PROJECTIVE GEOMETRY FOR GENERALIZABLE MULTI-VIEW PEDESTRIAN DETECTION

Anonymous authors
Paper under double-blind review

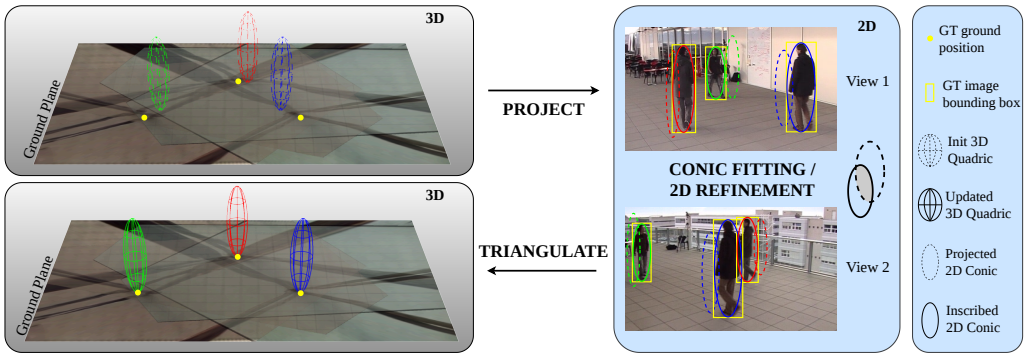


Figure 1: Overview of PGMVG: initial 3D quadrics are projected into each camera view as 2D conics, which are then refined by aligning conics with cross-view detections. The refined conics are triangulated back into 3D space and form an updated quadrics at correct ground positions.

ABSTRACT

Multi-view pedestrian detection (MVPD) seeks to localize pedestrians on the ground plane by gathering observations from multiple cameras. Recent MVPD methods achieve strong in-domain performance by projecting per-view features onto a shared Bird's-Eye-View (BEV) plane, where localization is performed. Since BEV construction couples learned representation with camera geometry, their spatial reasoning becomes strongly biased toward scene-specific occupancy patterns, limiting generalization to unseen environments. Rather than expanding the training distribution, we address generalization by integrating projective geometry into a scene-agnostic framework. Specifically, each pedestrian is represented as a ground-anchored 3D quadric and its 2D representation in image view is modeled as a projected 2D conic, connected by an exact, invertible ($2D \leftrightarrow 3D$) mapping under the pinhole model. As 3D localization is fully lifted from multi-view 2D predictions through camera parameters, the framework learns no scene-specific priors and naturally inherits the cross-scene transferability of pretrained image detectors. Based on this formulation, we propose **PGMVG**, a query-based framework that iteratively refines per-view detections and yields pedestrian locations through constrained projective triangulation. Extensive experiments across multiple benchmarks show that PGMVG improves MODA by 5.1 on cross-scene and 8.3 on sim-to-real settings, while maintaining strong in-domain performance.

1 INTRODUCTION

Pedestrian detection in crowded environments is essential for a wide range of practical applications, from surveillance Zhang et al. (2021b) to crowd management Mo et al. (2024). To tackle occlusion in the single-view setting, multi-view pedestrian detection (MVPD) aims to localize all individuals in a scene from synchronized cameras with largely overlapping fields of view. Recent methods perform remarkably well on standard benchmarks, yet their success relies on homogeneous train–test conditions that share the same scene, camera configurations, and crowd density (Hou & Zheng, 2021; Engilberge et al., 2023). Deployment does not guarantee such condition and installation, exposing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison between the properties of PGMVG and existing MVPD paradigms.

Property	**Top-down** ($2D \rightarrow ground$)	**Bottom-up** ($2D \rightarrow BEV$)	**PGMVG** ($2D \leftrightarrow 3D$)
Fusion and Localization	Independent per-view detection; late-fused on ground plane	Early cross-view BEV fusion; occupancy prediction	Per-query inter-view attention; iterative quadric refinement
Scene Representation	No centralized scene model	Dense BEV feature map	Sparse query set; no scene model
Predicted Ground Point	Computed through 2D detections	Directly regressed from BEV map	Triangulated through 2D detections /conics
Projection Model	Approximate bounding box back-projection	Homography warping introduces distortion	Exact quadric–conic (projective geometry)
Trade-off	Transfers across scenes; accuracy bounded by single-view quality	Strong in-domain; fails on scene or configuration change	Generalizes without sacrificing accuracy

the generalization challenge for existing MVPD methods. In this paper, we focus on developing an MVPD framework that is intrinsically transferable across scenes and camera setups.

To motivate our design, we analyze the two major MVPD paradigms. Early approaches (Chavdarova et al., 2018; Yan et al., 2021) adopt a *top-down* structure, reasoning about pedestrian presence through explicit geometric modeling. A monocular detector first produces 2D bounding boxes in each view independently, and the per-view detections are projected onto the ground plane and fused by Euclidean proximity and cross-view visual agreement. Because cross-view fusion is governed by camera geometry and the 2D detector benefits from large-scale image pretraining, such methods generalize well to new scenes and camera layouts Qiu et al. (2024). Nonetheless, independent per-view perception leaves partial occlusion errors uncompensated, which propagate ambiguities into the fusion stage. Recent approaches (Hou & Zheng, 2021; Teepe et al., 2024a) address this limitation through *bottom-up* structure, projecting image features from all views onto a shared ground plane and jointly predicting an occupancy map on that Bird's-Eye-View (BEV) feature representation. Fusing all viewpoints before detection lets complementary observations reinforce each other and improves in-domain accuracy, but that same map entangles appearance with camera geometry. As localization is learned on that map, the model tends to overfit and fail on unseen scenes. A common resolution is to expand the training distribution with large-scale synthetic datasets (Vora et al., 2023; Zhang et al., 2026) of greater diversity, and transfer the knowledge from simulation to real-world data. However, without addressing the shared representation entanglement, data scaling alone offers limited generalization gains for bottom-up framework. This analysis raises a key question: "*Can we inherit the occlusion robustness of multi-view feature aggregation while achieving generalization through a scene-agnostic geometric formulation?*"

We address this by reformulating MVPD ground localization through the lens of projective geometry, which bridges the 2D image and 3D space analytically, as illustrated in Fig. 1. Each pedestrian is represented as a 3D quadric anchored on the ground plane, and each per-view detection as its projected 2D conic. Under the pinhole camera model, a quadric projects precisely onto a conic in each view, and conics from multiple views can be triangulated back to a unique quadric, forming an exact, invertible geometric bridge. Hence, refining the conics to agree with observed 2D detections across all views directly drives the triangulated 3D position toward the correct ground location. Building on this formulation, we propose a query-based framework **P**rojective **G**eometry for **M**ulti-**V**iew **G**eneralization (**PGMVG**). PGMVG maintains a sparse set of 3D quadric queries, each iteratively refined through three steps: project each quadric into every view as a 2D conic, refine the conics' position and shape across views, and triangulate the refined conics into an updated 3D quadric. By jointly refining 2D evidences across all views and lifting them to 3D through explicit geometry without reliance on a scene-specific BEV representation, PGMVG resolves the trade-off between occlusion robustness and generalization. Tab. 1 demonstrates the differences between PGMVG and existing paradigms. The main contributions of this paper are:

- We propose **PGMVG**, a novel query-based MVPD framework in which cross-view evidence is aggregated in image space and localization is governed by projective geometry. The geometric path holds no learned parameters and applies to any scene and camera configuration.
- We introduce **Supporting Constraints**, which make projective quadric triangulation viable under MVPD's weak supervision by resolving two distinct failure modes: a *physical shape prior* removes the *ill-posedness* of the unconstrained quadric, and a *structural decoupling* of translation from shape removes the *ill-conditioning* of the joint solve.

- On three benchmarks (WildTrack, MultiViewX, and GMVD) under **cross-scene**, **cross-configuration**, and **sim-to-real** settings, PGMVG consistently surpasses state-of-the-art in generalization while maintaining strong in-domain performance.

2 QUADRIC-CONIC FORMULATION FOR MVPD

2.1 FUNDAMENTAL CONCEPTS

We establish the geometric foundations of our method by reviewing the key properties of quadrics and conics under projective geometry. For a comprehensive understanding, we refer the reader to Hartley & Zisserman (2003) and Appendix B.

We consider a multi-view setting comprising V calibrated and synchronized cameras with overlapping fields of view, indexed by $v \in \{1, \dots, V\}$. Each camera is described by a projection matrix $\mathbf{P}_v \in \mathbb{R}^{3\times4}$, and they share a common ground plane. A *quadric* surface in 3D is the set of homogeneous points $\mathbf{x} \in \mathbb{R}^4$ satisfying $\mathbf{x}^\top \mathbf{Q} \mathbf{x} = 0$, where $\mathbf{Q} = \mathbf{Q}^\top \in \mathbb{R}^{4\times4}$ with 9 degrees of freedom. Its *dual* $\mathbf{Q}^* = \mathrm{adj}(\mathbf{Q})$ represents the same surface as the envelope of its tangent planes $\boldsymbol{\pi} \in \mathbb{R}^4$, satisfying $\boldsymbol{\pi}^\top \mathbf{Q}^* \boldsymbol{\pi} = 0$. Dual quadric $\mathbf{Q}^*$ is parameterized as follows:

$$\mathbf{Q}^* = \mathbf{Z}\,\mathrm{diag}(s_1^2,\, s_2^2,\, s_3^2,\, -1)\,\mathbf{Z}^\top, \mathbf{Z} = \begin{bmatrix} \mathbf{R}(\boldsymbol{\theta}) & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{1}$$

where $\mathbf{s} = (s_1, s_2, s_3) \in \mathbb{R}^3$ are the quadric's semi-axes and $\mathbf{Z}$ places the quadric in the 3D scene through a translation $\mathbf{t} \in \mathbb{R}^3$ and a rotation matrix $\mathbf{R}(\boldsymbol{\theta}) \in \mathbb{R}^{3\times3}$ parameterized by $\boldsymbol{\theta} \in \mathbb{R}^3$. Together these give $\mathbf{q} = (\boldsymbol{\theta}, \mathbf{t}, \mathbf{s}) \in \mathbb{R}^9$, the parameter representation of $\mathbf{Q}^*$. In its canonical form ($\mathbf{Z} = \mathbf{I}_4$), the quadric reduces to an axis-aligned ellipsoid centered at the scene origin. The translation $\mathbf{t}$, which encodes the quadric center position in 3D, is recovered directly from the last column of $\mathbf{Q}^*$:

$$\mathbf{t} = (t_x, t_y, t_z) = \big(\mathbf{Q}^*[1,4],\, \mathbf{Q}^*[2,4],\, \mathbf{Q}^*[3,4]\big)^\top \,/\, \mathbf{Q}^*[4,4]. \tag{2}$$

The first two components (t_x, t_y) yield the quadric's ground position, while the third component t_z encodes its height above the ground plane. The projection of a quadric onto an image plane is called *conic*. A conic is a curve defined by $\mathbf{C}_v \in \mathbb{R}^{3\times3}$ over homogeneous image points $\mathbf{u} \in \mathbb{R}^3$, $\mathbf{u}^\top \mathbf{C}_v\, \mathbf{u} = 0$, whose dual $\mathbf{C}_v^* = \mathrm{adj}(\mathbf{C}_v)$ envelopes the conic's tangent lines. Under the pinhole model, a dual quadric $\mathbf{Q}^*$ projects exactly to a dual conic $\mathbf{C}_v^*$ in the camera view v:

$$\begin{aligned} &\mathbf{C}_v^* \;\propto\; \mathbf{P}_v\, \mathbf{Q}^*\, \mathbf{P}_v^\top, && \text{(projection)} \\ &\mathbf{C}_v^* = \mathbf{H}_v\, \mathrm{diag}(a_v^2,\, b_v^2,\, -1)\, \mathbf{H}_v^\top, \mathbf{H}_v = \begin{bmatrix} \mathbf{R}(\phi_v) & (c_{v,x}, c_{v,y})^\top \\ \mathbf{0}^\top & 1 \end{bmatrix} && \text{(dual conic)} \end{aligned} \tag{3}$$

defining the forward relation ($3D \rightarrow 2D$) via dual space. $\mathbf{H}_v$ places the conic with semi-axes (a_v, b_v) in 2D image through center $(c_{v,x}, c_{v,y})$ and angle ϕ_v. The parameter representation of $\mathbf{C}_v^*$ is $\mathbf{c}_v = (c_{v,x}, c_{v,y}, a_v, b_v, \phi_v) \in \mathbb{R}^5$. A bounding box $\mathbf{b}_v$ and the axis-aligned conic $\mathbf{c}_v$ inscribed in it are accordingly interchangeable, both being specified by a center and two half-extents (see Appendix B.1). For the inverse relation ($2D \rightarrow 3D$), we adopt the closed-form solution of LfD Rubino et al. (2017) for quadric triangulation from multi-view conics. Eq. 3 imposes a linear constraint on the entries of $\mathbf{Q}^*$ up to scale. Stacking the vectorized projection equations across all visible views yields a homogeneous linear system $\mathbf{M}\,\mathbf{w} = \mathbf{0}$, where $\mathbf{w} \in \mathbb{R}^{10}$ collects the 10 independent entries of the symmetric matrix $\mathbf{Q}^*$ and $\mathbf{M}$ is assembled from the per-view camera matrices $\{\mathbf{P}_v\}$ and observed dual conics $\{\mathbf{C}_v^*\}$. The solution is the right singular vector of $\mathbf{M}$ associated with its smallest singular value, obtained via SVD. Given complete 3D and 2D annotation, the projective geometry above provides end-to-end dual-domain supervision for 3D localization:

$$\text{3D annotation} \overset{\text{3D loss}}{\leftrightarrow} \mathbf{q} \;\leftrightarrow\; \mathbf{Q}^* \underset{\text{triangulation}}{\overset{\mathbf{P}_v(\cdot)\mathbf{P}_v^\top}{\rightleftharpoons}} \mathbf{C}_v^* \;\leftrightarrow\; \mathbf{c}_v \overset{\text{2D loss}}{\leftrightarrow} \text{2D annotation} \tag{4}$$

2.2 SUPPORTING CONSTRAINTS FOR MVPD

Eq. 4 admits an alternative to direct 3D regression, in which localization is lifted from 2D detections and generalizability inherited from pretrained image detectors. We accordingly represent each pedestrian as $\mathbf{Q}^*$ and reformulate ground localization as iterative conic refinement, in which the

projected $\{\mathbf{C}_v^*\}$ are refined to fit the observed detections and triangulated back into an updated $\hat{\mathbf{Q}}^*$ that converges toward the correct position. While the 2D side provides $\mathbf{b}_v$ in every view where the pedestrian is visible ($\delta_v = 1$), MVPD datasets only annotate ground-plane position (t_x, t_y) in 3D, leaving orientation $\boldsymbol{\theta}$ and semi-axes $\mathbf{s}$ unconstrained. Estimating a full 9-parameter quadric is therefore *ill-posed*, admitting no unique solution. This underdetermination is further compounded by frequent occlusion, as a pedestrian visible in only a few views supplies too few constraints to determine a quadric. Fortunately, considering a standing person, horizontal semi-axes are nearly equal and the orientation is immaterial to a ground position. We therefore impose a **physical shape prior** on the quadric parametrization: fixing the rotation to $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{I}_3$ so that the quadric is axis-aligned, tying the two horizontal semi-axes to a single radius $s_1 = s_2 = r$ so that it is horizontally isotropic, and setting $s_3 = t_z$ so that its base touches $z = 0$ while its apex reaches $2t_z$. Substituting reduced parameters into Eq. 1 gives the new closed-form dual quadric:

$$\mathbf{Q}^* = \mathbf{Z}\,\mathrm{diag}(r^2,\, r^2,\, t_z^2,\, -1)\,\mathbf{Z}^\top, \mathbf{Z} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{5}$$

The **physical shape prior** yields an upright ellipsoid of revolution, reducing the parametrization from 9 to 4 parameters $\mathbf{q} = (t_x, t_y, t_z, r) \in \mathbb{R}^4$. $\mathbf{q}$ can be separated into a 3D center translation $\mathbf{t} = (t_x, t_y, t_z)$ and a single shape parameter r. As noted in LfD, the entries of $\mathbf{Q}^*$ couple translation and shape at very different scales. Since center translation typically far exceeds the pedestrian dimension ($t_x, t_y \gg r, t_z$), small errors in the dominant translation term during triangulation could corrupt shape estimation. We address this *ill-conditioned* joint estimation through **structural decoupling**, leveraging the fact that translation parameters appear independently of shape in the last column and row of $\mathbf{Q}^*$. This is made explicit by expanding Eq. 5:

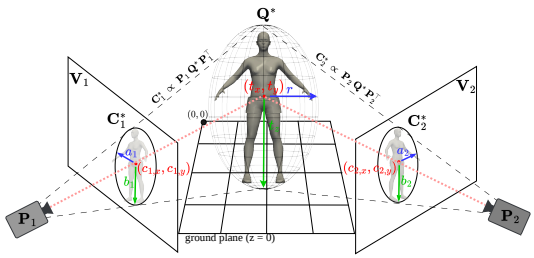


Figure 2: A pedestrian is modeled as an upright ellipsoid of revolution $\mathbf{Q}^*$ grounded at (t_x, t_y) with horizontal radius r and vertical semi-axis t_z. Each camera projects dual quadric $\mathbf{Q}^*$ onto its image plane as a dual conic $\mathbf{C}_v^*$, parameterized by center $(c_{v,x}, c_{v,y})$ and semi-axes (a_v, b_v).

$$\mathbf{Q}^* = \begin{bmatrix} \mathrm{diag}(r^2,\, r^2,\, t_z^2) - \mathbf{t}\mathbf{t}^\top & -\mathbf{t} \\ -\mathbf{t}^\top & -1 \end{bmatrix}. \tag{6}$$

While the upper-left 3×3 block entangles translation and shape, the last column and row depend solely on $\mathbf{t}$, allowing it to be recovered independently of r. **Structural decoupling** thus decomposes quadric triangulation into two sequential steps: *center translation* and *radius estimation*. First, we triangulate the quadric center $\mathbf{X} = (t_x, t_y, t_z, 1)^\top$ from projected conic centers $\{(c_{v,x}, c_{v,y})\}_{v:\delta_v=1}$ via DLT, where each visible view contributes two linear constraints:

$$(c_{v,x},\, c_{v,y},\, 1)^\top \times (\mathbf{P}_v\mathbf{X}) = \mathbf{0}. \tag{7}$$

Stacking over all visible views gives $\mathbf{A}\mathbf{X} = \mathbf{0}$, solved via SVD to yield $\mathbf{t}$. Secondly, since $\mathbf{Z}$ carries only the recovered translation $\hat{\mathbf{t}}$, absorbing it into the translated camera matrix $\mathbf{P}'_v = \mathbf{P}_v\mathbf{Z}$ centers the quadric at the origin and reduces Eq. 3 to:

$$\mathbf{C}_v^* \;\propto\; \mathbf{P}'_v\,\mathrm{diag}(r^2,\, r^2,\, t_z^2,\, -1)\,\mathbf{P}'_v{}^\top. \tag{8}$$

Centering sets $t_x = t_y = 0$, leaving the radius r and the vertical semi-axis t_z as the only unknowns, both on the scale of human body dimensions rather than of scene extent. This eliminates the scale disparity that made joint estimation ill-conditioned. With rotation fixed to $\mathbf{I}_3$ and t_z recovered in the first step, vectorizing Eq. 8 and stacking across visible views yields:

$$\mathbf{M}'\,\mathbf{w}' = \mathbf{0}, \qquad \mathbf{w}' = (r^2, t_z^2, -1)^\top, \tag{9}$$

where $\mathbf{M}'$ is assembled from $\{\mathbf{P}'_v\}$ and $\{\mathbf{C}_v^*\}_{v:\delta_v=1}$. The solution follows from SVD as described in the above section. Rescaling so that $[\mathbf{w}']_3 = -1$ directly yields $r = \sqrt{[\mathbf{w}']_1}$, unlike a generic quadric whose orientation must be additionally extracted from eigenvectors. Fig. 2 depicts the **physical shape prior**, where a pedestrian is enclosed in an upright ellipsoid of revolution $\mathbf{Q}^*$ parameterized by ground position (t_x, t_y), horizontal radius r, and vertical semi-axis t_z, projecting to dual conics $\mathbf{C}_v^*$ in each view. The conic centers $(c_{v,x}, c_{v,y})$ are used to recover the full translation $\mathbf{t}$ in the first step of **structural decoupling**. *Following differentiable triangulation Iskakov et al. (2019), we*

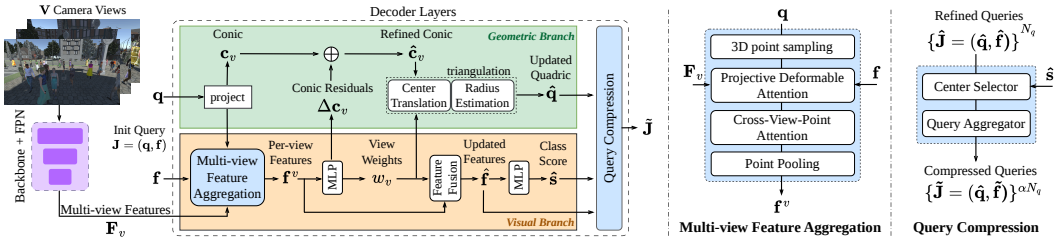


Figure 3: The overall architecture of the proposed PGMVG framework.

optimize the quadric–conic projection and both steps of triangulation as batched matrix operations that are end-to-end differentiable, so gradients propagate from the ground loss back to the per-view features through exact geometry. This path contains no learned parameters and depends only on camera matrices $\mathbf{P}_v$, so it applies to any number of cameras in any configuration.

3 PGMVG FRAMEWORK

Overview. Grounded in the above establishment, PGMVG adopts a DETR-based architecture Carion et al. (2020), in which each decoder layer comprises two branches (Fig. 3): a *Visual branch* that produces object scores, per-view reliability weights, and conic residuals from multi-view appearance features, and a *Geometric branch* that enforces projective geometry to update the 3D quadric.

Multi-view Backbone. Standard DETR encoder applies self-attention Vaswani et al. (2017) over all spatial tokens of an image to produce globally context-aware representations. However, in multi-view settings, its cost scales quadratically and compounds V-fold across views, making it a computational bottleneck. Following D^{2}ETR Lin et al. (2022), we omit the encoder entirely, as a strong backbone paired with a Feature Pyramid Network (FPN) Lin et al. (2017) already yields context-enriched representations. PGMVG is therefore built as a decoder-only DETR, in which a lightweight detector backbone extracts multi-scale feature maps $\mathbf{F}_v$ independently for each view.

Quadric-embedded Query. To jointly serve the two-branch design, each query $\mathbf{J} = \{\mathbf{f}, \mathbf{q}\}$ decouples appearance from geometry: the feature component $\mathbf{f} \in \mathbb{R}^d$ carries cross-view visual features refined by the Visual branch, while the geometric component $\mathbf{q} = (t_x, t_y, t_z, r)$ parameterizes $\mathbf{Q}^*$, updated by the Geometric branch through conic triangulation. N_q queries are initialized with $\mathbf{f}$ drawn randomly and placed on a uniform grid spanning the ground-plane extent, with t_z and r set to typical standing-human dimensions, transferable to scenes of any size without retraining.

Visual Branch. For each query, we construct N_p 3D sampling points within the ellipsoidal volume of $\mathbf{q}$ to capture appearance cues across the full body extent. $\mathbf{f}$ drives an offset MLP to adaptively predict 3D offsets $(\Delta x_p, \Delta y_p, \Delta z_p)$ from the quadric center based on the current appearance estimate, each constrained inside the ellipsoidal surface $\frac{\Delta x_p^2 + \Delta y_p^2}{r^2} + \frac{\Delta z_p^2}{t_z^2} \leq 1$. The resulting sample points are $(x_p, y_p, z_p) = (t_x + \Delta x_p,\ t_y + \Delta y_p,\ t_z + \Delta z_p)$. Each point is projected onto all $\mathbf{V}$ views via $[x_p^v, y_p^v, z_p^v, 1]^\top = \mathbf{P}_v[x_p, y_p, z_p, 1]^\top$, and multi-scale features are gathered at the projected locations from $\mathbf{F}_v$ via projective deformable attention Zhang et al. (2021a), yielding a per-view per-point feature $\mathbf{f}_p^v \in \mathbb{R}^d$. To aggregate evidence across sampling points and viewpoints, we introduce *Cross-View-Point Attention* with two sequential multi-head attention (MHA) operations. Point attention first mixes features across the N_p samples within each view, giving each point global body context, then View attention exchanges features across all V views at each location, letting visible views compensate for occluded ones:

$$\mathbf{h}_p^v = \mathrm{MHA}_{\mathrm{point}}\big(\mathbf{f}_p^v,\ \{\mathbf{f}_{p'}^v\}_{p'=1}^{N_p},\ \{\mathbf{f}_{p'}^v\}_{p'=1}^{N_p}\big), \quad \hat{\mathbf{f}}_p^v = \mathrm{MHA}_{\mathrm{view}}\big(\mathbf{h}_p^v,\ \{\mathbf{h}_p^{v'}\}_{v'=1}^{V},\ \{\mathbf{h}_p^{v'}\}_{v'=1}^{V}\big). \quad (10)$$

The enriched features are then compressed into a single per-view query feature $\mathbf{f}^v$ by learned point pooling. From $\mathbf{f}^v$, two MLP heads operate in parallel: one estimates a view confidence $w_v = \mathrm{sigmoid}(\mathrm{MLP}_{\mathrm{view}}(\mathbf{f}^v))$ that suppresses occluded or ambiguous views whose noisy features would corrupt the aggregated appearance; the other predicts a conic residual $\Delta\mathbf{c}_v = \mathrm{MLP}_{\mathrm{conic}}(\mathbf{f}^v)$ that corrects projection errors from early quadric estimates, letting the two branches progressively refine

each other. Point pooling and the view-weighted update of $\mathbf{f}$ are then given by

$$w_p^v = \mathrm{softmax}_p\big(\mathrm{MLP}_{\mathrm{point}}(\hat{\mathbf{f}}_p^v)\big), \quad \mathbf{f}^v = \sum_{p=1}^{N_p} w_p^v\, \hat{\mathbf{f}}_p^v, \quad \hat{\mathbf{f}} = \mathrm{LN}\Big(\mathbf{f} + \mathrm{FFN}\big(\textstyle\sum_{v=1}^{V} w_v\, \mathbf{f}^v\big)\Big), \quad (11)$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization. A classification score is predicted as $\hat{s} = \mathrm{MLP}_{\mathrm{class}}(\hat{\mathbf{f}})$, made robust to occlusion through cross-view feature synthesis.

Geometric Branch. Building on the formulation from Supporting Constraints in Sec. 2.2, the Geometric branch applies the derived projection and triangulation directly within the decoder. At each decoder step, $\mathbf{Q}^*$ is constructed from $\mathbf{q}$ via Eq. 5 and projected to per-view dual conics $\mathbf{C}^v$ via Eq. 3, which are decomposed into 2D conic parameters $\mathbf{c}_v$. $\mathbf{c}_v$ are then refined using the residuals supplied by the Visual branch: $\hat{\mathbf{c}}_v = \mathbf{c}_v + \Delta\mathbf{c}_v$. Following structural decoupling, $\mathbf{q}$ is updated from $\{\hat{\mathbf{c}}_v\}_{v:\delta_v=1}$ in two steps. First, the refined conic centers $(\hat{c}_{v,x}, \hat{c}_{v,y})$ and view weights w_v are fed into the confidence-weighted DLT Eq. 7, where each view's contribution is scaled by w_v to reduce errors introduced by unreliable views, recovering $\hat{\mathbf{t}} = (\hat{t}_x, \hat{t}_y, \hat{t}_z)$. Second, the recovered translation $\hat{\mathbf{t}}$ is substituted into the radius solver Eqs. 8 and 9 to yield $\hat{r}$, fully updating $\hat{\mathbf{q}}$.

Query Compression (QC). Although denser query initialization ensures better scene coverage, more queries inevitably surround the same pedestrian, sampling overlapping visual evidence and collapsing into near-duplicate predictions. Since only one query per pedestrian is assigned as the positive match via Hungarian matching, all remaining near-duplicate queries, despite their high confidence, are forcibly labeled as negatives. As shown in DDQ-DETR Zhang et al. (2023), the gradient from each forced negative weakens or even reverses that of the matched positive, dramatically suppressing convergence. In multi-view settings, occlusion further exacerbates the gradient conflict as most views contribute noise rather than genuine appearance, making positive and near-duplicate queries increasingly indistinguishable and thus reducing detection recall. To counter these compounding effects, we propose a **QC** module with two complementary operations: a *Center Selector (CS)* that compresses the dense queries into αN_q representatives, and a *Query Aggregator (QA)* that consolidates complementary information into compact, discriminative representations. The **CS** ranks all N_q queries by their classification score $\hat{s}$ and applies soft-NMS to select αN_q cluster centers, where $\alpha \in (0, 1)$ is the compression ratio. Unlike hard-NMS used in DDQ-DETR, which suppresses all queries within a fixed radius, soft-NMS progressively down-weights nearby candidates, preserving genuine detections in dense crowds. Subsequently, the **QA** enriches each center by casting it as an attention query over all N_q detection queries as keys and values. Since each center captures only a local context, the full query set supplies the broader scene context it alone cannot model. The αN_q centers first coordinate through self-attention, then cross-attend to all queries:

$$\mathbf{f}_k^{(1)} = \mathrm{LN}\Big(\mathbf{f}_k + \mathrm{MHA}_{\mathrm{self}}\big(\mathbf{f}_k,\, \{\mathbf{f}_{k'}\}^K,\, \{\mathbf{f}_{k'}\}^K\big)\Big), \quad (12)$$

$$\tilde{\mathbf{f}}_k = \mathrm{LN}\Big(\mathbf{f}_k^{(1)} + \mathrm{MHA}_{\mathrm{cross}}\big(\mathbf{f}_k^{(1)},\, \{\mathbf{f}_i\}^{N_q},\, \{\mathbf{f}_i\}^{N_q}\big)\Big), \quad (13)$$

where $K = \alpha N_q$ is the number of compressed queries. **QC** is applied at the last decoder layer, once classification scores are sufficiently converged to enable reliable center selection. The compressed queries $\tilde{\mathbf{J}}$ constitute the final predictions of PGMVG, each decoded into a ground-plane position $(\hat{t}_x, \hat{t}_y)$, per-view conics $\{\hat{\mathbf{c}}_v\}$, and a classification score $\hat{s}$.

Training Objectives. Predictions are first matched to ground-truth annotations via Hungarian matching, with a cost matrix combining classification and ground-plane localization terms. The classification loss $\mathcal{L}_{\mathrm{class}}$ applies focal loss to $\hat{s}$. The ground loss $\mathcal{L}_{\mathrm{ground}}$ supervises the triangulated BEV position $(\hat{t}_x, \hat{t}_y)$ with smooth-ℓ_1 against the ground-truth position. The conic loss $\mathcal{L}_{\mathrm{image}}$ converts each refined conic $\hat{\mathbf{c}}_v$ to its axis-aligned bounding box $\hat{\mathbf{b}}_v$ and supervises it with smooth-ℓ_1 and GIoU losses, summed over visible views $\{v : \delta_v = 1\}$. All three terms are applied at the output of each decoder layer: $\mathcal{L} = \lambda_{\mathrm{ground}}\mathcal{L}_{\mathrm{ground}} + \lambda_{\mathrm{class}}\mathcal{L}_{\mathrm{class}} + \lambda_{\mathrm{image}}\mathcal{L}_{\mathrm{image}}$.

4 Experiments

Datasets and Metrics. We evaluate on three MVPD datasets. *WildTrack* Chavdarova et al. (2018) and *MultiViewX* Hou et al. (2020) serve as standard in-domain benchmarks, covering real-world

and synthetic settings, respectively. *GMVD* Vora et al. (2023) is a large-scale multi-scene multi-configuration synthetic dataset used to assess cross-scene generalization and sim-to-real transfer. For multi-view detection, we report MODA, MODP, Precision (Pre), Recall (Rec).

Implementation Details. Our model is trained on a single NVIDIA RTX 4090. We adopt ResNet50 He et al. (2016) backbone pre-trained on CrowdHuman dataset Shao et al. (2018), paired with a 4-scale FPN and a 4-layer DETR decoder with N_q of 600. Input images are resized to 1280×720. The model is optimized with AdamW at a learning rate of 10^{-4}, with the backbone and FPN using a reduced rate of 10^{-5}. λ_{ground}, λ_{class}, λ_{image} are set as $1, 0.25, 0.5$, respectively.

Full dataset statistics, metric definitions, and additional implementation details are provided in Appendices C, D, and E, respectively.

Baselines include MVDet Hou et al. (2020), SHOT Song et al. (2021), MVDetr Hou & Zheng (2021), 3DROM Qiu et al. (2022), MVAug Engilberge et al. (2023), GMVD Vora et al. (2023), Earlybird Teepe et al. (2024a), TrackTacular Teepe et al. (2024b), Booster-SHOT Hwang et al. (2024), MVFP Aung et al. (2024), TMVD Qiu et al. (2024), MSASW Li et al. (2025), MCIVA Li et al. (2026), MVUDA Brorsson et al. (2026) and MVDGC Phan et al. (2026). Approaches are listed chronologically in below tables, with the best highlighted in **bold** and the second-best underlined.

Table 2: *Cross-scene generalization* on GMVD dataset.

Method	MODA	MODP	Pre	Rec	FPS
MVDet	50.5	72.8	83.6	64.7	**18.7**
GMVD	68.2	76.3	91.5	75.5	18.5
Booster-SHOT	73.0	76.5	92.8	79.1	18.0
MVFP	73.3	76.5	93.0	79.2	–
MSASW	73.1	76.8	93.2	78.8	18.4
MVDGC	71.2	79.6	93.0	76.9	3.8
MCIVA	75.9	78.6	93.8	81.2	18.5
PGMVG(ours)	**81.0**	**87.1**	**95.8**	**84.7**	8.6

Table 3: *Cross-configuration generalization* on WildTrack dataset.

Method	**Exp. A** Train: C2, C4, C5, C6 Test: C1, C3, C5, C7				**Exp. B** Train: C1, C3, C5, C7 Test: C2, C4, C5, C6			
	MODA	MODP	Pre	Rec	MODA	MODP	Pre	Rec
MVDet	43.2	68.2	94.6	45.8	27.8	68.7	90.8	31.0
SHOT	51.4	72.5	94.4	54.6	15.3	62.9	89.2	17.4
MVDeTr	41.7	73.7	92.0	45.7	5.6	65.5	62.4	14.0
GMVD	75.1	71.1	94.3	79.9	62.6	67.4	86.7	73.9
TMVD	78.7	73.2	95.2	82.9	70.4	68.3	91.2	77.9
MVUDA	79.2	78.0	96.2	82.5	**81.2**	68.8	**95.7**	85.0
PGMVG(ours)	**81.8**	**81.1**	**97.2**	**84.2**	81.1	**74.1**	94.6	**86.0**

Scene Generalization. Tab. 2 presents cross-scene generalization results on GMVD, where PGMVG outperforms all baselines across every metric, achieving 81.0 MODA with a margin of $+5.1$ over the second-best MCIVA and $+12.8$ over the GMVD method itself. The margin ($+7.5$) is larger in MODP, where PGMVG reaches 87.1 against 79.6 for the closest baseline. The high detection precision comes from the position computed through triangulation instead of regressing from BEV map. The superiority in all metrics indicates that the projective geometry framework generalizes effectively across varying environments and configurations. Regarding runtime, bottom-up baselines operate at comparable throughput ($\sim$18 FPS) owing to the analogy in their architecture. PGMVG runs at 8.6 FPS but remains practical for real-world deployment.

Cross-configuration Generalization. Tab. 3 evaluates generalization to near-disjoint camera subsets on WildTrack. In Exp. A, PGMVG achieves the highest MODA (81.8) and MODP (81.1), outperforming MVUDA by $+2.6$ MODA and $+3.1$ MODP. In Exp. B, PGMVG (81.1 MODA) matches MVUDA's performance (81.2) without utilizing any target-scene adaptation. Across both experiments, PGMVG remains stable with only 0.7 MODA variation, while standard bottom-up methods, including GMVD, collapse under configuration shift, exposing their reliance on scene-specific visual patterns. TMVD also shows high sensitivity at MODA, suggesting that cylinder-based back-projection does not generalize well across camera subsets. Moreover, PGMVG runs substantially faster than TMVD and MVUDA, which process one frame in 15.7s (0.06 FPS) and 1s respectively.

Table 4: *Sim-to-real scene generalization*: trained on synthetic datasets, tested on real scenes. NV_t and NV_i are numbers of views for training and inference.

Method	MODA	MODP	Pre	Rec
MultiViewX → WildTrack ($NV_t = 6$, $NV_i = 6$)				
MVDet	17.0	65.8	60.5	48.8
SHOT	53.6	72.0	75.2	79.8
MVDeTr	50.2	69.1	74.0	77.3
MVAug	26.3	58.0	71.9	50.8
3DROM	67.5	65.6	**94.5**	71.7
GMVD	66.1	72.2	82.0	84.7
TMVD	71.5	75.5	89.3	81.3
MVFP	76.7	74.9	85.2	**92.8**
PGMVG(ours)	**85.0**	**81.1**	92.4	92.6
MultiViewX → WildTrack ($NV_t = 6$, $NV_i = 7$)				
GMVD	70.7	73.8	89.1	80.6
TMVD	74.9	76.9	90.4	83.8
MVFP	82.6	76.2	89.6	**93.4**
MVUDA	**85.4**	75.3	**96.5**	88.7
PGMVG(ours)	85.1	**82.2**	92.1	92.7
GMVD → WildTrack (NV_t = multi, $NV_i = 7$)				
MVDet	69.0	71.1	88.4	79.5
GMVD	80.1	75.6	90.9	89.1
MVFP	85.6	78.0	91.8	**94.0**
PGMVG(ours)	**87.2**	**83.3**	**93.2**	**94.0**

Sim-to-real Generalization. Tab. 4 evaluates sim-to-real transfer across three settings, all testing on the real-world WildTrack dataset. In the first group, PGMVG achieves 85.0 MODA, outperforming the second-best MVFP by 8.3 MODA, demonstrating strong domain transfer. In the second group with full configuration, PGMVG maintains high MODA of 85.1. While MVUDA achieves a slightly higher MODA, it localizes far less precisely (75.3 vs 82.2 MODP) and requires unlabeled target-domain data for adaptation, a privilege our method foregoes. In the third group where training benefits from a larger dataset, PGMVG further improves to 87.2 MODA, surpassing all baselines. The MODP margin is nearly constant across the three settings (+5.6, +5.3, +5.3), indicating that the precision advantage follows from the geometric formulation. As pinhole cameras behave identically in simulation and reality, the geometric path carries no domain gap and only the visual branch confronts the appearance shift. Among methods that never access the target scene, PGMVG therefore makes the best use of synthetic training data.

In-domain Result. We report the average of MODA and MODP (Avg) as a holistic measure of detection and localization quality. As shown in Tab. 5, PGMVG achieves the highest overall average on MultiViewX (95.0 Avg) with the best MODP (94.1). On WildTrack, PGMVG remains competitive with other methods. Notably, methods that achieve strong in-domain performance, such as MCIVA and MVDGC, fall significantly below PGMVG on the cross-scene GMVD benchmark. In contrast, PGMVG maintains competitive in-domain performance while preserving cross-scene generalizability, demonstrating that geometry integration provides a strong scene-agnostic framework.

Table 5: In-domain performance on WildTrack and MultiViewX datasets.

Method	WildTrack					MultiViewX				
	MODA	MODP	Pre	Rec	Avg	MODA	MODP	Pre	Rec	Avg
MVDet	88.2	75.7	94.7	93.6	82.0	83.9	79.6	96.8	86.7	81.8
SHOT	90.2	76.5	96.1	94.0	83.4	88.3	82.0	96.6	91.5	85.2
MVDetr	91.5	82.1	97.4	94.0	86.8	93.7	91.3	**99.5**	94.5	92.5
3DROM	93.5	75.9	97.2	96.2	84.7	95.0	84.9	99.0	96.1	90.0
MVAug	93.2	79.8	96.3	97.0	86.5	95.3	89.7	99.4	95.9	92.5
GMVD	86.7	76.2	95.1	91.4	81.5	88.2	79.9	96.8	91.2	84.1
Earlybird	91.2	81.8	94.9	96.3	86.5	94.2	90.1	98.6	95.7	92.2
TrackTacular	93.2	77.5	97.3	95.8	85.4	**96.5**	75.0	99.4	97.1	85.8
Booster-SHOT	92.8	84.9	**97.5**	95.3	88.9	94.4	92.0	99.4	94.9	93.2
MVFP	94.1	78.8	96.4	97.7	86.5	95.7	85.1	98.4	97.2	90.4
TMVD	89.4	73.8	95.9	93.4	81.6	89.8	80.6	97.1	92.6	85.2
MSASW	89.8	79.7	95.9	94.2	84.8	91.8	83.8	98.4	93.3	87.8
MCIVA	93.4	**85.5**	95.3	**98.2**	**89.5**	96.4	93.3	99.4	97.0	94.9
MVDGC	**94.2**	83.1	97.4	96.9	88.7	95.5	93.0	98.2	**97.3**	94.3
PGMVG(ours)	94.0	83.7	**97.5**	96.4	88.9	95.9	**94.1**	99.4	96.5	**95.0**

Table 6: The impact of Geometry branch on generalizability of PGMVG (in-domain (left) and transfer (right)) under sim-to-real transfer and changes of camera configurations.

Geometry	In-domain				Transfer			
	MODA	MODP	Pre	Rec	MODA	MODP	Pre	Rec
	MultiViewX				**MultiViewX** → **WildTrack**			
✓	95.9	94.1	99.4	96.5	**85.1**	**82.2**	**92.1**	**92.7**
×	96.2	93.7	99.3	96.9	60.5 (↓-24.6)	58.7 (↓-23.5)	86.0 (↓-6.1)	72.3 (↓-20.4)
	WildTrack (C2, C4, C5, C6)				**WildTrack** (C2, C4, C5, C6) → (C1, C3, C5, C7)			
✓	90.0	79.2	96.1	93.9	**81.8**	**81.1**	**97.2**	**84.2**
×	90.9	79.7	96.0	94.9	58.3 (↓-23.5)	63.3 (↓-17.8)	91.4 (↓-5.8)	71.1 (↓-13.1)
	WildTrack (C1, C3, C5, C7)				**WildTrack** (C1, C3, C5, C7) → (C2, C4, C5, C6)			
✓	84.2	81.5	95.9	88.7	**81.1**	**74.1**	**94.6**	**86.0**
×	85.5	81.9	96.5	88.6	50.7 (↓-30.4)	52.3 (↓-21.8)	80.7 (↓-13.9)	70.1 (↓-25.9)

The Impact of Geometry Integration. Tab. 6 validates the claim above by isolating the contribution of the Geometric branch, replacing it with an MLP head that regresses $\hat{\mathbf{q}}$ directly from $\hat{\mathbf{f}}$ so that localization rests entirely on appearance. In every in-domain setting, the regressed variant (×) is the slightly better of the two, by 0.3 to 1.3 MODA, yet it collapses under transfer, falling 23.5 to 30.4 MODA behind. The stronger in-domain performance indicates not a better model but the overfitting that undermines generalizability. With no access to the 2D–3D geometry, the regressed variant can only learn the correlation between multi-view appearance and ground position that holds for one scene under one camera configuration. By removing the 3D lifting from the learning problem, PGMVG leaves the visual branch a purely 2D task that image pretraining already transfers.

Ablation on Supporting Constraints. Tab. 7 validates our adaptation of projective geometry for MVPD through two proposed constraints. The unconstrained 9-parameter quadric yields the weakest result, at 70.1 MODA. The shape prior, which resolves the ill-posedness of that parametrization,

Table 7: Ablation on Supporting Constraints on WildTrack dataset.

Shape Prior	Structural Decoupling	MODA	MODP	Pre	Rec
×	×	70.1	65.6	85.7	78.8
✓	×	78.9	71.3	90.4	85.6
×	✓	88.8	79.9	97.0	91.8
✓	✓	**94.0**	**83.7**	**97.5**	**96.4**

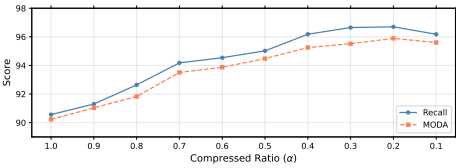


Figure 4: Ablation on Query Compression module on MultiViewX dataset.

gains +8.8 MODA, while structural decoupling, which resolves the ill-conditioning of the joint solve, gains +18.7 MODA. Together, the two constraints lift MODA by +23.9 points, with each addressing a distinct aspect: structural decoupling improves estimation conditioning and the shape prior constrains the solution to physically plausible dimensions.

Ablation on Query Compression. Fig. 4 ablates the compression ratio α of the **QC** module, where $\alpha = 1.0$ corresponds to removing QC entirely. Both Recall and MODA are lowest at $\alpha = 1.0$, confirming that near-duplicate gradient conflict is severe in the dense crowd setting of MultiViewX. As α decreases, both metrics rise steadily, with Recall and MODA gaining approximately 6 units at $\alpha = 0.2$. Performance peaks around $\alpha = 0.2$–0.3 before declining slightly at $\alpha = 0.1$, suggesting that too aggressive compression begins discarding genuine detections in crowded regions.

Qualitative Results. Fig. 5 presents qualitative results under the sim-to-real generalization setting. There are three failure modes of the appearance-only baseline explicitly highlighted: imprecise localization, missed detections, duplicate detections, which is consistent with the −24.6 MODA drop in Tab. 6. In contrast, PGMVG produces clean, well-aligned conics for each individual pedestrian across the full scene. The quadric–conic projection enforces geometric consistency across views, eliminating imprecise localization and duplicate detections. The 3D BEV visualization confirms that each pedestrian is represented by a distinct, well-separated quadric, demonstrating that our geometric formulation for MVPD provides the structural regularity necessary for cross-scene transfer that appearance features alone cannot supply.

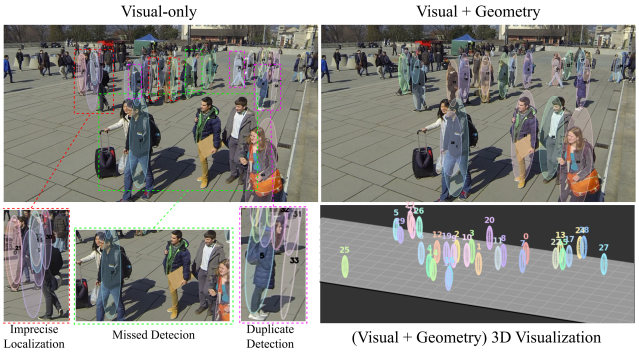


Figure 5: Illustration of PGMVG under sim-to-real generalization settings (MultiViewX → WildTrack) with (right column) and without (left column) the support of multi-view geometry reasoning.

Discussion. PGMVG may degrade in extremely dense crowds when all views are heavily occluded and provide no reliable cross-view correspondence. Occlusion can distort the predicted conics, making triangulation ill-conditioned and shifting feature sampling toward occluders or background, which reduces detection confidence and recall. Future work could address this limitation through part-aware geometric representations that triangulate visible body regions independently, together with temporal modeling that propagates reliable evidence from less-occluded frames. Moreover, because the geometric path carries no domain gap by construction, the 6.8 MODA residual under sim-to-real transfer is mostly the visual branch's appearance gap. Photometric and style augmentation, or human-background disentanglement could be composed into PGMVG.

5 Conclusion

We presented PGMVG, a query-based framework for generalizable MVPD that replaces scene-specific BEV reasoning with explicit projective geometry, made viable under MVPD's weak supervision by the Supporting Constraints we impose. By decoupling the appearance and geometric learning, PGMVG achieves strong cross-scene, cross-configuration, and sim-to-real transfer without target-scene adaptation, while remaining competitive in-domain. The key takeaway is that robust MVPD generalization depends not only on increasing training diversity, but also on designing the framework whose spatial reasoning is not tied to a particular camera layout.

AI USE STATEMENT

In this work, we used generative AI tools (i.e. ChatGPT) for revising passages, and correcting grammar.

All research ideas, methods, experiments, analyses, and conclusions are the authors' own, and the authors reviewed and take full responsibility for all content of this paper.

ETHICS STATEMENT

This work studies pedestrian localization using the existing benchmarks such as WildTrack, MultiViewX, and GMVD whereas backbone pre-trained on CrowdHuman dataset. Those datasets aim to estimate pedestrian positions rather than personal identities. This study does not perform facial recognition, determine personal identities, or infer sensitive attributes such as gender, ethnicity, or other demographic characteristics.

REPRODUCIBILITY STATEMENT

All the source code, models, and weights will be publicly available. All derivations supporting our projective geometry formulation, including the constrained quadric parametrization and the two-step decoupled triangulation, are given in full in Appendix B of the supplementary material.

REFERENCES

Alexandre Alahi, Laurent Jacques, Yannick Boursier, and Pierre Vandergheynst. Sparsity driven people localization with a heterogeneous network of cameras. *Journal of Mathematical Imaging and Vision*, 41(1):39–58, 2011.

Sithu Aung, Haesol Park, Hyungjoo Jung, and Junghyun Cho. Enhancing multi-view pedestrian detection through generalized 3d feature pulling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1196–1205, 2024.

Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 271–279, 2017.

Erik Brorsson, Lennart Svensson, Kristofer Bengtsson, and Knut Åkesson. Mvuda: Unsupervised domain adaptation for multi-view pedestrian detection. *Machine Vision and Applications*, 37(1): 6, 2026.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5030–5039, 2018.

Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12214–12223, 2020.

Martin Engilberge, Haixin Shi, Zhiye Wang, and Pascal Fua. Two-level data augmentation for calibrated multi-view detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 128–136, 2023.

Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):267–282, 2008.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1673–1682, 2021.

Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2020.

Jinwoo Hwang, Philipp Benz, and Pete Kim. Booster-shot: Boosting stacked homography transformations for multiview pedestrian detection with attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 363–372, 2024.

Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7718–7727, 2019.

He Li, Taiyu Liao, and Weihang Kong. Msasw: Multi-view scale-aware pedestrian detection with shared-weight supervision. *Advanced Engineering Informatics*, 68:103653, 2025.

He Li, Taiyu Liao, Weihang Kong, and Xingchen Zhang. Mciva: A multi-view pedestrian detection framework with central inverse nearest neighbor map and view adaptive module. *Information Fusion*, pp. 104142, 2026.

Junyu Lin, Xiaofeng Mao, Yuefeng Chen, Lei Xu, Yuan He, and Hui Xue. Dˆ 2etr: decoder-only detr with computationally efficient cross-scale attention. *arXiv preprint arXiv:2203.00860*, 2022.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Hong Mo, Xiong Zhang, Jianchao Tan, Cheng Yang, Qiong Gu, Bo Hang, and Wenqi Ren. Countformer: Multi-view crowd counting transformer. In *European conference on computer vision*, pp. 20–40. Springer, 2024.

Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 4 (1):1–8, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Peixi Peng, Yonghong Tian, Yaowei Wang, Jia Li, and Tiejun Huang. Robust multiple cameras pedestrian detection with multi-view bayesian network. *Pattern Recognition*, 48(5):1760–1772, 2015.

Thinh Phan, Hao Vo, Khoa Vo, Thanh Ngo, Cuong Pham, and Ngan Le. Mvdgc: Joint 3d and 2d multi-view pedestrian detection via dual geometric constraints. *arXiv preprint arXiv:2607.00273*, 2026.

Rui Qiu, Ming Xu, Yuyao Yan, Jeremy S Smith, and Xi Yang. 3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization. In *European Conference on Computer Vision*, pp. 695–710. Springer, 2022.

Rui Qiu, Ming Xu, Yuchen Ling, Jeremy S Smith, Yuyao Yan, and Xinheng Wang. A deep top-down framework towards generalisable multi-view pedestrian detection. *Neurocomputing*, 607:128458, 2024.

Cosimo Rubino, Marco Crocco, and Alessio Del Bue. 3d object localisation from multi-view image detections. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1281–1294, 2017.

Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Stacked homography transformations for multi-view pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6049–6057, 2021.

Torben Teepe, Philipp Wolters, Johannes Gilg, Fabian Herzog, and Gerhard Rigoll. Earlybird: early-fusion for multi-view tracking in the bird's eye view. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 102–111, 2024a.

Torben Teepe, Philipp Wolters, Johannes Gilg, Fabian Herzog, and Gerhard Rigoll. Lifting multi-view detection and tracking to the bird's eye view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 667–676, 2024b.

Rui Tian, Yunzhou Zhang, Yonghui Feng, Linghao Yang, Zhenzhong Cao, Sonya Coleman, and Dermot Kerr. Accurate and robust object slam with 3d quadric landmark reconstruction in outdoors. *IEEE Robotics and Automation Letters*, 7(2):1534–1541, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jeet Vora, Swetanjal Dutta, Kanishk Jain, Shyamgopal Karthik, and Vineet Gandhi. Bringing generalization to deep multi-view pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 110–119, 2023.

Yuyao Yan, Ming Xu, Jeremy S Smith, Mo Shen, and Jin Xi. Multicamera pedestrian detection using logic minimization. *Pattern Recognition*, 112:107703, 2021.

Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in neural information processing systems*, 34:13153–13164, 2021a.

Qi Zhang, Wei Lin, and Antoni B Chan. Cross-view cross-scene multi-view crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 557–567, 2021b.

Qi Zhang, Daijie Chen, Yunfei Gong, and Hui Huang. Synmvcrowd: A large synthetic benchmark for multi-view crowd counting and localization: Q. zhang et al. *International Journal of Computer Vision*, 134(4):191, 2026.

Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7329–7338, 2023.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A Related Work

Top-down methods infer ground-plane occupancy from per-view evidence gathered independently in each camera. Classical methods ground this in background-subtracted foreground silhouettes. Fleuret et al. (2008) introduce the Probabilistic Occupancy Map (POM), selecting the occupancy configuration whose silhouette projections best explain the per-view foreground. This line is later extended with sparsity constraints Alahi et al. (2011), occlusion modeling Peng et al. (2015), and logic-based consistency Yan et al. (2021). However, their reliance on clean foreground silhouettes limits applicability to controlled indoor settings, motivating deep learning approaches that

replace silhouettes with strong per-view 2D detectors. Baqué et al. (2017) integrate a CNN predicting pedestrian presence and body-part visibility into a CRF for occlusion-aware cross-view fusion, while Chavdarova et al. (2018) replace the generative POM entirely with an end-to-end model that crops and fuses sub-images at ground-plane cells. Independent per-view fusion leaves localization bounded by what each camera observes in isolation. *PGMVG instead jointly refines per-view conics through inter-view evidence before triangulation, enabling complementary views to compensate for partial occlusion.*

Bottom-up methods aggregate visual evidence from all views into a centralized BEV and predict pedestrian presence at each grid cell, demoting per-view localization to a feature-enhancement role. MVDet Hou et al. (2020) establishes this paradigm by projecting multi-view feature maps onto BEV and aggregating them. MVDeTr Hou & Zheng (2021) refines this by substituting deformable attention and introducing view-coherent data augmentation to mitigate perspective-warping artifacts. SHOT Song et al. (2021) and Booster-SHOT Hwang et al. (2024) address the flat-projection assumption by warping features onto a height-stacked set of homographies. 3DROM Qiu et al. (2022) tackles robustness through 3D random occlusion during training. TrackTacular Teepe et al. (2024b) extends BEV early fusion to joint detection and tracking. These methods advance occlusion robustness, yet remain bound to scene-specific patterns that fail to generalize. *PGMVG preserves inter-view aggregation while replacing the BEV model with projective geometry, generalizing by construction.*

Scene Generalization remains an open challenge for MVPD methods. GMVD Vora et al. (2023) exposes this limitation and adapts MVDet with randomly dropped views and average pooling to reduce dependence on a fixed camera count. MVAug Engilberge et al. (2023) counters scene overfitting through a two-level augmentation strategy that randomizes view geometry and BEV feature transformations during training. MVFP Aung et al. (2024) lifts 2D features into a 3D volume with dedicated foreground extraction modules, isolating pedestrian semantics from scene-specific background patterns. MVUDA Brorsson et al. (2026) adopts an unsupervised domain adaptation approach, finetuning a multi-view detector to a target rig via mean-teacher self-training and pseudo-labeling. TMVD Qiu et al. (2024) improves the top-down paradigm by back-projecting a cylinder per ground-plane cell, cropping per-view sub-features with RoIAlign, and fusing them via learned weights. *Unlike TMVD's exhaustive BEV grid search via fixed-scale rectangular cylinder projections, PGMVG employs sparse queries that iteratively adapt their geometry to each pedestrian through projective geometry. Moreover, unlike MVUDA, which requires unlabeled target-scene data and per-scene adaptation, PGMVG requires only camera parameters and generalizes to unseen configurations out of the box.*

B Projective Geometry for PGMVG: Details

This section provides the details on how projective geometry is adopted and accommodated for MVPD task, which is condensed in Sec. 2: "Quadric-Conic Formulation for MVPD".

B.1 Connecting MVPD Annotation to Quadric and Conic Representations

This subsection establishes the concrete correspondence between the projective objects (dual quadric, dual conic) and the annotations actually available in MVPD datasets (ground-plane positions, 2D bounding boxes), and shows how the two are connected.

MVPD supervision. Multi-view pedestrian detection provides two kinds of annotation for each pedestrian i:

- A *ground-plane position* $\mathbf{g} = (x, y)$, the foot-contact point on the ground plane $z = 0$.
- Per-view *boxes* $\mathbf{b}_v = (x_{\min}, y_{\min}, x_{\max}, y_{\max}) \in \mathbb{R}^4$, the axis-aligned rectangles enclosing the pedestrian in each view v where the pedestrian is visible ($\delta_v = 1$).

While the 3D supervision is coarse, with no metric depth, orientation, or body-size label, 2D bounding boxes adequately constrain the pedestrian's location and size across multiple views. Our goal is to transfer 3D localization to 2D views to capitalize on large-scale image-pretrained knowledge and eliminate the reliance on the BEV representation, which is prone to overfitting. This motivates us to find a representation that jointly encodes both $\mathbf{g}$ and $\mathbf{b}_v$ and be interchangeable between the 2D and 3D domains.

Quadric and Conic as 3D-2D domain bridge. For the 3D localization task, a 3D quadric and a 3D bounding box carry the same amount of object information (position, orientation, and size). However, the quadric offers greater flexibility: by independently adjusting its three semi-axes, it can conform more tightly to an object and reduce redundant bounding volume. More importantly, because a dual quadric is represented as a 4×4 matrix $\mathbf{Q}^*$, its projection under any pinhole camera $\mathbf{P}_v$ collapses to a single closed-form expression, $\mathbf{C}_v^* \propto \mathbf{P}_v \mathbf{Q}^* \mathbf{P}_v^\top$ (Eq. 3 in the main paper), yielding a 3×3 matrix that directly describes the conic observed in the image. This is a clean 3D-to-2D bridge that a 3D bounding box cannot provide due to lack of projective geometry representation. Based on this interchangeability, a quadric object conveniently encodes both forms of supervision from MVPD simultaneously.

It is worth noting that both the quadric and the conic admit three representations: the primal/point form ($\mathbf{Q}$ and $\mathbf{C}_v$), which describes the surface as a locus of points; the dual form ($\mathbf{Q}^*$ and $\mathbf{C}_v^*$), which describes it as an envelope of tangent planes or lines; and the parameter form ($\mathbf{q}$ and $\mathbf{c}_v$), which directly lists its geometric attributes. We work entirely in the dual space and use parameter representation, for two reasons:

- **The dual form carries the geometry.** Projection and triangulation are defined only in dual space, through Eq. 3, and there is no direct relation between $\mathbf{Q}$ and $\mathbf{C}_v$.
- **The parameter form carries the supervision.** The ground position is read from $\mathbf{q}$, and per-view refinement operates on $\mathbf{c}_v$ against bounding-box annotations. Both parameter sets are recovered directly from the dual matrices, so the primal form is never required.

The connection chain we adapt for MVPD is:

$$\underbrace{\text{ground point}}_{\text{3D annotation}} \leftrightarrow \mathbf{q} \leftrightarrow \mathbf{Q}^* \overset{\mathbf{P}_v(\cdot)\mathbf{P}_v^\top}{\underset{\text{triangulation}}{\rightleftharpoons}} \mathbf{C}_v^* \leftrightarrow \mathbf{c}_v \leftrightarrow \underbrace{\text{bounding box}}_{\text{2D annotation}} \tag{14}$$

The Conversion between Conic and Bounding Box. The center of the dual quadric $\mathbf{Q}^*$ in 3D space gives the ground position directly: $\mathbf{g} = (t_x, t_y)$ is simply the first two components of the translation $\mathbf{t}$. Its projection $\mathbf{C}_v^*$ from Eq. 3 in the main paper is a tilted dual conic characterized by five parameters: center $(c_{v,x}, c_{v,y})$, semi-axes (a_v, b_v), and orientation ϕ_v (the angle of the major axis from the image x-axis). To connect the conic to the bounding box supervision, we derive the tight axis-aligned bounding box of the conic analytically:

$$\begin{aligned}
\Delta x &= \sqrt{(a_v \cos\phi_v)^2 + (b_v \sin\phi_v)^2}, \\
\Delta y &= \sqrt{(a_v \sin\phi_v)^2 + (b_v \cos\phi_v)^2}, \\
\mathbf{b}_v &= \big(c_{v,x} - \Delta x,\ \ c_{v,y} - \Delta y,\ \ c_{v,x} + \Delta x,\ \ c_{v,y} + \Delta y\big),
\end{aligned}$$

where Δx and Δy are the half-width and half-height of the bounding box. This conversion allows the conic loss $\mathcal{L}_{\text{image}}$ to directly supervise the conic prediction against the ground-truth bounding boxes.

B.2 Quadric Projection & Triangulation

The forward relation ($3D \to 2D$) projects $\mathbf{Q}^*$ into each view to obtain $\mathbf{C}_v^*$, as shown in Eq. 3 in the main paper. The full projection is given by:

$$\mathbf{C}_v^* = \beta_v \, \mathbf{P}_v \, \mathbf{Q}^* \, \mathbf{P}_v^\top, \tag{15}$$

where $\beta_v > 0$ is a per-view scale factor arising because both $\mathbf{C}_v^*$ and $\mathbf{Q}^*$ are projective quantities defined only up to scale. We also have the inverse direction ($2D \to 3D$) that recovers $\mathbf{Q}^*$ from a set of $\mathbf{C}_v^*$ across multiple views, which is called triangulation. We detail the computation of $\mathbf{Q}^*$ in the following paragraphs.

Since $\mathbf{Q}^*$ is a 4×4 symmetric matrix ($Q_{ij} = Q_{ji}$), it has 4 diagonal entries and $\binom{4}{2} = 6$ independent off-diagonal entries, giving 10 independent entries in total. This follows from the general formula $\frac{n(n+1)}{2}$, where the factor $n+1$ accounts for including the diagonal ($i = j$) in the count, yielding $\frac{4 \times 5}{2} = 10$. The vectorization of a symmetric matrix is defined as $\mathbf{q}^* = \text{vech}(\mathbf{Q}^*) \in \mathbb{R}^{10}$, collecting

the lower-triangular entries column by column. Similarly, $\mathbf{c}_v^* = \mathrm{vech}(\mathbf{C}_v^*) \in \mathbb{R}^6$ with $\mathbf{C}_v^*$ as a 3×3 symmetric matrix. Since Eq. 15 is linear in the entries of $\mathbf{Q}^*$, the vectorized observation $\mathbf{c}_v^*$ can be expressed as a linear function of $\mathbf{q}^*$. Exploiting the symmetry of $\mathbf{Q}^*$, the 6 independent entries of $\mathbf{C}_v^*$ can be assembled into

$$\beta_v \mathbf{c}_v^* = \mathbf{G}_v \mathbf{q}^*, \tag{16}$$

where $\mathbf{G}_v \in \mathbb{R}^{6\times 10}$ is the matrix that encodes the linear mapping from the vectorized dual quadric $\mathbf{q}^* \in \mathbb{R}^{10}$ to the vectorized dual conic $\mathbf{c}_v^* \in \mathbb{R}^6$. Following LfD Rubino et al. (2017), $\mathbf{G}_v$ is assembled from the Kronecker product $\mathbf{P}_v \otimes \mathbf{P}_v$ via duplication matrices that exploit the symmetry of $\mathbf{Q}^*$. Our goal is to recover $\mathbf{q}^* \in \mathbb{R}^{10}$ from the observed conics $\{\mathbf{c}_v^*\}$. Rearranging Eq. 16 as $\mathbf{G}_v\mathbf{q}^* - \mathbf{c}_v^*\beta_v = \mathbf{0}$ and stacking across all V views yields the homogeneous linear system:

$$\underbrace{\begin{bmatrix} \mathbf{G}_1 & -\mathbf{c}_1^* & & & \\ \mathbf{G}_2 & & -\mathbf{c}_2^* & & \\ \vdots & & & \ddots & \\ \mathbf{G}_V & & & & -\mathbf{c}_V^* \end{bmatrix}}_{\mathbf{M} \in \mathbb{R}^{6V\times(10+V)}} \underbrace{\begin{pmatrix} \mathbf{q}^* \\ \beta_1 \\ \vdots \\ \beta_V \end{pmatrix}}_{\mathbf{w}\in\mathbb{R}^{10+V}} = \mathbf{0}, \tag{17}$$

which is the $\mathbf{M}\mathbf{w} = \mathbf{0}$ system. The solution is obtained via SVD as the right singular vector of $\mathbf{M}$ corresponding to its smallest singular value. The first 10 entries of $\mathbf{w}$ give $\mathbf{q}^*$, from which the dual quadric is reconstructed as $\mathbf{Q}^* = \mathrm{vech}^{-1}(\mathbf{q}^*)$.

B.3 Ill-Conditioned Coupled Triangulation & Proposed Structural Decoupling

With the proposed *physical shape prior* ($\mathbf{R}(\boldsymbol{\theta}) = \mathbf{I}_3; s_1 = s_2 = r; s_3 = t_z$) of the main paper, with $\mathbf{t} = (t_x, t_y, t_z)^\top$ and $\mathbf{D} = \mathrm{diag}(r^2, r^2, t_z^2)$, we expand $\mathbf{Q}^*$ step by step:

$$\begin{aligned} \mathbf{Q}^* &= \mathbf{Z}\, \underbrace{\mathrm{diag}(r^2,\, r^2,\, t_z^2,\, -1)}_{\mathbf{D}'}\, \mathbf{Z}^\top, \mathbf{Z} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \\ \mathbf{Z}\mathbf{D}' &= \begin{bmatrix} \mathbf{I}_3 & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0}^\top & -1 \end{bmatrix} = \begin{bmatrix} \mathbf{D} & -\mathbf{t} \\ \mathbf{0}^\top & -1 \end{bmatrix}, \\ \mathbf{Q}^* &= (\mathbf{Z}\mathbf{D}')\mathbf{Z}^\top = \begin{bmatrix} \mathbf{D} & -\mathbf{t} \\ \mathbf{0}^\top & -1 \end{bmatrix} \begin{bmatrix} \mathbf{I}_3 & \mathbf{0} \\ \mathbf{t}^\top & 1 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{D} - \mathbf{t}\mathbf{t}^\top & -\mathbf{t} \\ -\mathbf{t}^\top & -1 \end{bmatrix}, \\ \text{where } \mathbf{D} - \mathbf{t}\mathbf{t}^\top &= \begin{pmatrix} r^2 - t_x^2 & -t_x t_y & -t_x t_z \\ -t_x t_y & r^2 - t_y^2 & -t_y t_z \\ -t_x t_z & -t_y t_z & t_z^2 - t_z^2 \end{pmatrix}. \end{aligned} \tag{18}$$

The last entry vanishes identically under $s_3 = t_z$. This is the ground-tangency condition in algebraic form: substituting $\boldsymbol{\pi} = (0, 0, 1, 0)^\top$, the plane $z = 0$, into $\boldsymbol{\pi}^\top \mathbf{Q}^* \boldsymbol{\pi} = 0$ leaves exactly $\mathbf{Q}^*[3, 3]$, so the vanishing entry states that $z = 0$ belongs to the envelope of tangent planes.

The upper-left block $(\mathbf{D} - \mathbf{t}\mathbf{t}^\top)$ exposes exactly why coupled triangulation fails (as shown in Tab. 7 of the main paper). The block mixes shape and translation in every entry, where the diagonal entries are $(r^2 - t_x^2)$ and $(r^2 - t_y^2)$ with no separate r^2. In an MVPD setup, pedestrians appear in the scene at ranges up to $\sim$40 m with human radius $r \approx 0.3$ m, so t_x^2 and t_y^2 can exceed r^2 by multiple orders of magnitude. The shape (radius) signal is too small and negligible compared to the translation term, making r^2 unrecoverable from the coupled system. Moreover, as shown in prior research Tian et al. (2021), errors in shape and translation corrupt each other in the joint system. Tian et al. (2021) demonstrate this empirically: when translation noise exceeds 15% or bounding-box noise exceeds 2%, the unconstrained joint system Q-SLAM Nicholson et al. (2018) fails to construct valid ellipsoids, while the translation-decoupled method remains stable up to 5$\times$ larger noise levels.

Our proposed *structural decoupling* fixes this by estimating the translation $\mathbf{t}$ independently via DLT center triangulation. Since $\mathbf{Q}^* = \mathbf{Z}\mathrm{diag}(r^2, r^2, t_z^2, -1)\mathbf{Z}^\top$ as shown in Eq. 5 of the main paper, the

projection expands as:

$$\begin{aligned}\mathbf{C}_v^* &= \beta_v \mathbf{P}_v \mathbf{Q}^* \mathbf{P}_v^\top \\ &= \beta_v \mathbf{P}_v \mathbf{Z} \mathrm{diag}(r^2, r^2, t_z^2, -1) \mathbf{Z}^\top \mathbf{P}_v^\top .\end{aligned} \tag{19}$$

Defining $\mathbf{P}'_v = \mathbf{P}_v \mathbf{Z}$ absorbs the translation entirely into the camera, leaving:

$$\mathbf{C}_v^* = \beta_v \mathbf{P}'_v \mathrm{diag}(r^2, r^2, t_z^2, -1) \mathbf{P}'^\top_v . \tag{20}$$

This corresponds to Eq. 8 in the main paper. With the translation absorbed into $\mathbf{P}'_v$, no $\mathbf{t}$ term remains in the system: r^2 and t_z^2 are the only unknowns, and both are recovered directly from the image observations. *Center translation* and *radius estimation* of the proposed structural decoupling are explained in the next sections.

B.4 Confidence-weighted Center Triangulation

In a multi-view setup, not every camera observes the same object with equal reliability. Occlusion or truncation at image boundaries in a given view produce noisy conic center estimates that corrupt the DLT solution for all views. Confidence weighting suppresses these unreliable observations by scaling each view's constraint rows (the pair of linear equations contributed by view v to $\mathbf{A}$) by a learned per-view score before solving, so that high-confidence views dominate the triangulation.

For the center translation step of the structural decoupling, we use the conic centers $\{(c_{v,x}, c_{v,y})\}_{v:\delta_v=1}$ in visible views for quadric center triangulation $\mathbf{X} = (t_x, t_y, t_z, 1)^\top$. Under the pinhole camera model, the conic center is the image of $\mathbf{X}$:

$$(c_{v,x}, c_{v,y}, 1)^\top \sim \mathbf{P}_v \mathbf{X}, \tag{21}$$

where $\sim$ denotes equality up to scale. Since two homogeneous vectors are proportional if and only if their cross product vanishes, Eq. 21 is equivalent to:

$$(c_{v,x},\ c_{v,y},\ 1)^\top \times (\mathbf{P}_v \mathbf{X}) = \mathbf{0} \tag{22}$$

which is exactly Eq. 7 of the main paper. Expanding this cross product eliminates the unknown scale and yields the two scalar DLT constraints used in the next section.

Stacking the cross-product constraints from Eq. 7 across all visible views gives the homogeneous system $\mathbf{A}\mathbf{X} = \mathbf{0}$, where $\mathbf{A} \in \mathbb{R}^{2V \times 4}$. The stacking is performed as:

$$\begin{gathered}\mathbf{P}_v = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 \\ p_5 & p_6 & p_7 & p_8 \\ p_9 & p_{10} & p_{11} & p_{12} \end{bmatrix} = \begin{bmatrix} \mathbf{p}_1^\top \\ \mathbf{p}_2^\top \\ \mathbf{p}_3^\top \end{bmatrix} \\ \underbrace{\begin{bmatrix} c_{1,y}\mathbf{p}_{1,3}^\top - \mathbf{p}_{1,2}^\top \\ \mathbf{p}_{1,1}^\top - c_{1,x}\mathbf{p}_{1,3}^\top \\ \vdots \\ c_{V,y}\mathbf{p}_{V,3}^\top - \mathbf{p}_{V,2}^\top \\ \mathbf{p}_{V,1}^\top - c_{V,x}\mathbf{p}_{V,3}^\top \end{bmatrix}}_{\mathbf{A} \in \mathbb{R}^{2V \times 4}} \underbrace{(t_x\ t_y\ t_z\ 1)}_{\mathbf{X} \in \mathbb{R}^4} = \mathbf{0}.\end{gathered} \tag{23}$$

To suppress unreliable views, both rows of view v are scaled by a predicted confidence $w_v \in (0, 1)$, giving the weighted system:

$$\begin{aligned}\tilde{\mathbf{A}}\mathbf{X} = \mathbf{0}, \qquad \tilde{\mathbf{A}} = \mathbf{W}\mathbf{A}, \\ \mathbf{W} = \mathrm{diag}(w_1, w_1, \ldots, w_V, w_V).\end{aligned} \tag{24}$$

Since $\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} = \mathbf{A}^\top \mathbf{W}^2 \mathbf{A}$, a low-confidence view contributes a near-zero outer product to the normal equations, reducing its influence proportionally to w_v^2. The solution to $\tilde{\mathbf{A}}\mathbf{X} = \mathbf{0}$, computed via SVD, is the right singular vector of $\tilde{\mathbf{A}}$ corresponding to the smallest singular value, dehomogenized so $[\mathbf{X}]_4 = 1$, yielding $(t_x, t_y, t_z)^\top$ directly. w_v is validated in the ablation study in Appendix F.

B.5 Radius Estimation

From Eq. 20, the six independent entries of $\mathbf{C}_v^*$ are each linear in $(r^2, t_z^2, -1)$. Each $\mathbf{C}_v^*$ is normalized to settle the per-view projective scale β_v. Stacking these constraints across all views and confidence-

weighting each view's rows by w_v (as in the center triangulation) yields:

$$\mathbf{M}'\mathbf{w}' = \mathbf{0}, \qquad \mathbf{w}' = (r^2, t_z^2, -1)^\top, \tag{25}$$

where $\mathbf{M}' \in \mathbb{R}^{6|\mathcal{V}|\times 3}$ is assembled by stacking, for each visible view v, six rows built from the columns of $\mathbf{P}'_v$ and the corresponding entries of $\mathbf{C}^*_v$. The solution is obtained via SVD as the right singular vector of $\mathbf{M}'$ corresponding to the smallest singular value. Rescaling so that $[\mathbf{w}']_3 = -1$ directly yields $r = \sqrt{[\mathbf{w}']_1}$. This is in contrast to a generic unconstrained quadric, where the recovered $\mathbf{Q}^*$ is a full 4×4 matrix whose semi-axes and orientation must be additionally disentangled via eigen decomposition. Our physical shape prior reduces the problem to three unknowns.

B.6 Summary of the Two-Step Decoupled Quadric Triangulation

Collecting the results of Appendices. B.4 and B.5, the two triangulation steps replace one ill-conditioned $6V \times (10 + V)$ SVD of $\mathbf{M}\mathbf{w} = \mathbf{0}$ with two well-conditioned, closed-form linear systems:

Step	Homogeneous system	Matrix size
1. Center (DLT)	$\tilde{\mathbf{A}}\,\mathbf{X} = \mathbf{0}$ $\mathbf{X} = (t_x, t_y, t_z, 1)^\top$	$2\|V\| \times 4$
2. Radius	$\mathbf{M}'\,\mathbf{w}' = \mathbf{0}$, $\mathbf{w}' = (r^2, t_z^2, -1)^\top$	$6\|V\| \times 3$

Together they recover all four parameters $\mathbf{q} = (t_x, t_y, t_z, r)$ from multi-view 2D conic observations.

C Datasets

We evaluate our method on three established MVPD benchmarks that together validate complementary aspects of detection performance and generalization, as summarized in Tab. 8. WildTrack and MultiViewX serve as the standard in-domain benchmarks, following the protocol adopted by prior work. GMVD, a large-scale multi-scene dataset, further enables evaluation of cross-scene generalization, directly validating the key capability of our method as well as the framework `PGMVG`.

Tab. 8 summarizes the key properties of the three benchmarks across scene diversity and dataset capacity. The three datasets differ substantially in scale and diversity. WildTrack and MultiViewX each comprise a single short sequence in a fixed camera configuration, with 313 and 350 annotated individuals respectively, while GMVD contains 2,800 individuals across 53 sequences. In terms of crowd density, MultiViewX is the densest at 38.73 individuals per frame. WildTrack averages 23.8 and GMVD spans a wider range of 20.71–36.16 depending on the scene. Camera coverage also varies. WildTrack provides an average of 3.74 cameras per ground-plane location, MultiViewX 4.41, and GMVD 2.76–6.4 across its diverse configurations. Notably, WildTrack and MultiViewX share a nearly identical train/test split of 360/40 frames drawn from the same continuous recording, a setup that can lead to overfitting to

	WildTrack	MultiViewX	GMVD
Scene Diversity			
# Scenes	1	1	7
# Videos	1	1	53
Environment	Real	Synthetic	Synthetic
Setting	Outdoor	Outdoor	Indoor + Outdoor
Camera ConFigure	7, fixed	6, fixed	3–8, varying
Varied Time / Weather	×	×	✓
Dataset Capacity			
Scene Area (m^2)	12×36	16×25	$12-31 \times 20-33$
Train Frames	360	360	4,983
Test Frames	40	40	1,012
# Pedestrians	313	350	2,800
Cam Coverage	3.74	4.41	$2.76-6.4$
Avg. Crowd Density	23.8	38.73	$20.71-36.16$
Min-Max Crowd Density	13–40	33–44	5–58

Table 8: Comparison of MVPD benchmark datasets. Scene diversity assesses variety in environment and configuration while dataset capacity measures scale and coverage. Avg. coverage is the mean number of cameras observing each ground-plane location. Avg. Crowd Density is the mean number of pedestrians on the scene.

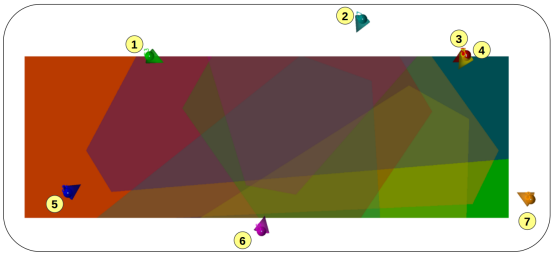


Figure 6: Camera placement and coverage on WildTrack dataset.

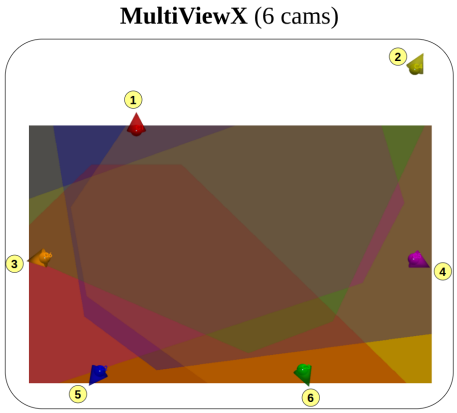


Figure 7: Camera placement and coverage on MultiViewX dataset.

scene-specific visual patterns. In contrast, GMVD's training and test splits are drawn from entirely disjoint scenes, providing a proper assessment of cross-scene generalization. Figs. 6, 7, and 8 depict the camera coverage and camera configurations on all three datasets. All MVPD datasets provide calibrated camera matrices. The details of three datasets are given as follows:

- *WildTrack* Chavdarova et al. (2018) is a real-world outdoor dataset captured with 7 synchronized, calibrated cameras at a resolution of 1920×1080. The monitored area spans $12\text{m} \times 36\text{m}$ and is discretized into a 480×1440 ground-plane grid with 2.5cm spatial resolution. Each frame contains on average around 20 pedestrians, each visible from an average of 3.74 cameras. The dataset comprises 400 frames with the first 360 used for training and the last 40 for testing.
- *MultiViewX* Hou et al. (2020) is a synthetic outdoor counterpart to WildTrack with 6 calibrated cameras. It covers a $16\text{m} \times 25\text{m}$ area discretized into a 640×1000 grid at 2.5cm resolution, and contains a considerably denser crowd of on average around 40 pedestrians per frame, with each individual visible from an average of 4.41 cameras. Following the same protocol as WildTrack, 360 frames are used for training and 40 for testing.
- *GMVD* Vora et al. (2023) is a large-scale synthetic dataset spanning both indoor and outdoor environments, designed explicitly to evaluate the generalizability of MVPD models. It spans 7 distinct scenes and 53 sequences with varying camera counts (3 to 8 per scene), diverse ground-plane dimensions (12–31m $\times$ 20–33m), and camera coverage ranging from 2.76 to 6.4. Temporal and weather conditions (including morning, afternoon, evening, and night lighting as well as sunny, cloudy, rainy, and snowy weather) vary across sequences, preventing models from exploiting scene-specific appearance statistics. The training split contains 4,983 frames across 6 scenes and the test split contains 1,012 frames across the entirely other scene.

D Evaluation Metrics

To evaluate the performance of MVPD, we adopt four detection metrics as other existing methods. The equations for the four metrics are defined as follows:

$$\begin{aligned} MODA &= 1 - \frac{FP + FN}{N}, \\ MODP &= \frac{\sum(1 - dist[dist < thres]/thres]}{TP}, \\ Precision &= \frac{TP}{FP + TP}, \\ Recall &= \frac{TP}{N}. \end{aligned} \tag{26}$$

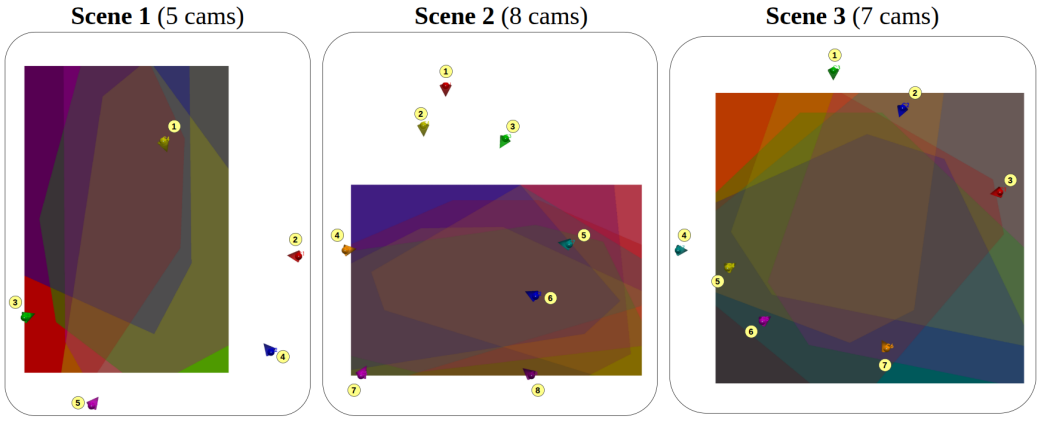


Figure 8: Camera placement and coverage on some scenes of GMVD dataset.

where TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives, N is the number of ground-truth objects. A TP detection is within a fixed distance from the corresponding ground-truth point, which is approximately the radius of a human body. Specifically, $dist$ is the distance from each detection to its corresponding ground-truth, and $thres$ is set to $0.5m$ as a threshold for all benchmarks. The $MODA$ metric measures detection accuracy while $MODP$ quantifies the spatial precision of localized detections.

E Implementation Details

Our model is implemented in PyTorch Paszke et al. (2019) and trained on a single NVIDIA RTX 4090 GPU. We adopt a ResNet-50 He et al. (2016) backbone pre-trained on CrowdHuman (Shao et al., 2018; Chu et al., 2020) via Deformable DETR Zhu et al. (2020), providing strong pedestrian-aware feature initialization. A 4-scale feature pyramid is extracted from the backbone and passed to the decoder. Input images from all camera views are resized to 1280×720 prior to feature extraction. The decoder follows the DETR design comprising 4 layers with an embedding dimension of 256. We initialize 600 object queries, each represented as a joint pair $\mathbf{J} = \{\mathbf{f}, \mathbf{q}\}$, where $\mathbf{f} \in \mathbb{R}^{256}$ is the feature vector and $\mathbf{q} = (t_x, t_y, t_z, r)$ encodes the 3D geometry. The height t_z and radius r are initialized to 0.8m and 0.25m, respectively, following the dimensions of an average standing adult. Query Compression is applied at the last layer, retaining the top $K = \alpha N_q = 120$ queries for final prediction. The model is trained using AdamW with a base learning rate of 10^{-4} for 50 epochs with a batch size of 1. The backbone and FPN are assigned a reduced learning rate of 10^{-5} to preserve the pre-trained pedestrian representations. PGMVG is adaptive to any scene-specific spatial layout without retraining, which is essential for generalizing across the diverse scene geometries of the GMVD benchmark or cross configurations. During training, image augmentations are applied independently to each camera view to simulate per-view variability, including random translation, scale jitter, and color jitter with a factor sampled uniformly from $[0.8, 1.2]$. We also adopt the Drop-View regularization Vora et al. (2023) that randomly discards one camera view per training iteration. For loss weights, we set λ_{ground} as 1, λ_{class} as 0.25, and λ_{image} as 0.5, with concentration on the ground localization.

F Additional Ablation Studies

Ablation study on Encoder-free DETR Framework.

In this section, we provide an additional ablation study of Encoder module defined in our framework as illustrated in Fig. 9. Tab. 9 justifies the encoder-free design of PGMVG. As shown in Fig. 10, the original DETR framework has three blocks. Removing the Transformer encoder and relying solely on ResNet-FPN introduces a negligible drop of 0.1 MODA while reducing parameters from 42.5M to 38.5M and nearly tripling inference speed from 3.2 to 8.6 FPS. The minimal accuracy loss suggests that the encoder's global attention is largely redundant when a strong backbone with FPN

Table 9: Ablation study on Encoder-free framework of PGMVG on MultiViewX.

Encoder	MODA ↑	MODP↑	Params↓	FPS↑
Backbone + Transformer Encoder (baseline)	**96.0**	94.0	42.5	3.2
Backbone + FPN (ours)	95.9	**94.1**	**38.5**	**8.6**



Figure 9: Ablation study on Encoder-free DETR Framework.

already provides rich multi-scale feature representations, consistent with the finding of D^{2}ETR Lin et al. (2022).

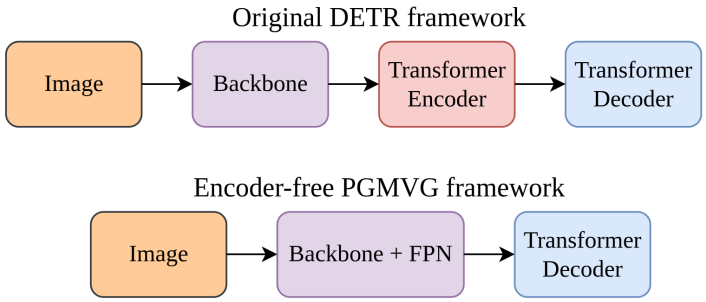


Figure 10: Original DETR design is with three blocks: CNN backbone, Transformer Encoder, and Transformer Decoder (top). PGMVG replaces the Transformer Encoder with backbone + FPN (bottom).

Ablation on Cross-View-Point-Attention.

Table 10: Ablation study on Cross-View-Point-Attention of Multi-view Feature Aggregation module on MultiViewX.

Cross-View-Point-Attention		MODA↑	MODP↑	Pre↑	Rec↑
$\mathbf{h}_p^v$	$\hat{\mathbf{f}}_p^v$				
×	×	93.7	93.2	98.4	93.6
✓	×	94.2	93.4	98.7	94.1
×	✓	95.0	93.8	98.9	94.9
✓	✓	**95.9**	**94.1**	**99.4**	**96.5**

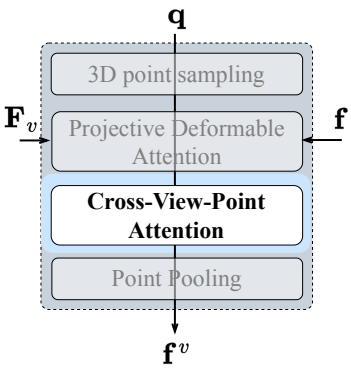


Figure 11: Ablation on Cross-View-Point-Attention component.

In this section, we provide an additional ablation study of Cross-View-Point Attention, a key component of the Multi-view Feature Aggregation module illustrated in Fig. 11. This module is originally presented in Fig. 3 and formally defined in Eq. 10 of the main paper.

Tab. 10 evaluates the two complementary operations in Cross-View-Point Attention: Point Attention $\mathbf{h}_p^v$, which exchanges information among the N_p sampled points within each view, and View Attention $\hat{\mathbf{f}}_p^v$, which aggregates evidence for the corresponding point across camera views. Without either operation, directly pooling the independently sampled features yields the weakest performance, with 93.7 MODA and 93.6 Recall. Introducing point attention $\mathbf{h}_p^v$ improves MODA by 0.5 points by en-

riching each sampling point with full-body context from the same view. View attention $\hat{\mathbf{f}}_p^v$ provides a larger gain of 1.3 MODA, confirming that cross-view interaction is particularly important for compensating for partial occlusion and ambiguous observations in individual cameras. Combining both operations achieves the best performance across all metrics, reaching 95.9 MODA, 94.1 MODP, 99.4 precision, and 96.5 recall. Notably, the combined MODA improvement of 2.2 points exceeds the sum of the individual gains, indicating that the two operations are complementary: Point attention $\mathbf{h}_p^v$ first forms coherent body-level representations within each view, while view attention $\hat{\mathbf{f}}_p^v$ subsequently transfers reliable evidence across views. The substantial 2.9-point recall improvement further demonstrates that this hierarchical interaction effectively recovers pedestrians that are only partially visible in individual viewpoints.

Ablation on Query Aggregator.

Table 11: Ablation study on Query Aggregator of Query Compression module on MultiViewX with $\alpha = 0.2$.

QA	MODA ↑	MODP↑	Pre↑	Rec↑
×	94.3	94.0	98.1	96.2
✓	**95.9**	**94.1**	**99.4**	**96.5**

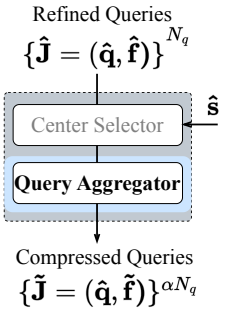


Figure 12: Ablation on Query Aggregator component.

This section provides additional ablation on Query Aggregator(QA), a key component of the Query Compression module illustrated in Fig. 12. This module is originally presented in Fig. 3 and formally defined in Eq.13 of the main paper.

Tab. 11 validates the contribution of QA within the Query Compression module. In our framework, the Center Selector distills the full query population down to K representative centers. Without QA, these K centers carry only local context derived from their own features, limiting their ability to distinguish between nearby pedestrians or resolve ambiguous detections. QA addresses this by cross-attending each selected center to the full query population, allowing it to incorporate global scene context before making a final prediction. Adding QA improves MODA by 1.6, with the most gain in precision (+1.3), confirming that this global context enrichment sharpens classification confidence and reduces false positives caused by locally ambiguous features.

Ablation on View Weights for quadric triangulation and feature fusion.

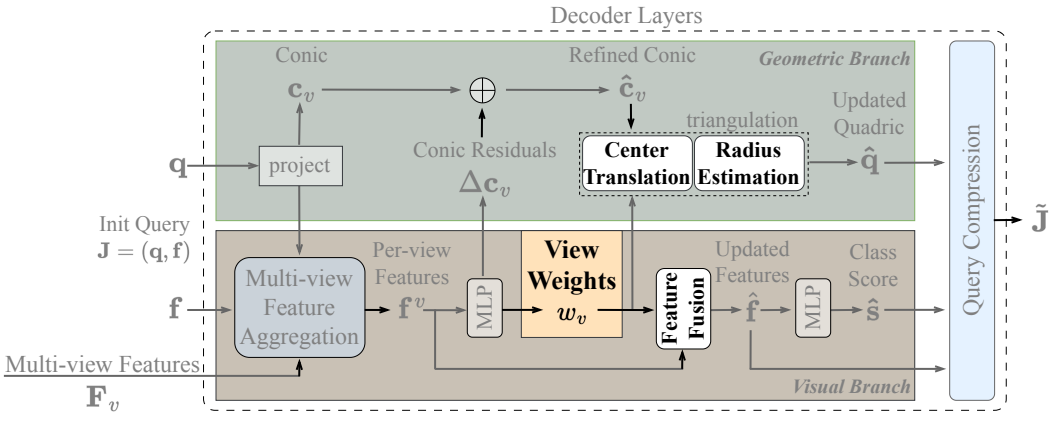


Figure 13: Ablation on View Weights for quadric triangulation and feature fusion.

This section provides an additional ablation study of View Weights for quadric triangulation and feature fusion as illustrated in Fig. 13. This module is originally presented in Fig. 3.

Table 12: Ablation study on the impact of view weights w_v on Geometric and Visual branches on MultiViewX.

view weights w_v	MODA↑	MODP↑	Pre↑	Rec↑
×	91.2	90.4	98.6	93.5
✓(ours)	**95.9**	**94.1**	**99.4**	**96.5**

Tab. 12 validates the role of view confidence w_v in weighted fusion. Without weighting, all views contribute equally to both branches. In the Visual branch, occluded views introduce corrupted appearance features that pollute the cross-view representation and weaken classification confidence. In the Geometric branch, views with heavy occlusion produce inaccurate conic estimates whose centers are displaced from the true pedestrian projection, directly biasing the triangulation and degrading 3D localization. Adding weighted fusion yields substantial gains across all metrics, with MODA improving by 4.7 and MODP by 3.7. The considerable MODP gain reflects more accurate 3D localization when occluded views are down-weighted in triangulation, reducing the influence of unreliable conic estimates on the recovered quadric center. The improvement in recall (+3.0) further confirms that suppressing noisy view contributions in feature fusion produces cleaner cross-view representations, preventing occluded views from corrupting the classification signal. Together, these results demonstrate that view confidence plays a critical role in both branches, and that its dual effect is essential for robust detection and precise localization.

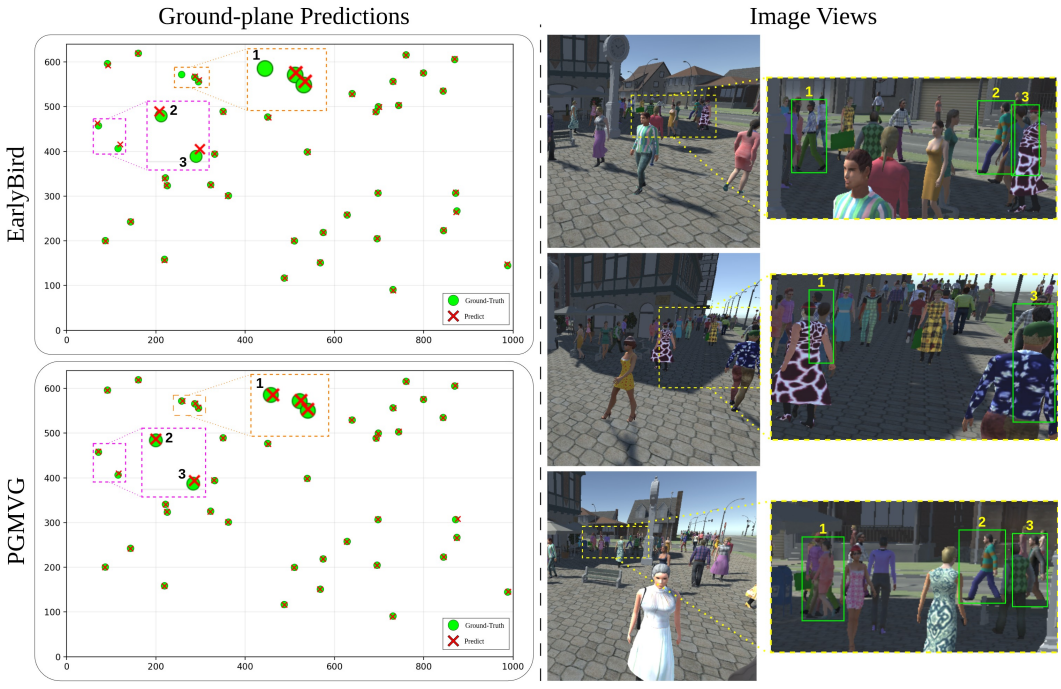


Figure 14: Qualitative comparison between EarlyBird and PGMVG on the dense MultiViewX dataset in the in-domain setting. Left: predicted pedestrian positions (red crosses) overlaid on the ground plane alongside ground-truth positions (green circles). Right: cropped image regions from the corresponding camera views, with ground-truth bounding boxes overlaid on the identified pedestrians. Orange boxes indicate missed detections and pink boxes indicate imprecise localizations. Each case is indexed by a number shared across the ground-plane view and the camera crops to facilitate failure analysis. Camera views are shown for all views in which the pedestrian falls within the field of view, including occluded appearances.

G QUALITATIVE RESULTS

In this section, we illustrate the qualitative comparison between PGMVG and existing method. To further validate our performance and quadric localization in 3D-2D domains, we present qualitative detection results of PGMVG on the testing set across all three benchmarks. For each sample, we show the projected conic overlays on each camera view alongside the recovered 3D quadric layout on the ground plane.

Qualitative Comparison (Fig. 14): We compare PGMVG against EarlyBirdTeepe et al. (2024a) on MultiViewX scene. EarlyBird is chosen due to open source code and available model weights. Overall, the two methods perform adequately in most of the scene regions. However, EarlyBird has cases of missed detection and low-precision localization at the scene border. In region 1 (orange box), EarlyBird suffers from missed detection with pedestrian ID 1. In the corresponding image views that could capture it, ID 1 is only seen by 1 camera and occluded by other pedestrians in the others. This validates the occlusion robustness of PGMVG where object is visible in one view. In region 2 (pink box), EarlyBird misaligns the two pedestrians (ID 2 & 3) that are partially occluded across views. EarlyBird and other BEV-based methods depends on the foot region of each object for localization. EarlyBird's imprecise prediction of ID 3 is because the foot region of ID 3 is occluded by other pedestrians and the bench. PGMVG's full-body multi-view feature extraction with 3D point feature sampling provides the predictions tightly aligned to the ground-truth positions.

WildTrack (Fig. 15): In this real-world outdoor scene with 7 cameras, PGMVG successfully handles the appearance diversity of real pedestrians while maintaining accurate conic alignment across views. The 3D view shows well-separated quadric ellipsoids despite significant occlusion in certain viewpoints, particularly Cams 3 and 6.

MultiViewX (Fig. 16): In this dense synthetic scene, PGMVG produces well-fitted conics across all 6 views under heavy mutual occlusion. The 3D layout confirms that the triangulated quadric centers are spatially consistent and uniformly distributed across the scene, with no duplicate detections in crowded regions.

The GMVD test scenes share the same location but differ in both camera configuration and environmental conditions: a nighttime outdoor setting captured with 8 cameras, and a rainy outdoor setting captured with 6 cameras.

GMVD – Night Scene (Fig. 17): This scene demonstrates robustness under challenging nighttime illumination with 8 cameras. Despite low ambient light, PGMVG reliably detects and localizes over 30 pedestrians, with conics correctly fitted even for partially lit individuals at scene boundaries.

GMVD – Rainy Scene (Fig. 18): Captured under rainy conditions with wet pavement and reduced visual contrast across 6 cameras, PGMVG successfully detects and localizes the dense crowd despite ground reflections and degraded appearance cues, demonstrating robust generalization to adverse weather conditions.



Figure 15: Qualitative results of PGMVG on real-world dataset Wildtrack.

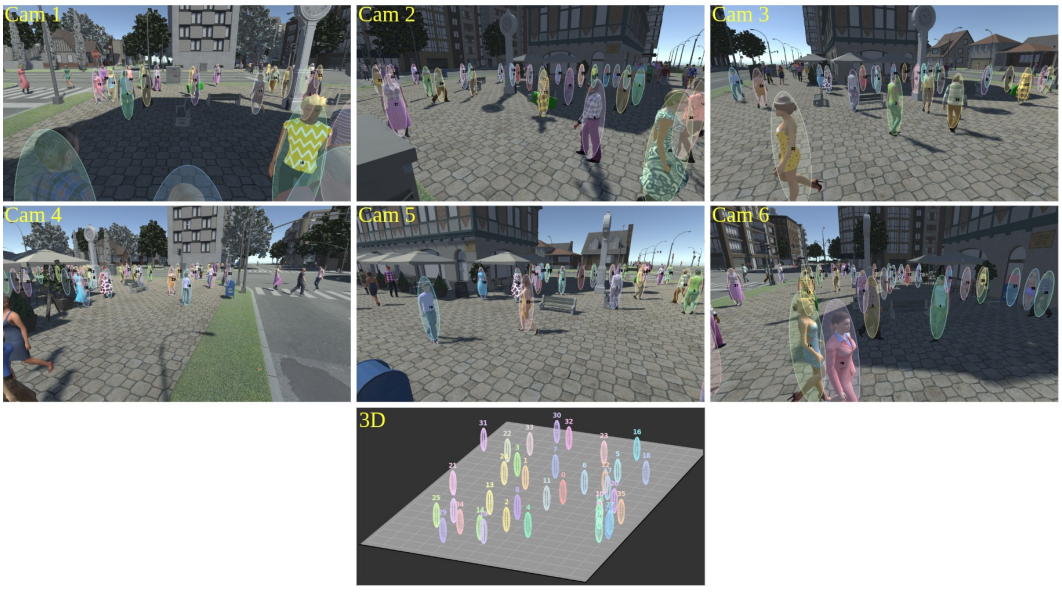


Figure 16: Qualitative results of PGMVG on the dense synthetic dataset MultiViewX.

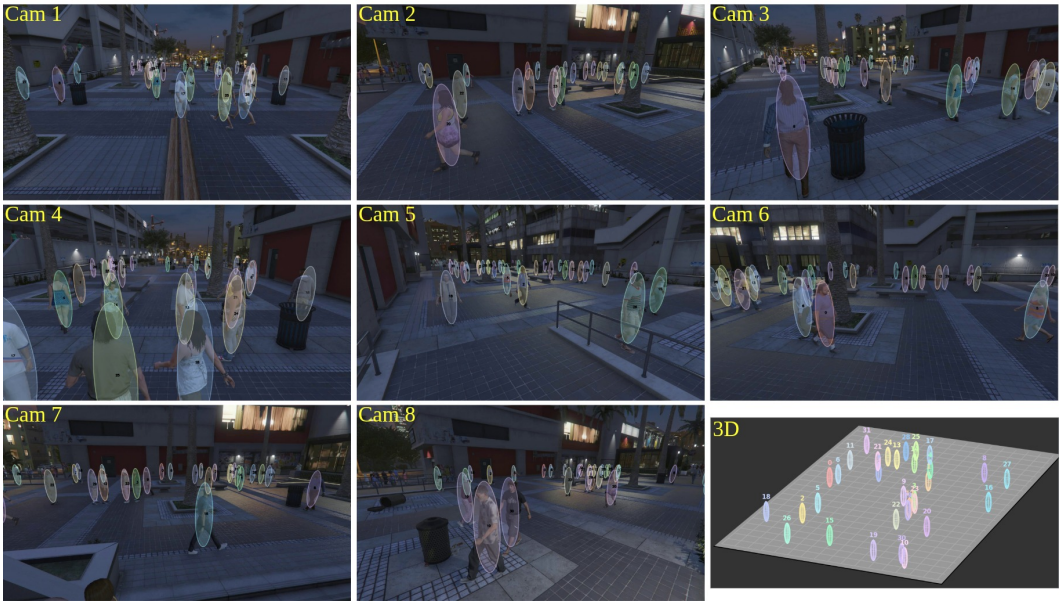


Figure 17: Qualitative results of PGMVG on the nighttime condition of GMVD testing set.

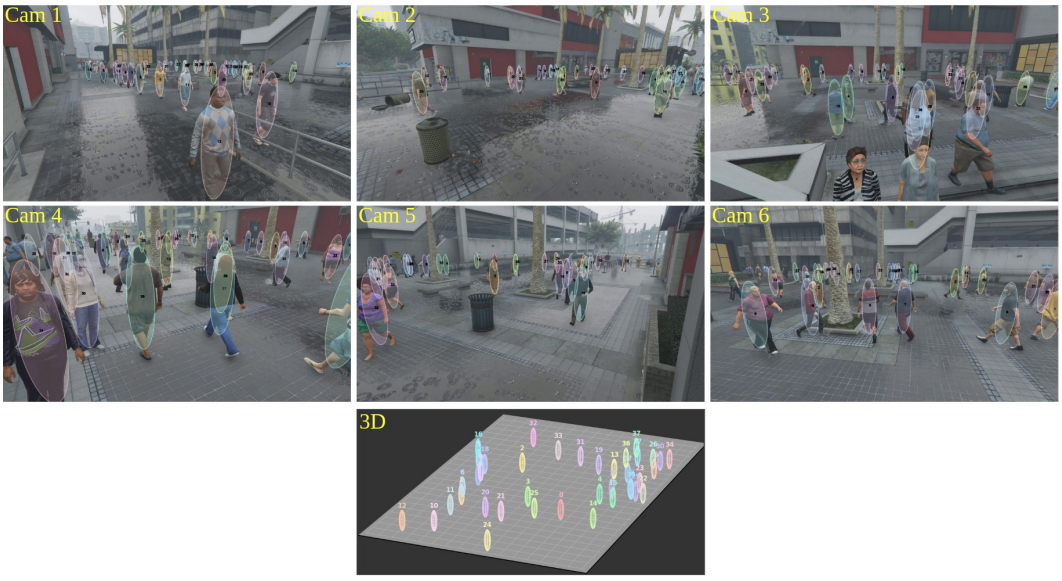


Figure 18: Qualitative results of PGMVG on the rainy condition of GMVD testing set.