acquisition environment changes. Studies of contextual bias and distribution shift show that an overall held-out score can hide reliance on unstable shortcuts[6–9]. Tiny-object detection therefore has two linked requirements: keep useful nearby cues and constrain the use of distant background. Fixed context windows and conventional feature-space RoI operations make that distinction difficult because deep features already aggregate information over a broad receptive field.

We introduce a proposal-level detector that explicitly constructs two views from raw pixels. A small local crop retains the candidate and close neighborhood; a larger crop becomes a context ring when its center is masked before resizing. Separate convolutional encoders produce embeddings that meet only at an evidence-conditioned residual fusion point. The gate observes the proposal size, uncertainty of a local-only classifier, amount of valid ring area, and disagreement between the views. This is a testable design for context allocation, not a guarantee that the resulting model has learned a causal decision rule.

We also construct a paired training view by modifying only distant unprotected background pixels. A shared proposal set makes the two final-head predictions directly comparable at fixed locations, while each image retains independent RPN supervision. This separates, within the loss definition, candidate-level head sensitivity from the effect of the intervention on proposal generation. At inference, the detector receives only an RGB image, predicts its own proposals, and does not require annotations, masks, source labels, or background exemplars.

The present empirical record contains one lightweight implementation and one training run. It reports development and test subset metrics, proposal recall, gate statistics, proxy-source variation, and runtime.

We use those measurements as a transparent pilot evaluation, without converting missing baselines or proposed audits into completed findings. The main contributions are the explicit spatial definition of local and ring evidence, evidence-gated residual fusion, and a protected paired-view objective, accompanied by a diagnostic assessment of the observed run.

2 Related work

2.1 Tiny-object detection and scale

Two-stage detectors separate candidate generation from candidate refinement[10]. FPN connects semantically rich and high-resolution features across scales[2]; SNIP and SNIPER address object-size mismatch with scale-selective training[3,4]. AutoFocus and image slicing concentrate computation on small image regions[5,11]. TinyPerson highlights extreme pedestrian scale and proposes scale-matched pretraining[1]. Anchor-free detection, focal-loss classification, and Mask R-CNN provide alternative candidate or refinement mechanisms[12–14], while Soft-NMS changes post-processing[15]. These lines of work chiefly improve feature resolution, coverage, or ranking. Our head instead studies the spatial source and conditional contribution of context for a candidate, and we report RPN recall because no head can classify a missed proposal.

2.2 Spatial context and shortcut dependence

Context-aware detectors can use surrounding objects and layout to disambiguate an object[16,17]. That benefit may come with a dependence on co-occurrences and scene features: contextual-bias experiments explicitly test object recognition when usual backgrounds are changed[6], and background challenge datasets reveal similar sensitivities[7]. Invariant-risk and group-robust formulations emphasize performance across environments[8,9,18]. These works motivate reporting source-stratified outcomes, but filename-derived groups are not equivalent to audited acquisition environments. Unlike attention applied after a shared backbone, our two crop encoders separately ingest raw pixel neighborhoods. The proposal stage itself still uses the image-wide backbone and is not spatially isolated by this choice.

2.3 Interventions, adaptive fusion, and consistency

Cut-and-paste, Copy-Paste, CutMix, and mixup perturb object/background associations for augmentation[19–22]. For extremely small people, generic mixing can erase or relabel the target; a protection mask defines where alteration is permitted. Channel attention and gating offer adaptive feature combination[23,24], but the magnitude or distribution of a gate is not direct evidence of useful adaptation. Prediction consistency across perturbed views has been studied extensively[25,26]. We use a symmetric Jensen–Shannon term for class distributions[27] and a positive-proposal localization term. The paired prediction is evaluated at the same original proposal locations, making the question narrower and more interpretable than a comparison between two independently sampled candidate sets.

3 Proposed method

3.1 Detection pipeline and implementation boundary

Let $X \in \mathbb{R}^{3\times H\times W}$ be an image, B its annotated person boxes, and U its ignore regions. A multiscale backbone and shared region proposal network generate $q_i = (c_x, c_y, w, h)$ with objectness score s_i. Top-ranked candidates after filtering and suppression become the inputs to a local-and-context detection head. Fig. 1 is the supplied design diagram: it shows a ResNet-50-like stem[28], FPN levels P2–P6, a shared RPN, two crop encoders, gated residual fusion, and a paired-training branch. The measured experiment used a lightweight FPN-style proposal network rather than the full ResNet-50-FPN depicted; the result tables refer only to that measured implementation. The diagram should therefore be read as the method specification, with this capacity difference made explicit.

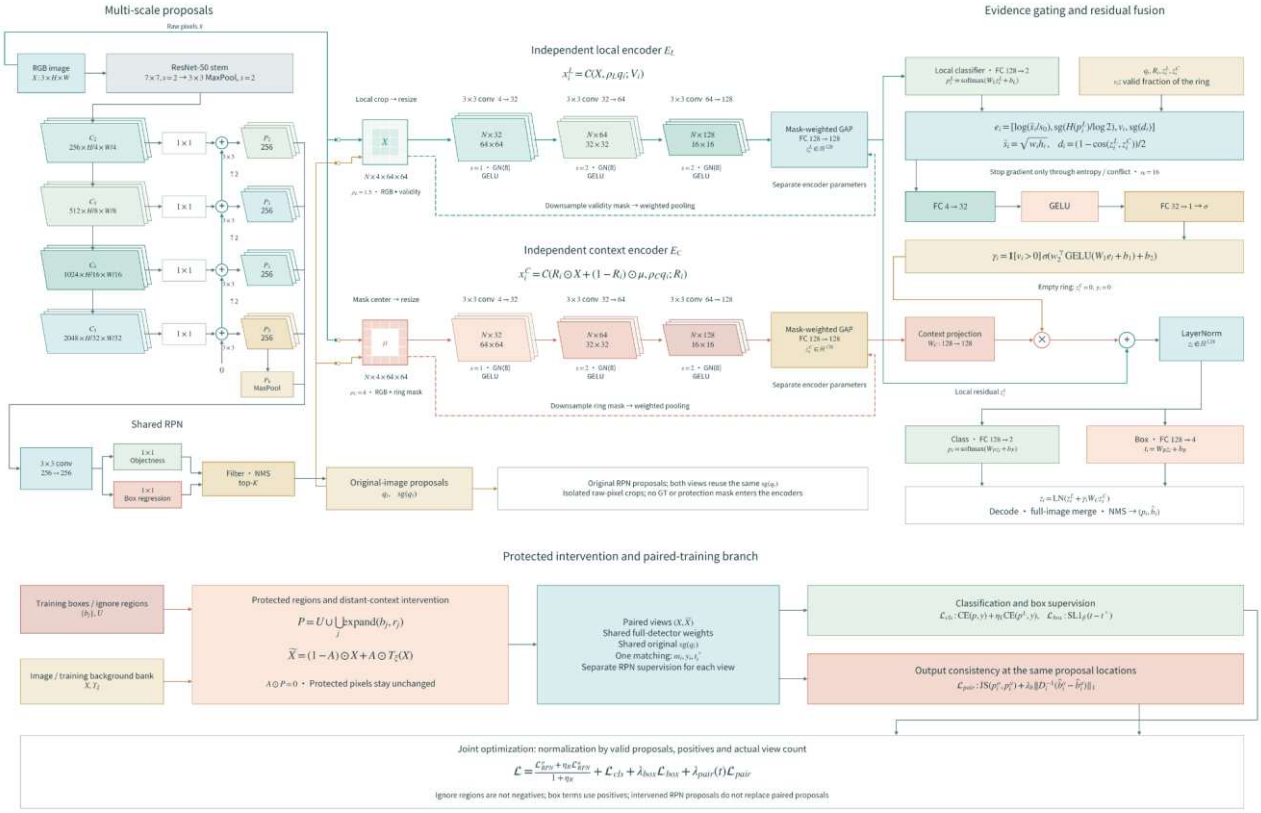

Fig. 1 Proposed architecture and training objective. The diagram depicts the intended ResNet-50-style pyramid; the reported pilot measurements use a lighter FPN-style proposal network. Training-only protected masks and paired views do not enter inference.

3.2 Raw-pixel local and ring views

For candidate q_i, let $\mathrm{crop}(X, q_i, \rho)$ extract a box centered at q_i with width ρw_i and height ρh_i, then resample it to 64 × 64. We use $\rho_{\mathrm{L}} = 1.5$ for the local neighborhood and $\rho_{\mathrm{C}} = 4.0$ for the outer context. The context crop has its central local region removed before resize, giving a ring rather than a second full patch containing the target. Each view concatenates an RGB image and a binary spatial-validity channel; out-of-image samples take a fixed RGB mean and validity zero. A ring mask also marks center pixels invalid. The local view therefore carries the person and close neighborhood, while the context encoder is prevented from directly observing those pixels.

$$x_i^{\mathrm{L}} = [\mathrm{crop}(X, q_i, \rho_{\mathrm{L}}), V_i^{\mathrm{L}}], \quad \rho_{\mathrm{L}} = 1.5,$$
$$x_i^{\mathrm{C}} = [\mathrm{crop}(X \odot (1 - R_i), q_i, \rho_{\mathrm{C}}), V_i^{\mathrm{C}}], \quad \rho_{\mathrm{C}} = 4.0. \tag{1}$$

Here R_i denotes the central local region in the outer window, and V_i^{L} and V_i^{C} encode pixels valid after boundary padding and ring masking. Because masking precedes resampling, interpolation cannot reintroduce central RGB information into the ring through the resize operation. The exact antialiasing and boundary convention must nevertheless be documented in an implementation release. Pixel-space separation applies to head inputs and not to the image-derived RPN proposals.

3.3 Separate encoders and valid-mask pooling

Independent encoders E_{L} and E_{C} each contain three 3 × 3 convolutional blocks with channel widths 32, 64, and 128, group normalization, and GELU. The latter two blocks downsample spatially. Feature maps are aggregated with validity-weighted global average pooling: the downsampled mask sets the contribution of invalid samples to zero, and the denominator is the valid-mask mass with a safe empty-mask convention. A learned 128-dimensional projection yields embeddings z_i^{L} and z_i^{C}. Separate parameters prevent the ring representation from being formed by simply reusing a shared feature tensor whose receptive field overlaps the local window. They also increase head computation, which must be accounted for when comparing to a local-only baseline.

$$z_i^k = W^k \frac{\sum_u V_i^k(u)\, E^k(x_i^k)(u)}{\max(\sum_u V_i^k(u), \varepsilon)}, \qquad k \in \{\mathrm{L}, \mathrm{C}\}. \tag{2}$$

3.4 Evidence-conditioned residual fusion

A supervised auxiliary classifier operating on z_i^{L} produces two-class probability p_i^{L}. Its normalized entropy h_i measures uncertainty of the target-neighborhood prediction. Let a_i denote normalized proposal log scale, v_i the fraction of valid ring pixels, and $d_i = (1 - \cos(z_i^{\mathrm{L}}, z_i^{\mathrm{C}}))/2$ a feature-disagreement score. A two-layer MLP maps $e_i = [a_i, h_i, v_i, d_i]$ into a scalar gate. We stop gradients through h_i and d_i when constructing the gate input, and force $\gamma_i = 0$ for an empty ring. Context is projected into the local embedding dimension and added as a gated residual, followed by layer normalization.

$$\gamma_i = \mathbf{1}[v_i > 0]\, \mathrm{sigmoid}(w_2^{\mathrm{T}} \mathrm{GELU}(W_1 e_i + b_1) + b_2),$$
$$z_i = \mathrm{LN}(z_i^{\mathrm{L}} + \gamma_i W^{\mathrm{C}} z_i^{\mathrm{C}}). \tag{3}$$

Classification and four-parameter box regression use z_i. The auxiliary local prediction receives its own classification supervision, so the local branch retains a direct label signal if the fusion path learns shortcuts. A larger γ_i does not necessarily mean "more useful context": the projected feature norm, its direction, and downstream classifier weights also determine influence. Interpreting the gate requires controlled perturbation and matched fixed-gate and input-feature ablations, which were not completed for the present run.

3.5 Protected distant-background pairing

Training first applies common geometric transforms so box coordinates stay aligned. Construct a protected set P from all training person boxes expanded by a neighborhood margin and from ignore regions U. An editable binary mask A is sampled only outside P. A transformation T_ξ, potentially using an image or background bank, changes the editable pixels while leaving protected pixels equal to those in the original image. The intervention is intended to vary distant context, not the object and its immediate neighbors. The recorded pixel-level check found zero differences inside the protected set. That check cannot establish photorealism of boundaries, correctness of missing annotations, or semantic plausibility of replacements.

$$P = U \cup \bigcup_j \mathrm{expand}(b_j, r_j), \qquad A \odot P = 0,$$
$$\tilde{X} = (1 - A) \odot X + A \odot T_\xi(X). \tag{4}$$

The RPN is run and supervised on each view. For the paired head comparison, proposals generated from X are detached and reused at the same image coordinates for $\tilde{X}$. Quality checks reject invalid pairs. Let p_i and $\tilde{p}_i$ be class distributions at matched coordinates, and let $\hat{b}_i$ and $\hat{\tilde{b}}_i$ be decoded boxes for a matched positive. The paired loss combines symmetric distribution consistency with a Smooth L1 box term; no localization consistency term is imposed on negative candidates. The loss formulation deliberately does not substitute the intervened image's proposals for the fixed paired proposals. If background change alters the intervened RPN, that sensitivity remains visible in the separate RPN task rather than being hidden by head alignment.

$$\mathcal{L}_{\mathrm{pair}} = \operatorname*{mean}_i \mathrm{JS}(p_i, \tilde{p}_i) + \lambda_b \operatorname*{mean}_{i \in \mathrm{positive}} \mathrm{SmoothL1}(\hat{b}_i - \hat{\tilde{b}}_i). \tag{5}$$

The full objective combines original and transformed-view RPN losses, supervised head classification and positive-only regression, auxiliary local classification, and a warmed-up paired term. Counts of valid proposals, positive boxes, and retained pairs are used to normalize their corresponding terms. Ignore areas do not supply background negatives. The exact numerical weights of the auxiliary, intervention-view, box, and paired terms, as well as transformation sampling probabilities, are missing from the supplied record and should be disclosed before attempting replication.

$$\mathcal{L} = \frac{\mathcal{L}_{\mathrm{RPN}}(X) + \eta_R \mathcal{L}_{\mathrm{RPN}}(\tilde{X})}{1 + \eta_R} + \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{box}} \mathcal{L}_{\mathrm{box}}$$
$$+ \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{local}} + \lambda_{\mathrm{pair}}(t) \mathcal{L}_{\mathrm{pair}}. \tag{6}$$

3.6 Inference

At inference, the model takes only X, forms its own multiscale proposals, extracts local and ring crops around them, computes the gate and final prediction, decodes boxes, merges tile predictions to the parent

image, and applies final suppression. Neither the ground-truth protection mask nor a background-bank image is needed. This distinction is operationally important because the intervention mechanism is a training regularizer rather than a deployment input.

4 Experimental design

4.1 Dataset and split construction

The pilot used TinyPerson images and annotations[1]. Training images were drawn from the annotated training split; a development subset and a test subset were evaluated with window-level inference followed by parent-image aggregation. Parent filenames were used as a proxy for source identity when grouping derivatives, and all windows from one parent were kept together. The supplied record does not state the number of parent images, windows, or ground-truth people in each subset, the exact tile size and overlap, or the official evaluator version. Consequently, the reported numbers are measurements from this subset protocol, not directly comparable to published TinyPerson leaderboard results. A filename is only a proxy for acquisition source; it cannot verify source independence.

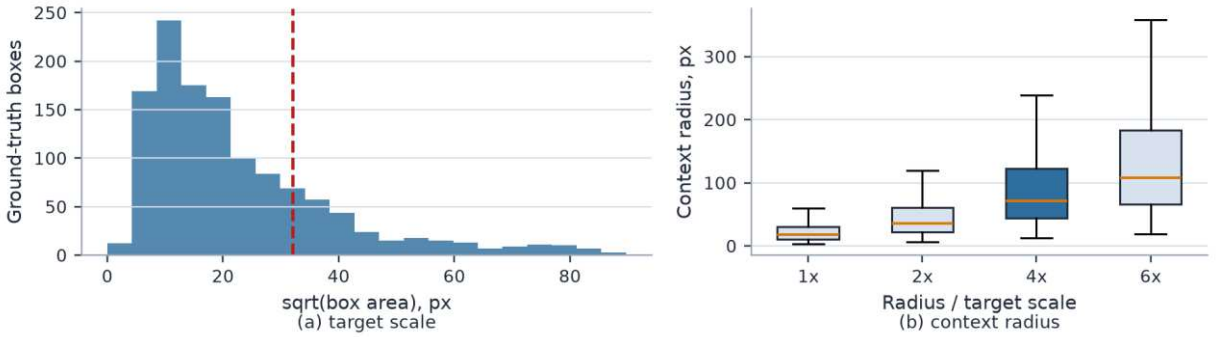

Fig. 2 Recorded target-size audit (left) and context radii relative to target scale (right). The vertical marker in the size histogram is at approximately 32 pixels; distributions are descriptive for the analyzed sample and are not an evaluation of alternative context radii.

Fig. 2 visualizes why context scale may matter: the observed target-scale histogram concentrates toward small box sizes while a long tail remains. The right panel reports geometric context radii at several multiples of target scale, not detection outcomes for those multiples. We therefore do not infer an optimal ring radius from this plot. The pilot trained only the 4.0× context configuration.

4.2 Training settings and reproducibility

The reported parameter count is approximately 0.4107 million for the measured lightweight model. The architecture drawing depicts a substantially larger backbone and should not be used to explain that count. The loss warm-up is visible in the recorded trace, but the full learning-rate schedule and checkpoint-selection rule are not specified. There is one run, so sampling and training-seed variability cannot be estimated.

Table 1 Available implementation settings and remaining provenance gaps

Component	Recorded pilot setting
Hardware	NVIDIA GeForce RTX 4090 D with 24 GB memory
Training budget	24 epochs; 1,536 optimizer steps; 512 training windows per epoch
Batch and candidates	Batch size 8; 192 training candidates; RPN 1,000 and detection head 384 at evaluation
Optimizer	SGD, initial learning rate 0.01, momentum 0.9, weight decay 0.0001
Head views	64 × 64 RGB plus validity channel; 1.5× local and 4.0× outer crop; 128-dimensional encoders
Missing provenance	Seed(s), precise tile rules, pretrained checkpoint, intervention operator probabilities, loss coefficients, evaluator revision

The absent entries are not inferred from model diagrams or software defaults.

4.3 Metrics and interpretation

The source report lists AP over multiple IoU thresholds, AP50, AP75, Tiny AP, Tiny AP50, RPN Recall@K, and false positives per parent image after tile aggregation. AP50/AP75 compare predictions with annotations at IoU 0.50/0.75. Recall@K concerns whether one of the top K proposals matches each annotated target under the report's recall criterion; it is a ceiling on the ability of a proposal-only second stage to recover targets at that threshold and candidate budget. The exact "tiny" box definition, AP averaging protocol, score cutoff used for FP/image, and recall-matching details were not preserved in the supplied summary. We retain the reported labels and values and refrain from official benchmark comparison.

5 Results

5.1 Optimization traces

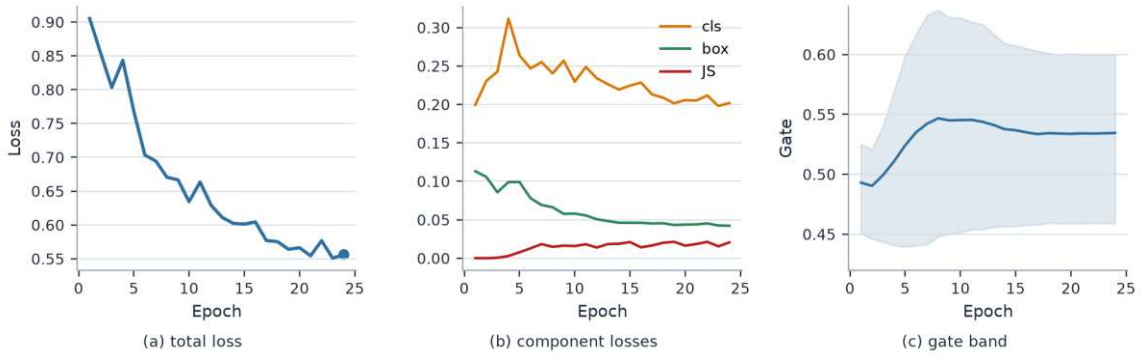

Fig. 3 Recorded 24-epoch pilot traces: total loss, classification/box/Jensen–Shannon components, and gate mean with a reported percentile band. The curves describe optimization behavior for one run and do not demonstrate comparative robustness.

The total logged loss changed from 0.9052 at epoch 1 to 0.5509 at the best recorded epoch, a decrease of 0.3543 in the reported objective units. The classification and regression terms trended downward; the pairwise JS term remained finite after its weight ramp. Because loss weights change during warm-up, the component curves should not be compared as if they had the same fixed coefficients, and a lower training loss is not evidence of better generalization. The gate mean moved from approximately 0.4931 to 0.5345. The percentile band remained broad enough to show proposal-level variation, although its center remained near 0.5. Without frozen-gate controls, this trace establishes stable optimization rather than an advantage of dynamic gating.

5.2 Detection accuracy and false detections

Table 2 Measured single-run subset metrics

Subset	AP	AP50	AP75	Tiny AP	Tiny AP50	R@1000	FP/image
Development	0.0057	0.0299	0.0003	0.0048	0.0266	0.1366	286.7292
Test	0.0129	0.0420	0.0101	0.0110	0.0334	0.1124	282.5703

R@1000 is candidate recall and not final-head detection recall. No matched baseline or confidence interval is available.

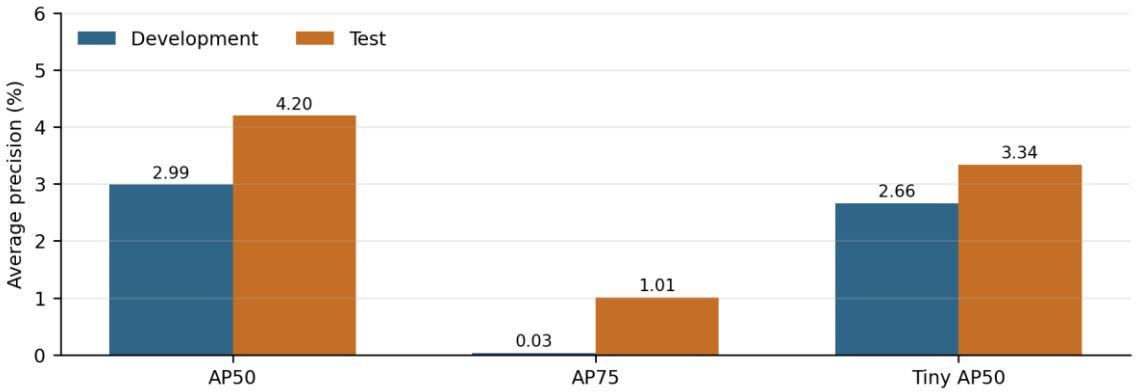

Fig. 4 Measured AP50, AP75, and Tiny AP50 in percent on the two subsets. These within-run comparisons contain no matched detector baseline.

Test AP50 was 0.0420 and AP75 was 0.0101; the 0.0319 gap indicates that high-IoU localization is especially difficult in this run. Test Tiny AP50 was 0.0334, below overall AP50 by 0.0086. On development the corresponding AP50/Tiny AP50 values were 0.0299/0.0266. Differences between the two subsets cannot be attributed to domain shift, model improvement, or sampling uncertainty without the subsets' object counts and composition. Test AP50 happened to exceed development AP50 by 0.0121 while candidate recall was lower by 0.0242, illustrating that ranking, box quality, object distribution, and candidate coverage contribute differently to each summary metric.

The recorded false-positive averages of 286.7292 and 282.5703 per image are substantial. Because the score threshold, tile overlap, final NMS, and parent-image accounting are not specified, these values cannot be translated into a deployment false-alarm rate or compared with another detector. Nonetheless, they show that the measured operating point would require meaningful precision and post-processing work before practical use. Reporting AP together with an explicit false-positive operating point is particularly important when the people of interest are extremely sparse.

5.3 Candidate recall as a limiting factor

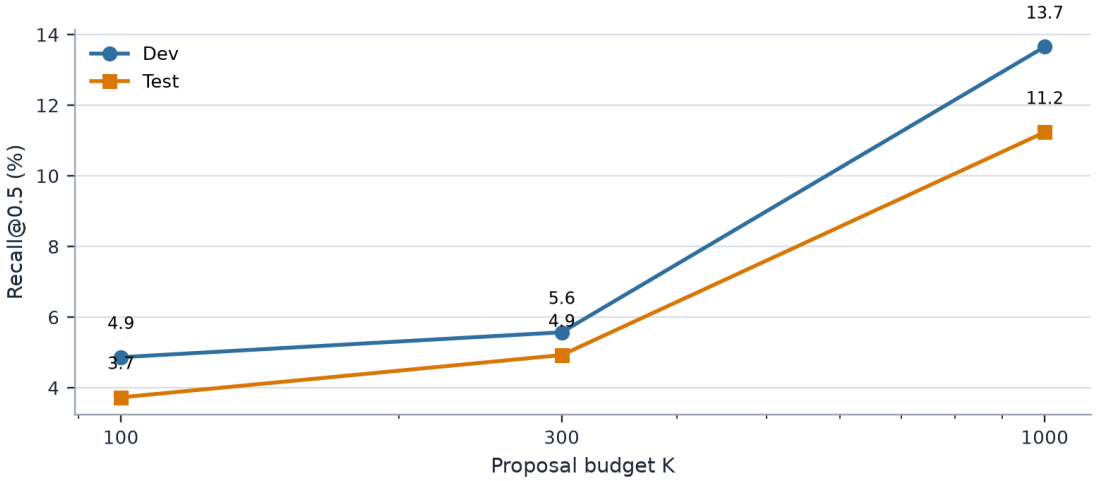

Fig. 5 Reported proposal Recall@0.5 at K=100, 300, and 1,000 candidates, expressed in percent. The graph records development and test measurements for the same pilot run; it is not an ablation of the detection head.

The plotted recall grows with candidate budget: approximately 4.9%, 5.6%, and 13.7% for development, and 3.7%, 4.9%, and 11.2% for test at $K = 100$, 300, and 1,000, respectively. The plot rounds the unrounded R@1000 values in Table 2 (13.66% and 11.24%). More than 86% of annotated objects on development and 88% on test were not covered at the 1,000-proposal budget under this recall definition. This directly restricts the opportunity for the local/ring head to improve true-positive count. Improving multi-scale proposals, scale matching, tile coverage, and candidate ranking deserves priority. The head could still improve precision among covered objects, but the current record has no local-only comparator to measure that increment.

5.4 Gate distribution and source-proxy variation

Table 3 Gate statistics from the supplied development-tile analysis

Gate statistic	Development-tile value	Interpretation
Mean	0.5329	Observed average gate value
10th percentile	0.5110	Reported lower decile across tile means
90th percentile	0.5567	Reported upper decile across tile means

These are tile-level summaries and must not be read as per-object confidence intervals.

The development-tile gate values cluster near 0.53. Their nonzero and nonunit values rule out a literal numerical collapse to the endpoints, but the narrow range could also arise from a near-constant gate. The distribution alone does not establish evidence conditioning. A controlled comparison should use local-only ($\gamma = 0$), unrestricted context ($\gamma = 1$), a learnable constant, scale-only, evidence-only, and complete gate variants under identical optimization and seeds. The source report lists these variants as pending; no ablation numbers are included here.

Across proxy groups with nonzero valid-target counts, the recorded extremes were AP50=0.1481 for bb_V0002 and AP50=0.0000 for bing_P000_tooopen. The chart shows other groups between these extremes, but the record omits their target and image counts; a zero AP may reflect few examples or extreme difficulty. Neither the range nor the average provides an unbiased estimate of cross-source generalization. To test that hypothesis, independent acquisition metadata, group sizes, repeated seeds, and confidence intervals are needed. SeaDronesSee[29] would provide a distinct maritime data source, but external evaluation was not run.

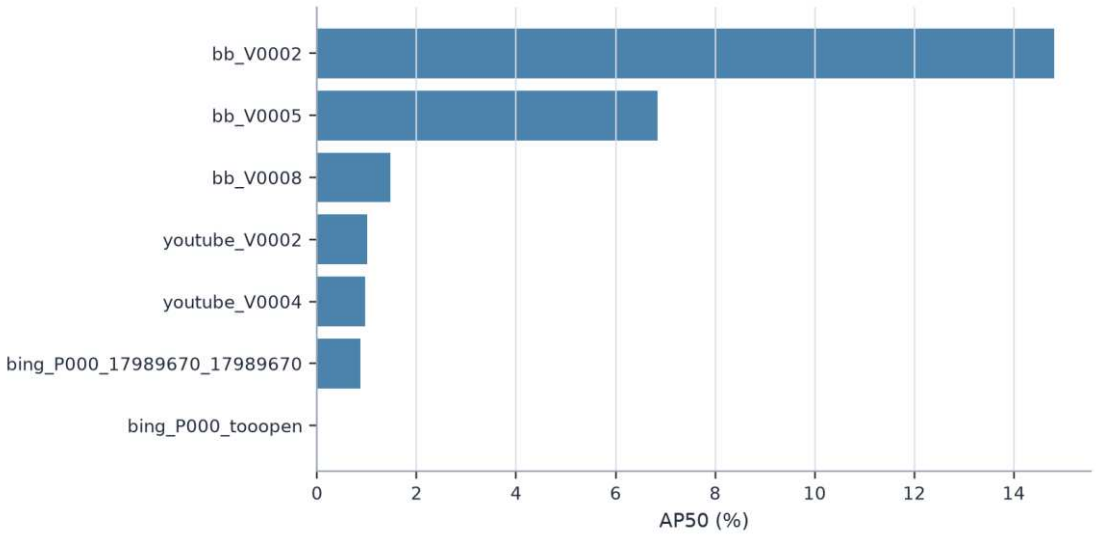

Fig. 6 AP50 by filename-derived source-proxy group in the supplied subset evaluation. The groups are descriptive, may have very different sample counts, and are not confirmed acquisition domains.

5.5 Computational profile

Table 4 Available efficiency measurements

Measured quantity	Pilot value	Scope
Parameters	approximately 0.4107 M	Measured lightweight implementation
Average inference time	48.0685 ms per window	Reported evaluation-window average, not full parent-image latency
Training wall time	approximately 8.0 min	Single GPU, 24 epochs / 1,536 steps

Window timing excludes an auditable end-to-end pipeline breakdown and has no same-hardware comparator.

The fast run provides a useful systems feasibility check for crop extraction and dual encoders, but it should not be used to call the method efficient relative to published systems. The number of windows per parent image, pre/post-processing costs, crop batching strategy, and all image transfers are required for an end-to-end throughput estimate. The illustrated ResNet-50-FPN version would have different parameter and latency profiles from Table 4.

6 Discussion

6.1 What the measured run establishes

The two-view head, valid-mask pooling, gate, intervention construction, and paired loss can be trained together without apparent numerical divergence in one recorded run. The observed detector has nonzero AP but low absolute accuracy, and proposal recall at $K = 1{,}000$ is a clear constraint on maximum second-stage coverage. These are observations about a specific subset and lightweight implementation. They cannot isolate whether improvements or harms arise from the ring, the gate, the background intervention, the auxiliary head, or the paired loss.

6.2 Threats to validity

First, matched baselines are absent. An architectural description plus a working training trace does not establish that context gating outperforms either a local-only head or ordinary feature fusion. Second, the proposed ResNet-50-FPN stem in Fig. 1 differs materially from the measured lightweight RPN. Model capacity, proposal quality, and latency claims must refer to the actual evaluated configuration; a full-backbone result requires a new, separately reported run. Third, split size, "tiny" threshold, tile overlap, confidence threshold for false positives, exact evaluator, weights, checkpoints, and seeds are missing. These omissions constrain reproduction and preclude direct benchmark comparison.

Fourth, the protected intervention verifies equality within the mask but not label preservation outside it. Unannotated persons, plausible seam artifacts, and shifts in global illumination can each modify the task. A blind image-quality review and annotation audit would separate these possibilities. Fifth, copying original proposals tests prediction stability at fixed locations; it does not force proposal recall to remain stable under a distant-background change. A complete robustness analysis should compare both fixed and independently generated candidates, with object size and nearest-neighbor distance stratification. Finally, source identifiers inferred from filenames are only proxy groups, and SeaDronesSee evaluation was planned rather than observed.

6.3 Required comparative protocol

A defensible claim of added value should freeze the same training/development/test parents, evaluator, backbone, pretraining, tile budget, and compute allocation across a local-only baseline, always-on ring context, constant gate, full gate, and full gate without pairing. Each variant should be run across at least three predeclared random seeds and reported with dispersion, per-size and per-source metrics, RPN recall, AP, false positives at specified thresholds, and runtime. A radius sweep should report both model performance and distance from other annotated people, because removing near-neighbor evidence may have a different effect from removing far-background context. An external dataset requires a frozen person-class mapping and scale protocol before test inspection. These experiments are future work, not results imputed into the present paper.

7 Conclusion

We specify a tiny-person detector with explicit local and outer-ring RGB crops, independently learned embeddings, proposal-level evidence gating, and protected distant-background paired training. Its initial TinyPerson subset run is trainable end to end, but low RPN recall and low AP leave substantial room for improvement. The current evidence supports a concrete mechanism and diagnostic observations; controlled comparisons and complete reproducibility metadata are required before claiming superior detection or cross-scene generalization.

Statements and Declarations

Funding

No specific funding was received for this work.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable. This study consists of computational model development and evaluation using publicly available TinyPerson images and their existing annotations. No new participants were recruited, and no interventions on human participants were performed. The background interventions described in the method are image transformations used during training.

Data availability

The TinyPerson benchmark and its existing annotations are available through the dataset authors' TinyBenchmark repository (https://github.com/ucas-vg/PointTinyBenchmark/tree/TinyBenchmark). Aggregate training, detection, proposal-recall, gate, and runtime summaries are reported in this article. The local split manifests, image-level predictions, and detailed training and evaluation logs are not included with the manuscript.

Authors' contributions

Peipei Song, Fengxiang Pan, and Jian Li contributed to the conception and design of the study, analysis and interpretation of the computational results, and preparation and revision of the manuscript. All authors read and approved the final manuscript. Jian Li serves as the corresponding author.

Use of AI-assisted technologies

OpenAI ChatGPT was used during manuscript revision to assist with language editing, manuscript formatting, citation formatting, and the wording of administrative declarations. It was not used to generate experimental data or alter the reported numerical results.

References

1. Yu, X., Gong, Y., Jiang, N., Ye, Q., Han, Z.: Scale Match for Tiny Person Detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1257–1265 (2020). https://doi.org/10.1109/WACV45572.2020.9093394
2. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 936–944 (2017). https://doi.org/10.1109/CVPR.2017.106
3. Singh, B., Davis, L.S.: An Analysis of Scale Invariance in Object Detection – SNIP. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3578–3587 (2018). https://doi.org/10.1109/CVPR.2018.00377
4. Singh, B., Najibi, M., Davis, L.S.: SNIPER: Efficient Multi-Scale Training. Advances in Neural Information Processing Systems 31 (2018).
5. Akyon, F.C., Altinuc, S.O., Temizel, A.: Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. In: IEEE International Conference on Image Processing, pp. 966–970 (2022). https://doi.org/10.1109/ICIP46576.2022.9897990
6. Singh, K.K., Mahajan, D., Grauman, K., Lee, Y.J., Feiszli, M., Ghadiyaram, D.: Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11067–11075 (2020). https://doi.org/10.1109/CVPR42600.2020.01108
7. Xiao, K., Engstrom, L., Ilyas, A., Madry, A.: Noise or Signal: The Role of Image Backgrounds in Object Recognition. In: International Conference on Learning Representations (2021).
8. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant Risk Minimization. arXiv preprint arXiv:1907.02893 (2019).
9. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In: International Conference on Learning Representations (2020).
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems 28 (2015).
11. Najibi, M., Singh, B., Davis, L.S.: AutoFocus: Efficient Multi-Scale Inference. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9745–9755 (2019).
12. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017).
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017).

14. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully Convolutional One-Stage Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9627–9636 (2019).
15. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-NMS — Improving Object Detection With One Line of Code. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5561–5569 (2017).
16. Bell, S., Zitnick, C.L., Bala, K., Girshick, R.: Inside-Outside Net: Detecting Objects in Context With Skip Pooling and Recurrent Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2874–2883 (2016).
17. Chen, X., Gupta, A.: Spatial Memory for Context Reasoning in Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4086–4096 (2017).
18. Koh, P.W., et al.: WILDS: A Benchmark of in-the-Wild Distribution Shifts. In: Proceedings of the International Conference on Machine Learning. Proceedings of Machine Learning Research 139, 5637–5664 (2021).
19. Dwibedi, D., Misra, I., Hebert, M.: Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1301–1310 (2017). https://doi.org/10.1109/ICCV.2017.146
20. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6023–6032 (2019). https://doi.org/10.1109/ICCV.2019.00612
21. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2918–2928 (2021).
22. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond Empirical Risk Minimization. In: International Conference on Learning Representations (2018).
23. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
24. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. In: Computer Vision – ECCV. Lecture Notes in Computer Science, vol. 11211, pp. 3–19 (2018). https://doi.org/10.1007/978-3-030-01234-2_1
25. Laine, S., Aila, T.: Temporal Ensembling for Semi-Supervised Learning. In: International Conference on Learning Representations (2017).
26. Tarvainen, A., Valpola, H.: Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. Advances in Neural Information Processing Systems 30 (2017).
27. Lin, J.: Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory 37(1), 145–151 (1991). https://doi.org/10.1109/18.61115
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016).
29. Varga, L.A., Kiefer, B., Messmer, M., Zell, A.: SeaDronesSee: A Maritime Benchmark for Detecting Humans in Open Water. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2260–2270 (2022).