# Occlusion-Aware Image and Video Perception for Vulnerable Road User Detection: Methods, Benchmarks, and Deployment Challenges

**Jiahao Feng[1], Xu Zhang[2], Zhiliang Liu[3,*], Jiacheng Wang[1]**
[1]College of Computer and Information Engineering, Xiamen University of Technology, No. 600 Ligong Road, Jimei District, Xiamen 361024, Fujian, China
[2]College of Software Engineering, Xiamen University of Technology, No. 600 Ligong Road, Jimei District, Xiamen 361024, Fujian, China
[3]School of Cyberspace Security, No. 26 Wang'an Road, Dongxihu District, Wuhan, Hubei, China

Corresponding author: Zhiliang Liu (e-mail: liuzhiliang@csuc.edu.cn)

**ABSTRACT** Occlusion remains one of the main failure points in traffic-scene perception, and the errors it causes do not follow a single, predictable pattern. In a still frame, a camera may capture only a pedestrian's head or upper torso. In video, a tracker can lose that person for several frames and assign a different identity when the person reappears. Cyclists and riders create a related problem because the body parts that remain visible may overlap visually with cars, trucks, or nearby road users. This review examines how image- and video-based methods handle such incomplete observations. The review covers 96 studies published between 2009 and 2025. These studies investigate visible-to-full-body localization, interactions among neighboring instances, feature completion, transformer-based detection, RGB-thermal fusion, tracking, and the less frequently studied problem of inferring fully hidden pedestrians. Because these tasks produce different outputs, we did not combine their results. Pedestrian and crowd detection, multispectral detection, hidden-target prediction, and multi-object tracking are discussed under their original evaluation protocols. CityPersons miss rate measures frame-level detection; MOT17 HOTA includes association across frames; hidden-pedestrian F1 measures prediction quality. We also recorded runtime, hardware constraints, model size, and sensor requirements when the source papers provided those details. The published evidence is strongest for pedestrians who remain partly visible. Much less evidence is available for cyclists, riders, and fully hidden targets. Many deployment claims are also difficult to assess because the papers do not report results for specific occlusion levels or provide enough detail about computational cost.

**INDEX TERMS** Applied computer vision, occlusion-aware visual perception, vulnerable road user detection, image and video analysis, pedestrian detection, object tracking, multimodal fusion, deployment challenges

## I. INTRODUCTION

TRANSPORT and infrastructure systems now rely routinely on image and video streams, as do robotics and surveillance applications. Traffic scenes expose a particular weakness of these systems: vulnerable road users may occupy only a few pixels, change appearance with pose, overlap, or be reduced to a fragment [1]–[4]. In some frames, the camera records little more than a head, an upper torso, part of a wheel, or a handlebar. Even so, the perception stack must decide that a road user is present and maintain a useful state estimate over time.

Average benchmark accuracy can conceal the failure that matters in an occluded scene. If the legs or torso are hidden, the detector loses localization cues; if two people overlap, pixels from one person enter the region assigned to the other. A brief disappearance in video can split one trajectory or cause the returning person to receive a new identity. Performance on a mixed validation set does not reveal how often these failures occur in the heavily occluded subset. Interpreting a result consequently requires the visibility range and box convention, as well as the runtime setting and the operational cost of a false warning.

Recent image and video systems draw on convolutional backbones, feature pyramids, anchor-free detection, transformers, multispectral fusion, segmentation cues, and end-to-end tracking [5]–[11]. Their occlusion results, however, come from protocols that ask different questions. CityPersons and Caltech stratify pedestrian detection by visibility [12]–[14]; CrowdHuman stresses overlapping people [15]; KAIST-style evaluation adds a thermal stream [16]–[18];

and MOT benchmarks use IDF1, HOTA, and related measures to test identity continuity [19]–[24]. These results all inform applied computer vision, but they are not interchangeable evidence.

Existing surveys provide useful background on generic object detection, pedestrian detection, autonomous-driving perception, occlusion handling, VRU safety, and multi-object tracking [1]–[7], [25]–[31]. They do not yet provide a single account that follows an occlusion mechanism from the available visual evidence, through the benchmark used to test it, to the constraints of a deployed system. Instead, studies are commonly arranged by task, architecture, or application, with occlusion discussed inside each group. That organization makes two practical questions hard to answer: which mechanisms have evidence at each visibility level, and which published results are sufficiently specified to support a deployment claim?

This review is guided by five research questions:

1) What occlusion mechanisms degrade visual evidence for pedestrian and VRU perception in image and video systems?
2) Which mechanisms have been evaluated at partial, heavy, near-full, and full occlusion?
3) What conclusions about occlusion robustness are actually supported by current datasets, annotations, and metrics, and where do those protocols stop being informative?
4) When detection, segmentation-assisted methods, tracking, multimodal fusion, and hidden-target inference use different protocols, on what basis can their results be compared?
5) Which 2023-2025 advances are most relevant for real-time and resource-constrained visual perception systems?

The review therefore contributes in four respects.

- It develops a mechanism-based taxonomy for occlusion-aware image and video perception, linking visible-region modelling, part/full-body localization, attention, context, feature completion, transformer detection, multimodal fusion, temporal reasoning, and hidden-target inference.
- For every numerical result, we retain the evaluation setting in which it was obtained. Benchmark subsets and label definitions are stated explicitly, and we avoid direct comparisons when changes in the metric, backbone, or computing platform would distort the interpretation.
- Detection, segmentation cues, tracking, multimodal sensing, and hidden-target prediction are compared by the evidence each task retains, while their incompatible metrics remain separate.
- Practical evidence is examined separately from accuracy. We check whether runtime, FLOPs, parameter count, memory, input resolution, sensor setup, uncertainty, and false-warning cost are reported before treating a method as deployable.

## II. LITERATURE SEARCH AND REVIEW METHODOLOGY

The evidence base is limited to deep learning studies that process images or video from traffic scenes and address pedestrians or other VRUs under occlusion. Detection, tracking, and hidden-target inference are included because all three operate on incomplete visual observations. Transportation policy and non-visual warning systems fall outside the review, except when a system description is needed to interpret deployment requirements.

Searches were run in IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Web of Science, Scopus, the MDPI database, and Google Scholar. An occlusion expression (for example, occluded pedestrian, visible box, full-body detection, or hidden pedestrian) was paired with a road-user or task expression. The second group included pedestrian, cyclist, rider, VRU, crowded detection, multi-spectral detection, and tracking. After the database search, we followed references from benchmark papers and reviews and used citation searches to check the methods retained for analysis.

The review window runs from 2009 to 2025. We selected 2009 because Caltech-USA Pedestrian and the benchmark work that followed established much of the modern pedestrian-detection evaluation practice [44], [45]. Work predating 2009 appears only when a later comparison depends on its dataset, metric, or baseline. Searches for 2023-2025 work focused on context-aware detection, transformer tracking, feature completion, hidden-target reasoning, RGB-thermal transformers, occlusion-aware association, and real-time detectors [32]–[39], [46]–[48].

A paper entered the corpus when occlusion was part of either the method or the evaluation. This covered visual detection of pedestrians, cyclists, riders, and related VRUs; dense-scene pedestrian detection; visible-region or full-body annotation; compensation for weak RGB evidence with another modality; and inference of a target that was partly or fully hidden. Video work was retained when disappearance and reappearance were evaluated. For tracking and re-identification papers, we required an explicit connection to missed observations, identity switches, or track fragmentation rather than a generic MOT result.

Studies were left out when occlusion was incidental to a generic detector, or when the work concerned non-visual traffic systems, communication-only warnings, policy, or a perception task unrelated to detection, segmentation, tracking, or hidden-target reasoning. Generic detectors and foundation models are mentioned only when they explain an architecture, pretraining route, or deployment baseline used by an occlusion-aware method.

Title and abstract screening, followed by full-text review, left 96 studies in the final corpus. We did not attempt a statistical meta-analysis because the papers differed substantially

in dataset, split, backbone, input resolution, metric, sensing modality, and hardware. Numerical results are juxtaposed only when these conditions are reasonably aligned; otherwise, we discuss the evidence qualitatively and state the protocol mismatch.

Most numerical evidence comes from pedestrian benchmarks, principally Caltech, CityPersons, CrowdHuman, and pedestrian tracks in MOT datasets [12], [44], [15], [19]–[21]. Although this review uses VRU to include cyclists and riders, those groups are poorly represented in the reported occlusion results. We include category-specific results when they exist and do not extrapolate pedestrian scores to other road users. This asymmetry sets a clear limit on the review's conclusions.

## III. BACKGROUND: OCCLUSION IN APPLIED IMAGE AND VIDEO PERCEPTION

We tell apart four situations based on what the camera still manages to show. First, some portion of the target stays visible. Second, adjacent instances start to overlap each other. Third, the target you saw a bit earlier goes missing for a few video frames. Fourth, the system goes on predicting a road user even though no current target pixels are present. The first pair are mostly detection problems, the third one is the third one introduces association, and the fourth is more correctly handled as prediction. Any of these can delay or spoil a warning for a pedestrian or cyclist. Figure 1 connects the observation states to the failure modes and method families we look at next below.

### A. VRU CATEGORIES AND EVIDENCE BOUNDARIES

The benchmark record is kinda dominated by pedestrians, as in, that's what shows up the most. Caltech, CityPersons, CrowdHuman, and MOTChallenge give established evaluation for that category [12], [44], [15], [19]–[21]. Cyclists together with riders appear in Tsinghua-Daimler Cyclist, EuroCity Persons, BDD100K, nuScenes and Waymo [49]–[55], however these datasets seldom release standardized slices for cyclist or rider occlusion, so it's not really consistent.

Evidence from CityPersons or CrowdHuman concerns people, mostly pedestrians, under each dataset's own annotation rules. It is not evidence of cyclist performance. Conversely, a driving dataset that contains riders but lacks visibility labels can test transfer to traffic imagery, yet it cannot isolate robustness at a defined occlusion level. The tables keep those two types of evidence separate.

### B. OCCLUSION TYPES AND SEVERITY LEVELS

Occlusion can be described by occluder source, visibility loss, and temporal behavior. We use inter-class for a road user blocked by a vehicle or by infrastructure such as a pole, tree, sign, or building. Intra-class refers to pedestrians, cyclists, or riders blocking one another. Dynamic occlusion is caused by moving agents; static occlusion is caused by stationary objects; mutual occlusion is common in crowds and mixed traffic.

Because visibility categories differ across datasets, Table I uses a review-level descriptive scale. It is a common vocabulary for discussing how methods may behave as less of the target remains visible, not a replacement for the benchmark definitions. Every numerical result is still read using the source dataset's thresholds and subset rules.

### C. FAILURE MECHANISMS UNDER OCCLUSION

Occlusion changes both the visual input and the detector's decisions. A pedestrian may lose visible legs or a cyclist may lose visible wheels and handlebars; at the same time, the crop can contain features from the object in front. The visible extent then, differs from the annotated full-body box , which kind of makes localization ambiguous. In a crowd, the detector might put out two decent hypotheses and lose one, only later during matching or NMS [56]–[61]. In video, a streak of missed observations often shows up in the evaluation as a fragmented trajectory or even an identity switch [62], [19]–[21], [63].

What the useful mechanism is, really depends on what evidence still remains. Visible-part models, and attention, can still work when a recognizable fragment is there. But with fewer target pixels, a prediction leans more heavily on body priors, nearby objects, or a second sensor. And at zero current visibility, there is no observed region for a conventional detector, to localize. So the output then comes from temporal memory or scene context and should be treated as an uncertain prediction, including those false warnings.

## IV. DATASETS, ANNOTATIONS, AND METRICS

The benchmark picked for assessment sorta sets the edge of the final claim. In some datasets you can clearly see full body boxes, others lean on heavy overlap or identity continuity and yeah, multispectral datasets add a whole extra sensor stream. These kinds of evidence, they usually do not show up together in one single benchmark. So, Table II ends up listing what questions each dataset can answer, and also which questions its own protocol just cannot resolve.

The clearest occlusion labels in Table II come from CityPersons, Caltech and CrowdHuman, all centered on pedestrians, or people i guess. Datasets that have more road user categories make the scenes in which a method can be checked a bit wider, but their cyclist and rider instances are rarely separated into typical visibility groups. So they bring more category variety without really giving the severity controlled, kind of evidence that you get for pedestrians.

### A. VISIBLE-REGION, FULL-BODY, AND TRACKING ANNOTATIONS

Occluded-person datasets mostly depend on two kinds of box definitions. On one hand there is the visible box which sticks to only the observed pixels, on the other hand there is the full-body box that somehow goes behind the occluder

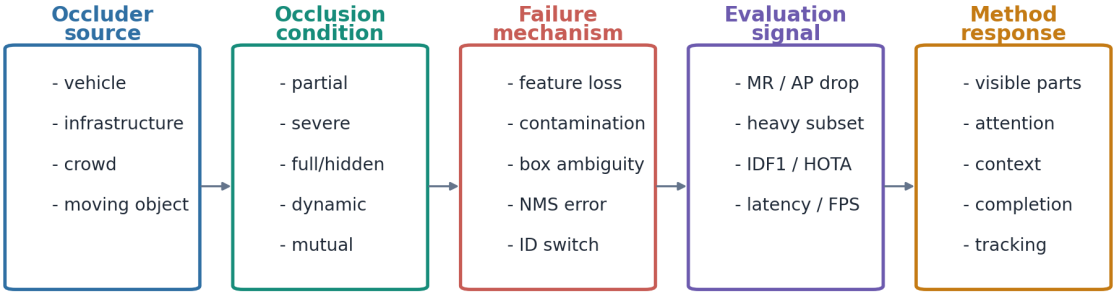


**Fig. 1.** From traffic-scene occlusion to system error. The diagram relates occluder sources and visibility states to common perception failures, the measurements used to observe them, and the method families intended to reduce them. The links organize the review and do not represent measured causal effects.

**TABLE I.** Review-level conceptual definition of occlusion severity. The visibility ranges are approximate; quantitative claims retain the definitions used by Caltech, CityPersons, CrowdHuman, KAIST, or the relevant source benchmark.

| Level | Visibility ratio | Typical traffic case | Main failure | Suitable metric |
|---|---|---|---|---|
| Bare | >80% visible | Pedestrian or cyclist mostly unobstructed | Ordinary scale, pose, or appearance variation | AP, MR, recall |
| Partial | 50-80% visible | Legs hidden by vehicle, cyclist partly blocked by traffic | Local evidence is lost; visible and full-body extents diverge | Subset MR/AP, visible/full-body localization error |
| Heavy | 20-50% visible | Pedestrian mostly behind parked car or dense crowd | Weak discriminative cues, feature contamination, box ambiguity | Heavy-subset MR/AP, false negative rate |
| Near-full | 0-20% visible | Only head, wheel, handlebar, or small body fragment visible | Severe uncertainty, context dependence, high false-alarm risk | Hidden/near-hidden recall, uncertainty, false positives |
| Full | 0% visible | Pedestrian hidden behind vehicle before entering road | No current pixels to detect; the task becomes risk prediction | Hidden-target detection, prediction F1, warning precision/recall |

to reach an estimated body boundary [12], [13], [15]. These labels end up supervising slightly different tasks: the first one points to evidence that is actually there, while the second one tells the model to guess a span that is not directly seen, and that guess gets more vague the more the person is blocked . CrowdHuman also includes a head box. That extra annotation can split neighboring people in crowded scenes, yet it does not retrieve the hidden body extent [15].

Frame-level labels kind of say not much about identity when theres a disappearance. Track annotations do the real work, they show if one pedestrian gets sort of split into a few short trajectories, or picks up a new identity when they reappear, or even gets wrongly linked to someone else inside a crowd [19]–[21] [22]–[24]. These differences matter quite a lot, especially when motion or even intention is inferred from a whole trajectory instead of just a single box.

### *B. METRICS AND PROTOCOL LIMITS*

Pedestrian-detection studies often report log-average miss rate, reflecting the importance of missed detections in safety-related settings [44], [45]. General and crowded-scene detectors more commonly use AP, while CrowdHuman also reports the Jaccard Index to expose duplicate and merged predictions [15]. Video benchmarks add an association dimension: MOTA, IDF1, HOTA, AssA, and identity-switch counts evaluate temporal continuity rather than treating every frame independently [22]–[24].

A single composite score would blur the distinctions among these metrics. CityPersons MR, CrowdHuman AP/JI, KAIST multispectral MR, and MOT HOTA evaluate different outputs and failure modes. Accordingly, we interpret each reported value together with its dataset, split, target definition, backbone, image scale, metric, and hardware rather than placing all values in one overall ranking.

## V. MECHANISM-BASED TAXONOMY FOR APPLIED OCCLUSION-AWARE VISUAL PERCEPTION

Section 5 groups methods by the evidence used when the target observation is incomplete or unreliable. Fig. 2 presents the taxonomy, while Fig. 3 shows representative mechanisms. Table III is the corresponding comparison matrix.

### *A. VISIBLE-PART AND BODY-STRUCTURE MODELING*

This family presumes that a recognizable body fragment survives the occlusion. DeepParts models parts explicitly, whereas OR-CNN, Bi-box regression, V2F-Net, and PR-Net++ combine visible-region cues with full-body localiza-

**TABLE II.** Occlusion-related evidence available from pedestrian and VRU image/video datasets. "Evidence strength" refers only to the occlusion claim listed in the table; it is not a judgment of a dataset's overall quality.

| Dataset | VRU category covered | Occlusion evidence strength | What it can support | What it cannot prove |
|---|---|---|---|---|
| Caltech-USA [44], [45] | Pedestrians | Moderate | Classical driving-scene pedestrian detection and visibility-related subsets | Modern dense traffic, cyclists, riders, full hidden targets |
| CityPersons [12] | Persons, riders in Cityscapes context | Strong for pedestrians; moderate for riders | Visible-region/full-body annotation, partial/heavy pedestrian evaluation | Dedicated non-pedestrian VRU occlusion robustness |
| CrowdHuman [15] | Humans in dense scenes | Strong for dense human occlusion | Mutual occlusion, full/visible/head boxes, NMS and duplicate errors | Traffic-specific VRU behavior and road context |
| WiderPerson [64] | Persons and riders | Moderate | Outdoor diversity, dense persons, broader person categories | Standardized occlusion severity comparison |
| EuroCity Persons [49] | Pedestrians, riders, related person classes | Moderate for traffic generalization | European traffic diversity, weather and city variation | Dedicated occlusion benchmark for cyclists/riders |
| Tsinghua-Daimler Cyclist [50] | Cyclists | Moderate for cyclist detection; weak for occlusion | Cyclist-specific visual detection in traffic | Systematic cyclist occlusion severity evaluation |
| BDD100K / KITTI / Cityscapes [51-53] | Pedestrians, riders, vehicles, traffic agents | Indirect | Traffic-domain generalization, scene diversity, segmentation/detection context | Occlusion-specific VRU ranking |
| nuScenes / Waymo / Argoverse [54], [55], [65] | Multimodal traffic agents | Indirect to moderate | Sensor complementarity, 3D/tracking context, broader traffic agents | 2D occlusion-aware pedestrian/cyclist benchmark comparison |
| JAAD / PIE [66], [67] | Pedestrians in driving video | Indirect to moderate | Behavior, intention, context, temporal cues | Large-scale detector robustness under occlusion |
| MOT17 / MOT20 [19], [20] | Pedestrian tracking | Strong for temporal continuity | Identity switches, fragmentation, dense tracking | Single-frame visible/full-body occlusion severity |
| DanceTrack [21] | Humans in motion | Moderate for association | Similar appearance and complex motion association | Traffic-specific occlusion and VRU semantics |
| KAIST / CVC-14 / LLVIP [16-18] | Pedestrians, multispectral | Moderate to strong for weak RGB evidence | RGB-thermal/infrared compensation, night/low visibility | General VRU categories and unified occlusion labels |
| OccluRoads [38] | Hidden pedestrians in road scenes | Strong for hidden reasoning | Partially/fully hidden pedestrian prediction | Direct comparison with visible-box detectors |

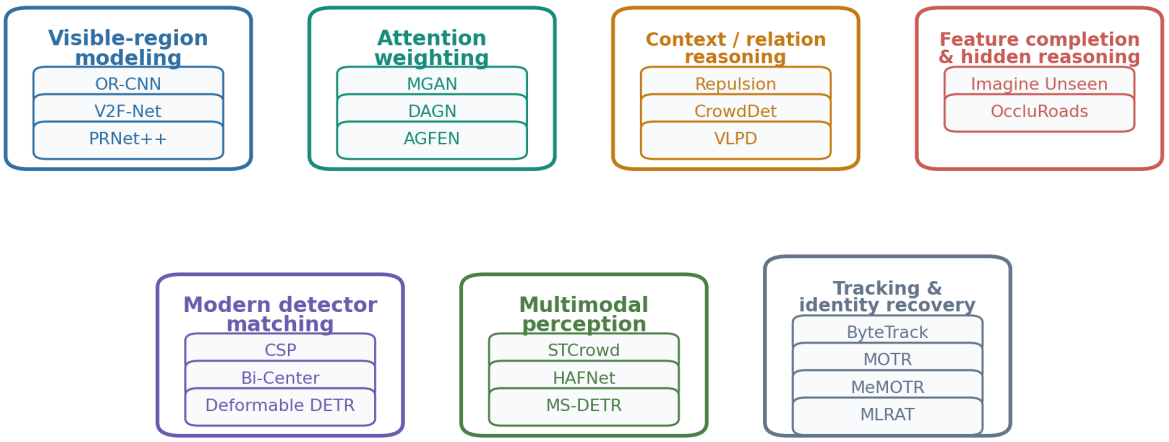


**Fig. 2.** Taxonomy used in this review. The branches follow the information or prior available under occlusion, not the publication date of a paper. A system may occupy more than one branch; the diagram identifies its mechanisms rather than assigning it a unique historical category.

tion [13], [68], [69]–[71]. The supporting evidence therefore comes mainly from partial and heavy occlusion, where the head, torso, or another meaningful fragment remains visible. Caltech, CityPersons, and CrowdHuman provide the

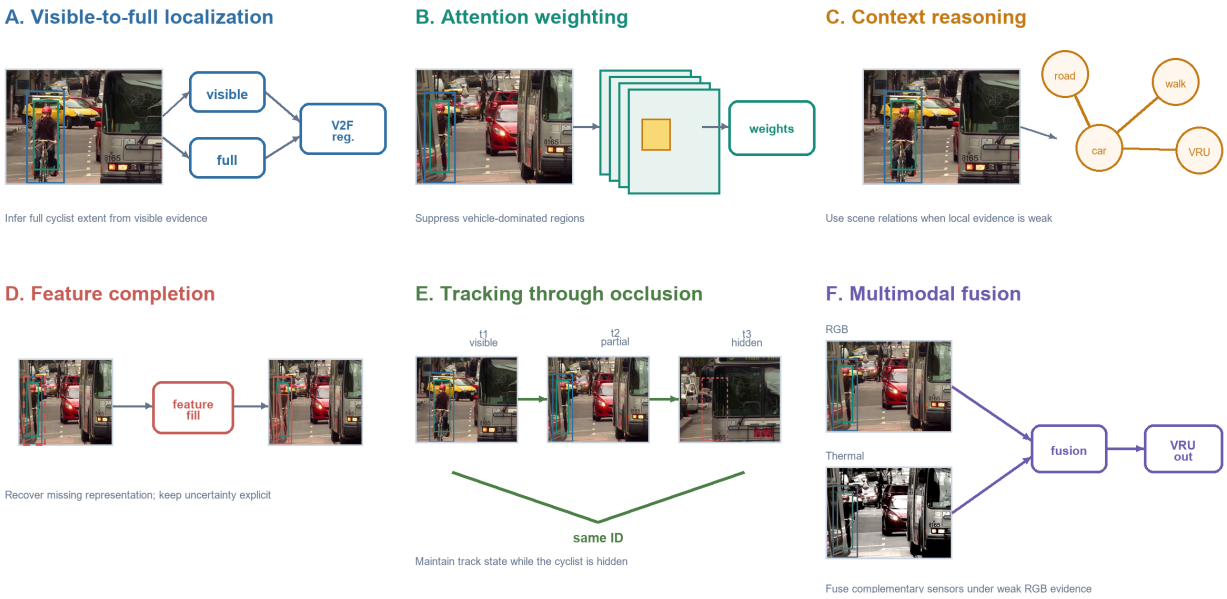


**Fig. 3.** Representative inputs and cues used across the taxonomy. The panels show visible and full-body boxes, attention on an observed region, neighboring instances, feature completion, earlier video observations, and a paired sensor stream. The crops illustrate the input conditions; the overlaid boxes and arrows explain the mechanism and are not presented as dataset annotations.

**TABLE III.** Mechanism-based taxonomy for applied occlusion-aware visual perception, with evidence and deployment implications. Rows may overlap because a modern system can combine several of the listed mechanisms.

| Method family | Core assumption | Representative methods | Best-suited occlusion level | Evidence dataset | Main limitation | Deployment cost |
|---|---|---|---|---|---|---|
| Visible-part and body-structure modeling | Observed body fragments are used to estimate the hidden extent | DeepParts, Bi-box, OR-CNN, V2F-Net, PRNet++ [13], [68], [69-72] | Partial to heavy | Caltech, CityPersons, CrowdHuman | Needs visible evidence and often visible/full-body labels | Moderate; extra branches or part heads |
| Attention and visible-region weighting | Reliable visible regions can be emphasized while occluders are suppressed | MGAN, DAGN, AGFEN [14], [73], [46] | Partial to heavy | CityPersons, Caltech, ECP, CrowdHuman | Cannot infer fully absent targets | Moderate; attention modules add cost |
| Context and relational reasoning | Scene relations and neighboring objects provide cues when local evidence is weak | Repulsion Loss, Adaptive NMS, CrowdDet, VLPD [32], [56-58] | Heavy and dense mutual occlusion | CrowdHuman, CityPersons, traffic scenes | Context bias and false positives | Moderate to high depending on relation modeling |
| Feature completion and generative reconstruction | Learned priors fill in features that the image does not show | Imagine the Unseen, completion-based Re-ID, V2F-style regression [37], [70], [74] | Heavy to near-full | CityPersons, CrowdHuman, Re-ID datasets | Hallucination and uncertainty control | Often high; adversarial or generative modules |
| Anchor-free and transformer-based detection | Global attention and alternative matching may reduce ambiguity | CSP, Bi-Center, Deformable DETR, OBhunter [75-78] | Partial to heavy, if trained accordingly | CityPersons, Caltech, COCO/OCHuman transfer | Not automatically occlusion-aware | Varies; transformers may be heavy |
| Multimodal and simulation-assisted perception | Non-RGB cues compensate for weak visible evidence | KAIST baselines, STCrowd, HAFNet, MS-DETR [16-18], [79-80, 47] | Partial to near-full under weak RGB | KAIST, CVC-14, LLVIP, STCrowd | Sensor alignment, missing modality, domain shift | Hardware and calibration cost |
| Temporal tracking and identity preservation | Temporal memory and association bridge short disappearance | ByteTrack, MOTR, MOTRv2, MeMOTR, StrongSORT, OC-SORT, OccluTrack [33], [34], [62], [63], [48], [81], [82] | Dynamic occlusion and reappearance | MOT17, MOT20, DanceTrack | Depends on detector recall; long occlusion remains hard | Low to high depending on tracker complexity |

principal evaluations.

The same dependence on remaining pixels defines the boundary of the approach. Near-full or full occlusion leaves body-structure priors weakly constrained, and several methods also require visible-region labels or an additional regression head. The extra computation is often moderate, but older two-stage designs may exceed an edge-device latency budget without further optimization.

### *B. ATTENTION AND VISIBLE-REGION WEIGHTING*

Attention-based designs attempt to distinguish target evidence from features contributed by the occluder or background. MGAN uses a visibility mask to guide one branch [14]; DAGN changes the sampled locations through deformable attention for heavily occluded pedestrians [73]; and AGFEN enhances features in crowded scenes [46]. Published gains are concentrated in partial and heavy occlusion, where the module still has some target pixels to emphasize.

Most evaluations use CityPersons, Caltech, EuroCity Persons, or CrowdHuman. Feature weighting can suppress contamination and strengthen a surviving cue, but it cannot create evidence when the target contributes no pixels. Full occlusion therefore lies outside the usual scope of these methods. Their computational cost is also hard to compare because FPS, parameter count, FLOPs, and measurement hardware are not reported consistently.

### C. CONTEXT AND RELATIONAL REASONING

Relational approaches compensate for a weak target crop by considering nearby instances or scene context. Repulsion Loss changes the training interaction between neighboring pedestrian boxes [56]; Adaptive NMS adjusts suppression according to local crowd density [57]; and CrowdDet permits several person hypotheses to arise from a single proposal [58]. VLPD draws on a different contextual signal by adding vision-language self-supervision [32].

Dense mutual occlusion is the main setting in which relational cues have been tested. Their benefit can reverse after a change of city, viewpoint, or traffic layout if the learned scene prior no longer holds. The risk is sharper for hidden-target inference: a plausible context can trigger a pedestrian prediction even when the scene contains none. Deployment cost ranges from low post-processing changes to heavier relation or language-guided modules.

### D. FEATURE COMPLETION AND GENERATIVE RECONSTRUCTION

Imagine the Unseen reconstructs pedestrian traits that were lost to occlusion [37] kinda, like not fully but enough to be tricky. The visible-to-full regression, and the occluded-person re-identification, end up yielding different outputs. Still, both try to extend the estimate beyond what you can actually see, that observed fragment [70], [74]. As the visible area gets smaller, the learned prior starts to matter a lot more, while the actual image evidence contributes less. Under near-full occlusion, there is basically little direct evidence left to verify if the filled in representation is correct.

A done feature can help out when the guessed content matches the hidden target, not the other way around. The same branch can also build a wrong body representation, or even just further entrench a target that is actually absent. For traffic scenes, detection accuracy alone is, consequently, not enough ; you still need uncertainty and false-positive behavior in order to reveal where it went wrong. The extra objective or generative branch may also raise the training AND inference cost, so that cost needs to be measured, not merely assumed.

### E. ANCHOR-FREE AND TRANSFORMER-BASED DETECTION

CSP and Bi-Center try to dodge anchor matching, via formulations that are more center oriented [75], [76]. Deformable DETR keeps the attention within a small bunch of sampled spatial spots [77], and OBhunter comes with a transformer representation, meant more for occluded things [78]. So these architectural tweaks shift how matching happens, and how features get gathered, but they do not, on their own, prove anything about robustness to occlusion. The whole conclusion still needs an evaluation that is built for occlusion, using a specific protocol.

Evidence is most convincing for partial and heavy occlusion because some target pixels remain available to context modules or alternative matching strategies. Many studies, however, report only generic detection scores. A strong COCO result is useful as a baseline, but it does not establish occlusion robustness. Such a claim needs visibility-stratified evaluation, visible/full-body localization analysis, dense-scene error measures, or identity continuity across video.

### F. MULTIMODAL AND SIMULATION-ASSISTED PERCEPTION

Multimodal methods supplement unreliable RGB imagery with non-RGB observations. The RGB-thermal literature has progressed from the KAIST baselines to cross-modal representation learning, weak-alignment methods, modality-imbalance correction, HAFNet, and MS-DETR [16]–[18], [80], [47], [83]–[86]. STCrowd extends this direction to crowded multispectral pedestrian perception [79]. These methods are especially relevant at night, under low contrast, during partial occlusion, and in other situations where RGB evidence is weak.

Those gains assume a more demanding sensing setup. Both devices must be available, their streams synchronized, and their calibration maintained as weather and operating domains change. Consequently, a result on KAIST cannot be transferred directly to a low-cost monocular roadside camera. Deployment-oriented reporting should specify the sensor arrangement, alignment assumptions, hardware, runtime, and performance when one modality is absent.

### G. TEMPORAL TRACKING AND IDENTITY PRESERVATION

Temporal methods assume that short-term disappearance can be handled by motion, appearance, memory, or association. The evidence ranges from SORT, DeepSORT, Tracktor, CenterTrack, and FairMOT to ByteTrack and OC-SORT. Query-based and memory-based alternatives include MOTR, MOTRv2, MeMOTR, StrongSORT, and OccluTrack [62], [63], [87-91, 81], [92, 82, 93]. Tracking is not a secondary issue for VRU image/video perception because warning and prediction depend on stable trajectories.

Tracking methods are most useful when a target reappears after a short dynamic occlusion. ByteTrack retains low-confidence detections for association; OC-SORT revises motion-based matching; transformer trackers propagate track queries; and memory modules preserve a longer history. None of these mechanisms can bridge an extended, complete occlusion if the detector supplies no credible observation. Detection and tracking metrics must therefore be reported together, because an improvement in one does not imply an improvement in the other.

## VI. PROTOCOL-AWARE COMPARATIVE ANALYSIS FOR APPLIED VISUAL SYSTEMS

### A. RELATION TO EXISTING SURVEYS

Table IV clarifies how this review differs from prior reviews. The goal is not to claim that earlier surveys are weak; rather,

it is to specify why they do not fully address occlusion-aware pedestrian and VRU detection in image/video traffic scenes.

### B. PROTOCOL-AWARE DETECTION COMPARISON UNDER PEDESTRIAN AND CROWDED OCCLUSION BENCHMARKS

Table V separates single-frame pedestrian and crowded-scene detection from multimodal reasoning and tracking. These rows should be interpreted within their own benchmark protocols and not as a single leaderboard. For compactness in Tables V-VII, NR means that hardware or FPS was not reported in the source paper or was not summarized in the table.

The strongest detection evidence concerns occluded pedestrians, not all VRUs. CityPersons supports partial/heavy pedestrian occlusion; CrowdHuman supports dense mutual occlusion; Caltech provides historical driving-scene continuity. None of these fully resolves cyclist or rider occlusion.

### C. MULTIMODAL IMAGE/VIDEO FUSION AND HIDDEN-TARGET REASONING EVIDENCE

Multimodal and hidden-target methods address a different problem from ordinary visible-box detection. They should therefore be separated from Table V.

Table VI is not an extension of the RGB-only detector ranking. An RGB-thermal system observes an extra signal and entails a second sensor installation. A hidden-target method may output a risk estimate even when no visible box exists, so both the output and the cost of error change. Putting either result on a visible-box leaderboard would obscure those differences.

### D. TRACKING AND IDENTITY PRESERVATION UNDER OCCLUSION

Table VII , is about identities not so much isolated detections. Like the main question i guess is whether a target stays the same target after a spell of disappearance. So the tracking outputs are deliberately kept separate , from the frame level comparisons, sort of like different boxes.

Tracking evidence is a bit operational, like it helps in practice but ,doesn't really answer the same question as occlusion-aware detection. A tracker might reconnect an identity after a short gap you know ,and it may still "work", but the result really hangs on detector recall and also on how solid those remaining observations are.

### E. SEVERITY-AWARE AND DEPLOYMENT-ORIENTED IMAGE/VIDEO SYNTHESIS

The evidence you can use is not spread evenly across the visibility levels, it kind of depends. Partial occlusion is handled most thoroughly, mainly because visible body cues stay, and several benchmarks spell out visibility subsets too. Heavy occlusion, on the other hand, has drawn more attention, with body-structure models, crowd reasoning, contextual hints, and multimodal sensing all showing up. Near-full and full occlusion are way less standardized though. For those settings, hidden target prediction, uncertainty modeling, and false warning analysis tend to be a better fit, rather than AP or MR by itself. Figure 4 pulls together selected results inside their original protocols, but Figures 5-7 go further, basically they merge robustness, deployment constraints, failure coverage and benchmark support.

Runtime comparisons get, kind of weakened because the reporting is not fully there. Like, a FPS number is hard to repeat or verify unless the input resolution, the hardware, and the whole software stack are stated. Also FLOPs and parameter counts, they really say little about memory behavior or sensor related overhead. And the resource limits themselves differ a lot between roadside servers, inside the vehicle processors, and those small edge devices. So any claim of "real time" should be taken mostly as true only within the exact conditions where it was measured, not as a general promise.

## VII. RECENT TRENDS IN DEPLOYMENT-AWARE OCCLUSION-ROBUST VISION

Table VIII is kinda centered on how the evidence changes for a system, not just publication year, period. The rows, they separate the semantic context stuff, memory that spans across frames, extra sensing additions, hidden-target forecasting, and lower detector latency. Figure 8 then lines up these more technical turns on a 2023–2025 timeline so it feels like, well, a rough ordering.

### A. CONTEXT, LANGUAGE, AND FOUNDATION-MODEL CUES

Recent context-aware work suggests that local appearance is insufficient when a VRU is heavily occluded. VLPD uses vision-language semantic self-supervision to improve pedestrian detection through contextual semantics [32]. SAM, Grounding DINO, YOLO-World, and other foundation or open-vocabulary models add region-level and language-conditioned cues [40]–[43]. In this problem, their most plausible uses are pseudo-label generation, weakly supervised scene analysis, open-set handling of VRU categories, and context-aware pretraining.

That potential should not be confused with direct evidence on occluded VRUs. Segmenting the visible fragment of a pedestrian does not show that a generic foundation model can recover the full-body extent, maintain an identity, or infer a fully hidden road user. Calling such a model occlusion-aware requires evaluation on visibility subsets and on traffic protocols for pedestrians, cyclists, or riders.

### B. FROM VISIBLE ENHANCEMENT TO HIDDEN-TARGET PREDICTION

Most detector families still want at least some kind of visible fragment. Feature completion kind of reduces that reliance, even so its reconstruction stays tied to what we can actually observe, those cues. [37]. OccluRoads handles a different

**TABLE IV.** Comparison with representative prior reviews. The last column records the part of occlusion-aware traffic-scene VRU detection that remains outside each review's main scope.

| Prior survey | Main coverage | What it contributes | What it misses for this topic | Why insufficient for occlusion-aware VRU detection |
|---|---|---|---|---|
| Object detection surveys [5], [7] | Generic object detection architectures | Detector evolution, backbones, losses, benchmarks | Traffic VRU scope, occlusion severity, visible/full-body annotations | Too broad; occlusion and VRU are not organizing axes |
| Autonomous-driving perception surveys [2-4] | Driving datasets, sensors, perception tasks | Broad traffic-scene context and multimodal sensing | Pedestrian occlusion subsets, full-body/visible boxes, tracking continuity under occlusion | Useful background but not focused on occlusion-aware detection evidence |
| Pedestrian detection surveys [6], [30] | Pedestrian detectors, datasets, generalization | Strong pedestrian benchmark discussion | Recent hidden-target, multimodal, and tracking-centered occlusion analysis | Pedestrian-centered and not deployment/occlusion-severity driven |
| Generic occlusion handling review [25] | Occlusion in object detection broadly | Treats occlusion as central problem | Traffic-scene VRUs, pedestrian benchmarks, deployment feasibility | Occlusion-focused but not VRU- or traffic-specific |
| Occluded pedestrian review [26] | Occluded pedestrian detection | Close to visible/occluded pedestrian scope | 2023-2025 trends, broader VRUs, protocol-aware deployment analysis | Important precursor, but less current and less deployment-oriented |
| VRU safety survey [1] | VRU safety, sensing, traffic applications | Broad safety motivation and VRU relevance | Detailed occlusion mechanisms, datasets, quantitative comparison | Safety-oriented rather than detection-protocol-oriented |
| MOT surveys [27-29] | Multi-object tracking | Association metrics, tracking pipelines, temporal modeling | Single-frame visible/full-body detection, pedestrian occlusion subsets | Tracking-centered; detector-side occlusion remains secondary |
| This review | Occlusion-aware pedestrian and VRU detection | Scope, severity, taxonomy, protocol-aware comparison, deployment, recent trends | Acknowledges limited cyclist/rider evidence | Directly targets occlusion-aware traffic-scene detection and its evidence gaps |

**TABLE V.** Protocol-aware detection comparison under pedestrian and crowded occlusion benchmarks. Results are not directly comparable across datasets, splits, backbones, and metrics.

| Method | Year | Dataset | Split/subset | Backbone | Metric | Reported result | Hardware/FPS | Evidence type | Main limitation |
|---|---|---|---|---|---|---|---|---|---|
| OR-CNN [13] | 2018 | CityPersons | val; Reasonable/Heavy/Partial/Bare | VGG-16, x1.3 | MR^-2 | 11.0 / 51.3 / 13.7 / 5.9 | NR | Primary fixed-protocol result | Pedestrian-only; two-stage design |
| MGAN [14] | 2019 | CityPersons, Caltech | heavy occlusion | VGG-style detector | MR gain | 9.5-point CityPersons gain; 5.0-point Caltech gain | NR | Primary subset-gain result | Improvement is protocol-specific and should not be compared across settings |
| V2F-Net [70] | 2021 | CrowdHuman, CityPersons | full-body / val | FPN-based | AP, MR^-2 | +5.85 AP on CrowdHuman; -2.24 MR^-2 on CityPersons vs FPN | NR | Primary baseline-relative result | Read protocol details in the context of the original paper |
| DAGN [73] | 2021 | Caltech, CityPersons, ECP | heavy subsets | Deformable attention network | MR^-2 gain | 12.44-point Caltech reduction; about 5-point gains on CityPersons/ECP | NR | Primary reported improvement | Not directly comparable across datasets |
| Repulsion Loss [56] | 2018 | CityPersons / dense pedestrian settings | crowded pedestrians | Faster R-CNN-style | MR/AP depending protocol | Source-protocol dense-interaction result; CrowdHuman figures require caution | NR | Primary method; some CrowdHuman values are secondary reported results | Secondary reported results should not be used for unified ranking |
| Adaptive NMS [57] | 2019 | CrowdHuman | validation full-body | FPN | MR^-2, AP | MR 52.35 to 49.73; AP 83.07 to 84.71 | NR | Primary CrowdHuman result | Dense-human benchmark; not directly traffic-specific |
| **Method** | **Year** | **Dataset** | **Split/subset** | **Backbone** | **Metric** | **Reported result** | **Hardware/FPS** | **Evidence type** | **Main limitation** |
| CrowdDet [58] | 2020 | CrowdHuman | validation | ResNet-50/FPN | AP, MR, JI | AP 90.7, MR 41.4, JI 82.3 | NR | Primary CrowdHuman result | Dense-human evidence; road context absent |
| AGFEN [46] | 2024 | CrowdHuman | validation | attention-guided detector | AP, MR, JI | AP 92.52, MR 40.54, JI 83.44 | NR | Primary recent dense-scene result | Transfer and deployment cost require separate evaluation |

circumstance , by leaning on contextual relations, it helps infer a pedestrian who could be entirely hidden. [38].When the current frame contains no target pixels, the output is better interpreted as a risk estimate than as conventional detection. Uncertainty and false alarms therefore become part of the primary evaluation.

### C. TEMPORAL AND MULTIMODAL REASONING

Video provides evidence that is unavailable in a single frame. When a VRU passes behind a vehicle, a tracker can preserve its previous state and compare the returning detection with that history [33], [34], [63], [48]. IDF1, HOTA, and identity-switch counts reflect this continuity; AP and MR do not. RGB-thermal methods compensate for weak visible imagery through a paired thermal or infrared observation [16]–[18], [80], [47]. The two solutions incur different costs: temporal memory adds association state, whereas multispectral fusion adds a sensor and its calibration.

### D. DEPLOYMENT-AWARE DETECTION

YOLOv10, D-FINE, RT-DETR, YOLOX, and YOLOv7 are relevant real-time detector baselines [35], [36], [39], [94]–[96]. Their speed-accuracy results make them possible

**TABLE VI.** Multimodal image/video fusion and hidden-target reasoning evidence. These rows are not directly comparable with RGB-only pedestrian detection because modalities, tasks, and metrics differ.

| Method | Year | Dataset | Split/subset | Backbone/modality | Metric | Reported result | Hardware/FPS | Evidence type | Main limitation |
|---|---|---|---|---|---|---|---|---|---|
| HAFNet [80] | 2023 | KAIST | occlusion subsets | RGB-thermal hierarchical attention | MR | partial-occlusion MR 30.10 vs BAANet 34.07 | 0.017 s/frame on GTX 1080Ti reported | Primary multispectral evidence | A registered thermal stream is required alongside RGB |
| MS-DETR [47] | 2024 | KAIST-style multispectral benchmarks | multispectral pedestrian detection | RGB-thermal DETR fusion | MR/AP depending protocol | Source-paper multispectral detection results; not reproduced numerically in this review table | NR | Primary transformer-fusion evidence | Sensor-dependent; not directly comparable with RGB-only detectors |
| Imagine the Unseen [37] | 2024 | occluded pedestrian settings | feature completion | adversarial feature completion | detection metrics in source paper | Source-paper occluded-pedestrian gains; not reproduced numerically in this review table | NR | Feature-completion evidence | arXiv status; uncertainty and hallucination risk |
| OccluRoads [38] | 2025 | road hidden-pedestrian benchmark | partially/fully hidden pedestrians | knowledge graph/KGE + Bayesian reasoning | F1 / task-specific metrics | F1 0.91; up to 42% over traditional ML baselines | Real-time cost not established | Hidden-target reasoning evidence | Different task from visible-box detection; false-warning control needed |

**TABLE VII.** Video tracking and identity-preservation evidence under occlusion. Tracking evidence should be interpreted with detection quality and association protocol.

| Method | Year | Dataset | Split/subset | Backbone / association | Metric | Reported result | Hardware/FPS | Evidence type | Main limitation |
|---|---|---|---|---|---|---|---|---|---|
| ByteTrack [62] | 2022 | MOT17 | test | YOLOX + low-score association | MOTA, IDF1, HOTA | MOTA 80.3, IDF1 77.3, HOTA 63.1 | 29.6 FPS on V100 reported | Primary tracking benchmark | Tracker depends on detector recall |
| ByteTrack [62] | 2022 | MOT20 | test, dense scenes | YOLOX + association | MOTA, IDF1, HOTA | MOTA 77.8, IDF1 75.2, HOTA 61.3 | NR | Dense tracking result | Pedestrian tracking, not detector occlusion subsets |
| MOTR [63] | 2022 | MOT17 | tracking benchmark | transformer track queries | HOTA, MOTA, IDF1 | HOTA 57.8, MOTA 73.4, IDF1 68.6 | NR | Primary end-to-end tracking result | More costly than simple association; not detector-only evidence |
| MOTRv2 [33] | 2023 | MOT benchmarks | proposal-bootstrapped MOT | detector proposals + track queries | HOTA/MOTA/IDF1 | Source-paper tracking results; not reproduced numerically in this review table | Depends on detector and implementation | Recent transformer tracking evidence | Strong detector proposals are still needed |
| MeMOTR [34] | 2023 | DanceTrack / MOT | memory-augmented tracking | long-term memory | HOTA, AssA | +7.9 HOTA and +13.0 AssA over prior SOTA on reported setting | NR | Primary memory-tracking evidence | Memory can propagate wrong associations |
| StrongSORT [82] | 2023 | MOT benchmarks | tracking-by-detection | stronger Re-ID and association | MOTA/IDF1/HOTA | Source-paper MOT results; not reproduced numerically in this review table | NR | Association baseline evidence | Not specifically occlusion-aware |
| OC-SORT [81] | 2023 | MOT17 | test | observation-centric motion association | HOTA, MOTA, IDF1 | HOTA 63.2, MOTA 78.0, IDF1 77.5 | NR | Primary motion-centric tracking result | Short-term motion cannot solve long full occlusion |
| OccluTrack [48] | 2025 | MOT/DanceTrack-style settings | occlusion-aware pedestrian tracking | motion suppression + pose/Re-ID association | IDF1, ID switches, AssA/AssR | Source-paper identity-continuity improvements; exact values not reproduced in this review table | NR | Recent occlusion-aware tracking evidence | Primarily improves post-detection association |

**TABLE VIII.** Recent technical trends in occlusion-aware pedestrian and VRU image/video perception. The table highlights what changed technically, not just which papers were published.

| Trend | Representative work | What changed | Relevance to occlusion | Remaining concern |
|---|---|---|---|---|
| Context and semantic priors | VLPD [32] | Uses vision-language semantic self-supervision | Local visible evidence is supplemented by scene semantics | Scene regularities may not transfer to unfamiliar locations |
| Part-guided transformers | Parts-based attention transformers [72] | Injects body-part bias into transformer attention | Combines structure and global modeling | Still needs visible body evidence |
| Feature completion | Imagine the Unseen [37] | Completes missing occluded-pedestrian features | Moves beyond visible-only enhancement | Hallucination and uncertainty must be evaluated |
| Hidden-target reasoning | OccluRoads [38] | Predicts partially/fully hidden pedestrians | Treats full occlusion as risk inference | Metrics and false-warning costs remain immature |
| Transformer tracking and memory | MOTRv2, MeMOTR [33], [34] | Uses detector proposals and long-term memory | Maintains identity through disappearance/reappearance | Tracking evidence differs from detection evidence |
| Occlusion-aware association | OccluTrack [48] | Models occlusion in pedestrian tracking | Reduces identity fragmentation | Depends on detector and Re-ID reliability |
| Multispectral transformer fusion | HAFNet, MS-DETR [80], [47] | Uses RGB-thermal hierarchy or DETR fusion | Compensates for weak RGB under occlusion/low visibility | Sensor cost, calibration, and domain shift |
| Deployment-aware real-time detection | YOLOv10, D-FINE, RT-DETR [35], [36], [39] | Improves speed-accuracy trade-off and NMS-free deployment | Provides deployable detector backbones | Needs occlusion-specific VRU evaluation |

backbones for a traffic system, but do not demonstrate an occlusion solution. Evidence for that stronger claim would need visibility-stratified results and, from the same implementation, FPS, input size, hardware, FLOPs, parameter

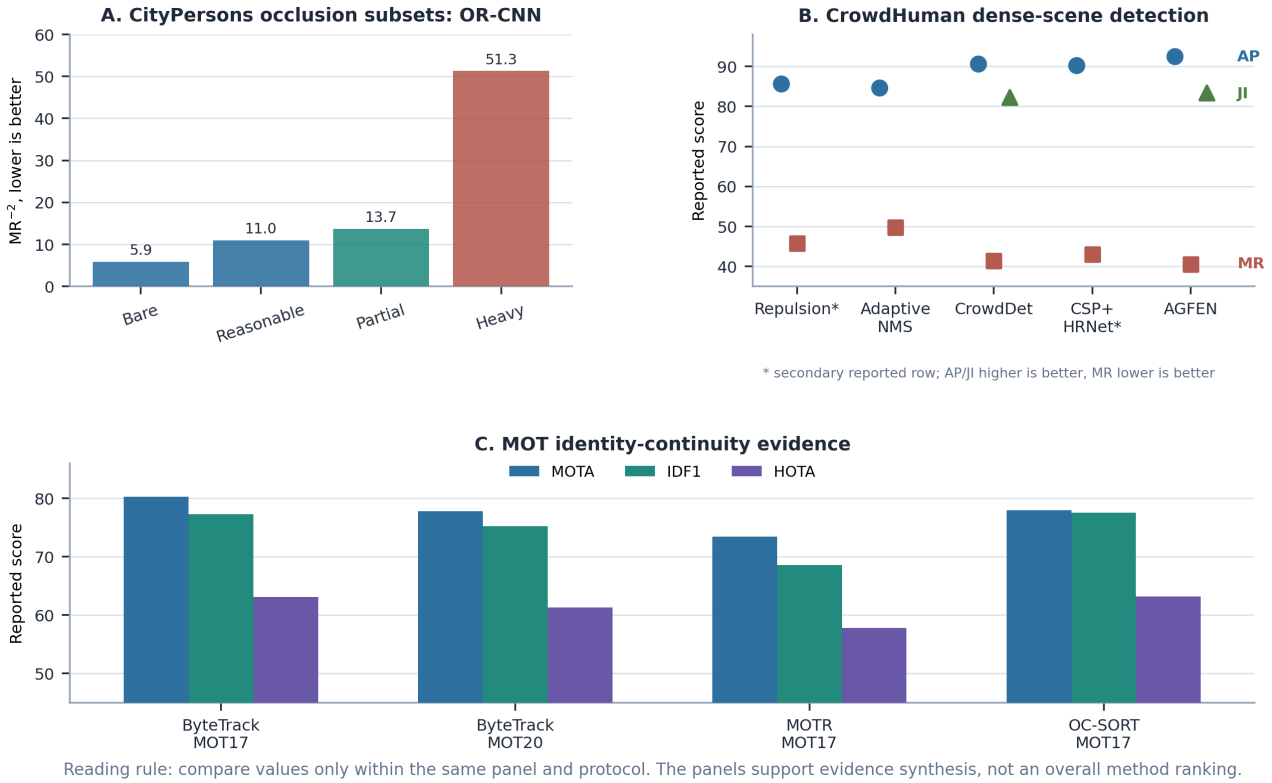


**Fig. 4.** Protocol-aware quantitative evidence from representative benchmarks. The panels summarize selected protocol-specific results for pedestrian occlusion subsets, dense human detection, and MOT identity-continuity evaluation. The CrowdHuman panel uses unconnected points because the methods are categorical rather than temporally ordered. The figure is not a unified leaderboard.

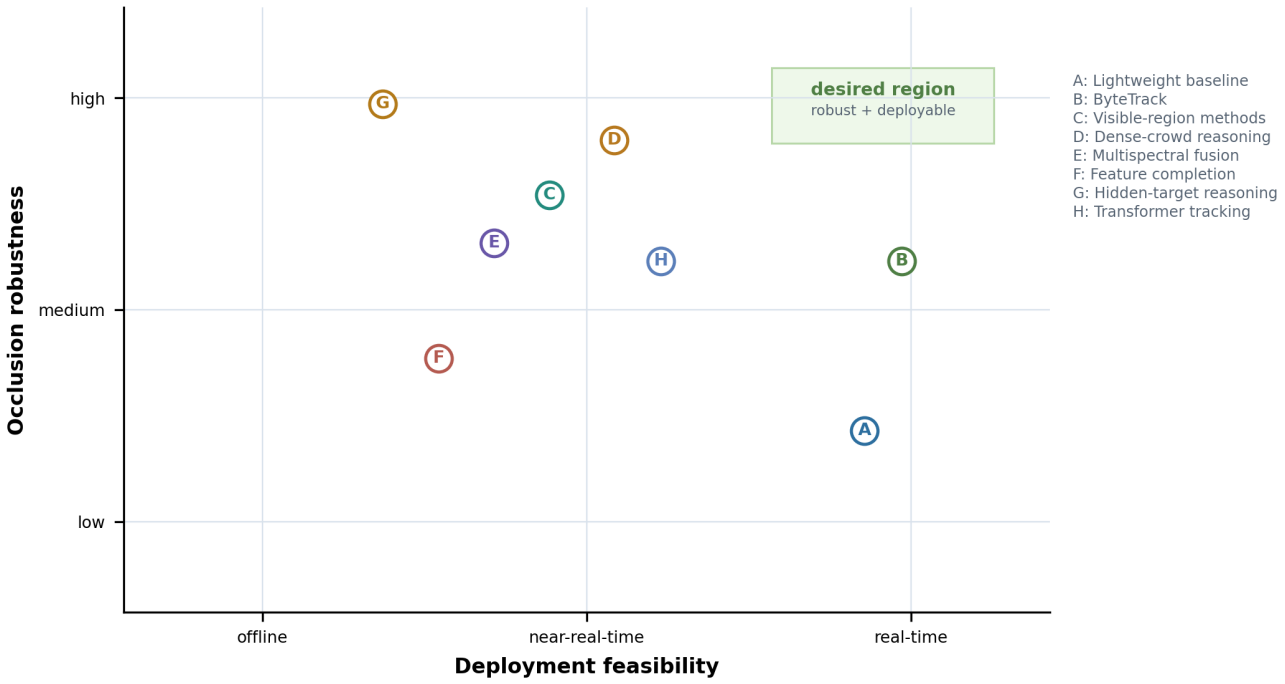


**Fig. 5.** Qualitative robustness-deployment map of method families for applied visual systems. The map summarizes conceptual tendencies derived from Tables V-VII. Positions are qualitative and should not be interpreted as measured coordinates.

count, and memory.


**Fig. 6.** Failure-mode coverage by method family. The heatmap indicates which method families conceptually address feature loss, feature contamination, localization ambiguity, NMS or assignment failure, full hidden targets, and identity switches. It is a qualitative synthesis, not a benchmark ranking.

## VIII. LIMITATIONS OF THIS REVIEW

The paper's VRU scope is wider than its quantitative evidence base. Caltech, CityPersons, CrowdHuman, and MOTChallenge provide relatively strong results for pedestrians, but cyclists and riders seldom appear in dedicated occlusion subsets. Claims about those categories must therefore remain cautious, and comparable visibility annotations are still unavailable.

Cross-paper comparison is constrained by changes in backbone, input size, training data, and evaluation split. The metrics also refer to different outputs. CityPersons MR and CrowdHuman AP/JI score detections; KAIST MR is obtained with paired sensing; OccluRoads F1 concerns hidden-target inference; and MOT HOTA includes association. The corresponding results remain in separate protocol groups rather than a common leaderboard. Values are taken from the source papers because the code, training configuration, and evaluation environment needed for a unified rerun were not consistently available.

Evidence about computational cost is notably sparse. An accuracy table may omit memory use, input resolution, hardware, or even the software stack used to obtain an FPS value. The omission is consequential because feature-completion branches, transformers, memory-based trackers, and multispectral systems add cost at different points in the pipeline.

Visible-box detection and hidden-target prediction are not the same task. A detector localizes a currently observed region, and a tracker associates such regions over time. By contrast, a hidden-target model estimates risk when no current target pixels are available. Near-full and full-occlusion benchmarks need to state which output is expected and evaluate uncertainty and false warnings as well as recall.

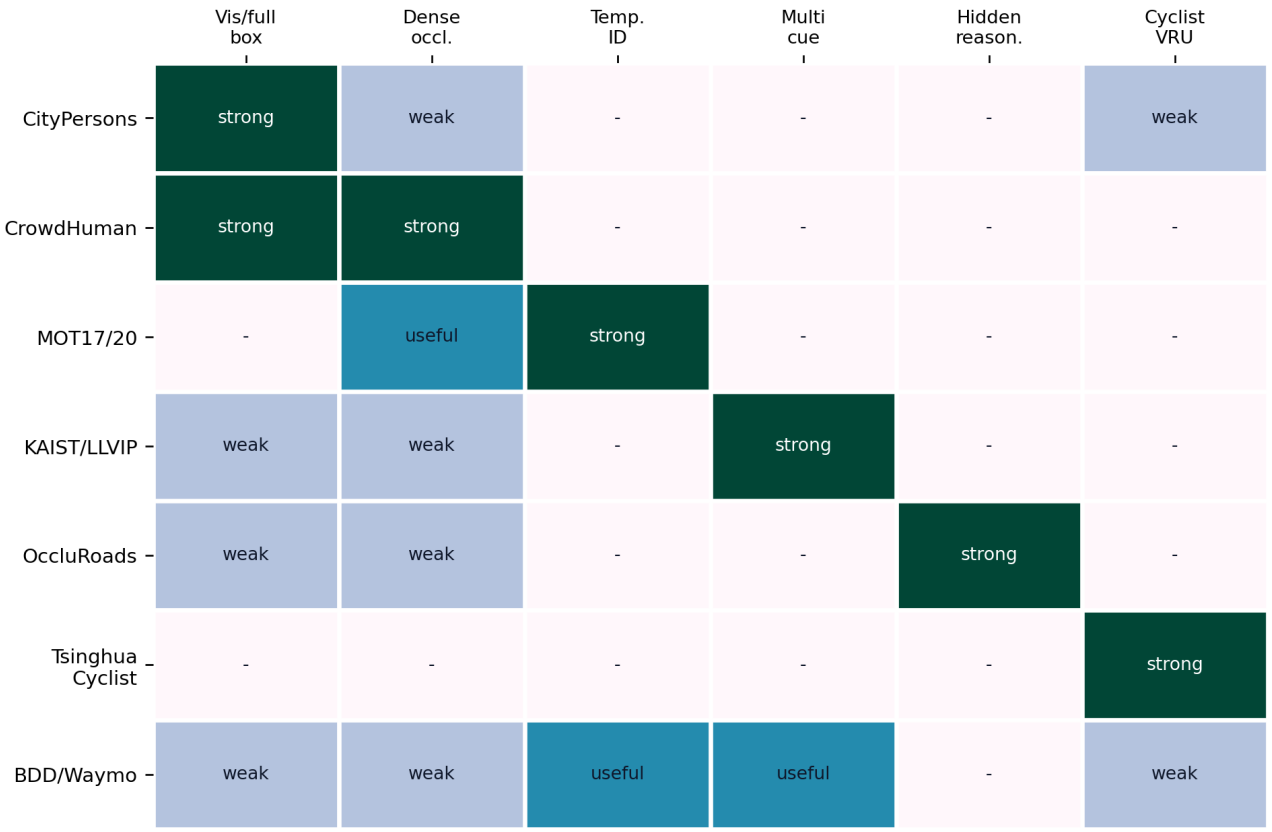


**Fig. 7.** Evidence supplied by the benchmarks used in this review. The map marks whether a dataset supports analysis of visible/full-body boxes, dense overlap, temporal identity, paired sensing, hidden-target inference, or non-pedestrian VRUs. It describes dataset suitability, not method performance.

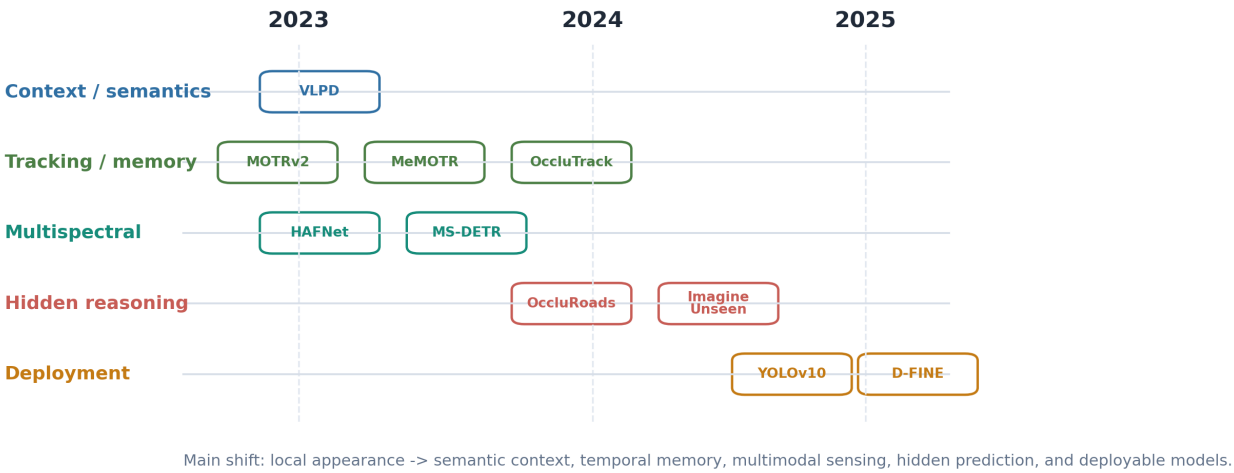


**Fig. 8.** Recent technical shifts in occlusion-aware pedestrian and VRU image/video perception. The timeline groups 2023-2025 representative studies by technical direction, including context-aware detection, part-guided transformers, multispectral fusion, transformer tracking, hidden-pedestrian reasoning, feature completion, and deployment-aware real-time detection.

## IX. CONCLUSION

Tables V-VII are complementary evidence summaries rather than a unified ranking. A low miss rate on a reasonable pedestrian subset does not resolve behavior under dense, heavy overlap. Likewise, strong frame-level detection may

coexist with identity failure after reappearance, and an RGB-thermal improvement presumes hardware absent from a monocular roadside installation. AP and MR are informative only when read with the visibility definition, task protocol, and sensing configuration that produced them.

For a detector, the informative combination is a stated visible/full-body annotation rule, an occlusion subset, and the runtime conditions. A tracking study must additionally show whether identity survives disappearance. If the system infers a fully hidden pedestrian, its uncertainty and false-warning cost are more appropriate than the evaluation language of visible-box detection. Existing benchmarks permit these analyses far more often for pedestrians than for cyclists or riders.

Recent studies obtain more evidence than the current RGB crop alone can provide. They kind of draw from scene context , earlier frames, or maybe even a second sensor; while completion and hidden-target methods also estimate content that the camera can't really observe directly. In our view the immediate benchmark gap is not just another overall ranking. It's more like a clearly specified image or video protocol, that separates visibility regimes and then reports enough computational detail so you can actually identify the hardware where a method is practical.

## REFERENCES

[1] R. M. Silva, G. F. Azevedo, M. V. V. Berto, J. R. Rocha, E. C. Fidelis, M. V. Nogueira, P. H. Lisboa, and T. A. Almeida, "Vulnerable road user detection and safety enhancement: A comprehensive survey," Expert Systems with Applications, vol. 292, Art. no. 128529, 2025, https://doi.org/10.1016/j.eswa.2025.128529.

[2] J. Janai, F. Guney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," Foundations and Trends in Computer Graphics and Vision, vol. 12, no. 1-3, pp. 1–308, 2020.

[3] D. Feng, C. Haase-Schuetz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 3, pp. 1341–1360, 2021.

[4] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," Journal of Field Robotics, vol. 37, no. 3, pp. 362–386, 2020.

[5] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikainen, "Deep learning for generic object detection: A survey," International Journal of Computer Vision, vol. 128, pp. 261–318, 2020, https://doi.org/10.1007/s11263-019-01247-4.

[6] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Pedestrian detection: Domain generalization, CNNs, transformers and beyond," arXiv:2201.03176, 2022, https://doi.org/10.48550/arXiv.2201.03176.

[7] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212–3232, 2019.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017, https://doi.org/10.1109/TPAMI.2016.2577031.

[9] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944, https://doi.org/10.1109/CVPR.2017.106.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proc. IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007, https://doi.org/10.1109/ICCV.2017.324.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Proc. European Conference on Computer Vision (ECCV), 2020, pp. 213–229, https://doi.org/10.1007/978-3-030-58452-8_13.

[12] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4457–4465, https://doi.org/10.1109/CVPR.2017.474.

[13] S. Zhang, J. Yang, and B. Schiele, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in Proc. European Conference on Computer Vision (ECCV), 2018, pp. 657–674, https://doi.org/10.1007/978-3-030-01219-9_39.

[14] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network for occluded pedestrian detection," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4966–4974, https://doi.org/10.1109/ICCV.2019.00507.

[15] 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:1805.00123, 2018.

[16] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1037–1045, https://doi.org/10.1109/CVPR.2015.7298706.

[17] A. Gonzalez, Z. Fang, Y. Socarras, J. Serrat, D. Vazquez, J. Xu, and A. M. Lopez, "Pedestrian detection at day/night time with visible and FIR cameras: A comparison," Sensors, vol. 16, no. 6, Art. no. 820, 2016.

[18] X. Jia et al., "LLVIP: A visible-infrared paired dataset for low-light vision," in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021.

[19] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," arXiv:1603.00831, 2016.

[20] P. Dendorfer et al., "MOT20: A benchmark for multi object tracking in crowded scenes," arXiv:2003.09003, 2020.

[21] P. Sun et al., "DanceTrack: Multi-object tracking in uniform appearance and diverse motion," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[22] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," EURASIP Journal on Image and Video Processing, vol. 2008, Art. no. 246309, 2008.

[23] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in Proc. European Conference on Computer Vision Workshops (ECCVW), 2016, pp. 17–35.

[24] J. Luiten et al., "HOTA: A higher order metric for evaluating multi-object tracking," International Journal of Computer Vision, vol. 129, pp. 548–578, 2021.

[25] K. Saleh, S. Szenasi, and Z. Vamossy, "Occlusion handling in generic object detection: A review," in Proc. IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), 2021, pp. 477–484, https://doi.org/10.1109/SAMI50585.2021.9378657.

[26] J. Li et al., "Occlusion handling and multi-scale pedestrian detection based on deep learning: A review," IEEE Access, 2022.

[27] Z. Guan, Z. Wang, G. Zhang, L. Li, M. Zhang, Z. Shi, and N. Jiang, "Multi-object tracking review: Retrospective and emerging trend," Artificial Intelligence Review, vol. 58, Art. no. 235, 2025, https://doi.org/10.1007/s10462-025-11212-y.

[28] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple object tracking: A literature review," Artificial Intelligence, vol. 293, Art. no. 103448, 2021.

[29] G. Ciaparrone, F. L. Sanchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," Neurocomputing, vol. 381, pp. 61–88, 2020.

[30] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," Neurocomputing, vol. 300, pp. 17–33, 2018.

[31] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 2872–2893, 2022.

[32] W. Liu et al., "VLPD: Context-aware pedestrian detection via vision-language semantic self-supervision," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[33] Y. Zhang, T. Wang, and X. Zhang, "MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors," in Proc. IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22056–22065, https://doi.org/10.1109/CVPR52729.2023.02112.

[34] R. Gao and L. Wang, "MeMOTR: Long-term memory-augmented transformer for multi-object tracking," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[35] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, 2024, https://doi.org/10.52202/079017-3429.

[36] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement," in Proc. International Conference on Learning Representations (ICLR), 2025. Available: https://openreview.net/forum?id=MFZjrTFE7h.

[37] S. Zhang, M. Ji, Y. Li, and J. Yang, "Imagine the unseen: Occluded pedestrian detection via adversarial feature completion," arXiv:2405.01311, 2024, https://doi.org/10.48550/arXiv.2405.01311.

[38] A. N. Melo, S. M. Serrano, C. Salinas, and M. A. Sotelo, "Prediction of occluded pedestrians in road scenes using human-like reasoning: Insights from the OccluRoads dataset," in Proc. IEEE Intelligent Vehicles Symposium (IV), 2025, pp. 385–391, https://doi.org/10.1109/IV64158.2025.11097510.

[39] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16965–16974, https://doi.org/10.1109/CVPR52733.2024.01605.

[40] A. Kirillov et al., "Segment anything," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4015–4026, https://doi.org/10.1109/ICCV51070.2023.00371.

[41] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in Proc. European Conference on Computer Vision (ECCV), 2024, pp. 38–55, https://doi.org/10.1007/978-3-031-72970-6_3.

[42] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-World: Real-time open-vocabulary object detection," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16901–16911, https://doi.org/10.1109/CVPR52733.2024.01599.

[43] M. Minderer et al., "Simple open-vocabulary object detection with vision transformers," in Proc. European Conference on Computer Vision (ECCV), 2022, pp. 728–755.

[44] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[45] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 743–761, 2012.

[46] S. Tang, Y. Zhou, J. Li, C. Liu, and J. Shi, "Attention-guided sample-based feature enhancement network for crowded pedestrian detection using vision sensors," Sensors, vol. 24, no. 19, Art. no. 6350, 2024.

[47] Y. Xing, S. Yang, S. Wang, S. Zhang, G. Liang, X. Zhang, and Y. Zhang, "MS-DETR: Multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization," IEEE Transactions on Intelligent Transportation Systems, vol. 25, pp. 20628–20642, 2024, https://doi.org/10.1109/TITS.2024.3450584.

[48] J. Gao, Y. Wang, K.-H. Yap, K. Garg, and B. S. Han, "OccluTrack: Rethinking awareness of occlusion for enhancing multiple pedestrian tracking," IEEE Transactions on Intelligent Transportation Systems, vol. 26, no. 7, pp. 9852–9866, 2025, https://doi.org/10.1109/TITS.2025.3565334.

[49] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "The EuroCity Persons dataset: A novel benchmark for object detection," arXiv:1805.07193, 2018.

[50] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila, "A new benchmark for vision-based cyclist detection," in Proc. IEEE Intelligent Vehicles Symposium (IV), 2016.

[51] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3354–3361, https://doi.org/10.1109/CVPR.2012.6248074.

[52] M. Cordts et al., "The Cityscapes dataset for semantic urban scene understanding," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3213–3223.

[53] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2633–2642, https://doi.org/10.1109/CVPR42600.2020.00271.

[54] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11618–11628, https://doi.org/10.1109/CVPR42600.2020.01164.

[55] P. Sun et al., "Scalability in perception for autonomous driving: Waymo Open Dataset," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2443–2451, https://doi.org/10.1109/CVPR42600.2020.00252.

[56] X. Wang et al., "Repulsion loss: Detecting pedestrians in a crowd," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7774–7783, https://doi.org/10.1109/CVPR.2018.00811.

[57] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6452–6461, https://doi.org/10.1109/CVPR.2019.00662.

[58] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: One proposal, multiple predictions," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12211–12220, https://doi.org/10.1109/CVPR42600.2020.01223.

[59] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS: Improving object detection with one line of code," in Proc. IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5561–5569.

[60] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4507–4515.

[61] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3588–3597.

[62] Y. Zhang et al., "ByteTrack: Multi-object tracking by associating every detection box," in Proc. European Conference on Computer Vision (ECCV), 2022, pp. 1–21, https://doi.org/10.1007/978-3-031-20047-2_1.

[63] F. Zeng et al., "MOTR: End-to-end multiple-object tracking with transformer," in Proc. European Conference on Computer Vision (ECCV), 2022.

[64] S. Zhang et al., "WiderPerson: A diverse dataset for dense pedestrian detection in the wild," arXiv:1909.12118, 2019.

[65] M.-F. Chang et al., "Argoverse: 3D tracking and forecasting with rich maps," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8740–8749.

[66] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in Proc. IEEE International Conference on Computer Vision Workshops (ICCVW), 2017.

[67] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[68] Y. Tian, P. Luo, X. Wang, and X. Tang, "DeepParts: Occlusion handling in deep learning based pedestrian detection," in Proc. IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3258–3266.

[69] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in Proc. European Conference on Computer Vision (ECCV), 2018, pp. 138–154, https://doi.org/10.1007/978-3-030-01246-5_9.

[70] C. Chi et al., "Pedestrian detection via visible-to-full body regression," arXiv:2104.03106, 2021.

[71] X. Song, B. Chen, P. Li, B. Wang, and H. Zhang, "PRNet++: Learning towards generalized occluded pedestrian detection via progressive refinement network," Neurocomputing, vol. 482, pp. 98–115, 2022, https://doi.org/10.1016/j.neucom.2022.01.056.

[72] K. N. A. Shastry, J. Chaudhari, D. Thapar, A. Nigam, and C. Arora, "Parts-based attention for highly occluded pedestrian detection with transformers," in Proc. IEEE International Conference on Image Processing (ICIP), 2023.

[73] J. Kim et al., "DAGN: Deformable attention-guided network for occluded pedestrian detection," Applied Sciences, vol. 11, no. 13, 2021.

[74] G. Lin, Z. Bao, Z. Huang, Z. Li, W.-S. Zheng, and Y. Chen, "A multi-level relation-aware transformer model for occluded person re-identification," Neural Networks, vol. 177, Art. no. 106382, 2024.

[75] W. Liu et al., "High-level semantic feature detection: A new perspective for pedestrian detection," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[76] Q. Li, Y. Bi, R. Cai, and J. Li, "Occluded pedestrian detection through bi-center prediction in anchor-free network," Neurocomputing, vol. 507, pp. 199–207, 2022.
[77] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," in Proc. International Conference on Learning Representations (ICLR), 2021.
[78] J. Zhang, K. Xia, Z. Huang, S. Wang, and R. G. Akindele, "OBhunter: An ensemble spectral-angular based transformer network for occlusion detection," Expert Systems with Applications, vol. 248, Art. no. 123324, 2024, https://doi.org/10.1016/j.eswa.2024.123324.
[79] P. Cong et al., "STCrowd: A multimodal dataset for pedestrian perception in crowded scenes," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[80] P. Peng, T. Xu, B. Huang, and J. Li, "HAFNet: Hierarchical attentive fusion network for multispectral pedestrian detection," Remote Sensing, vol. 15, no. 8, Art. no. 2041, 2023, https://doi.org/10.3390/rs15082041.
[81] J. Cao, X. Weng, R. Khirodkar, J. Pang, and K. Kitani, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9686–9696.
[82] Y. Du et al., "StrongSORT: Make DeepSORT great again," IEEE Transactions on Multimedia, vol. 25, pp. 8725–8737, 2023.
[83] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in Proc. British Machine Vision Conference (BMVC), 2016.
[84] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5363–5371.
[85] L. Zhang, Z. Liu, X. Zhu, Z. Song, X. Yang, Z. Lei, and H. Qiao, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5127–5137.
[86] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in Proc. European Conference on Computer Vision (ECCV), 2020.
[87] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in Proc. IEEE International Conference on Image Processing (ICIP), 2016, pp. 3464–3468.
[88] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in Proc. IEEE International Conference on Image Processing (ICIP), 2017, pp. 3645–3649.
[89] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 941–951.
[90] X. Zhou, V. Koltun, and P. Krahenbuhl, "Tracking objects as points," in Proc. European Conference on Computer Vision (ECCV), 2020, pp. 474–490.
[91] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," International Journal of Computer Vision, vol. 129, pp. 3069–3087, 2021, https://doi.org/10.1007/s11263-021-01513-4.
[92] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "TrackFormer: Multi-object tracking with transformers," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8844–8854.
[93] F. Seidenschwarz, G. Braso, V. C. Frias, I. Possegger, and H. Bischof, "Simple cues lead to a strong multi-object tracker," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13813–13823.
[94] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv:1804.02767, 2018, https://doi.org/10.48550/arXiv.1804.02767.
[95] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," arXiv:2107.08430, 2021, https://doi.org/10.48550/arXiv.2107.08430.
[96] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475.