Object-Based Camera Pose Estimation from a Single Object Detection and Gravity Vector

Szilard Molnar[1][0009−0002−8295−6702], Zita Amstadt[2][0009−0003−2483−0623], Levente Tamas[1][0000−0002−8583−8296], and Zoltan Kato[2][0000−0002−9328−4254]

[1] Department of Automation, Technical University of Cluj-Napoca, Romania.
<first name>.<last name>@aut.utcluj.ro
[2] Institute of Informatics, University of Szeged, Hungary.
{amstadt,kato}@inf.u-szeged.hu

Abstract. Recent results on pose estimation from ellipsoid-ellipse correspondences, which can be readily obtained from an object detector, allow a direct computation of the camera pose from object-level correspondences. Unfortunately, standard bounding boxes (either horizontal or minimal enclosing boxes) are symmetric, which introduces an inherent ambiguity in the correspondence, yielding multiple or even infinite solutions. Furthermore, the current state of the art requires minimum two such correspondences to provide sufficient constraints for camera rotation. Our contributions make object-based pose estimation efficient in practice: First, a novel object detection method is proposed, called Directional Object Bounding Box (DOBB), which is capable of detecting the object's own direction together with its minimal enclosing box (OBB), yet independently from it, which not only breaks the symmetry of OBBs, but also provides the necessary additional geometric information for our pose estimation method. Second, a novel object-based robust camera pose estimation pipeline is proposed where a minimal solution can be obtained from a single object for outlier filtering when vertical direction and the object orientation w.r.t. that axis are known; followed by a closed-form least squares solution for multiple inlier objects to compute the camera pose. Comparative tests confirm the state-of-the-art performance of the proposed DOBB-based pose estimation method on the standard KITTI360 and 7-Scenes datasets.

Keywords: Object direction · camera pose · object detection.

1 Introduction

Camera pose estimation refers to the fundamental problem of computing the orientation and position of the camera in a world coordinate frame in which the image is captured. Classical approaches rely on direct matching of geometric primitives such as points or lines, as well as object level 2D-3D matches [42, 43]. Recent results on object-based camera pose estimation from an ellipsoid-ellipse correspondence provide an efficient solution for a minimum of two correspondences [49, 10, 8, 11, 9], which can be readily obtained from an object detector.

A related dual problem is object pose estimation, where the 3D position and orientation of an object in the camera coordinate system are estimated based on its 2D image. Although object pose could be used to estimate camera pose, when correspondences are available, these methods [27] are often object specific, while our goal is the abstraction of any detected object (regardless of its object class) into a minimal enclosing ellipsoid in 3D and ellipse in 2D equivalent to the bounding box representation of a general object detector (see Fig. 1) which can be later on used for camera pose estimation in an object-independent unified way.

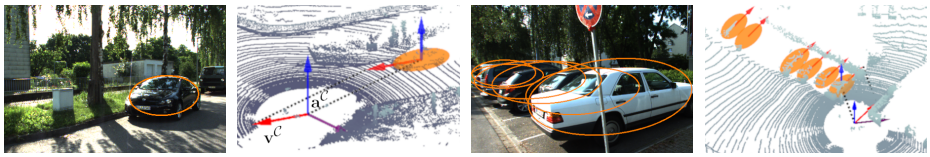

Fig. 1. Detection of parking cars (2D ellipses in images), while also predicting their direction ($\mathbf{v}^{\mathcal{C}}$ as red vector) w.r.t. the vertical axis ($\mathbf{a}^{\mathcal{C}}$ as blue vector) in the 3D camera coordinate frame (we used KITTI360 [24] Lidar data for the 3D visualizations).

State of the art object detectors will predict a *minimal enclosing box*, called oriented bounding box (OBB), for precise localization [15, 18, 19, 29, 45, 46]. Current OBB methods do not capture the object direction like the front of a car, ship, or airplane. In many applications, however, this information is often needed, such as predicting a vehicle's displacement direction from a single image [35], or as initialization for 3D object pose estimation [32], and in particular, as we will discuss it in Section 4, for an efficient object-based camera pose estimation. While many objects have their natural direction, what DOBB detects as direction depends merely on the training data annotations, hence an arbitrary, but consistently labeled direction can be trained and used for pose estimation. Of course, symmetric objects may challenge direction detection: *e.g.* an n-fold symmetry introduces n possible directions of which only one is detected by DOBB. For pose estimation, that can be handled by adding the remaining $n-1$ possible directions and let outlier filtering choose the right one similar to [49].

The core contribution of this paper is thus two-fold: First, we introduce *Directional Object Bounding Box* (DOBB), which simultaneously detects OBBs and a *primary direction* of the detected object. Second, based on DOBB, we derive a minimal solver that gives the camera pose from a single ellipse-ellipsoid pair, which is then used to get a robust camera pose least squares estimate directly from n such correspondences as the roots of a single cubic polynomial.

2 Related work

Camera pose estimation requires feature matches (*e.g.* point correspondences) from which the camera pose is computed [33, 26]. Other features include lines [2,

30, 1, 17], planar regions [7, 4], radiance field-based [38], or object-based [49, 11, 8, 36] correspondences. While descriptors are critical for low level feature correspondences, object level matches are easier to establish as during detection and recognition a rich set of features are already available within the detector network (see [49] for an overview of object-level descriptors and matching). However, unlike keypoints or lines, object-based methods often suffer from direction ambiguities due to the symmetrical shapes of bounding boxes [49] predicted by 2D detectors, where correspondences are then used as an ellipsoid-ellipse pair, which we are investigating in this work for camera pose estimation.

3D coherent ellipses (3DCE) [49] is an ellipse detector particularly designed for camera pose estimation, composed of a traditional object detector based on *Faster R-CNN* [31], followed by an ellipse predictor using a VGG-19 backbone [34]. The ellipse detector uses a 2D embedding to resolve the boundary discontinuity problem. From these ellipses various approaches are used to obtain the camera pose: 1) **P2E** approach [11] requires 2 detected ellipses and the assumption that the camera roll is null; 2) **P3P** approach [21] actually reduces to classical point correspondences using the centroids of 3 ellipsoid-ellipse pairs, which is inherently imprecise as these are *not* corresponding under perspective projection. A third approach mentioned but not used in [49] is **1 ellipse-ellipsoid** approach [8], which estimates only the translation for a known rotation which they obtain from an IMU sensor or vanishing points. Herein, we make use of the same method for translation, but directly obtain the rotation from DOBB.

PGD [40] is a 3D object estimator using monocular images based on the FCOS3D method [39]. This method reduces the 3D object detector to a depth estimation problem, uncertainty modelling, and a graph-based depth propagation. This method assumes that the objects share the same height in the camera coordinates, *e.g.* KITTI dataset [12], which is a widespread assumption in a man-made environment where objects are on a ground plane.

Object detection with *heading detection* aims to predict the frontal direction of detected objects within the 2D image plane. In OHDet [44], the authors encode the BB rotation range of $\pm 90^\circ$ in discrete angles. They use various angle encoding to perform the OBB detection, while an additional module selects one of the four BB edges to be the heading of the object, therefore, the heading is strictly bound to the 4 sides of the OBB, which is useful primarily for remote sensing imagery. HDDet [5] predicts the angle for a regression-based bounding box detector but it only returns discrete angle values on the full (0°, 360°) range, treating heading as classification problem.

The state of the art solution for camera pose estimation from a single ellipsoid-ellipse pair is [9], which provides a theoretical derivation of the infinite set of camera poses and the separation of the rotational and translational components. Solutions exist when at least two object-correspondences are available [11] or the rotation is readily available from other sources like an IMU [8]. Herein, we will show that knowing the vertical direction *e.g.* from a gravity sensor, our DOBB detector provides sufficient constraint to obtain an accurate camera pose from a *single object* correspondence (see Fig. 1), which can be used for efficient outlier

filtering in RANSAC to obtain a robust maximum likleihood camera pose from several object correspondences.

3 DOBB: Directional Bounding Box Detector

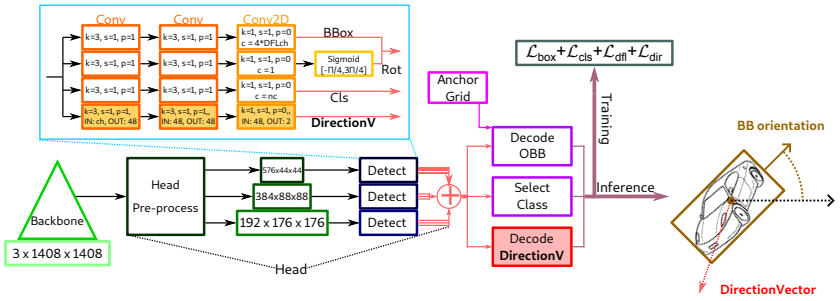

Fig. 2. Architecture overview of the proposed DOBB method, which is based on *CSP-Darknet53* [37] backbone with the extra detector branch for direction vectors.

State of the art object detectors use 5 parameters for OBB representation: orientation angle in the range of $\pm 90°$, center, width, and height, where the angle describes the necessary rotation of the bounding box around its centroid to obtain the minimal enclosing box or equivalently an *oriented ellipse* in the image with axis lengths width and height. For object-based camera pose estimation, DOBB also learns an object's direction in the camera coordinate frame as a rotation around the vertical direction. In a man-made environment, most of the objects are on a ground plane, and their direction can be interpreted as an orientation on this plane using the gravity vector too. This allows obtaining a full camera pose from a single DOBB object detection. Thus, in addition to the OBB parameters, we also predict an extra parameter: the rotation around a known vertical axis within the camera coordinate frame. A natural choice would be an extra angle in the full range $\pm 180°$, but this would make our angle representation susceptible to the boundary discontinuity problem [41, 47], which would hinder the model convergence. Therefore, we propose to represent direction via a *direction vector* $\mathbf{d} = [d_x, d_y]^T$, which uniquely identifies the direction of the object.

Herein, we propose to implement DOBB on top of an optimized version of the *CSPDarknet53* [37] and *feature pyramid networks* (FPN) [25] backbone feature extraction network (see Fig. 2 for the architecture). Starting from the YOLOv8-OBB object detector [20], the Head module has 3 branches at 3 different feature resolutions, and each branch has a *Detect* module, which in itself contains 4 sub-branches for the class label, OBB centroid and size, the OBB orientation angle, and the newly created module, which adds the prediction of the *direction*

vector by estimating its x and y coordinates. The OBB is put on a uniform grid of anchor points covering the input image, while the direction vector is independent of the anchor grid, thus allowing a direction representation independent from the orientation of the enclosing bounding box.

The new model's loss function for OBBs inherits from the basic method, composed of three terms: The *bounding box loss*, $\mathcal{L}_{box}$, is a *probabilistic intersection over union (ProbIoU)* between the ground truth (GT) and the candidate minimal enclosing boxes [28, 48]. The *class loss*, $\mathcal{L}_{cls}$, is a *binary cross entropy* between the ground truth and predicted classes. The *distribution focal loss (DFL)*, $\mathcal{L}_{dfl}$, is based on the *generalized focal loss* [23]. Furthermore, a new *object direction loss*, $\mathcal{L}_{dir}$ is proposed, which is responsible for learning the 3D direction of detected objects. The total loss is a weighted sum of these components, with an appropriate, experimentally tuned weight factor for each component:

$$Loss = \omega_{box}\mathcal{L}_{box} + \omega_{cls}\mathcal{L}_{cls} + \omega_{dfl}\mathcal{L}_{dfl} + \omega_{dir}\mathcal{L}_{dir} \tag{1}$$

where $\omega_{box} = 7.5$, $\omega_{cls} = 0.5$, and $\omega_{dfl} = 1.5$ are OBB gains, while $\omega_{dir} = 20$ is the object direction loss gain set experimentally.

The 3D direction vector $\mathbf{d} = (d_x, d_y)$ is decoupled from the image plane; however, our experiments show that this vector can be encoded as a point on the image related to its OBB centroid (useful if the direction is a significant point of the object). For simplicity, we show tests where the 3D direction is interpreted as a 2D unit vector $\mathbf{v}^{\mathcal{C}}$ within the 3D ground plane perpendicular to the gravity vector $\mathbf{g}^{\mathcal{C}}$ (see Fig. 1 for the visual representation of these vectors). Indeed, the meaning of this direction vector purely depends on the training data and loss, which we will discuss next. Of course, a predicted direction $\mathbf{d}$ could be naturally represented by the coordinates expressed as sin and cos of its angle. However, such an explicit trigonometric representation may lead to a boundary discontinuity problem [41, 47]. Therefore, we let a predicted direction vector $\hat{\mathbf{d}}_i$ be a point whose distance from the head of the GT unit *direction vector* $\mathbf{d}_i$ defines the loss for the detected object number i. However, instead of the classical (isotropic) Euclidean measure, we construct an anisotropic Mahalanobis distance as

$$m(\mathbf{d}_i, \hat{\mathbf{d}}_i) = (\hat{\mathbf{d}}_i - \mathbf{d}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\hat{\mathbf{d}}_i - \mathbf{d}_i) \tag{2}$$

where $\boldsymbol{\Sigma}_i$ is the covariance matrix representing the base of the normal distribution used by the Mahalanobis distance calculation. Setting the longer axis of this ellipse to be aligned with the GT direction allows for larger shifts along this direction, while errors in other directions are quickly increasing this distance. The center of this ellipse is on the head of the GT direction vector, which is a unit lenght away from the origin, while the length of the major semi axis is 0.9, therefore, the ellipse does not include the origin, which ensures that the direction vector can be near the origin but not exactly on it hence providing a well-defined direction. The short axis defines the strictness of the direction estimation, which we choose to be 0.2 in our experiments.

Actually, $\boldsymbol{\Sigma}_i$ can be easily set when it is diagonal (*i.e.* its axes are aligned with the image coordinate system). Then rotating this diagonal covariance matrix by

$\mathbf{R}(\theta_\mathbf{i})$ corresponding to the object's GT direction θ_i gives the covariance matrix aligned with the direction of the object. In fact, we can directly set the inverse of the diagonal covariance matrix (denoted by $\mathbf{S}_i$) as well by simply taking the inverse of the diagonal elements. This way, we can control the actual range of the numerical values used by the loss calculation. The full $\boldsymbol{\Sigma}_i^{-1}$ used in Eq. (2) is then obtained as

$$\boldsymbol{\Sigma}_i^{-1} = \mathbf{R}(\theta_\mathbf{i})\mathbf{S}_i\mathbf{R}(\theta_\mathbf{i})^\top \tag{3}$$

The direction loss is then defined using the probability of the predicted *direction vector* $\hat{\mathbf{d}}_i$ according to a Normal distribution with mean $\mathbf{d}_i$ and covariance $\boldsymbol{\Sigma}_i$:

$$\mathcal{L}_{dir} = 1 - \frac{\sqrt{|\boldsymbol{\Sigma}_i^{-1}|}}{2\pi} \exp\left(-\frac{1}{2}(\hat{\mathbf{d}}_i - \mathbf{d}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\hat{\mathbf{d}}_i - \mathbf{d}_i)\right) \tag{4}$$

Note that neither our direction representation nor the above loss function uses angles explicitly. This is important from the point of view of the boundary discontinuity problem studied extensively in [41, 47], where authors argue that the fundamental issue is directly performing regression on the rotation angle. Our direction vector representation allows the regression of a *location* (pointed by the direction vector) instead of an angle and our direction loss in Eq. (4) is based on a specially constructed Mahalanobis distance of the predicted direction vector, which -by construction- encodes the GT direction of the object used also for camera pose estimation.

4 Object-Based Camera Pose

Camera pose can be obtained up to a single unknown scalar parameter from one ellipsoid-ellipse correspondence using the direct method presented in [8, 10, 11, 9]. Note that this P1E pose estimation method [9] reduces to 1 Degree-of-Freedom (1DoF), which is the solution of a generalized eigenvalue problem. Thus, P1E provides the pose as a function of a scalar unknown. This is a very interesting theoretical result, but in a practical application, a unique solution is needed. Unfortunately, a minimum of two ellipsoid-ellipse pairs are required [10, 11] to obtain a unique solution. Furthermore, it is also shown in [9] that P1E formalism enables to reduce the pose estimation problem to a translation-only or rotation-only estimation problem in which the remaining unknowns can be exactly computed in closed form. When rotation is known, then we get two solutions for the translation (as a solution of a quadratic equation, see [9] for details), from which the one that satisfies the chirality constraint (ellipsoid located in front of the camera) is easily selected [9] – herein we will use this formula to calculate the translation $\hat{\mathbf{t}}$ once the rotation $\hat{\mathbf{R}}$ is obtained using our novel method. When translation is known, then there are minimum 4 solutions for $\mathbf{R}$, but one can end up with infinitely many solutions as well, depending on the ellipsoid symmetries. Therefore the case of known $\mathbf{R}$ has a more significant practical interest, which was already presented in [10, 9, 49], but they rely on either 2 or more object matches or external estimate of the full rotation $\mathbf{R}$ from IMU and

vanishing points and do not make use of the object detection itself to extract $\mathbf{R}$. Therefore, we will show first how to obtain $\mathbf{R}$ from a *single* object detection when the object direction and vertical direction are known. This opens the way to a very efficient robust pose estimation using one object correspondence for outlier filtering and a closed-form least squares (LSE) estimate of $\mathbf{R}$ when more than one object correspondence is available. DOBB detects object OBB within the image as a 2D ellipse along with the object direction in 3D camera coordinate frame, which provides the necessary constraints for our pose estimation method: the 2D ellipse parameters are used to calculate $\hat{\mathbf{t}}$ as in [9] and the direction is used by our novel rotation estimation method outlined next.

4.1 Minimal solution for a single object detection with direction

Using DOBB, we get the direction $\mathbf{v}^{\mathcal{C}}$ of each detected object in the camera coordinate frame $\mathcal{C}$ around the vertical axis represented by the unit vector $\mathbf{a}^{\mathcal{C}}$ obtained *e.g.* from a gravity sensor (see Fig. 1). Assuming the corresponding object direction in the world coordinate frame $\mathcal{W}$ is $\mathbf{v}^{\mathcal{W}}$ around the vertical axis $\mathbf{a}^{\mathcal{W}}$, we can directly construct the rotational component $\mathbf{R} : \mathcal{C} \rightarrow \mathcal{W}$ of the camera pose. First, let us align the vertical axes with the Z axis and the direction vectors with the X axis of the respective coordinate systems via $\mathbf{R}^{\mathcal{C}}$ and $\mathbf{R}^{\mathcal{W}}$:

$$(0,0,1)^\top \equiv (\mathbf{R}^{\mathcal{C}})^\top \mathbf{a}^{\mathcal{C}} \equiv (\mathbf{R}^{\mathcal{W}})^\top \mathbf{a}^{\mathcal{W}} \tag{5}$$

$$(1,0,0)^\top \equiv (\mathbf{R}^{\mathcal{C}})^\top \mathbf{v}^{\mathcal{C}} \equiv (\mathbf{R}^{\mathcal{W}})^\top \mathbf{v}^{\mathcal{W}} \quad \text{where} \tag{6}$$

$$\mathbf{R}^{\mathcal{C}} = \left[\mathbf{v}^{\mathcal{C}}, \frac{\mathbf{v}^{\mathcal{C}} \times \mathbf{a}^{\mathcal{C}}}{\|\mathbf{v}^{\mathcal{C}} \times \mathbf{a}^{\mathcal{C}}\|}, \mathbf{a}^{\mathcal{C}}\right] \tag{7}$$

$$\mathbf{R}^{\mathcal{W}} = \left[\mathbf{v}^{\mathcal{W}}, \frac{\mathbf{v}^{\mathcal{W}} \times \mathbf{a}^{\mathcal{W}}}{\|\mathbf{v}^{\mathcal{W}} \times \mathbf{a}^{\mathcal{W}}\|}, \mathbf{a}^{\mathcal{W}}\right] \tag{8}$$

The rotational component $\mathbf{R} : \mathcal{C} \rightarrow \mathcal{W}$ is then obtained as

$$\mathbf{R} = \mathbf{R}^{\mathcal{W}} (\mathbf{R}^{\mathcal{C}})^\top \tag{9}$$

Knowing $\mathbf{R}$, we can easily compute translation $\mathbf{t}$ using the closed form solution of [9]. In our experiments, we have used the object annotations in the KITTI360 Lidar data as the object's ellipsoid in the world coordinate frame $\mathcal{W}$, but in a real life application one can reconstruct an object ellipsoid in $\mathcal{W}$ along with its direction $\mathbf{v}^{\mathcal{W}}$ from DOBB detections in 3 views [49, 16].

4.2 Robust solution for n object detections

Since we can obtain $\mathbf{R}$ from a single DOBB object, it allows for a very efficient outlier filtering: as opposed to the usual iterative RANSAC [6] solution, we can simply obtain a putative pose for each object using Eq. (9) and filter the outliers using the following equations:

$$\mathbf{R}\mathbf{v}_i^{\mathcal{C}} \times \mathbf{v}_i^{\mathcal{W}} = \mathbf{0}, \;\; i = 1, \ldots, n \tag{10}$$

which should be satisfied up to a small error threshold $\tau = sin^2(\theta_{max})$ ($\theta_{max} = 5°$ in our experiments) for the inliers. Then we take the largest inlier set as the input for our direct LSE solver.

Without loss of generality, we can assume, that the Z axis is the vertical axis, *i.e.* $\mathbf{a}^{\mathcal{C}} = \mathbf{a}^{\mathcal{W}} = (0,0,1)^\top$ like in Eq. (5), which also implies that the Z coordinate of the unit direction vectors $\mathbf{v}_i^{\mathcal{C}}$ and $\mathbf{v}_i^{\mathcal{W}}$ are 0 and the vectorial equation in Eq. (10) simplifies to a single equation as the rotations around X and Y are 0 in $\mathbf{R}$. Hence, only the remaining rotation $\mathbf{R}_Z$ around Z has to be calculated. In order to eliminate $\sin(\alpha)$ and $\cos(\alpha)$ in $\mathbf{R}_Z$, we can use the substitution $q = \tan(\alpha/2)$ [22, 3, 17], for which $\cos(\alpha) = (1-q^2)/(1+q^2)$ and $\sin(\alpha) = 2q/(1+q^2)$. Substituting $\mathbf{R}_Z$ into Eq. (10) and multiplying both sides by $(1+q^2)$, we get the following quadratic polynomial equation in the single unknown q and coefficients a_i for a particular object direction pair $(\mathbf{v}^{\mathcal{C}}, \mathbf{v}^{\mathcal{W}})$:

$$\sum_{i=0}^{2} a_i q^i = \left(v_1^{\mathcal{C}} v_2^{\mathcal{W}} - v_2^{\mathcal{C}} v_1^{\mathcal{W}}\right) q^2 + \left(2\, v_1^{\mathcal{C}} v_1^{\mathcal{W}} + 2\, v_2^{\mathcal{C}} v_2^{\mathcal{W}}\right) q - v_1^{\mathcal{C}} v_2^{\mathcal{W}} + v_2^{\mathcal{C}} v_1^{\mathcal{W}} = 0 \quad (11)$$

Having $n \geq 2$ equations of the above form yields an overdetermined polynomial system of equations in q which can be solved in the least squares sense, *i.e.* we take the sum of the squared errors of the equations which becomes a quartic function in q:

$$\sum_{k=1}^{n} \left(a_{2,k}^2 q^4 + 2a_{2,k}a_{1,k}q^3 + (2a_{2,k}a_{0,k} + a_{1,k}^2)q^2 + 2a_{1,k}a_{0,k}q + a_{0,k}^2\right), \quad (12)$$

whose minima are the roots of its derivative

$$\sum_{k=1}^{n} \left(4\, a_{2,k}^2 q^3 + 6\, a_{1,k}\, a_{2,k}\, q^2 + \left(4\, a_{0,k}\, a_{2,k} + 2\, a_{1,k}^2\right) q + 2\, a_{0,k}\, a_{1,k}\right) = 0 \quad (13)$$

where $a_{i,k}$ denotes the coefficients a_i from Eq. (11) of the k^{th} equation. There are maximum 3 roots of this cubic polynomial that can be calculated in a closed form. At least one of them is real, which provides the least squares solution for q from which we get $\mathbf{R}_Z$. In case of more than one real solution, we simply select the geometrically correct one satisfying the chirality constraint. Once the LSE estimate of the rotation $\hat{\mathbf{R}}$ is obtained, we can get a robust estimate of the translation $\hat{\mathbf{t}}$ using the closed form solution of [9] and choose the one with the smallest average backprojection error which is calculated as the ProbIoU error between the backprojected 3D ellipsoid and its corresponding 2D ellipse detected by DOBB. Let us emphasize that the least squares solution minimizes the *geometric error* of the solution as Eq. (10) (hence Eq. (11)) expresses the sin of the angle between the directions $\mathbf{v}_i^{\mathcal{C}}$ and $\mathbf{v}_i^{\mathcal{W}}$.

5 Experiments

Training of DOBB and quantitative evaluations are done on two standard benchmark datasets:

KITTI360 [24] is a benchmark dataset of 2D images and 3D point clouds for autonomous driving containing 3D oriented annotations of typical roadside objects like *cars*, *buildings*, *trucks*, etc. In KITTI360, 3D object annotations are given via a 3D transformation (rotation, translation, scaling) from the object local coordinate system $\mathcal{O}$ into the world $\mathcal{W}$, such that the X axis in $\mathcal{O}$ is the object direction [24]. Thus training data is easily generated as $\mathbf{v}^{\mathcal{W}} = \mathbf{R}_{\mathcal{O}\to\mathcal{W}}(1,0,0)^\top$. In $\mathcal{W}$, the Z axis is the vertical direction (and it is aligned physically with the camera's vertical direction [24]) with negligible error, so in $\mathbf{v}^{\mathcal{W}}$ we can safely set the Z coordinate to 0 and make it unit length. Assuming that the vertical axis $\mathbf{a}^{\mathcal{C}}$ is also Z in the camera $\mathcal{C}$ coordinate system, $\mathbf{v}^{\mathcal{C}} = (v_1^{\mathcal{C}}, v_2^{\mathcal{C}}, 0)^\top = \mathbf{R}_Z\mathbf{v}^{\mathcal{W}}$ with $\mathbf{R}_Z$ beeing the ground truth pose rotation around the Z (vertical) axis (see Section 4.1).

7-Scenes dataset [14] is an indoor dataset, where the camera rotation around each axis is significant, hence for the default DOBB detector, we rectified the images to a single-axis rotation estimation by a homography transformation of the images and the annotations. This rectifying homography is easily obtained by letting the vertical direction upward along the $-Y$ camera axis: $\mathbf{H} = \mathbf{K}\mathbf{R}_Y(\phi_Y)\mathbf{R}_X(\phi_X)\mathbf{K}^{-1}$, where $\mathbf{K}$ is the camera calibration matrix and the decomposition of camera rotation matrix $\mathbf{R} = \mathbf{R}_Z(\phi_Z)\mathbf{R}_Y(\phi_Y)\mathbf{R}_X(\phi_X)\mathbf{R}_X(-90°)$.

The DOBB prediction provides ellipses in the rectified image plane, which means that the resulting conic matrix $\mathbf{C}'$ corresponds to the image in the transformed coordinate frame. To retrieve the conic $\mathbf{C}$ in the original image -which is needed for the pose estimation-, we apply the inverse of the homography transformation: $\mathbf{C} = \mathbf{H}^T\mathbf{C}'\mathbf{H}$.

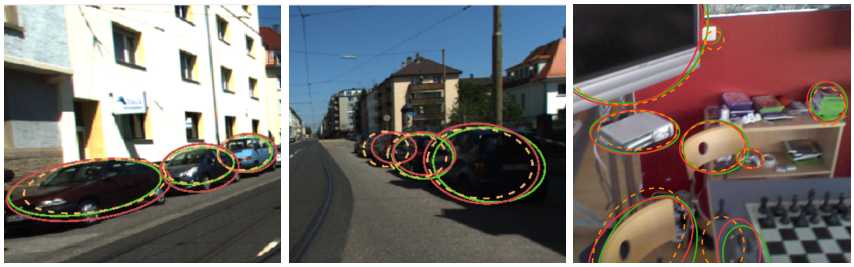

Fig. 3. Example of the GT projected ellipse (Green), the detection of DOBB (Red), and the reprojected ellipsoid using the estimated camera pose (Orange). Columns 1-2: KITTI360 dataset [24], last column: 7-Scenes dataset [14]

5.1 Indoor Camera Pose Estimation

Using the 7-Scenes indoor dataset [14], a quantitative comparison is done with the 3DCE-P3P method [49]. The sequence contained a total of 3000 images on which our minimal solver obtained 2758 absolute camera pose estimates,

the robust LSE estimator (denoted by DOBB-LSE) successfully calculated the pose for 2721 frames, while 3DCE-P3P [49] managed to compute the pose of the camera in 2601 frames only. Considering only the subset of images where all methods successfully calculated a pose, the average rotation and translation errors are 2.54° and 34 cm for our minimal solver, 1.97° and 12 cm for DOBB-LSE, and 4.78° and 12 cm for 3DCE-P3P. The average relative rotation error between consecutive frames is lower, at 0.84°. This comparison is shown in more detail in Fig. 4.

5.2 Outdoor Camera Pose Estimation

Comparative results presented in Fig. 4 confirm the state-of-the-art performance of the DOBB object-based camera pose estimation method. As a baseline, we also show results with recent deep line detector-based methods L2D2 [2] and SOLD2 [30]; and the point-based SiLK [13], which significantly outperforms all the other methods.

For a fair comparison in Fig. 4, we only show results on 112 images (from 749) where *all* methods obtained sufficiently many correspondences to compute a robust pose estimate. SiLK demonstrates the lowest mean rotation error of 0.25° and a median of 0.15°, followed by DOBB with 5.06° and a median of 2.92°. SOLD2 and L2D2 follow with mean errors of 30.84° and 46.76°, and medians of 1.67° and 2.56°. In translation estimation, SiLK again achieves the lowest errors, with a mean of 0.18 m and a median of 0.1 m. DOBB attains the next best performance with a mean error of 2.81 m and a median of 1.25 m, while L2D2 and SOLD have higher mean errors of 33.9 m and 88.72 m, and medians of 2.35 m and 0.72 m. Point- or line-based methods need 3 correspondences, while only 1 correspondence is sufficient in our method for a minimal solution. Although SiLK provides more accurate pose estimates, in complex real-life applications object detection may be needed anyway and -thanks to DOBB- it can also provide accurate pose estimates *without* running a keypoint detector.

DOBB can also be used for relative rotation estimation between cameras without any 3D ellipsoid correspondence (only the detector direction predictions are needed). This is useful for reliable direction estimation during navigation. A translation can be obtained either from other correspondences, sensors, or a GPS. For the evaluation on a sequence of about 100 frames, the method achieves a mean rotation error of 1.43° and a median of 1.21°.

Finally, we conducted a systematic comparison with object-based pose estimation methods. For this purpose, a total of 805 object instances with confirmed detections from each method in the selected image set were used for evaluating the minimal pose solution. The results, summarized in terms of rotation error, show that DOBB achieves the lowest mean error of 10.54°, followed by FCOS3D with 13.7°, and PGD with 17.75°. In terms of median error, DOBB achieves 3.45°, while FCOS3D and PGD obtain 7.66° and 6.7°, respectively. Similarly, in translation estimation, DOBB attains the lowest errors, with a mean of 3.35 m and a median of 1.63 m. PGD and FCOS3D have mean errors of 5.89 m and 13.67 m, and medians of 3.73 m and 12.4 m.

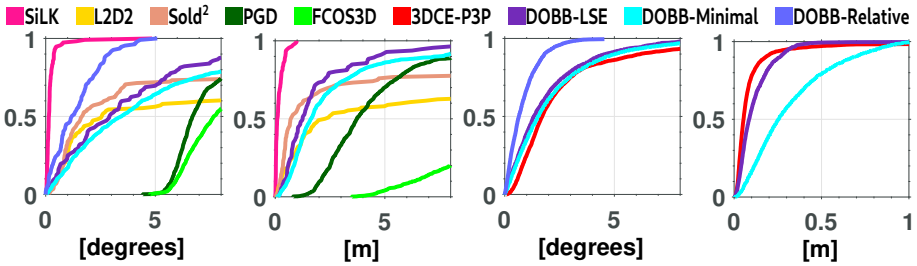

Fig. 4. Rotation and translation errors in the pose estimation comparing our methods (LSE, Minimal, Relative) with the following: SiLK [13], L2D2 [2], SOLD2 [30], PGD [40], FCOS3D method [39], 3DCE method [49] (using the P3P approach). Columns 1-2: KITTI360 dataset [24], Columns 3-4: 7-Scenes dataset [14]. The plots represent the *Cumulative Distribution Function (CDF)* of the results.

6 Conclusions

A novel approach is proposed for object-based camera pose estimation. Objects are reduced to an ellipse, thus providing a unified, object-independent representation for establishing ellipsoid-ellipse correspondences. We have shown that knowing the vertical direction and training an object detector for predicting the object direction in 3D space around this vertical axis, we can obtain the camera pose from a single object correspondence - which was not possible with traditional object detectors. Furthermore, we have also proposed a least squares formulation which, together with the minimal solver, can be used for robust pose estimation. A novel directional object detector, DOBB is proposed where not only a minimal enclosing bounding box is returned but also a primary direction specific to the detected object is predicted. Experiments confirmed that our method delivers state-of-the-art bounding box detection, which can be successfully used for pose estimation on real data with accuracy comparable to state of the art point and line-based methods. Code and test data will be made publicly available for reproducibility.

Acknowledgments. This work was supported by Romanian National Authority for Scientific Research, project nr. PN-IV-P7-7.1-PTE-2024-0105; by the ATLAS project funded by the EU CHIST-ERA programme (CHIST-ERA-23-MultiGIS-02) and the Hungarian National Research, Development and Innovation Fund under grants 2024-1.2.2-ERA-NET-2025-00020, TKP2021-NVA-09, and K135728 and HAS Domus.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdellali, H., Frohlich, R., Kato, Z.: Robust absolute and relative pose estimation of a central camera system from 2D-3D line correspondences. In: Proc. of ICCV

Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving. IEEE, Seoul, Korea (Oct 2019)
2. Abdellali, H., Frohlich, R., Vilagos, V., Kato, Z.: L2D2: Learnable line detector and descriptor. In: Proc. of International Conference on 3D Vision. pp. 442–452. IEEE, London, United Kingdom (Dec 2021)
3. Albl, C., Kukelova, Z., Pajdla, T.: Rolling shutter absolute pose problem with known vertical direction. In: Proc. of Conference on Computer Vision and Pattern Recognition. pp. 3355–3363. Las Vegas, NV, USA (Jun 2016)
4. Csurka, G., Kato, Z., Juhasz, A., Humenberger, M.: Estimating low-rank region likelihood maps. In: Proc. of International Conference on Computer Vision and Pattern Recognition. pp. 1–10. IEEE, Seattle, Washington, USA (Jun 2020)
5. Ding, S., Liu, J., Yang, F., Xu, M.: HDDet: A More Common Heading Direction Detector for Remote Sensing and Arbitrary Viewing Angle Images. Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
7. Frohlich, R., Tamas, L., Kato, Z.: Absolute pose estimation of central cameras using planar regions. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(2), 377–391 (Feb 2021)
8. Gaudilliere, V., Simon, G., Berger, M.O.: Camera relocalization with ellipsoidal abstraction of objects. In: International Symposium on Mixed and Augmented Reality. IEEE, Beijing, China (2019)
9. Gaudilliere, V., Simon, G., Berger, M.O.: Perspective-1-ellipsoid: Formulation, analysis and solutions of the camera pose estimation problem from one ellipse-ellipsoid correspondence. International Journal of Computer Vision **131**, 2446–2470 (2023)
10. Gaudilliere, V., Simon, G., Berger, M.O.: Camera pose estimation with semantic 3D model. In: International Conference on Intelligent Robots and Systems. IEEE, Macau SAR China (Nov 2019)
11. Gaudilliere, V., Simon, G., Berger, M.O.: Perspective-2-ellipsoid: Bridging the gap between object detections and 6-DoF camera pose. IEEE Robotics and Automation Letters pp. 5189–5196 (2020)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
13. Gleize, P., Wang, W., Feiszli, M.: Silk: Simple learned keypoints. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22442–22451 (2023)
14. Glocker, B., Izadi, S., Shotton, J., Criminisi, A.: Real-Time RGB-D Camera Relocalization. In: International Symposium on Mixed and Augmented Reality (ISMAR). IEEE (October 2013)
15. Han, J., Ding, J., Li, J., Xia, G.: Align Deep Features for Oriented Object Detection. Transactions on Geoscience and Remote Sensing **60**, 1–11 (2022)
16. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, UK (2004)
17. Horanyi, N., Kato, Z.: Generalized pose estimation from line correspondences with known vertical direction. In: Proc. of International Conference on 3D Vision. pp. 1–10. IEEE, Qingdao, China (Oct 2017)
18. Hou, L., Lu, K., Yang, X., Li, Y., Xue, J.: G-Rep: Gaussian Representation for Arbitrary-Oriented Object Detection. Remote Sensing **15**(3), 757 (2023)

19. Huang, Z., Li, W., Xia, X.G., Tao, R.: A General Gaussian Heatmap Label Assignment for Arbitrary-Oriented Object Detection. Transactions on Image Processing **31**, 1895–1910 (2022)
20. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics (2023)
21. Kneip, L., Scaramuzza, D., Siegwart, R.: A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In: CVPR 2011. pp. 2969–2976 (2011)
22. Kukelova, Z., Bujnak, M., Pajdla, T.: Closed-form solutions to minimal absolute pose problems with known vertical direction. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) Proc. of Asian Conference on Computer Vision, Part II. LNCS, vol. 6493, pp. 216–229. Springer, Queenstown, New Zealand (Nov 2010)
23. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In: 34th International Conference on Neural Information Processing Systems. pp. 21002–21012. No. 1763 in NIPS'20, Vancouver, BC, Canada (Dec 2020)
24. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. Transactions Pattern Analysis and Machine Intelligence (PAMI) **45**(3), 3292–3310 (2022)
25. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 936–944. IEEE Computer Society (2017)
26. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17627–17638 (October 2023)
27. Liu, J., Sun, W., Yang, H., Zeng, Z., Liu, C., Zheng, J., Liu, X., Rahmani, H., Sebe, N., Mian, A.: Deep learning-based object pose estimation: A comprehensive survey. CoRR **abs/2405.07801** (2024)
28. Murrugarra-Llerena, J., Kirsten, L.N., Zeni, L.F., Jung, C.R.: Probabilistic Intersection-Over-Union for Training and Evaluation of Oriented Object Detectors. Transactions on Image Processing **33**, 671–681 (2024)
29. Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C., Xu, C.: Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In: Conference on Computer Vision and Pattern Recognition. pp. 11204–11213. Seattle, WA, USA (Jun 2020)
30. Pautrat, R., Juan-Ting, L., Larsson, V., Oswald, M.R., Pollefeys, M.: SOLD2: Self-supervised occlusion-aware line description and detection. In: Proc. of Conference on Computer Vision and Pattern Recognition (2021)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28, pp. 91–99. Curran Associates, Inc. (2015)
32. Shan, M., Feng, Q., Jau, Y., Atanasov, N.: ELLIPSDF: Joint Object Pose and Shape Optimization with a Bi-level Ellipsoid and Signed Distance Function Description. In: International Conference on Computer Vision. pp. 5926–5935. Montreal, QC, Canada (Oct 2021)
33. Shavit, Y., Ferens, R., Keller, Y.: Learning single and multi-scene camera pose regression with transformer encoders. Computer Vision and Image Understanding **243**, 103982 (2024)

34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
35. Suttiponpisarn, P., Charnsripinyo, C., Usanavasin, S., Nakahara, H.: Detection of Wrong Direction Vehicles on Two-Way Traffic. In: 13th International Conference on Knowledge and Systems Engineering. pp. 1–6. Bangkok, Thailand (Dec 2021)
36. Vavra, V., Sattler, T., Kukelova, Z.: Camera pose estimation from bounding boxes. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5535–5542 (2024)
37. Wang, C., Liao, H.M., Wu, Y., Chen, P., Hsieh, J., Yeh, I.: CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In: Conference on Computer Vision and Pattern Recognition Workshops. pp. 1571–1580. Seattle, WA, USA (Jun 2020)
38. Wang, J., Zhou, K., Markham, A., Trigoni, N.: WSCLoc: Weakly-supervised sparse-view camera relocalization via radiance field. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 8414–8420 (2024)
39. Wang, T., Zhu, X., Pang, J., Lin, D.: FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection . In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 913–922. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2021)
40. Wang, T., Zhu, X., Pang, J., Lin, D.: Probabilistic and Geometric Depth: Detecting objects in perspective. In: Conference on Robot Learning (CoRL) 2021 (2021)
41. Xu, H., Liu, X., Xu, H., Ma, Y., Zhu, Z., Yan, C., Dai, F.: Rethinking boundary discontinuity problem for oriented object detection. In: Conference on Computer Vision and Pattern Recognition. pp. 17406–17415. Seattle, WA, USA (Jun 2024)
42. Xu, M., Wang, Y., Xu, B., Zhang, J., Ren, J., Huang, Z., Poslad, S., Xu, P.: A critical analysis of image-based camera pose estimation techniques. Neurocomputing **570**, 127125 (2024)
43. Xu, M., Wang, Y., Xu, B., Zhang, J., Ren, J., Huang, Z., Poslad, S., Xu, P.: A critical analysis of image-based camera pose estimation techniques. Neurocomputing **570** (feb 2024)
44. Yang, X., Yan, J.: On the arbitrary-oriented object detection: Classification based approaches revisited. International Journal of Computer Vision **130**(5), 1340–1365 (2022)
45. Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., Tian, Q.: Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. In: Meila, M., Zhang, T. (eds.) 38th International Conference on Machine Learning. vol. 139, pp. 11830–11841. Virtual Event (Jul 2021)
46. Yi, J., Wu, P., Liu, B., Huang, Q., Qu, H., Metaxas, D.N.: Oriented Object Detection in Aerial Images with Box Boundary-Aware Vectors. In: Winter Conference on Applications of Computer Vision. pp. 2149–2158. Waikoloa, HI, USA (Jan 2021)
47. Yu, Y., Da, F.: On Boundary Discontinuity in Angle Regression Based Arbitrary Oriented Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(10), 6494–6508 (2024)
48. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In: AAAI Conference on Artificial Intelligence. pp. 12993–13000. New York, NY, USA (Feb 2020)
49. Zins, M., Simon, G., Berger, M.O.: Object-Based Visual Camera Pose Estimation From Ellipsoidal Model and 3D-Aware Ellipse Prediction. International Journal of Computer Vision **130**(4), 1107–1126 (2022)