

Improving RT-DETRv4 Person and Head Detection on CrowdHuman with C-RADIOv4 Distillation

Nikhil Reddy

nikhileswara.reddy@awiros.com

Ravi Teja Palepu

ravi.teja@awiros.com

Abu Ossama Siddiqui

abu.siddiqui@awiros.com

Abstract

Real-time person and head detection in dense crowds remains fundamentally difficult, as detectors must localize partially visible bodies under severe inter-person occlusion while preserving the fine-grained semantics needed to separate tiny heads in heavily cluttered scenes. Recent real-time detection transformers such as RT-DETRv4 address feature degradation in lightweight detectors by distilling knowledge from vision foundation models (VFMs), yet the impact of teacher choice on detection performance is underexplored. In this paper, we benchmark RT-DETRv4-S against YOLO26-S, YOLOv8-S, and RF-DETR-S on CrowdHuman visible-person and head detection, and investigate whether replacing the default DINOv3-Base teacher with C-RADIOv4-SO400M can further improve detection quality. Rather than scaling the DINOv3 teacher from the default Base model to a much larger DINOv3-7B variant, we study an alternative design in which the student is supervised by C-RADIOv4, an agglomerative vision backbone that already distills knowledge from DINOv3-7B together with complementary teachers such as SAM3 and SigLIP2. RT-DETRv4-S already surpasses all three baselines under both teachers. With C-RADIOv4, it achieves the best mAP on both visible-person (0.8410) and head detection (0.7881), improving over DINOv3 by +0.91 pp and +1.71 pp respectively, outperforming YOLO26-S by +2.47 pp and +2.91 pp.

1. Introduction

Detecting individual people and their heads in dense crowds is a long-standing challenge in computer vision. While modern object detectors achieve strong performance on standard benchmarks, they often fail to resolve individual instances in heavily occluded scenes where dozens of people overlap within a single frame. The CrowdHuman benchmark [1, 2] targets precisely this regime, with an average of 23 annotated persons per image and annotation protocols that distinguish full-body, visible-body, and head bounding boxes.

For resource-constrained deployment, compact detector variants are of particular interest. Among recent small architectures, YOLO26-S [7] represents the latest Ultralytics real-time detector, while RF-DETR-S [9] offers a transformer-based alternative. However, the strongest results in our evaluation come from RT-DETRv4-S [4], a real-time DETR variant that distills knowledge from vision foundation models (VFMs) into a lightweight student built on the RT-DETR framework [3]. This raises a natural question: *does the choice of teacher model materially affect detection quality when the student architecture is held constant?*

We make the following contributions:

1. We show that RT-DETRv4-S is the strongest evaluated compact model by mAP on both visible-person and head detection, surpassing YOLO26-S, YOLOv8-S, and RF-DETR-S.
2. We demonstrate that replacing the default DINOv3-Base teacher with C-RADIOv4 further improves RT-DETRv4-S, showing that a compact agglomerative teacher distilled from a richer pool that includes DINOv3-7B can be more effective than directly using the smaller standalone DINOv3 teacher shipped with RT-DETRv4. The largest gain appears on head detection (+1.71 pp mAP).

The remainder of this paper is organized as follows. Section 2 introduces the CrowdHuman benchmark and the detection setting used in this study. Section 3 reviews the most relevant detector and teacher-model literature. Section 4 describes the distillation framework, training setup, and evaluation protocol. Section 5 presents quantitative and deployment-oriented results. Section 6 provides qualitative analysis. Section 7 concludes with a discussion of the main findings and future directions.

2. CrowdHuman Benchmark and Detection Setting

2.1. CrowdHuman

CrowdHuman [1] is a large-scale benchmark for human detection in dense, heavily occluded scenes. The dataset

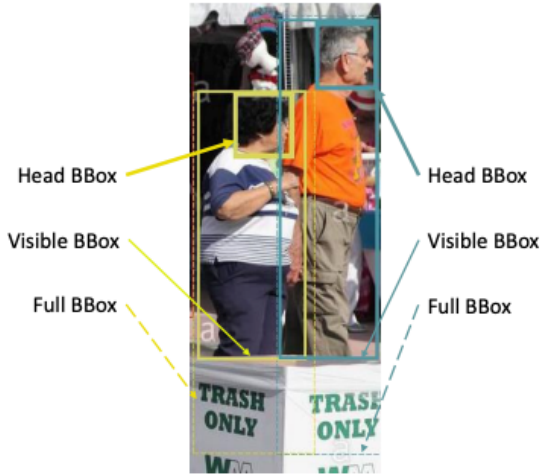


Figure 1. CrowdHuman provides full-body, visible-body, and head annotations; this example highlights the visible-person and head targets used throughout this study.

comprises 15,000 training images, 4,370 validation images, and 5,000 test images, totaling approximately 470K annotated instances [1, 2]. Each instance is annotated with three bounding-box types: full body, visible body, and head. The high instance density (~ 23 persons per image) and severe inter-person occlusion make CrowdHuman a challenging testbed for evaluating detector robustness.

2.2. Visible-Person and Head Detection Setting

This work adopts a visible-person and head detection protocol derived from CrowdHuman. Only the visible-body and head annotations are used, while the full-body boxes are excluded throughout training and evaluation. The visible-person target corresponds to the dataset’s visible-body annotations, which measure the observable extent of each individual and avoid rewarding predictions that extend through severe occlusion. The head target provides a complementary small-object signal that remains informative when torsos are heavily truncated or overlapped by nearby people.

All detectors are trained on the CrowdHuman training split and evaluated on the validation split. Performance is reported separately for visible-person and head detection using mean average precision (mAP). This separation is important in crowded scenes because the two targets stress different capabilities: visible-person detection emphasizes extent estimation under occlusion, whereas head detection emphasizes small-target localization in dense groups. Figure 1 shows a representative validation example with both annotation types.

3. Related Work

3.1. Real-Time DETR Variants

RT-DETR [3] introduced a real-time DETR framework that combines efficient hybrid encoding with end-to-end set prediction, eliminating hand-crafted post-processing such as non-maximum suppression. RT-DETRv4 [4] extends this line by incorporating VFM-based knowledge distillation: during training, a frozen teacher model provides dense semantic supervision to the student detector through an auxiliary distillation loss. This design decouples the teacher from the detector backbone, allowing different vision foundation models to serve as teachers without modifying the student architecture.

3.2. Vision Foundation Teachers

RT-DETRv4 uses frozen vision foundation models as teachers, so the quality and diversity of the teacher representation directly affect the supervision transferred to the detector. In this work, the teacher comparison is between DINOv3-Base and C-RADIOv4-SO400M.

C-RADIOv4 [5] is described by its authors as an *agglomerative vision backbone* built through multi-teacher distillation. Rather than imitating a single source model, it trains a unified student to absorb the complementary capabilities of several teachers while preserving strong downstream performance at similar computational scale. The official C-RADIOv4 release used in this report is distilled from a teacher set that includes DINOv3-7B, SigLIP2-g, and SAM3 [10], making it substantially richer than a single-teacher representation. This distinction is central to the present study. Instead of scaling RT-DETRv4 to a much larger standalone DINOv3 teacher, the report evaluates whether an agglomerative backbone distilled from a larger and more diverse teacher pool can provide a stronger distillation signal to the same lightweight detector. This motivation is especially relevant for crowded detection, where person and head localization in dense scenes demands both semantic discrimination and sensitivity to small, partially visible structures.

DINOv3 [6] is a large-scale self-supervised vision foundation model designed to produce strong transferable representations across a broad range of visual tasks. Its training recipe emphasizes careful scaling of both data and model size, and introduces *Gram anchoring* to mitigate the degradation of dense feature maps during long training schedules [6]. As a result, DINOv3 provides high-quality dense visual features without task-specific fine-tuning, which makes it a strong default teacher for RT-DETRv4 when the goal is to distill robust semantic structure into a lightweight detector. In the context of this paper, DINOv3 serves as the single-teacher baseline: it offers powerful dense rep-

representations, but unlike C-RADIOv4, it does not explicitly fuse supervision from multiple heterogeneous teacher families. The comparison in this work therefore isolates whether an agglomerative multi-teacher representation can provide a stronger distillation signal than a high-quality single-teacher self-supervised model.

3.3. Compact Detection Baselines

We include three compact baselines to contextualize the RT-DETRv4 results. **YOLO26-S** [7] is the most recent Ultralytics single-stage detector. **YOLOv8-S** [8] is a well-established Ultralytics baseline that remains competitive on many benchmarks. **RF-DETR-S** [9] provides a transformer-based reference point outside the RT-DETR family. All baselines use their respective small-model variants to ensure a fair comparison at a similar computational scale.

4. Method

4.1. RT-DETRv4 Distillation Architecture

All experiments keep the RT-DETRv4-S detector fixed and vary only the vision foundation model used for supervision. The student follows the standard RT-DETRv4 design [4]: an HGNetv2-S backbone feeds the hybrid encoder, whose deepest stride-32 feature is refined by the encoder transformer before multi-scale fusion and decoding. Following the RT-DETRv4 distillation design, supervision is attached to this deepest semantic feature, denoted here as the AIFI-processed F_5 map. This is the role of the **Deep Semantic Injector (DSI)**: rather than distilling into early backbone activations or directly into the decoder, DSI injects foundation-model semantics into the student feature with the highest semantic density, where the representation is already compact enough to align with a frozen teacher and still influential enough to shape the downstream detector.

In this formulation, DSI is implemented with a lightweight projection head attached to the deepest encoder output. The hybrid encoder operates at a hidden width of 256 channels, and the selected F_5 feature is projected into the teacher feature space before alignment. With **DINOv3-Base** [6] as the teacher, the projection maps $256 \rightarrow 768$, matching the DINOv3 embedding width. With **C-RADIOv4-SO400M** [5, 10], the projection instead maps $256 \rightarrow 1152$, matching the teacher feature dimensionality of C-RADIOv4. In both cases, the teacher remains frozen and is evaluated on a 2x downsampled image. DINOv3 produces normalized patch tokens that are reshaped into an NCHW spatial feature map, whereas C-RADIOv4 returns NCHW visual features directly. If the teacher and student resolutions differ, the teacher map is bilinearly resized to the student F_5 resolution, and distillation is performed by minimizing cosine distance between L2-normalized teacher and projected

student features at every spatial location.

This formulation isolates teacher quality from detector design, since the backbone, encoder, decoder, and detection losses are otherwise identical across both variants. The comparison is therefore not a simple teacher-scaling experiment inside the DINOv3 family. The RT-DETRv4 baseline uses DINOv3-Base directly as a single teacher, whereas the proposed variant replaces that branch with C-RADIOv4-SO400M, a compact agglomerative backbone distilled from a stronger teacher pool that includes DINOv3-7B, SigLIP2-g, and SAM3 [5, 10]. The experiment therefore tests whether a reduced all-domain agglomerative representation can provide richer supervision than a direct single-teacher pipeline without changing the detector itself.

RT-DETRv4 further stabilizes this transfer with **Gradient-guided Adaptive Modulation (GAM)** [4]. Instead of keeping the distillation loss weight fixed, GAM monitors the fraction of total gradient magnitude flowing through the encoder transformer and updates the distillation weight so that the encoder remains within a prescribed gradient band. For the **DINOv3-Base** teacher, this target is set to $11 \pm 1\%$; for **C-RADIOv4-SO400M**, it is adjusted to $10 \pm 2\%$ after ablations to better match the stronger 1152-dimensional representation. This adaptive control is important because it prevents semantic distillation from collapsing into a weak auxiliary signal, while also preventing it from dominating the detection losses. The teacher branch, projection layer, and cosine-alignment loss are used only during training, so inference remains the standard RT-DETRv4-S detector with no additional deployment cost.

4.2. Training Setup

Both configurations use the same CrowdHuman visible-person/head annotation files and are initialized from the default RT-DETRv4-S COCO pretrained weights. Relative to that default training recipe, the only intentional modification highlighted here is the visual-encoder gradient band in the C-RADIOv4 variant, which was tuned specifically for the C-RADIOv4 teacher after several ablation studies. This keeps the comparison centered on teacher quality while allowing the optimization to remain stable under the different teacher representation.

4.3. Evaluation Setup

All models are evaluated on the CrowdHuman validation split. Metric computation is performed with the Megvii CrowdDetection evaluation toolkit [11], following the standard CrowdHuman-style evaluation protocol. We report mAP separately for visible-person and head detection. We follow the CrowdHuman convention of using ignore regions, which excludes predictions that overlap annotated ignore boxes. All reported numbers therefore follow the

Table 1. Visible-person detection results on CrowdHuman validation. Best result in **bold**, second-best underlined (per metric).

Model	mAP	AR
RT-DETRv4-S (C-RADIOv4-SO400M)	0.8410	0.9081
RT-DETRv4-S (DINOv3 Base)	<u>0.8319</u>	<u>0.9048</u>
YOLOv8-S	0.8263	0.8886
YOLO26-S	0.8163	0.8940
RF-DETR-S	0.8006	0.8682

Table 2. Head detection results on CrowdHuman validation (ignore-region mode 1). Best in **bold**, second-best underlined (per metric).

Model	mAP	AR
RT-DETRv4-S (C-RADIOv4-SO400M)	0.7881	0.8178
RT-DETRv4-S (DINOv3 Base)	<u>0.7710</u>	<u>0.8025</u>
YOLO26-S	0.7590	0.7955
YOLOv8-S	0.7336	0.7514
RF-DETR-S	0.6990	0.7350

CrowdHuman evaluation style implemented in the Crowd-Detection repository [1, 11].

5. Experimental Results

5.1. Comparison with Compact Baselines

Visible-person detection. Table 1 compares visible-person performance across all five models. Both RT-DETRv4-S variants outperform the three baselines by a clear margin in mAP. The C-RADIOv4-teacher configuration leads at 0.8410, followed by the DINOv3 variant at 0.8319. Among baselines, YOLOv8-S is strongest at 0.8263, followed by YOLO26-S at 0.8163 and RF-DETR-S at 0.8006.

Head detection. Table 2 reveals an even wider separation. YOLO26-S outperforms YOLOv8-S on head mAP (0.7590 vs. 0.7336), making it the strongest non-RT-DETRv4 head detector. Nevertheless, both RT-DETRv4 variants surpass all baselines by a substantial margin, and the C-RADIOv4-teacher model achieves the best head mAP at 0.7881.

5.2. COCO-Pretrained Checkpoints on CrowdHuman

To separate the effect of CrowdHuman-specific training from the underlying detector initialization, Table 3 reports visible-person performance when the default COCO-pretrained checkpoints are evaluated directly on CrowdHuman validation. This comparison uses the off-the-shelf detector checkpoints and therefore complements, rather than replaces, the fine-tuned results in Tables 1 and 2.

These pretrained-checkpoint results are substantially lower than the fine-tuned results in Table 1, which is expected

Table 3. Visible-person mAP of default COCO-pretrained checkpoints evaluated on CrowdHuman validation. Best in **bold**, second-best underlined.

Model	mAP
RF-DETR-S	0.6956
YOLOv8-S	<u>0.6935</u>
RT-DETRv4-S	0.6846
YOLO26-S	0.6715

Table 4. TensorRT latency on NVIDIA T4 together with reported model complexity.

Model	Res.	T4 (ms)	Params	GFLOPs
RT-DETRv4-S [12]	640	3.66	10M	25
YOLO26-S [16]	640	2.5	9.5M	20.7
YOLOv8-S [8, 15]	640	6.97	11M	29
RF-DETR-S [13, 14]	512	3.50	32.1M	59.8

given the domain shift from generic COCO training to dense crowd scenes. Even so, the ranking is informative: RF-DETR-S and YOLOv8-S are strongest before adaptation, while RT-DETRv4-S remains competitive and becomes the strongest family after CrowdHuman-specific training and teacher-guided distillation.

5.3. Teacher Study: DINOv3-Base vs. C-RADIOv4-SO400M

A more targeted question is whether the teacher model materially changes detection quality when the student architecture is held constant. Relative to the DINOv3-Base teacher, replacing the supervision branch with C-RADIOv4-SO400M improves visible-person mAP from 0.8319 to 0.8410 and head mAP from 0.7710 to 0.7881. These consistent gains indicate that the agglomerative teacher provides a stronger training signal than the direct single-teacher configuration while leaving the detector architecture unchanged.

5.4. Deployment Efficiency

To contextualize the accuracy results, Table 4 summarizes deployment-oriented model statistics collected from the corresponding official papers and repositories. Because RT-DETRv4 uses distillation only during training, the DINOv3-Base and C-RADIOv4 variants share the same inference architecture and therefore the same deployment cost. These numbers should still be read with their published benchmarking settings in mind: RT-DETRv4-S and the YOLO baselines are reported at 640x640, whereas RF-DETR-S is reported at 512x512.

From a deployment perspective, YOLO26-S remains the lightest and fastest model in the comparison, with the lowest latency, parameter count, and FLOPs. The more informative comparison for this study, however, is between RT-DETRv4-S and the other transformer baselines. Relative to

YOLOv8-S, RT-DETRv4-S is both faster on T4 TensorRT (3.66 ms vs. 6.97 ms) and slightly more compact (10M vs. 11M parameters; 25 vs. 29 GFLOPs). Relative to RF-DETR-S, RT-DETRv4-S is substantially smaller and more compute-efficient (10M vs. 32.1M parameters; 25 vs. 59.8 GFLOPs), although RF-DETR-S reports a similar T4 latency at a lower 512x512 resolution. Overall, the accuracy gains reported in Tables 1 and 2 are not obtained by increasing deployment cost between the two RT-DETRv4 variants, since the teacher branch is used only during training and is absent at inference.

6. Qualitative Results

To complement the quantitative comparison, this section presents the qualitative comparison of the models. Each panel is arranged as a 2×3 grid: the original image appears first, followed by the five detectors in descending mAP order for the corresponding task. For visualization, predictions are filtered at a confidence threshold of 0.40, and true-positive versus false-positive assignments are determined by greedy one-to-one matching at an IoU threshold of 0.50. Some apparent false positives can therefore arise either from predictions whose overlap with the nearest ground-truth box falls below the 0.50 IoU threshold or from plausible instances that appear visually valid but are not annotated in the ground truth. Dashed blue boxes denote ground truth, green boxes denote true-positive predictions, and red boxes denote false positives. Full-resolution result images are available at <https://workdrive.zohoexternal.com/folder/r1mao3e3eb6a699414d7eb5fb8a562230f47d>.

6.1. Visible-Person Detection in Low-Density Scenes

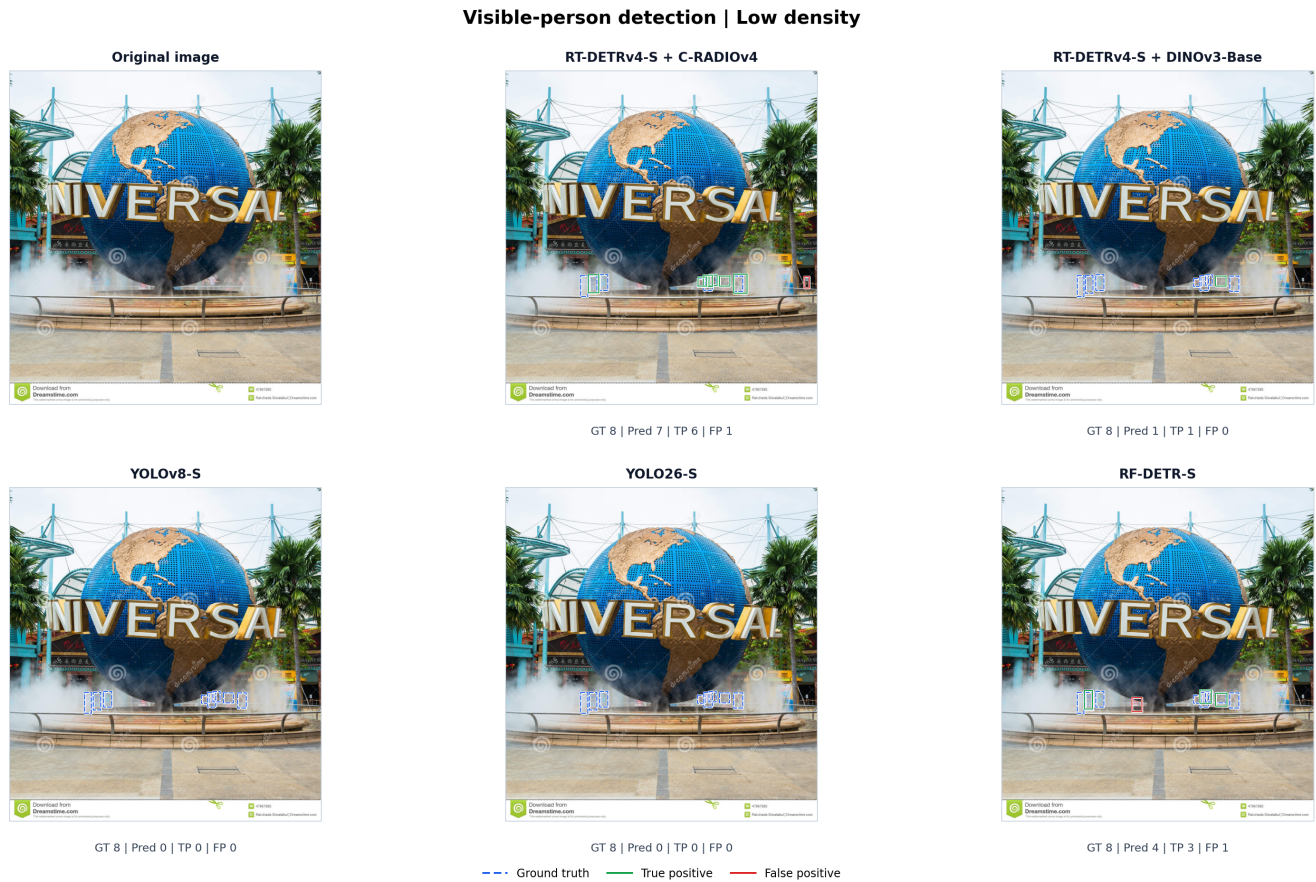


Figure 2. Qualitative comparisons for visible-person detection in low-density scenes.

Visible-person detection | Low density



Figure 3. Qualitative comparisons for visible-person detection in low-density scenes.

6.2. Visible-Person Detection in Medium-Density Scenes

Visible-person detection | Medium density



Figure 4. Qualitative comparisons for visible-person detection in medium-density scenes.



Figure 5. Qualitative comparisons for visible-person detection in medium-density scenes.



Figure 6. Qualitative comparisons for visible-person detection in medium-density scenes.

6.3. Visible-Person Detection in High-Density Scenes



Figure 7. Qualitative comparisons for visible-person detection in high-density scenes.

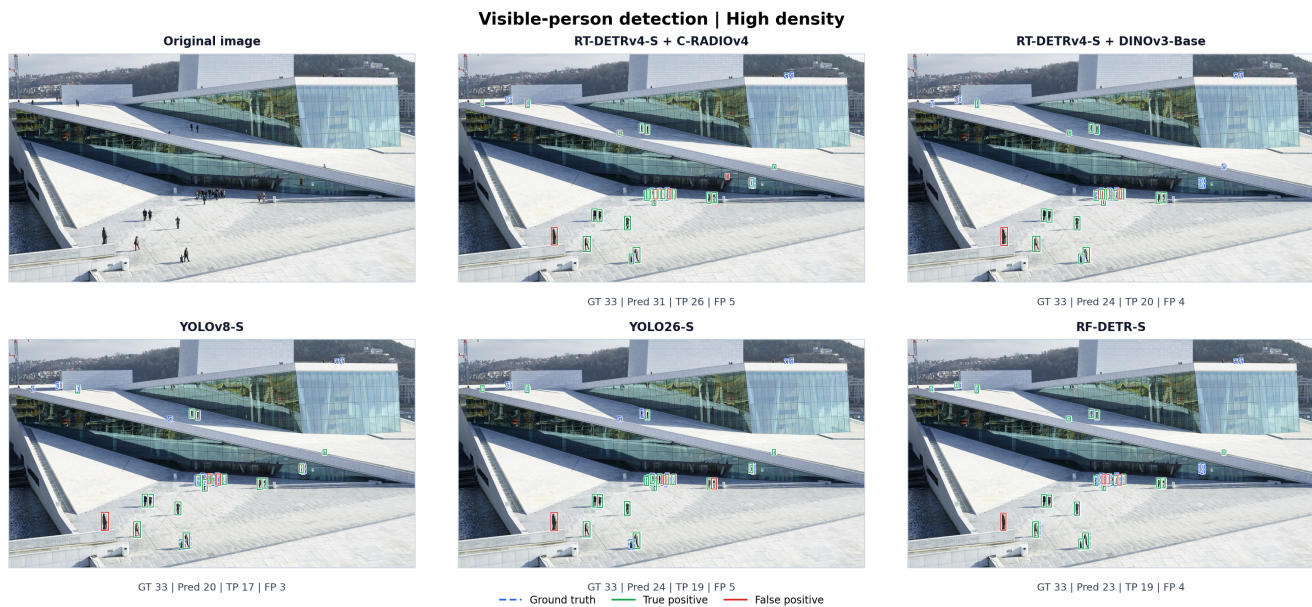


Figure 8. Qualitative comparisons for visible-person detection in high-density scenes.

6.4. Head Detection in Low-Density Scenes

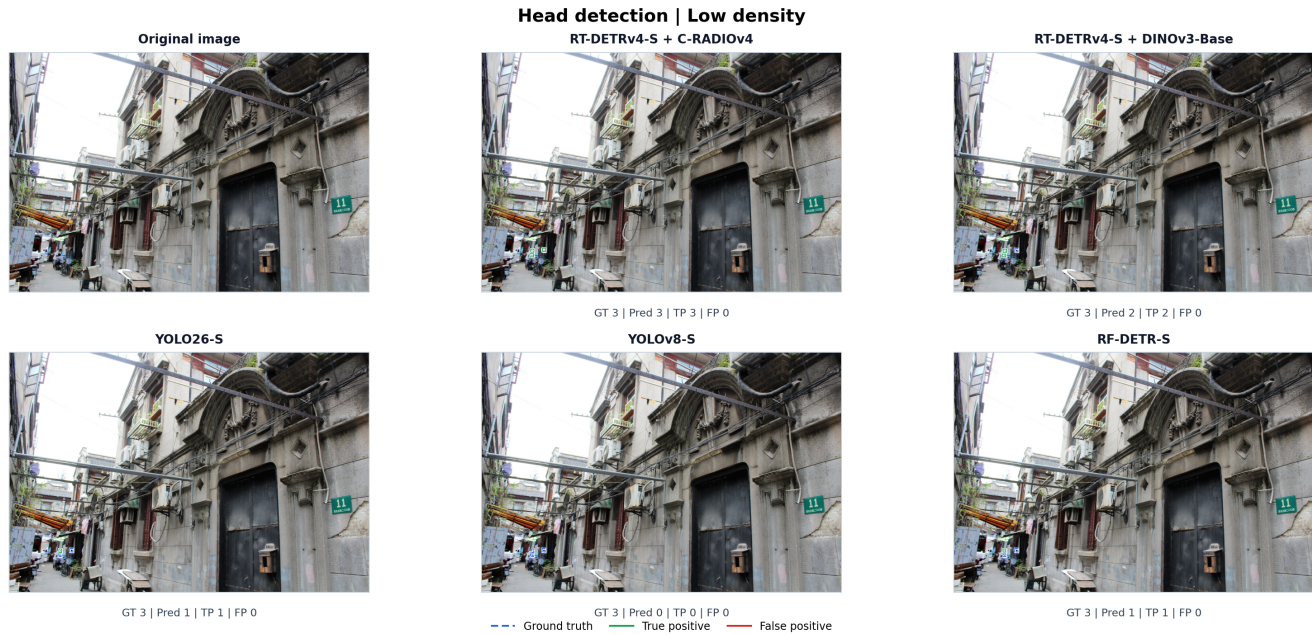


Figure 9. Qualitative comparisons for head detection in low-density scenes.



Figure 10. Qualitative comparisons for head detection in low-density scenes.

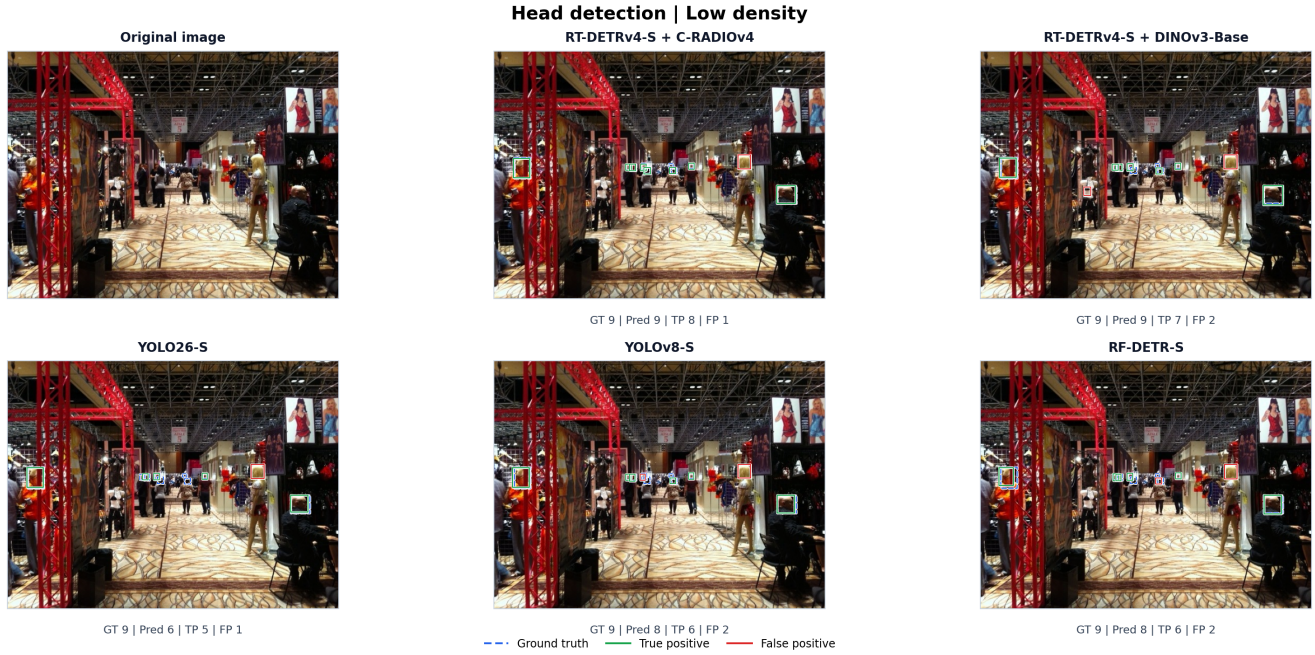


Figure 11. Qualitative comparisons for head detection in low-density scenes.

6.5. Head Detection in Medium-Density Scenes



Figure 12. Qualitative comparisons for head detection in medium-density scenes.



Figure 13. Qualitative comparisons for head detection in medium-density scenes.

6.6. Head Detection in High-Density Scenes



Figure 14. Qualitative comparisons for head detection in high-density scenes.

7. Conclusion and Future Work

Teacher selection emerges as a first-order design axis for VFM-distilled detectors. With the RT-DETRv4-S student held fixed, replacing DINOv3-Base with C-RADIOv4 yields +0.91 pp visible-person mAP and +1.71 pp head mAP, showing that the gains arise from the transferred representation rather than from changes in student capacity. Because C-RADIOv4 compresses supervision from DINOv3-7B, SigLIP2-g, and SAM3 into a compact agglomerative backbone [5, 10], the results indicate that teacher diversity can be as important as teacher scale.

Among the compact models evaluated, the C-RADIOv4-distilled RT-DETRv4-S variant achieves the best visible-person mAP (0.8410) and the best head mAP (0.7881), establishing RT-DETRv4 as the strongest compact detector in this comparison. Future work should test whether these gains persist on COCO, across larger RT-DETRv4 variants such as L and X, and under a direct teacher comparison between C-RADIOv4 and DINOv3-7B.

References

- [1] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "CrowdHuman: A Benchmark for Detecting Human in a Crowd," *arXiv preprint arXiv:1805.00123*, 2018.
- [2] CrowdHuman Dataset Card, Hugging Face Datasets. Available: <https://huggingface.co/datasets/sshao0516/CrowdHuman>
- [3] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs Beat YOLOs on Real-time Object Detection," *arXiv preprint arXiv:2304.08069*, 2023.
- [4] W. Lv, Y. Zhao, S. Xu, J. Wei, G. Wang, C. Cui, Y. Du, Q. Dang, and Y. Liu, "RT-DETRv4: Improved Baseline for Real-Time Detection Transformer," *arXiv preprint arXiv:2510.25257*, 2025.
- [5] A. Hakhamaneshi, M. Soltanolkotabi, and A. Anandkumar, "C-RADIOv4: Agglomerative Vision Foundation Models," *arXiv preprint arXiv:2601.17237*, 2026.
- [6] "DINOv3," *arXiv preprint arXiv:2508.10104*, 2025.
- [7] Ultralytics, "YOLO26 Documentation." Available: <https://docs.ultralytics.com/models/yolo26/>
- [8] Ultralytics, "YOLOv8 Documentation." Available: <https://docs.ultralytics.com/models/yolov8/>
- [9] "RF-DETR: Real-Time, Fast, Accurate Object Detection and Instance Segmentation," *arXiv preprint arXiv:2511.09554*, 2025.
- [10] NVIDIA, "C-RADIOv4-SO400M Model Card," Hugging Face. Available: <https://huggingface.co/nvidia/C-RADIOv4-SO400M>
- [11] Megvii, "CrowdDetection Evaluation Toolkit." Available: <https://github.com/megvii-model/CrowdDetection/tree/master/evaluate>
- [12] RT-DETRs, "RT-DETRv4 GitHub Repository." Available: <https://github.com/RT-DETRs/RT-DETRv4>
- [13] Roboflow, "RF-DETR Benchmarks." Available: <https://github.com/roboflow/rf-detr/blob/main/docs/learn/benchmarks.md>
- [14] Roboflow, "RF-DETR GitHub Repository." Available: <https://github.com/roboflow/rf-detr>
- [15] Ultralytics, "YOLOv10 Documentation." Available: <https://docs.ultralytics.com/models/yolov10/>
- [16] Ultralytics, "Ultralytics GitHub Repository." Available: <https://github.com/ultralytics/ultralytics>