arXiv:2608.07579v1 [cs.CV] 4 Aug 2026

Geometry Beats Estimated Depth: RGB-Only Multi-Camera 3D Tracking under Sim2Real

Abdullah Naeem[1], Ayon Dey[1$], Anav Katwal[1$], Md Tamjidul Hoque[1*], and Noman Khan[2]

[1] LSU New Orleans, New Orleans, USA
{anaeem,adey,akatwal,thoque}@lsuneworleans.edu
[2] PinPark, Inc.
noman@pinpark.co

Abstract. The AI City Challenge 2026 Track 1 evaluates multi-camera 3D perception in large indoor warehouses under a synthetic-to-real (Sim2Real) setting; depth is available only for training and validation, so inference is RGB-only. We use two RGB-only routes as a controlled test of one hypothesis: that *cross-view geometric consistency*, not monocular depth accuracy, governs performance under Sim2Real. The first is a *geometry-first* pipeline: YOLO11x detection, homography lifting to the world frame, class-level 3D size priors, multi-camera fusion, world-coordinate tracking, and offline tracklet stitching. The second is *estimated-depth pseudo-LiDAR*: monocular depth (D4RT, Metric3D v2) back-projected into a fused point cloud and passed to a 3D detector (V-DETR), mirroring prior point-cloud winners that used *provided* depth. The gap is decisive: geometry-first reaches 13.0 3D HOTA (51.6 LocA), whereas pseudo-LiDAR collapses to 0.12 (9.2 LocA). We trace the collapse to cross-view inconsistency of monocular depth—scale correction is necessary but not sufficient—which domain-adaptation fine-tuning does not repair within budget. Within the geometry pipeline, offline stitching is the only intervention that helps; SAHI detection, appearance Re-ID, learned lifting, RT-DETR ensembling, test-time augmentation, and domain randomization all fail to beat the baseline detector. The bottlenecks are complementary: detection quality bounds the geometry route (DetA), localization consistency bounds pseudo-LiDAR (LocA). We release a complete, reproducible RGB-only pipeline and ablation.

Keywords: Multi-camera tracking · 3D perception · Sim2Real · AI City Challenge · Warehouse perception

1 Introduction

Multi-camera 3D perception in industrial indoor environments is a challenging setting for detection, localization, and identity association. AI City Challenge 2026 Track 1 requires participants to detect and track people and mobile

[$]Equal contribution. [*]Corresponding author.

objects, including forklifts, mobile robots, humanoids, transporters, and pallet trucks, across synchronized warehouse cameras. The output is a single text file containing world-coordinate 3D bounding boxes and object identities for each frame. The challenge is especially difficult because the training data is primarily synthetic while hidden test scenes include real-world videos and visual stressors. In addition, depth maps are available only for training and validation; inference must rely on RGB images.

Our goal was to build a complete end-to-end system quickly, evaluate the importance of each component, and identify the main bottleneck. We implemented a practical baseline around a strong 2D detector and explicit camera geometry. The system detects objects in each camera view, lifts each detection into a world coordinate system using calibration metadata and homographies, assigns class-prior 3D dimensions, fuses duplicate detections across cameras, and performs world-coordinate tracking. This pipeline produced a valid leaderboard submission and enabled systematic ablation.

The main conclusion from our experiments is that the current system is detection-limited. Our best submission achieved a stable localization score near 51 LocA and an association score higher than detection accuracy, but DetA remained low. Attempts to improve identity association or replace the detector did not improve final HOTA. This suggests that sparse 2D detection followed by heuristic lifting is not sufficient for high-ranking performance in this challenge. We therefore tested the natural alternative—estimated-depth pseudo-LiDAR, the route used by prior provided-depth winners—and found it performs far *worse*, isolating cross-view depth consistency rather than detection alone as the obstacle for learned-geometry substitutes.

Central hypothesis. We frame the two routes as a controlled test of a single question: for RGB-only multi-camera 3D tracking under Sim2Real conditions, is *cross-view geometric consistency*—agreement of the recovered geometry across cameras—more decisive than the *per-image accuracy* of monocular depth? The geometry-first lift maximizes cross-view consistency by construction (all cameras share one calibrated ground plane) while discarding fine per-pixel depth; estimated-depth pseudo-LiDAR maximizes per-image depth detail but recovers each camera's geometry independently, sacrificing cross-view consistency. Because both routes address the same task, data, and metric, comparing them isolates which factor governs performance. *Our hypothesis is that cross-view geometric consistency, not monocular depth accuracy, is the dominant factor*—and the experiments below support it: the consistency-preserving lift reaches 13.0 HOTA while the depth-accurate but cross-view-inconsistent pseudo-LiDAR route collapses to 0.12.

Our contributions are primarily *scientific findings*, supported by a reproducible system:

– **A systematic geometry-vs-depth comparison.** A controlled head-to-head of geometry-first lifting against estimated-depth pseudo-LiDAR for RGB-only Sim2Real 3D tracking, showing that explicit cross-view geometry

outperforms learned monocular depth by roughly two orders of magnitude (13.0 vs 0.12 HOTA), and identifying cross-view metric consistency—not per-image depth accuracy and not the 3D detector—as the governing factor.
- **A geometric-consistency diagnostic.** A simple, annotation-free *floor-coherence* metric that quantifies cross-view geometric consistency, separates scale error from consistency error, and predicts the pseudo-LiDAR collapse before any detector is run.
- **A design principle from ablation.** We group every intervention we tried into detector-side, geometry-side, and association-side categories and give the common reason each category fails to move HOTA: the geometry route is bounded by detection *quality*, so only association-side relinking (offline stitching) helps, while adding detections (detector-side) or replacing calibrated geometry with learned substitutes (geometry-side) hurts.
- **A reproducible baseline.** A complete, open RGB-only multi-camera 3D tracking pipeline for AI City Challenge 2026 Track 1 and its full ablation, as a reference point for Sim2Real 3D perception.

2 Related Work

Multi-camera 3D tracking. The AI City Challenge series has driven progress in multi-target multi-camera tracking and, more recently, scene-level 3D perception [1, 6]. Strong prior entries fuse per-camera observations into a common 3D frame and associate identities over time, typically via tracking-by-detection with motion and appearance cues [14, 19]; recent point-cloud winners voxelize fused depth into a scene representation for 3D detection and offline tracklet linking. Our setting differs in that depth is unavailable at inference, so geometry must be recovered from RGB and calibration alone.

Monocular depth and pseudo-LiDAR. Pseudo-LiDAR [13] back-projects estimated depth into a point cloud so that LiDAR-style 3D detectors can operate on images. Monocular depth estimation has advanced rapidly [8, 15]: recent models predict metric depth zero-shot (Metric3D [17]) or reconstruct dynamic scenes from video (D4RT [18]). Once back-projected, such geometry is consumed by point-cloud 3D detectors, whether voting-based [7] or transformer-based [5]. We test whether such *estimated* depth can replace the *provided* depth used by earlier winners, and find that cross-view inconsistency—not raw per-image depth accuracy—is the limiting factor (Section 7).

2D detection and association. We build on the YOLO family [9, 12], compare against transformer detectors (DETR [3], RT-DETR [20]), and evaluate sliced inference (SAHI) [2] for small-object recall. Association uses geometry-based world-coordinate tracking with an offline tracklet-stitching stage, and we evaluate with 3D HOTA [4], which factors detection, association, and localization accuracy.

3 Task and Evaluation

Each submission is a plain-text file with one row per predicted 3D detection:

$$(\text{scene}, \text{class}, \text{id}, \text{frame}, x, y, z, w, l, h, \theta). \tag{1}$$

The challenge evaluates seven classes: Person, Forklift, NovaCarter, Transporter, FourierGR1T2, AgilityDigit, and PalletTruck. The official metric is 3D HOTA [4], which balances detection accuracy, association accuracy, and localization accuracy. We report the leaderboard components HOTA, DetA, AssA, and LocA.

The challenge setting is Sim2Real. Synthetic training data is available with 2D/3D annotations, calibration, maps, and depth maps. The hidden test set contains additional real or stressed scenes. Since depth is not available at inference time, our system uses only RGB videos and calibration for test prediction.

4 Method

We compare two RGB-only routes (Fig. 1): a geometry-first pipeline (Sections 4.1–4.5) and a pseudo-LiDAR baseline (Section 4.6).

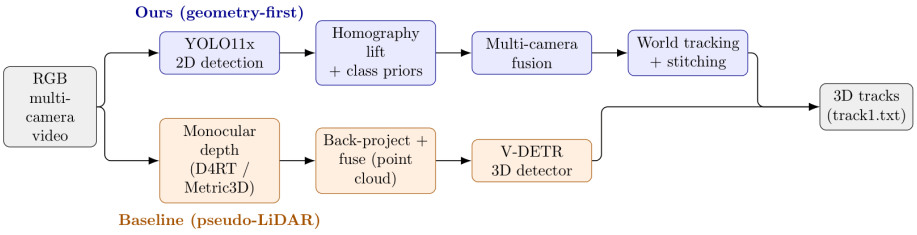

Fig. 1: The two RGB-only routes we compare. *Top (ours):* per-camera 2D detection, homography ground-plane lifting with class-size priors, same-frame multi-camera fusion, world-coordinate tracking, and offline tracklet stitching. *Bottom (baseline):* monocular depth, back-projection and multi-camera fusion into a scene point cloud, and a V-DETR 3D detector.

4.1 2D Detection

We trained an Ultralytics YOLO11x detector at 1280 input resolution. The model predicts per-camera 2D bounding boxes and class labels. Detections are cached as JSONL files per scene and camera. We use a low detector confidence threshold during inference to retain recall, then apply stronger filtering in the tracker and submission post-processing.

4.2 World-Frame Lifting

Each 2D detection is lifted into the world coordinate system using the camera calibration and a homography-based footpoint projection. For a bounding box, the bottom-center point is projected into the ground/world plane. We found the following projection setting to be best on validation diagnostics:

```
-invert-homography -footpoint-y 1.00 -world-plane xy
```

This configuration produced a median validation lift error of approximately 0.633 m in our alignment diagnostic. The projected point defines the object position in the world plane. The vertical coordinate and 3D box dimensions are assigned using class-level priors estimated from training and validation annotations. Yaw is estimated heuristically.

4.3 Multi-Camera Fusion

Since the same object can be visible in multiple synchronized camera views, detections are first fused per frame and class in world space. Detections within a class-specific spatial radius are merged into one scene-level observation. This reduces duplicate boxes before tracking and partially recovers detections from cameras with weaker views.

4.4 World-Coordinate Tracking

Tracking is performed separately for each class (Fig. 2). The tracker associates fused detections to active tracks using world-coordinate distance and simple motion prediction: each fused observation is matched to the nearest active track whose predicted position falls within a distance gate. We include optional coasting to bridge short gaps—when a track has no detection in a frame, its position is extrapolated from velocity for up to a fixed number of frames before termination. After tracking, short tracks and observations that match no track or violate class priors are filtered out, and object IDs are remapped to satisfy the submission constraints.

4.5 Offline Tracklet Stitching

The online tracker fragments identities whenever an object is missed for longer than the coasting window and later reappears, which caps association accuracy. We add an offline stitching stage that links same-class fragments: fragment B is joined to fragment A when B begins shortly after A ends (within a gap budget) and B's start position lies within a distance threshold of A's velocity-extrapolated endpoint. Links are matched one-to-one (cheapest residual first) and chained via union-find, and each merged chain receives a single scene-unique identity. Because stitching only relabels identities and never alters boxes, it improves association without changing detection or localization: in our experiments it raises AssA from 14.53 to 16.71 at essentially unchanged DetA and LocA. This stage makes the submission offline, since it uses future track evidence.

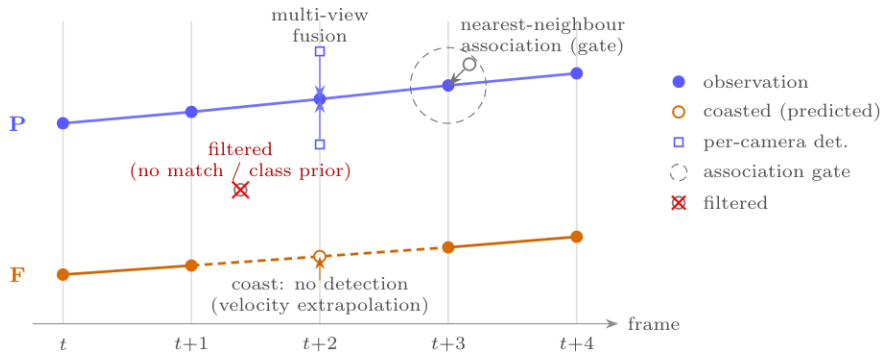

Fig. 2: World-frame association and track maintenance. Per-camera detections are lifted to the ground plane and fused across views (squares → dot); each fused observation is linked to the nearest active track whose predicted position lies within a distance gate (dashed circle). When a track has no detection in a frame it is *coasted* by extrapolating its velocity (open marker, dashed segment) for up to a fixed number of frames before termination; observations that match no track or violate class priors are filtered (×). Person (**P**) and Forklift (**F**) tracks are shown over five frames.

4.6 Pseudo-LiDAR Baseline

For comparison we implement the estimated-depth pseudo-LiDAR route. Per-camera depth is predicted from RGB with a monocular model—D4RT [18] or Metric3D v2 [17]—back-projected using calibration, and fused into a world-frame point cloud. A V-DETR 3D detector [10] trained on the provided-depth clouds then predicts 3D boxes; we additionally fine-tune it on estimated-depth clouds to reduce the train/test depth-domain gap. Section 7 reports the outcome.

5 Experiments

5.1 Detector Performance

Table 1 shows YOLO11x validation performance. The detector has high precision but limited recall, especially for PalletTruck. This recall weakness directly limits DetA and final HOTA.

On the training split, class-wise detector metrics were much higher, with recall above 0.87 for all evaluated classes and above 0.98 for PalletTruck. This indicates that the model can learn the object categories, but generalization to validation/test appearance is weak. We attribute this to domain shift, small or distant objects, real camera artifacts, illumination changes, and low object-background contrast.

Table 1: YOLO11x validation detector metrics.

Class	Precision	Recall	mAP50	mAP50-95
All	0.907	0.603	0.655	0.486
Person	0.817	0.723	0.753	0.577
Forklift	0.812	0.601	0.646	0.438
NovaCarter	0.989	0.806	0.840	0.709
Transporter	0.968	0.647	0.729	0.521
PalletTruck	0.950	0.239	0.309	0.185

5.2 Leaderboard Results and Ablations

Table 2 summarizes the main official leaderboard experiments. The YOLO11x geometry baseline with offline tracklet stitching is the best result: stitching raises AssA from 14.53 to 16.71 (and HOTA from 12.49 to 13.04) at unchanged DetA/LocA, since it only relinks fragmented identities. Re-ID, learned lifting, detector ensembling, test-time augmentation, and YOLO26/domain-randomized training all reduced final performance.

Table 2: Official leaderboard experiments.

Experiment	HOTA	DetA	AssA	LocA	Decision
YOLO11x + geometry baseline	12.4891	10.7900	14.5291	51.5784	Base
+ offline tracklet stitching	13.0413	10.7897	16.7105	51.5785	**Best**
Light Re-ID variant	10.5120	10.5182	13.0722	51.0577	Reject
Low-confidence recall bump	12.0249	10.2352	14.2158	51.5209	Reject
Learned lift MLP	0.9418	0.9679	0.8150	39.0014	Reject
YOLO + RT-DETR ensemble	10.8819	9.6225	12.5183	50.6654	Reject
YOLO11x TTA	11.1725	9.9193	12.2902	51.4127	Reject
SAHI sliced detection (+ stitching)	11.4943	7.2698	17.2804	49.8864	Reject
YOLO26 + stress augmentation	∼10	–	–	–	Reject

Rather than list the interventions individually, we group them by the pipeline stage they target—detector, geometry, or association. The outcomes cluster cleanly by category, and each category fails (or succeeds) for a single, generalizable reason.

Detector-side interventions (all reject). RT-DETR gave small validation recall gains for Forklift and PalletTruck but was worse overall; a YOLO+RT-DETR ensemble added noisy boxes and lowered HOTA; YOLO26 trained on a domain-randomized [11] stress-augmented set also underperformed the YOLO11x baseline; test-time augmentation and a low-confidence recall bump likewise did not help. Sliced inference (SAHI), which tiles each frame to recover small and distant

objects, more than doubled the detection count but *lowered* DetA (10.79→7.27) and HOTA (13.04→11.49). *Common reason:* under Sim2Real the deficit is detection *quality*, not quantity or detector family—adding boxes (SAHI, low threshold, ensembling) trades precision for recall and floods false positives, while swapping architectures does not transfer the missing real-domain generalization.

Geometry-side interventions (all reject). Replacing the calibrated homography lift with a learned bbox-to-3D MLP failed severely (HOTA 0.94, LocA 39.0), and the estimated-depth pseudo-LiDAR route of Section 7 collapsed entirely (HOTA 0.12). *Common reason:* both replace the one component that is already reliable—explicit, cross-view-consistent geometry from shared calibration—with a learned substitute that lacks metric grounding and cross-view agreement. A crop/bbox-only regressor carries too little information for metric 3D, and monocular depth is cross-view inconsistent (Section 7).

Association-side interventions (one helps). Appearance Re-ID [16] from object crops did not improve association and reduced HOTA, because real/stress scenes yield low-resolution, self-similar, domain-shifted crops that make appearance embeddings less reliable than geometry. The one intervention that helped is offline tracklet stitching (Section 4.5), which raises AssA 14.53→16.71 (HOTA 12.49→13.04) at unchanged DetA/LocA. *Common reason:* association is *not* the bottleneck, so adding an appearance signal only injects noise, whereas repairing fragmentation—relabeling identities without touching boxes—recovers the single kind of association error the online tracker actually makes.

Design principle. Across all three categories, one variable explains the outcomes: the geometry route is limited by detection quality, and its explicit geometry is already its most reliable part. The interventions that help are exactly those that improve association *without* disturbing detection or geometry; those that hurt either add low-quality detections or replace trustworthy geometry with a learned substitute. For RGB-only Sim2Real 3D perception this yields a concrete principle—*invest in detection quality and preserve explicit geometry; do not trade precision for recall, and do not replace calibrated geometry with learned depth or lifting*—which also predicts the pseudo-LiDAR result of Section 7.

6 Qualitative Analysis

We built video overlays that combine world-frame track boxes with the raw 2D detections (Fig. 4); they expose the 2D-to-3D lift as the weak link, and in domain-dependent ways. On the synthetic scene the world boxes *over-detect*, and their ground-plane bases are frequently mis-placed relative to the objects (lift error), producing false positives. On the real scene the failure inverts: the 2D detector still finds people, but those detections often fail to propagate into world tracks, so the world output *under-covers* real objects. In both cases the error lies in producing clean world-frame detections, not in association—where a correct world detection exists, the geometry tracker maintains a usable trajectory.

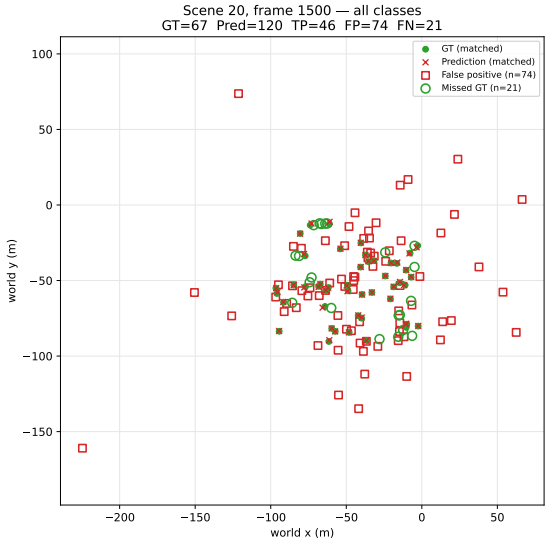

Fig. 3: World-coordinate predictions (red) vs. ground truth (green) for the final submission on a validation frame (Scene 20, frame 1500, all classes). Linked pairs are matches within 2 m; matched pairs align reasonably (moderate localization; LocA is the strongest of the three components), but predictions far outnumber ground truth (Pred = 120 vs GT = 67), with false positives (74) exceeding true positives (46)—the over-detection that bounds DetA. A few predictions are gross lift/coasting errors outside the scene (off-view). Best viewed in colour.

The score decomposition supports this qualitative finding. Our LocA is relatively stable near 51, and AssA is higher than DetA. Thus, the main limitation is not identity association alone but the availability and quality of scene-level detections. Figure 3 makes this concrete on a validation frame: matched predictions align reasonably with ground truth (LocA is the comparatively strongest component), yet predictions far outnumber ground truth (Pred = 120 vs GT = 67; 74 false positives against 46 true positives), the over-detection that bounds DetA.

Synthetic-to-real gap. The test set mixes synthetic scenes (Warehouses 23–25) and real-world scenes (Warehouses 26–27). Detection recall is markedly lower on the real scenes, consistent with the large train-vs-validation detector gap, so the real scenes contribute disproportionately to missed detections and to the DetA ceiling. This localizes the dominant error source to real-domain detection rather than to lifting or association.[3]

[3] All scores reported in this paper are those returned by the evaluation server's public leaderboard, which ranks submissions on approximately half of the test set; final scores on the full test set may differ. Per-scene HOTA and recall breakdowns are deferred to the camera-ready version.

Fig. 4: World-frame track boxes (projected back to the image) overlaid with the raw 2D detections, on a synthetic scene (top) and a real scene (bottom). *Top:* the projected boxes over-detect (false positives) and their ground-plane bases are mis-placed relative to the objects—homography-lift error. *Bottom:* the 2D detector correctly finds people, but these detections fail to propagate into world-frame tracks, so the world output *under-covers* real objects. The 2D→3D lift is the weak link in complementary ways—spurious, mis-placed boxes on synthetic and dropped detections on real—both bounding DetA. Best viewed in colour.

Where detection errors originate. Three axes localize the detection bottleneck using evidence already in the paper. *By class:* validation recall ranges from 0.81 (NovaCarter) down to 0.24 (PalletTruck) (Table 1); the low-profile, self-similar PalletTruck and the partially-occluded Forklift (0.60) dominate false negatives, whereas compact, high-contrast robots are detected reliably. *By error type:* on the validation frame of Fig. 3 false positives outnumber true positives (74 vs 46) while 46 ground-truth objects are simultaneously missed—the detector both over-fires (mis-lifted or duplicated boxes) and under-covers (small, distant, or occluded instances), both symptoms of detection *quality* rather than association. *By domain:* recall is markedly lower on real than on synthetic scenes, so real-domain instances contribute disproportionately to missed detections. The three views converge: DetA is bounded by detection quality for small, thin, and real-domain objects, not by lifting or association. A finer breakdown by absolute object distance and occlusion level would sharpen this further and is limited only by the availability of per-object distance/occlusion annotations.

Robustness and variability. Inference in our pipeline is deterministic, so run-to-run seed variance is not the relevant axis; the meaningful variability is *across classes and domains*, and it is already substantial. Per-class detector recall is 0.60 ± 0.20 across the five evaluated classes (Table 1)—from 0.81 (NovaCarter) down to 0.24 (PalletTruck)—so a single aggregate score hides which categories drive the DetA ceiling. Performance also splits sharply by domain, with markedly lower recall on real than on synthetic scenes. Per-scene 3D HOTA breakdowns across the validation warehouses are deferred to the camera-ready version.

7 Estimated-Depth Pseudo-LiDAR: A Negative Result

The natural RGB-only route to true 3D perception is *pseudo-LiDAR*: estimate per-pixel depth from RGB, back-project each camera into a common world frame, fuse into a scene point cloud, and run a 3D point-cloud detector—exactly the recipe used by prior point-cloud winners, but with *provided* depth. Because 2026 forbids depth at inference, we tested whether *estimated* depth can substitute. We report the outcome as a controlled negative result, since it is the most-assumed path to a stronger 3D representation.

Setup. We instantiate the pseudo-LiDAR baseline of Section 4.6 with D4RT and Metric3D v2 depth and the provided-depth-trained V-DETR detector, evaluated with and without domain-adaptation fine-tuning on estimated-depth clouds.

Findings. **(i) Scale.** Monocular depth is not metric out of the box: D4RT required a near-constant $\sim 4.3\times$ correction to reach metric scale, while Metric3D v2 is metric by construction but places clouds inconsistently—its reconstructed floor floats above the ground plane. **(ii) Cross-view inconsistency.** Even after scale correction, monocular depth disagrees across views, so multi-camera fusion yields warped, non-planar geometry (Fig. 5). As a proxy we measure *floor coherence*

on a validation scene: provided-depth clouds place 39% of points within ±0.3 m of the floor with 0% below it, whereas scale-corrected D4RT places only 20% in-band with 20% *below* the floor, and Metric3D leaves the floor essentially unreconstructed. **(iii) Result.** The provided-depth-trained V-DETR applied to estimated-depth clouds scored 0.12 HOTA with 9.2 LocA—roughly two orders of magnitude below the geometry lift—and domain-adaptation fine-tuning on estimated-depth clouds did not yield usable detections within our compute budget.

Why floor coherence? The warehouse floor is a large, planar surface at known elevation that is visible to most cameras, which makes it an ideal probe for *global* cross-view consistency: if per-camera depth agrees across views, every camera's floor points must land at the same world height, so their concentration in a thin band around the ground plane—and the absence of points below it—directly measures cross-view metric agreement, with no object annotation required. It is also a *necessary* condition for correct boxes, since every target rests on the floor, so a warped floor guarantees mis-placed objects. We set the band to ±0.3 m to match object scale: it is below the height of the shortest targets yet within the vertical tolerance at which a predicted box still overlaps ground truth, and the qualitative ordering (provided $\gg$ D4RT $\gg$ Metric3D) is insensitive to the exact value over 0.2–0.5 m, since Metric3D reconstructs essentially no floor and the provided cloud has essentially no sub-floor mass regardless of the cutoff. We prefer this proxy to a per-point error against ground-truth depth, which is unavailable at test time and, more importantly, measures *per-image* accuracy rather than cross-view consistency—monocular depth can be locally accurate yet globally warped, which is exactly the failure we observe. Plane-fit residuals or inter-view reprojection error are alternatives, but they require per-camera correspondence; floor coherence needs only the calibrated ground height and is therefore simpler and fully annotation-free.

Interpretation. The failure is one of *geometric consistency*, not detection: LocA collapses from 51.6 to 9.2 because fused monocular depth does not agree across cameras, so predicted boxes rarely overlap ground truth. Figure 5 makes the mechanism visual: the estimated-depth (D4RT) cloud is fragmented and non-planar (top) beside the flat, coherent ground-truth cloud (middle), while the floor-coherence bars (bottom) quantify the collapse—ground-truth points concentrate at the floor with none below it, whereas D4RT scatters an equal share above and below and Metric3D reconstructs no floor at all. Table 3 contrasts the two routes. For RGB-only multi-camera 3D tracking in this regime, an explicit geometric 2D→3D lift is far more reliable than learned monocular depth, which—unlike the provided depth used by earlier winners—lacks the cross-view metric consistency that point-cloud detectors require. This reframes the common recommendation to "move to 3D": the gap is not the 3D detector but the metric consistency of the geometry fed to it.

Table 3: Geometry-first lift vs. estimated-depth pseudo-LiDAR (test-set 3D HOTA components).

Approach	HOTA	DetA	AssA	LocA
Geometry-first lift (ours)	13.04	10.79	16.71	51.58
Estimated-depth pseudo-LiDAR	0.12	0.05	0.26	9.23

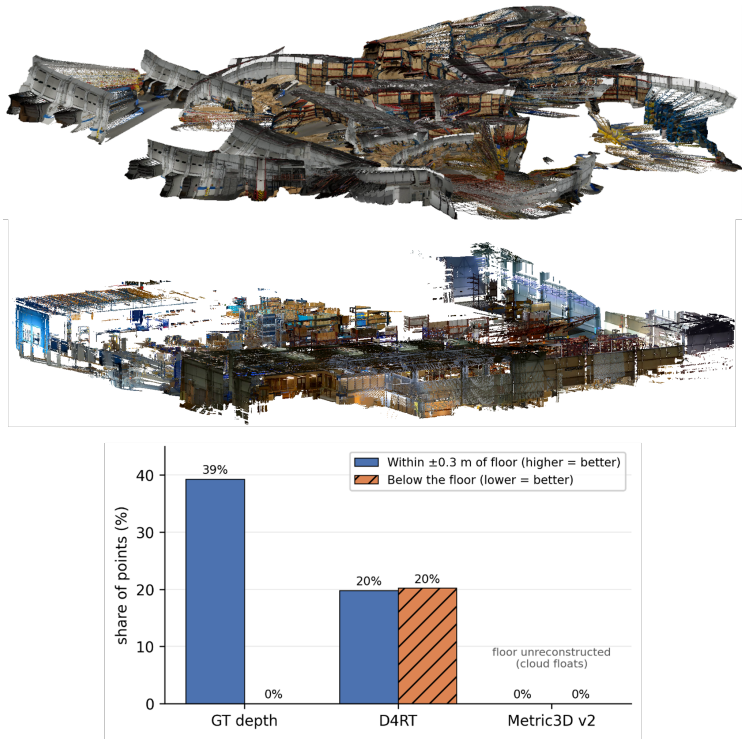

Fig. 5: Why estimated-depth pseudo-LiDAR fails. *Top:* a fused cloud from scale-corrected D4RT (estimated) depth—warped, fragmented, and non-planar. *Middle:* the provided-depth (ground-truth) cloud for the same scene, with a flat floor and coherent structure. *Bottom:* floor coherence—share of points within ±0.3 m of the ground plane, and share below it—for GT vs. D4RT vs. Metric3D v2, quantifying the cross-view inconsistency that collapses LocA.

8 Conclusion

We presented and compared two RGB-only routes to multi-camera 3D tracking for AI City Challenge 2026 Track 1. A geometry-first pipeline—YOLO11x detection, homography lifting, class priors, multi-camera fusion, world-coordinate tracking, and offline tracklet stitching—reaches 13.04 HOTA, while estimated-depth pseudo-LiDAR, the route suggested by prior provided-depth winners, collapses to 0.12 HOTA because monocular depth lacks cross-view metric consistency. Within the geometry pipeline, offline stitching is the only lever that helps, whereas sliced (SAHI) detection, Re-ID, learned MLP lifting, detector ensembling, TTA, and heavy domain randomization do not. The bottlenecks are complementary: detection quality under Sim2Real bounds the geometry pipeline, and localization consistency bounds pseudo-LiDAR. We conclude that, absent inference-time depth, explicit geometry is the more reliable foundation, and that closing the Sim2Real detection-quality gap—not learned monocular depth—is the highest-value next step.

References

1. AI City Challenge: Ai city challenge 2026 track 1: Multi-camera 3d perception. https://www.aicitychallenge.org/2026-track1/ (2026)
2. Akyon, F.C., Altinuc, S.O., Temizel, A.: Slicing aided hyper inference and fine-tuning for small object detection. In: IEEE International Conference on Image Processing (ICIP) (2022)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV) (2020)
4. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixe, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International Journal of Computer Vision **129**, 548–578 (2021)
5. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
6. Naphade, M., Wang, S., Anastasiu, D., Tang, Z., Chang, M.C., Yao, Y., Zheng, L., Shaiqur Rahman, M., Arya, M.S., Sharma, A., et al.: The 7th ai city challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2023)
7. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
8. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
9. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
10. Shen, Y., Geng, Z., Yuan, Y., Lin, Y., Liu, Z., Wang, C., Hu, H., Cao, N., Guo, B.: V-DETR: DETR with vertex relative position encoding for 3d object detection. In: International Conference on Learning Representations (ICLR) (2024)
11. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2017)
12. Ultralytics: Ultralytics yolo documentation. https://docs.ultralytics.com/ (2026)
13. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
14. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: IEEE International Conference on Image Processing (ICIP) (2017)
15. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
16. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

17. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
18. Zhang, C., Le Moing, G., Koppula, S., Rocco, I., Momeni, L., Xie, J., Sun, S., Sukthankar, R., Barral, J.K., Hadsell, R., Ghahramani, Z., Zisserman, A., Zhang, J., Sajjadi, M.S.M.: Efficiently reconstructing dynamic scenes one d4rt at a time. arXiv preprint arXiv:2512.08924 (2025)
19. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-object tracking by associating every detection box. In: European Conference on Computer Vision (ECCV) (2022)
20. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)