Geometry-Aware Risk Estimation for Vehicle Interactions in Monocular Traffic Video

Amaan Khan Anirudha Kyathsandra Dennis Binford Mingyao Wang

CS543: Computer Vision, University of Illinois Urbana–Champaign

Abstract

We present a computer vision pipeline that estimates the risk of vehicle interactions in monocular traffic video. Vehicles are detected with a pretrained YOLOv8 model, associated across frames with ByteTrack, and then scored using simple pairwise interaction features (distance, closing speed, and time-to-collision). To address the perspective distortion inherent in single-camera traffic footage, we compare an image-plane baseline that operates directly on pixel coordinates against a geometry-aware variant that first projects vehicle positions into a metric bird's-eye-view (BEV) frame via a normalized 8-DOF Direct Linear Transform. We evaluate on the MVI_40855 sequence from UA-DETRAC (1,090 frames, dense urban intersection traffic). The image-plane baseline flags 5,474 risky pair-frames across 830 unique vehicle pairs; the geometry-aware variant flags only 610 pair-frames across 108 unique pairs, and the Jaccard overlap between the two pair sets is only 0.114. A scatter plot of every flagged pair-frame in (image, BEV) distance space, plus a sensitivity sweep over the distance thresholds, shows that this gap is dominated by perspective compression near the camera horizon—exactly the failure mode that motivates the BEV projection. A track-age ablation that requires both tracks to be at least six frames old shifts the headline numbers proportionally without changing the qualitative picture.

1 Introduction

Estimating which vehicle interactions in a traffic scene are risky is a long-standing problem in transportation safety analytics and autonomous driving. A natural first attempt is to read off geometric quantities— inter-vehicle distance, relative velocity, and time-to-collision (TTC)— directly from the image plane of a fixed traffic camera. However, perspective distortion makes these measurements unreliable: two cars that appear close in the image may be far apart on the road, and a vehicle near the horizon traverses many fewer pixels per second than a vehicle near the camera even at identical real-world speeds.

In this project we build a four-stage pipeline that turns raw traffic video into a stream of pairwise risk events, and use it to compare two risk-scoring strategies that differ only in the geometric frame they operate in.

Contributions.

- A reproducible, modular pipeline (detection → tracking → calibration → risk) implemented in Python on top of YOLOv8, ByteTrack, and OpenCV, orchestrated by a single shell script.
- An image-plane baseline and a geometry-aware formulation of pairwise risk that share thresholds and feature definitions but differ in the coordinate frame they evaluate them in.

- A side-by-side evaluation on a UA-DETRAC subset showing that the two methods disagree on the majority of flagged events (Jaccard 0.11 between flagged-pair sets), with most of the disagreement explained by perspective distortion, and a track-age ablation that confirms the result is not an artifact of short track fragments.

Related work. Multi-object detection and tracking in traffic video are well-studied problems. YOLO [4] and its successors are widely used pretrained vehicle detectors, and ByteTrack [2] is a strong detection-based tracker for crowded scenes. UA-DETRAC [1] provides fixed-camera traffic video with bounding-box and identity annotations and has become a standard benchmark for vehicle tracking. Camera-to-ground homography estimation from manually annotated lane features is a classical approach to monocular calibration; the BrnoCompSpeed benchmark [3] formalizes this for traffic cameras and provides a useful reference for evaluating calibration quality. Surrogate safety measures including TTC and post-encroachment time are widely used in transportation engineering as collision-free proxies for crash risk.

2 Approach

Our system is organized as four loosely coupled modules that communicate through CSV / MOT-format files in a shared `outputs/` directory. The data flow is

$$\underbrace{\text{frames}}_{\text{UA-DETRAC}} \xrightarrow{\texttt{detect.py}} \text{YOLOv8} \xrightarrow{\texttt{tracking.py}} \text{ByteTrack} \xrightarrow{\texttt{export\_homography.py}} H \xrightarrow{\texttt{risk\_score.py}} \text{pairwise risk.}$$

2.1 Vehicle Detection

We use the pretrained YOLOv8n model from the Ultralytics package [4], restricted to the COCO classes car (2), bus (5), and truck (7). For each frame the detector outputs a list of axis-aligned bounding boxes (x_1, y_1, x_2, y_2) with confidence scores $c \in [0, 1]$. These are written to `outputs/detections.csv` and converted to MOT format $(\text{frame}, -1, x, y, w, h, c, -1, -1, -1)$ for the tracker.

2.2 Multi-Object Tracking

We associate detections across frames using ByteTrack [2], which performs two-stage matching that retains low-confidence boxes for recovery. The tracker is configured with activation threshold 0.2, a lost-track buffer of 60 frames, and matching threshold 0.8 at 30 Hz.

Tracks are exported to two derived artifacts:

- `outputs/tracks_mot.txt`: MOT-format $(\text{frame}, \text{id}, x, y, w, h, c, -1, -1, -1)$ for every frame in which the track is alive.
- `outputs/trajectory_features.csv`: a per-frame, per-track table adding the box center, the bottom-center reference point, a finite-difference pixel velocity (v_x, v_y), the speed magnitude $\sqrt{v_x^2 + v_y^2}$, and track-age and gap indicators.

The bottom-center of the box is used as the reference point for projection because, for upright objects on a flat ground plane, that point lies at (or very near) the road surface and is the most stable candidate for an image-to-ground homography.

2.3 Geometric Calibration and BEV Projection

The geometry-aware branch requires a 3×3 homography H mapping points on the image plane to a metric ground plane. For MVI_40855 we pick four points on the visible road surface near the intersection foreground and map them to a 20 m $\times$ 30 m metric rectangle in BEV (Table 1). We then solve for H with a normalized Direct Linear Transform (DLT): points are pre-normalized so that each set has its centroid at the origin and mean distance to the origin equal to $\sqrt{2}$ (Hartley normalization), each correspondence contributes two rows to a $2N \times 9$ system, the right singular vector of the smallest singular value of its SVD gives the homography up to scale, and the solution is denormalized at the end. The four-point exact case in our calibration set admits a zero-error solution; the implementation nevertheless generalizes to overdetermined sets of correspondences.

Table 1: Manual correspondences used for MVI_40855. BEV x runs across the road, BEV y runs away from the camera. These correspondences are baked into `calibration/export_homography.py` so the calibration step is reproducible without an interactive session.

u (px)	v (px)	BEV x (m)	BEV y (m)
350	250	0	30
700	250	20	30
900	500	20	0
60	500	0	0

The recovered homography is

$$H \approx \begin{bmatrix} -0.143 & -0.166 & 91.43 \\ 0.000 & 0.300 & -150.00 \\ 0.000 & -0.014 & 1.000 \end{bmatrix}, \tag{1}$$

with mean reprojection error 0 m on the calibration points (the minimal 4-point system admits an exact solution).

A point $\mathbf{p} = (x, y)^\top$ on the image plane is projected as

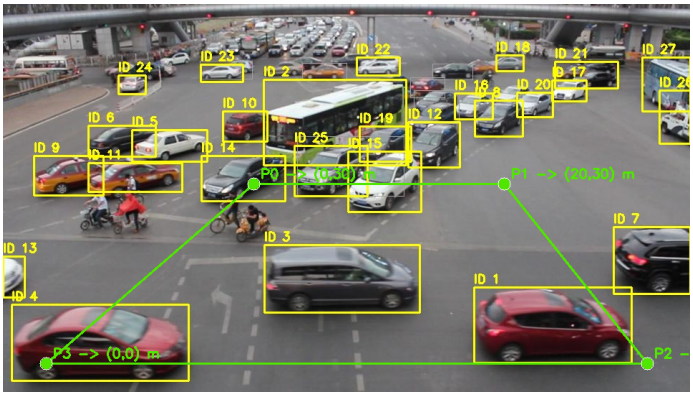
$$\tilde{\mathbf{p}}' = H\,\tilde{\mathbf{p}}, \qquad \mathbf{p}' = (\tilde{p}'_1/\tilde{p}'_3,\ \tilde{p}'_2/\tilde{p}'_3)^\top, \tag{2}$$

where $\tilde{\mathbf{p}} = (x, y, 1)^\top$ is the homogeneous form of $\mathbf{p}$. We convert a pixel-space velocity (v_x, v_y) at $\mathbf{p}$ into a BEV velocity by a small-step linearization,

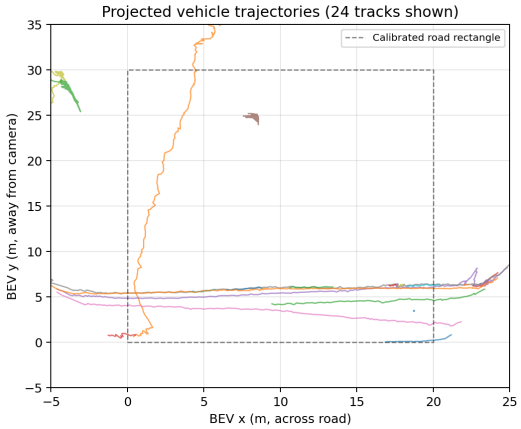
$$\mathbf{v}^{\mathrm{BEV}}(\mathbf{p}, \mathbf{v}) \ \approx\ \frac{1}{\Delta t}\big(\pi(\mathbf{p} + \mathbf{v}\Delta t) - \pi(\mathbf{p})\big), \tag{3}$$

where $\pi(\cdot)$ is the homographic projection and Δt is one frame period. This avoids estimating a full BEV trajectory history.

Figure 1 shows the four chosen image points overlaid on a frame of the sequence (left) and the resulting projected trajectories inside the calibration rectangle (right). The long-lived tracks inside the calibrated BEV window form approximately parallel streams, validating the homography in the foreground region.



(a) Image-plane correspondences P0–P3 on the road plane (annotated on a tracked frame).



(b) Tracked trajectories projected into BEV (metres). Parallel lanes appear approximately parallel after projection.

Figure 1: Calibration sanity check on MVI_40855.

2.4 Risk Estimation

Given the trajectory features—and, for the BEV variant, their homographic projections—we compute three pairwise quantities for every unordered pair (a, b) of vehicles co-present in a frame. With $d = \|\mathbf{p}_b - \mathbf{p}_a\|$ the inter-vehicle distance, $\hat{\mathbf{u}}_{ab} = (\mathbf{p}_b - \mathbf{p}_a)/d$ the unit vector from a to b, and $\mathbf{v}_{ab} = \mathbf{v}_b - \mathbf{v}_a$ the relative velocity, the closing speed and TTC are

$$s_{\text{close}} = -\mathbf{v}_{ab} \cdot \hat{\mathbf{u}}_{ab}, \qquad \tau = \begin{cases} d/s_{\text{close}} & s_{\text{close}} > 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{4}$$

A pair is flagged risky when both

$$d < d_{\text{thr}} \quad \text{and} \quad 0 < \tau < \tau_{\text{thr}}. \tag{5}$$

The image-plane baseline applies (5) with pixel-space thresholds $(d_{\text{thr}}, \tau_{\text{thr}}) = (120\,\text{px}, 1.5\,\text{s})$; the BEV variant applies the same predicate after projecting $\mathbf{p}_a, \mathbf{p}_b, \mathbf{v}_a, \mathbf{v}_b$ through H, with metric thresholds $(d_{\text{thr}}, \tau_{\text{thr}}) = (5\,\text{m}, 1.5\,\text{s})$. Algorithm 1 summarizes the per-frame scoring procedure.

3 Results

3.1 Data and protocol

We evaluate on the MVI_40855 sequence from UA-DETRAC [1] (1,090 frames at 30 Hz, ≈ 36 s of footage at 960×540 resolution). The clip is a busy urban intersection with high vehicle density. We use the pretrained YOLOv8n weights without any task-specific fine-tuning; ByteTrack runs with the parameters listed in Section 2.2; the image-plane risk thresholds are (120 px, 1.5 s) and the BEV thresholds are (5 m, 1.5 s).

3.2 Detection and tracking

The detector produces 31,158 vehicle boxes across the 1,090 frames (mean 28.6, range 17–39 per frame). Confidence is distributed around a median of 0.54, with 12.3% of boxes below 0.3; these low-

Algorithm 1 Per-frame pairwise risk scoring

Require: Frame f, tracks $T_f = \{(\mathbf{p}_i, \mathbf{v}_i)\}$, optional H, thresholds $(d^{\text{img}}_{\text{thr}}, \tau^{\text{img}}_{\text{thr}}, d^{\text{bev}}_{\text{thr}}, \tau^{\text{bev}}_{\text{thr}})$, minimum track age $a_{\min}$

$E \leftarrow \emptyset$
if H is given **then**
  Project each $(\mathbf{p}_i, \mathbf{v}_i)$ through H to obtain $(\mathbf{p}_i^{\text{bev}}, \mathbf{v}_i^{\text{bev}})$
end if
for each pair (i, j) with $i < j$ in T_f **do**
  if age$(i) < a_{\min}$ **or** age$(j) < a_{\min}$ **then**
    continue
  end if
  $d, \tau \leftarrow \textsc{PairFeatures}(\mathbf{p}_i, \mathbf{v}_i, \mathbf{p}_j, \mathbf{v}_j)$
  $r_{\text{img}} \leftarrow \mathbb{1}\left[d < d^{\text{img}}_{\text{thr}} \wedge 0 < \tau < \tau^{\text{img}}_{\text{thr}}\right]$
  if H is given **then**
    $d', \tau' \leftarrow \textsc{PairFeatures}(\mathbf{p}_i^{\text{bev}}, \mathbf{v}_i^{\text{bev}}, \mathbf{p}_j^{\text{bev}}, \mathbf{v}_j^{\text{bev}})$
    $r_{\text{bev}} \leftarrow \mathbb{1}\left[d' < d^{\text{bev}}_{\text{thr}} \wedge 0 < \tau' < \tau^{\text{bev}}_{\text{thr}}\right]$
  end if
  if $r_{\text{img}} = 1 \vee r_{\text{bev}} = 1$ **then**
    append $(f, i, j, d, \tau, r_{\text{img}}, d', \tau', r_{\text{bev}})$ to E
  end if
end for
return E

confidence boxes are still passed to the tracker, which filters them through its activation threshold.

ByteTrack produces 317 distinct track identities covering 28,376 track-frames (Table 2). The track length distribution is heavy-tailed: median 19 frames, but the longest track lasts 1,066 frames, reflecting vehicles that traverse the full scene. Roughly 27.8% of tracks are shorter than five frames and most likely correspond to fragments from missed detections or vehicles that briefly enter the field of view; the `--min_track_age` ablation in Section 3.4 quantifies the impact.

Table 2: Tracking summary on MVI_40855.

Metric	Value
Total tracks	317
Track-frames exported	28,376
Average track length (frames)	89.5
Median track length (frames)	19
Maximum track length (frames)	1,066
Short-track ratio (≤ 5 frames)	27.8%

3.3 Risk: image-plane vs. geometry-aware

Table 3 compares the two formulations on the same set of vehicle pairs. The image-plane baseline flags 5,474 pair-frames across 830 unique pairs (and at least one risky pair on 1,072/1,090 $\approx$ 98.3% of frames). The BEV variant flags 610 pair-frames across 108 unique pairs on 430 frames—almost

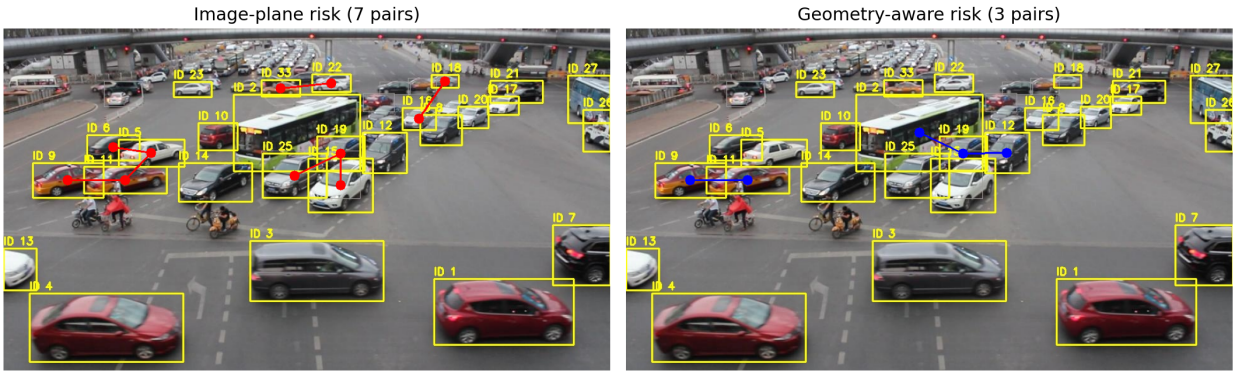


Figure 2: Risk overlays on frame 4 of MVI_40855, using identical feature definitions and TTC thresholds. The image-plane baseline (left, 7 pairs) over-flags pairs across the scene due to perspective compression; the BEV variant (right, 3 pairs) suppresses those false positives.

an order of magnitude fewer events. Only 385 pair-frames are flagged by both methods; the unique-pair Jaccard between the two flagged sets is 0.114, meaning the methods agree on roughly one in nine pairs they flag.

Table 3: Image-plane vs. geometry-aware risk on MVI_40855.

	Image plane	Geometry-aware	Both / overlap
Risky pair-rows	5,474	610	385
Unique pairs flagged	830	108	—
Frames with any risk	1,072	430	—
Row agreement among flagged		$385/5{,}699 = 0.068$	
Unique-pair Jaccard		0.114	

Figure 2 shows risk overlays on the same frame under each method. Coloured circles mark vehicles that participate in at least one flagged pair on this frame, and coloured lines connect risky pair members. In the image-plane view the baseline flags seven pairs spread across the entire visible scene, including vehicles separated by an entire lane that simply happen to overlap in projection. The BEV view retains only three flagged pairs, all concentrated in the actually-congested central queue around the bus, where vehicles are physically close in metres.

Figure 3 shows that this order-of-magnitude gap is consistent across the entire 1,090-frame sequence: image-plane flags peak at roughly 15 pairs per frame and average ≈ 5, while BEV flags rarely exceed 5 per frame and are zero on most frames.

Figure 4 plots every flagged pair-frame as a point in the plane of image-plane distance (px) against BEV distance (m). The four quadrants formed by the two dashed threshold lines tell the story directly: pair-frames in the top-left are flagged by image only (small pixel distance, large metric distance—*the perspective-compression false positives*), and pair-frames in the bottom-right

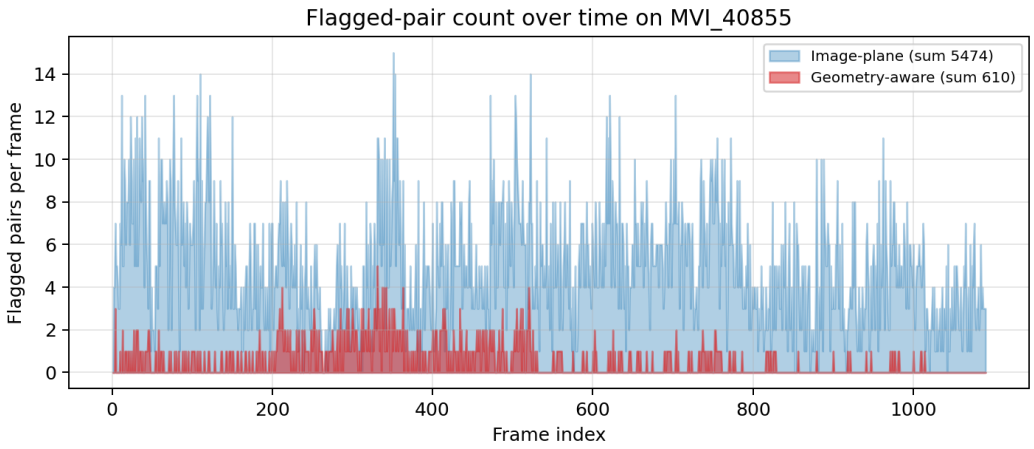


Figure 3: Number of flagged pairs per frame across the whole sequence under each method. The image-plane baseline (blue) fires on 98.3% of frames; the BEV variant (red) on 39.4%.

are flagged by BEV only (moderate pixel distance but small metric distance—*the true positives that the baseline misses*).

The TTC distribution of flagged pairs (Figure 5) confirms this. The image-plane baseline's TTC histogram is approximately uniform on $[0, 1.5]$ s, with the bulk above 0.5 s—these are slowly closing, far-apart-in-3D pairs that only look risky due to perspective. The BEV variant peaks sharply below 0.2 s, consistent with the small subset of pairs that are both physically close and actually closing fast.

3.4 Track-age ablation

To check that the disagreement between the two methods is not driven by noisy short tracks, we rerun the entire risk scoring with `--min_track_age 6`, which excludes any pair where either track has been alive for fewer than six frames. The result is consistent: the image-plane count drops from 5,474 to 4,859 (−11%), the BEV count from 610 to 470 (−23%), and the pair Jaccard moves from 0.114 to 0.092. The relative order of magnitude between the two methods is preserved, indicating that the perspective-distortion gap is a genuine property of the feature definitions rather than a tracking artefact.

Table 4: Track-age ablation. Filtering pairs in which either track has age < 6 frames shrinks both event sets but preserves the qualitative picture.

	Default	`--min_track_age 6`	Δ
Image-plane rows	5,474	4,859	−11.2%
BEV rows	610	470	−23.0%
Both rows	385	293	−23.9%
Image-plane pairs	830	648	−21.9%
BEV pairs	108	64	−40.7%
Pair Jaccard	0.114	0.092	—

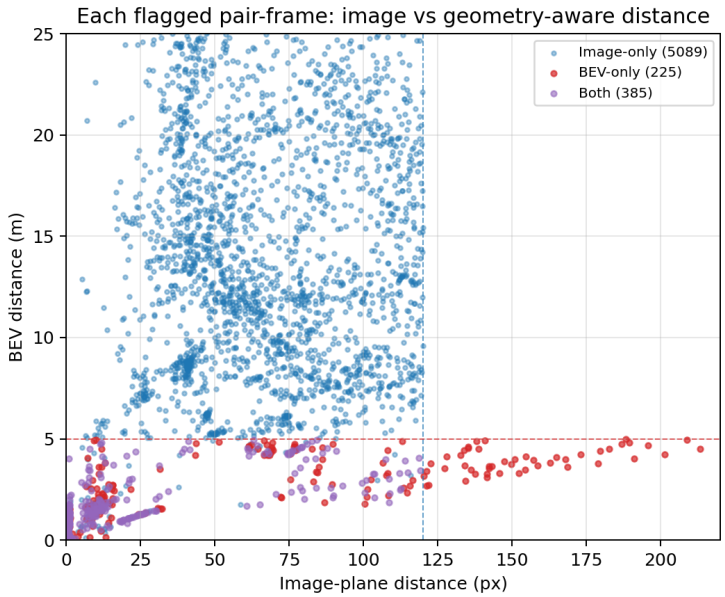


Figure 4: Each flagged pair-frame plotted in (image-plane distance, BEV distance) space. Dashed lines mark the chosen thresholds (120 px, 5 m). Image-only flags cluster well below the BEV threshold but with large metric distance; BEV-only flags sit above the image threshold but with small metric distance.

3.5 Threshold sensitivity

Figure 6 sweeps the distance threshold of each method while holding the other variables (TTC threshold, the other method) fixed. The image-plane count grows nearly linearly with the pixel threshold, reflecting how dense the scene is in pixel space. The BEV count grows slowly until about 7–8 m and then steeply, consistent with vehicles in adjacent lanes being separated by roughly that many metres. The default thresholds (120 px, 5 m) sit at corresponding points on each curve and broadly reflect a comparable operating point in each frame.

4 Discussion

The image-plane baseline over-fires as predicted. The image-plane baseline flags risk in 98.3% of frames. Two failure modes explain the over-firing. First, two vehicles that are far apart in 3D but near the camera horizon project to image-plane points that are only tens of pixels apart; the 120-pixel threshold therefore catches many pairs that are nowhere near each other on the road. Second, pixel-velocity TTC compresses near the camera and explodes near the horizon, so cars moving in parallel at constant real-world spacing show closing speeds that fluctuate sharply with image position. Both failure modes are direct consequences of perspective, and both are visible as the top-left cloud in Figure 4.

The BEV variant isolates a much smaller, plausible event set. The geometry-aware variant flags only 108 unique pairs, a factor of roughly 8× fewer than the baseline, and only 430 frames have at least one BEV-risky pair. Inspection of the BEV-only events (those flagged by the BEV variant but not the image-plane baseline) shows that they correspond to vehicles whose pixel

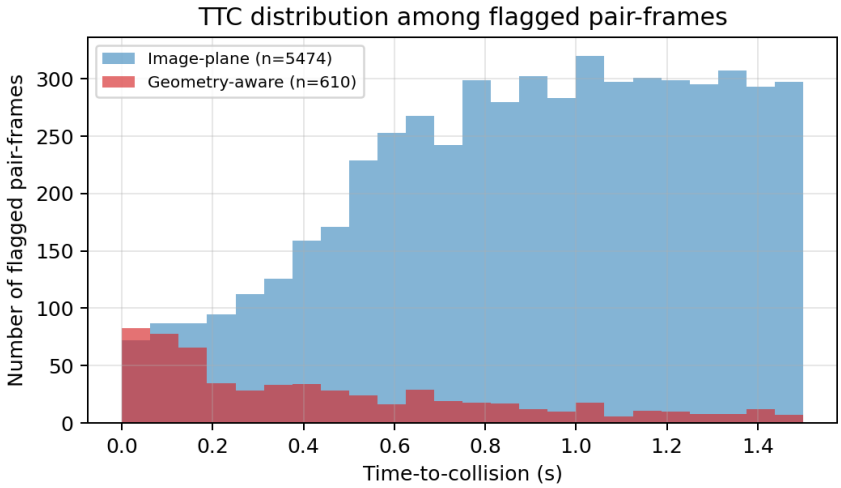


Figure 5: TTC distribution of flagged pair-frames under each method. The image-plane baseline (blue) flags many pairs with relatively large TTC; the BEV variant (red) concentrates near the truly urgent low-TTC regime.

separation is moderate (above the 120-pixel threshold) but whose metric distance is small because they are at the same depth from the camera—a class of true positives that the baseline misses. Conversely, the 5,089 image-only flagged pair-frames correspond mostly to vehicles separated by a lane or more in the world but only a few tens of pixels apart on the screen.

Track fragmentation is a secondary noise source. Roughly 28% of our tracks are ≤ 5 frames long. These fragments contribute pair events that survive a few frames before the track is lost, inflating the unique-pair count even after BEV projection. We expose a `--min_track_age` flag in the risk module that filters pairs in which either track has age below a threshold. With the threshold set to six frames the BEV unique-pair count drops by 41% while the qualitative gap between the two methods is preserved, indicating that track fragmentation accounts for some of the absolute count but not for the underlying disagreement.

Limitations. Our risk model is intentionally simple. It uses centroid-to-centroid distance rather than box-aware distance, treats vehicles as point masses with first-order kinematics, ignores headings derived from box orientation, and uses a single fixed homography assuming a flat ground plane. Our four-point calibration is exact rather than overdetermined, so it does not reflect any noise in the chosen image points; a more robust setup would pick six or more correspondences and use least-squares DLT. The thresholds are also a single global choice; in practice one would expect them to depend on vehicle size and approach geometry. These approximations are tolerable for the comparative analysis we are after but would each be worth revisiting in a more operational deployment.

5 Conclusion

We built a clean, four-stage monocular traffic risk pipeline and used it to quantify how much perspective distortion matters for downstream risk scoring. On dense urban footage the answer

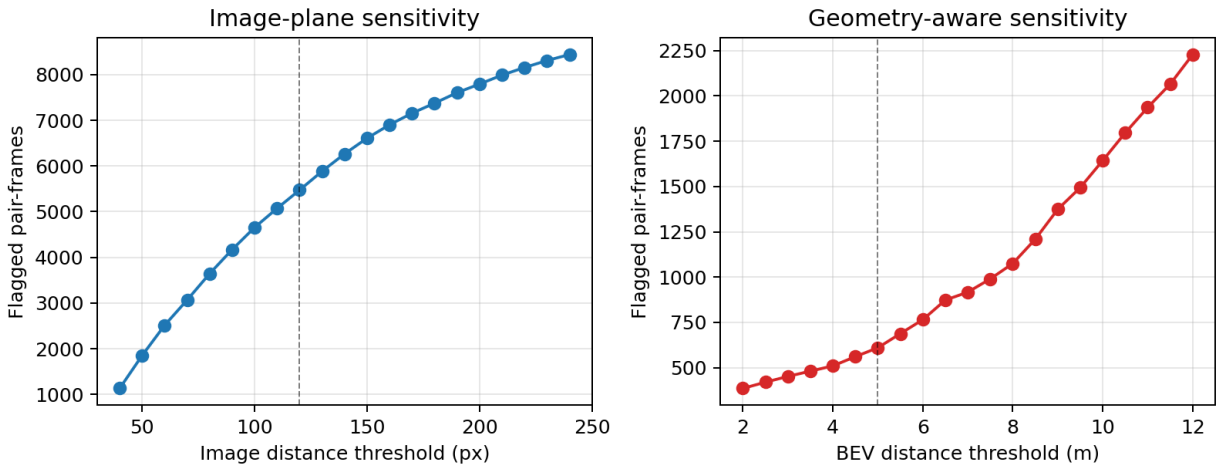


Figure 6: Flagged pair-frames as a function of the distance threshold of each method, computed over the full population of co-present pair-frames (not just the union of already-flagged pairs). Dashed lines mark our default thresholds.

is: a lot. The image-plane baseline and the geometry-aware variant share thresholds and feature definitions and yet agree on only $\approx 11\%$ of their flagged pairs (Jaccard), with BEV producing an event stream that is sparser, focused on genuinely close vehicles, and easier to interpret. A track-age ablation and a threshold sensitivity sweep both confirm that the gap is a property of the geometric reasoning, not of the particular cleanup choices or the particular threshold values we picked.

6 Statement of Individual Contributions

- **Amaan Khan**—System integration; pairwise trajectory-feature extraction; image-plane and geometry-aware risk formulation; ablation and sensitivity tooling; risk-event visualization; report writing.
- **Anirudha Kyathsandra**—Vehicle detection module with pretrained YOLOv8; CSV / MOT exporters; detection visualization.
- **Mingyao Wang**—Multi-object tracking with ByteTrack; trajectory and projection-point export; tracking visualization and summary statistics.
- **Dennis Binford**—Geometric calibration: manual correspondences on MVI_40855, normalized DLT homography estimation, and BEV trajectory visualization.

References

[1] L. Wen, D. Du, Z. Lei, S. Lyu, S. Z. Li, and Y. Wang, "UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking," 2015. https://detrac-db.rit.albany.edu/

[2] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," 2021.

[3] J. Sochor, J. Špaňhel, and A. Herout, "BrnoCompSpeed: Review of Traffic Camera Calibration and Comprehensive Dataset for Monocular Speed Measurement," 2017. https://github.com/JakubSochor/BrnoCompSpeed

[4] Ultralytics YOLO Documentation. https://docs.ultralytics.com/

[5] OpenCV Documentation. https://docs.opencv.org/

[6] Project source repository. https://github.com/AniRaoHK/cs543-traffic-risk