X3CALIB: SINGLE IMAGE LIDAR-CAMERA CALIBRATION AS CROSS-MODAL 3D REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Extrinsic calibration between a LiDAR and a camera is a prerequisite for sensor fusion, yet learned calibrators must match 3D geometry against 2D appearance, and most are validated only on the dataset they were trained on. We reformulate calibration from a single LiDAR scan and a single image as 3D-3D registration. A monocular metric depth network lifts the image into a metric point cloud, so the LiDAR scan and the image become two point sets in one geometric domain. An overlap-aware correspondence network, trained only on KITTI, matches the two clouds, and iterated least squares with a coarse-to-fine inlier threshold turns the matches into extrinsics, with either a PnP or a 3D-3D solver. Zero-shot on nuScenes, Argoverse, and PandaSet, X3Calib outperforms LCCNet and CalibNet, neither of which even reaches the recall of simply returning the identity transform. On nuScenes, we also evaluate the recent CMRNext, which was trained on the other three datasets: X3Calib exceeds its recall by a wide margin, and CMRNext's recall again falls below that of the identity transform. A simple pipeline built on 3D registration thus generalizes better than dedicated calibration networks, even one trained on more data. The code will be made public upon acceptance.

1 INTRODUCTION

Extrinsic calibration between a LiDAR and a camera underpins multimodal perception in autonomous driving and robotics (Zhang et al., 2025a). LiDAR-camera 3D detectors (Vora et al., 2020; Bai et al., 2022; Liu et al., 2023) and LiDAR-visual odometry and mapping (Zhang & Singh, 2015; Lin & Zhang, 2022) all consume the rigid transform between the two sensor frames, and errors in it degrade their output: perturbing the calibration matrices markedly lowers the accuracy of fusion-based 3D detectors (Dong et al., 2023). Mounts shift in use, so calibration must be repeated in the field (Levinson & Thrun, 2013), without targets and, ideally, from a single frame.

Targetless calibration has moved from hand-designed cues, such as edges (Levinson & Thrun, 2013) and mutual information (Pandey et al., 2015), to learned regressors such as RegNet, CalibNet, and LCCNet (Schneider et al., 2017; Iyer et al., 2018; Lv et al., 2021). These networks share a structural difficulty: they compare 3D LiDAR geometry with 2D image appearance. The network must learn perspective projection, resolve depth ambiguity, and bridge the gap between sparse geometry and dense photometry at once. Their evaluation shares a second weakness. Nearly all are trained and tested on KITTI, so how a learned correction behaves on a different sensor rig is rarely measured.

We take a different route and calibrate from a single frame: one LiDAR scan and one image, with no motion, sequence, or target. Monocular metric depth estimation (Hu et al., 2024; Yang et al., 2024a; Piccinelli et al., 2024) now recovers absolute depth from a single image well enough to serve as a geometric proxy for the camera. Lifting the image into a metric point cloud turns calibration into the registration of two 3D point sets, a problem with mature solutions (Huang et al., 2021; Qin et al., 2022). We fine-tune an overlap-aware registration network, PREDATOR (Huang et al., 2021), on pairs of LiDAR and depth-lifted point clouds from KITTI, without changing its architecture or losses. It yields cross-modal correspondences, and iterated least squares turns them into the extrinsic transform (Fig. 1): we keep the correspondences that agree with the current pose under a threshold that shrinks from coarse to fine, and refit. The refit can use a PnP solver, which uses pixel coordinates and intrinsics, or a 3D-3D solver, which uses the lifted depth of every matched pixel as a direct measurement of the scene geometry.

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.

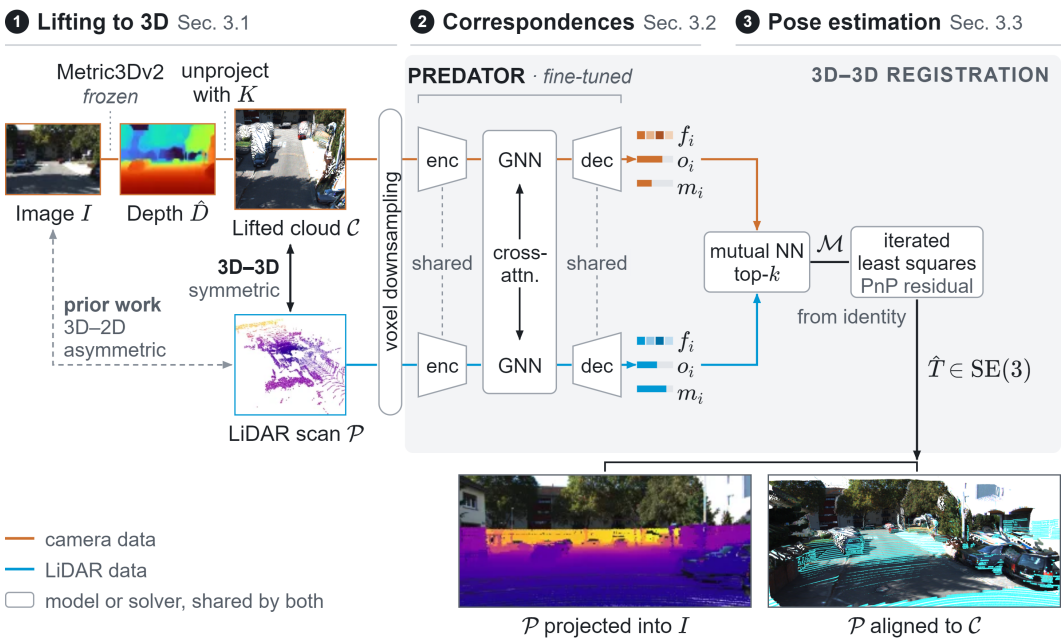

Figure 1: Read left to right. Stage 1 lifts the image to a metric point cloud with a frozen depth network, so both inputs are 3D. Stage 2 runs both clouds through the same PREDATOR network to get per-point descriptors, overlap and matchability. Stage 3 keeps mutual nearest neighbours and fits the pose by iterated least squares from the identity.

This simple pipeline is accurate and transfers to new sensors. On KITTI, X3Calib reaches 57.7% registration recall at $(1°, 0.25\,\mathrm{m})$, against 50.0% for LCCNet, trained on our test sequences. Zero-shot, at $(10°, 1\,\mathrm{m})$, it reaches 90.5% recall on Argoverse and 80.0% on PandaSet, where simply returning the identity transform reaches 51% and 47%. On nuScenes, it is the only method whose recall exceeds that of the identity transform (51.7% against 47.9%), well ahead of CMRNext (Cattaneo & Valada, 2025) (15.4%), which was trained on KITTI, Argoverse, and PandaSet.

Our main contribution is a new and simple direction for single-frame LiDAR-camera calibration: instead of designing a calibration network, we lift the image into metric 3D with an off-the-shelf depth model, match the two point clouds with an unmodified registration network, and fit the pose by iterated least squares. This simplicity is the benefit. The network learns local geometric correspondences rather than calibration parameters, so it transfers to sensor rigs it has never seen, where its recall exceeds that of both KITTI-trained calibrators and, on nuScenes, that of CMRNext, which was trained on many datasets. The pose solve is a few deterministic least-squares fits without hypothesis sampling, so the two networks take most of the 83 ms per frame, excluding preprocessing. Each component can be replaced by a stronger one as depth and registration models improve.

2 Related Work

Classical calibration. Target-based methods observe a known object, typically a checkerboard, in both sensors (Zhang & Pless, 2004; Geiger et al., 2012). They are accurate but need a dedicated procedure and cannot recalibrate online. Targetless methods align natural scene structure through edges (Levinson & Thrun, 2013), mutual information between LiDAR reflectance and image intensity (Pandey et al., 2015), or its normalized form (Taylor & Nieto, 2013). They need no infrastructure but depend on the initialization and on scenes rich in the chosen cue. X3Calib needs neither a target nor a hand-crafted cue: it learns correspondences between the two sensors from data and calibrates from a single frame. Zhang et al. (2025a) provides a recent survey on the field and its open problems.

Learned calibration. RegNet (Schneider et al., 2017) regresses the extrinsics from CNN features of the image and the projected depth, CalibNet (Iyer et al., 2018) adds geometric supervision through a spatial transformer, and LCCNet (Lv et al., 2021) builds a cost volume between image and projected LiDAR features and refines the estimate coarse to fine. Calib-Anything (Luo et al., 2024) exploits

segmentation foundation models, UniCal (Yang et al., 2024b) calibrates several sensors jointly, and CMRNext (Cattaneo & Valada, 2025), trained on three driving datasets, matches camera images to LiDAR projections. These methods compare 3D geometry with 2D appearance or projected depth. Closest to us, PseudoCal (Cocheteux et al., 2023b) works in 3D: it turns monocular depth into a pseudo-LiDAR cloud and regresses the extrinsics with a pillar-based network and two transformer regressors (Cocheteux et al., 2023a). We also lift the image with monocular depth, but we learn local correspondences between the two clouds and solve for the pose geometrically, so calibration parameters are never a network output or a training target.

Point cloud registration. ICP (Besl & McKay, 1992) alternates correspondence search and least-squares fitting and depends on its initialization; trimmed ICP (Chetverikov et al., 2002) discards a fraction of the worst residuals in each round. Learned descriptors, from 3DMatch (Zeng et al., 2017) and FCGF (Choy et al., 2019) to end-to-end matching in DCP (Wang & Solomon, 2019), replaced hand-crafted features. PREDATOR (Huang et al., 2021) added overlap attention so that two clouds with little overlap can still be matched, and GeoTransformer (Qin et al., 2022) matches superpoints with geometric self-attention. Robust fitting on putative correspondences proceeds either by hypothesize-and-verify, as in RANSAC (Fischler & Bolles, 1981) and its modern descendants (Barath, 2025), or by iterative reweighting, of which graduated non-convexity (Yang et al., 2020) is the principled form. Our pose estimation (Section 3.3) belongs to the second family. Unlike these methods, which align two scans from the same kind of sensor, we align a LiDAR scan with a point cloud predicted by a depth network, whose errors are structured rather than random.

Monocular metric depth. Affine-invariant models such as MiDaS (Ranftl et al., 2022) and DPT (Ranftl et al., 2021) generalize across datasets but leave scale and shift unknown. Metric models recover absolute depth: ZoeDepth (Bhat et al., 2023) adds metric heads to a relative backbone, Metric3D (Hu et al., 2024) normalizes images to a canonical camera so that one model serves many intrinsics, Depth Anything (Yang et al., 2024a) scales up unlabeled training data, and UniDepth (Piccinelli et al., 2024) predicts intrinsics and depth together; Zhang et al. (2025b) give a survey. Unlike these works, we do not evaluate depth itself: the lifted cloud is an intermediate representation, and what matters is whether its local geometry can be matched to a LiDAR scan. We use Metric3Dv2 as a fixed front end, and any metric model that accepts known intrinsics can take its place.

3 METHOD

The proposed X3Calib calibrates from one LiDAR scan and one image in three stages: lift the image into a metric 3D point cloud (Section 3.1), learn cross-modal correspondences between the lifted cloud and the LiDAR scan (Section 3.2), and estimate the rigid transform from them by iterated least squares fitting (Section 3.3).

3.1 CALIBRATION AS CROSS-MODAL 3D REGISTRATION

Let $\mathcal{P} = \{l_i \in \mathbb{R}^3\}_{i=1}^{N}$ be a single LiDAR scan in the LiDAR frame and I the image captured with it by a pinhole camera with known intrinsics K; no other frame is used. We seek the rigid transform $T^* = (\mathbf{R}^*, \mathbf{t}^*) \in \mathrm{SE}(3)$ that maps LiDAR points into the camera frame, $l_i^{\mathrm{cam}} = \mathbf{R}^* l_i + \mathbf{t}^*$. Prior learned calibrators estimate T^* by comparing $\mathcal{P}$ with image features, an asymmetric problem. We make it symmetric by building a camera-side point cloud $\mathcal{C} = \{c_j \in \mathbb{R}^3\}_{j=1}^{M}$. A monocular metric depth network predicts a depth map $\hat{D}$, and every pixel (u_j, v_j) with $\hat{D}(u_j, v_j) > 0$ is unprojected,

$$c_j = \hat{D}(u_j, v_j)\, K^{-1}[u_j, v_j, 1]^\top. \tag{1}$$

Metric depth, not relative depth, is essential: an unknown scale would resize the local structures that the correspondence network matches at the LiDAR's metric scale. We use Metric3Dv2 (Hu et al., 2024), which generalizes across cameras without fine-tuning. Both clouds are voxel-downsampled to a common resolution. Calibration then becomes the registration problem as follows:

$$T^* = \arg\min_{T \in \mathrm{SE}(3)} \sum_{(i,j) \in \mathcal{M}} r_{ij}(T)^2, \tag{2}$$

where $\mathcal{M}$ is a set of cross-modal correspondences and $r_{ij}(T)$ is a 3D distance or a reprojection error under T (Section 3.3). Two tasks remain: to find $\mathcal{M}$ (Section 3.2), and to solve Eq. (2) despite

the outliers in $\mathcal{M}$. Monodepth is imperfect. For correspondence learning this matters less than for reconstruction: the network learns to match local surface structure despite noise, and in distribution the robust solver absorbs the resulting outliers. Section 4.3 shows where this tolerance ends.

3.2 Cross-Modal Correspondence Learning

We build on PREDATOR (Huang et al., 2021), a network designed for registering point clouds with low overlap. This fits our setting, since the camera frustum covers only a fraction of the LiDAR field of view. PREDATOR is a two-stream encoder-decoder with a shared KPConv backbone. At the bottleneck, a graph neural network aggregates context within each cloud, and cross-attention between the clouds lets each point reason about whether its region is visible in the other. The decoder upsamples to full resolution and outputs, for every point, a descriptor f_i, an overlap probability o_i (does the point have a counterpart in the other cloud?), and a matchability score m_i (can the current descriptors match it reliably?). We use the architecture (Fig. 1) and its three losses unchanged: a circle loss on descriptor distances of positive and negative pairs defined by 3D proximity under the ground-truth transform, a binary cross-entropy loss on overlap, and a matchability loss whose labels are generated online from the current descriptors. We initialize from the public 3DMatch weights (Zeng et al., 2017) and fine-tune on KITTI pairs of LiDAR and depth-lifted clouds. Ground-truth correspondences come from projecting LiDAR points into the image with the true calibration before downsampling; points lost in downsampling are dropped. The only cross-modal ingredient is the data. One stream sees LiDAR geometry, the other sees depth-network geometry, and the losses force their descriptors into a common space.

3.3 Pose Estimation by Iterated Least Squares

Correspondences. We match descriptors by mutual nearest neighbours: a pair (l_i, c_j) survives only if each point is the nearest neighbour of the other, which removes most spurious matches. The survivors are ranked by a score that combines descriptor similarity, overlap probability, and matchability, and the top-k pairs form $\mathcal{M}$.

Two residuals, one correspondence set. A correspondence can be scored in two ways. The *3D-3D* residual $r_{ij}^{\mathrm{3D}}(T) = \|T(l_i) - c_j\|_2$, in metres, treats the pair as two 3D points; the rigid transform that minimizes its sum of squares over a set of pairs has a closed form (Kabsch, 1976; Umeyama, 1991). It uses the lifted depth and inherits its errors, and the intrinsics enter only through Eq. (1). The *PnP* residual $r_{ij}^{\mathrm{PnP}}(T) = \|\pi(K\,T(l_i)) - [u_j, v_j]^\top\|_2$, with $\pi([x, y, z]^\top) = [x/z,\ y/z]^\top$, is a reprojection error in pixels that discards the lifted depth of c_j. Minimizing it is a nonlinear least-squares problem, which we solve with PoseLib (Larsson & contributors, 2020). It does not use the lifted depth in the pose solve, although its correspondences still come from descriptors computed on the lifted cloud.

Iterated least squares. Given an initial pose T_0, round k keeps the correspondences whose residual under the current pose T_k is below a threshold ϵ_k and refits the pose on them,

$$w_{ij}^{(k)} = \mathbf{1}[\, r_{ij}(T_k) < \epsilon_k \,], \qquad T_{k+1} = \arg\min_{T} \sum_{(i,j)\in\mathcal{M}} w_{ij}^{(k)}\, r_{ij}(T)^2, \tag{3}$$

for $k = 0, \ldots, K-1$, with a fixed schedule $\epsilon_0 \geq \epsilon_1 \geq \cdots \geq \epsilon_{K-1}$ that shrinks from coarse to fine: early rounds accept many correspondences and remove gross error, late rounds keep only the accurate ones. Thresholds are in metres for 3D-3D and in pixels for PnP, whose first round compares 3D distances instead. The refit is the closed-form solution for the 3D-3D residual and PoseLib's nonlinear refinement for the PnP residual, applied once per round. Each round is deterministic, and robustness comes from the binary weights and the shrinking threshold. The procedure is truncated least squares with a decreasing truncation, the schedule of graduated non-convexity (Yang et al., 2020), and a relative of trimmed ICP (Chetverikov et al., 2002) with fixed correspondences. It is not iteratively reweighted least squares with a smooth kernel, since the weights are zero or one.

Training. We fine-tune PREDATOR end to end with SGD on one NVIDIA A100 (40 GB) on the KITTI Odometry dataset (Geiger et al., 2012), using synchronized LiDAR scans, monocular images, and the ground-truth calibration. To simulate miscalibration, we apply a random $\mathrm{SE}(3)$ perturbation to the LiDAR cloud after mapping it into the camera frame: a rotation $\exp(\theta[\mathbf{a}]_\times)$ about a random unit axis $\mathbf{a}$ with $\theta \sim \mathcal{U}(-20^\circ, 20^\circ)$, and a translation $\Delta\mathbf{t} \sim \mathcal{U}(-0.5, 0.5)^3$ m, whose norm is at most 0.87 m. The same range is used at test time; each test frame, one LiDAR scan and one image,

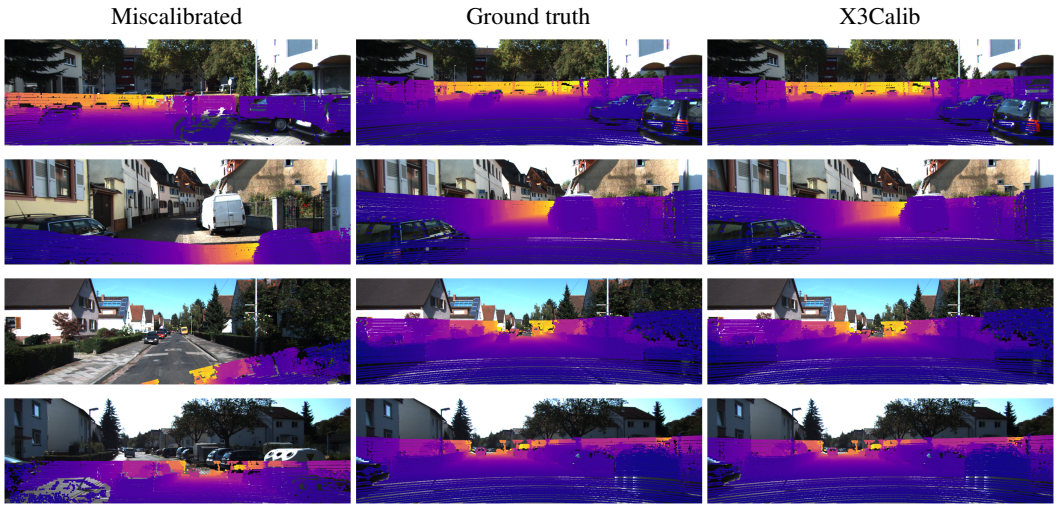

Figure 2: **Qualitative KITTI results.** Each row shows a KITTI (Geiger et al., 2012) frame with LiDAR points projected onto the image using the extrinsic calibration before (left) and after (right) correction by X3Calib, alongside the ground-truth projection (centre). Point colour encodes depth. After calibration, projected points align tightly with object boundaries and scene geometry, demonstrating the precision of the recovered $\mathrm{SE}(3)$ transformation.

receives one perturbation drawn from a fixed seed and shared across methods. In all experiments, iterated least squares starts from the identity. The range is comparable to the widest stage of LCCNet ($\pm 20°$, ± 1.5 m) and wider than CalibNet's.

4 Experiments

4.1 Setup

Datasets. KITTI (Velodyne HDL-64E, PointGrey Flea2) is the training set and the in-distribution test bed, with 1,522 test frames. Three datasets are used only for testing, without fine-tuning. nuScenes (Caesar et al., 2020): Velodyne HDL-32E, front camera, 601 frames from the validation split. Argoverse (Chang et al., 2019): two 32-beam LiDARs merged into one 64-beam sweep, ring front-centre camera, three held-out logs, 472 frames. PandaSet (Xiao et al., 2021): Pandar64, front camera, eight held-out scenes, 640 frames. They differ from KITTI and from one another in LiDAR make and beam count, camera intrinsics, and geography: KITTI was recorded in Germany, nuScenes in Boston and Singapore, Argoverse in Pittsburgh and Miami, and PandaSet in California.

Baselines. We compare with three single-frame, targetless, learned calibrators that release code and trained weights. LCCNet (Lv et al., 2021) and CalibNet (Iyer et al., 2018) are trained on KITTI, LCCNet also on our test sequences 18–20, so its KITTI row is gray. CMRNext (Cattaneo & Valada, 2025), the most recent, matches camera images to LiDAR projections and reports state-of-the-art calibration accuracy; its released weights were trained jointly on KITTI, Argoverse, PandaSet and other datasets. These training data overlap our test data: they include our KITTI test sequences 18–20, the Argoverse `train1`–`train3` logs, recorded with the same vehicle and sensors as our validation logs, and all PandaSet scenes except nine held-out ones. CMRNext is thus held out only on nuScenes, and we show its other results in gray. We do not compare with other methods since none of them can be run as a trained single-frame calibrator: PseudoCal and RegNet release no weights (Cocheteux et al., 2023b; Schneider et al., 2017), Calib-Anything and UniCal optimize the calibration for each scene or sequence instead of predicting it (Luo et al., 2024; Yang et al., 2024b). Since CMRNext is the state of the art among the methods that can be run, we consider the comparison representative. We add an *Identity* baseline, which always returns the identity transform, whatever the input. Under our protocol, its expected median errors are $10.0°$ and 0.49 m, and its

Table 1: **In-distribution results on KITTI.** All methods face the same perturbations (rotation up to 20°, translation up to 0.5 m per axis). *Identity* always returns the identity transform, whatever the input; we list its expected scores under the perturbation distribution. X3Calib rows use iterated least squares with the named solver. All methods are trained on KITTI. †Not held out: the weights of LCCNet (Lv et al., 2021) and CMRNext (Cattaneo & Valada, 2025) were trained on our KITTI test sequences (Section 4.1); gray rows are excluded from the ranking. **Bold**: best value in each column.

	Rot. err. (°) ↓		Trans. err. (m) ↓		Recall (%) ↑		
Method	Mean	Med.	Mean	Med.	1°, 0.25 m	5°, 0.5 m	10°, 1 m
Identity	10.00	10.00	0.48	0.49	0.3	13.1	50.0
CalibNet (Iyer et al., 2018)	10.27	9.93	0.49	0.50	0.0	10.2	50.4
LCCNet (Lv et al., 2021)†	0.79	0.51	0.26	0.29	50.0	50.0	50.0
CMRNext (Cattaneo & Valada, 2025)†	0.13	0.08	0.04	0.01	99.3	99.6	99.8
X3Calib (3D-3D)	**0.58**	**0.47**	**0.24**	**0.21**	**57.7**	**94.2**	**99.7**

expected recall is 0.3%, 13.1%, and 50.0% at the three thresholds below. We report the expectation for KITTI and the measured values for the other three datasets.

Metrics. We report the rotation error $e_R = arccos(\frac{tr(\hat{R}^T \cdot R^*)-1}{2})$, the geodesic error between estimated and true rotation, the translation error $e_t = \|\hat{\mathbf{t}} - \mathbf{t}^*\|_2$, and the registration recall $\mathrm{RR@}(\Delta r, \Delta t)$, the fraction of frames with $e_R < \Delta r$ and $e_t < \Delta t$, at $(1°, 0.25\,\mathrm{m})$, $(5°, 0.5\,\mathrm{m})$, and $(10°, 1\,\mathrm{m})$. No perturbation moves the translation by more than 0.87 m, so Identity's recall at $(10°, 1\,\mathrm{m})$ measures rotation alone. We give mean and median errors and treat the median as primary, since one catastrophic failure can dominate the mean. All X3Calib variants on a dataset read from one cache of network outputs computed in a single forward pass, so every fitting strategy and solver sees identical correspondences and identical perturbations on every test frame, and differences between rows come from the fitting alone.

4.2 In-Distribution Results

Table 1 compares all methods on KITTI. With the 3D-3D solver, X3Calib reaches median errors of 0.47° and 0.21 m and recall of 57.7%, 94.2%, and 99.7% at the three thresholds, ahead of both KITTI-trained baselines at every threshold: LCCNet reaches 50.0% at all three and CalibNet 0.0%, 10.2%, and 50.4%. LCCNet's flat 50.0% means that half of its estimates fall within the tightest threshold and the other half outside 10° or 1 m. We report the output of the public code as produced. LCCNet and CMRNext were trained on our test sequences, so their rows are gray; CMRNext is more accurate, with 99.3% at the tightest threshold. CalibNet, whose training range is narrower than our perturbations, does not improve on the Identity baseline, which always returns the identity transform: it is worse on five of the seven metrics, for example 10.2% against 13.1% at $(5°, 0.5\,\mathrm{m})$.

4.3 Cross-Dataset Generalization

The promise of registering in a shared geometric domain is that correspondences transfer across sensor rigs. Table 2 tests this on three unseen datasets. X3Calib, LCCNet, and CalibNet are trained only on KITTI; CMRNext is also trained on Argoverse and PandaSet and is held out only on nuScenes.

Identity is a strong baseline, and the published calibrators fall below it. Because the perturbation is bounded, always returning the identity transform already scores far from zero. Averaged over the three datasets (Table 2, last block), Identity has a median rotation error of 10.26° and recall of 12.9% at $(5°, 0.5\,\mathrm{m})$ and 48.6% at $(10°, 1\,\mathrm{m})$. Neither calibrator trained only on KITTI reaches this level. CalibNet averages 11.66°, 3.8%, and 38.9%, and it is never better than Identity except in median translation on nuScenes, where it is lower by 1 cm. LCCNet averages 13.86°, scores 0% at every threshold, and has median translation errors of 3.3 to 4.3 m, far beyond the 0.87 m the perturbation can produce. Both methods regress a correction learned on KITTI, and on a new sensor rig that correction moves the extrinsics away from the truth rather than towards it.

More training data does not close the gap. On Argoverse and PandaSet, whose sensor rigs it saw in training, CMRNext exceeds 99% recall at every threshold; this is what in-domain training buys, and why those gray rows are not a comparison. On nuScenes, the one dataset it has not seen, it has the lowest median rotation error of all methods (2.59°), but its median translation error of 1.61 m

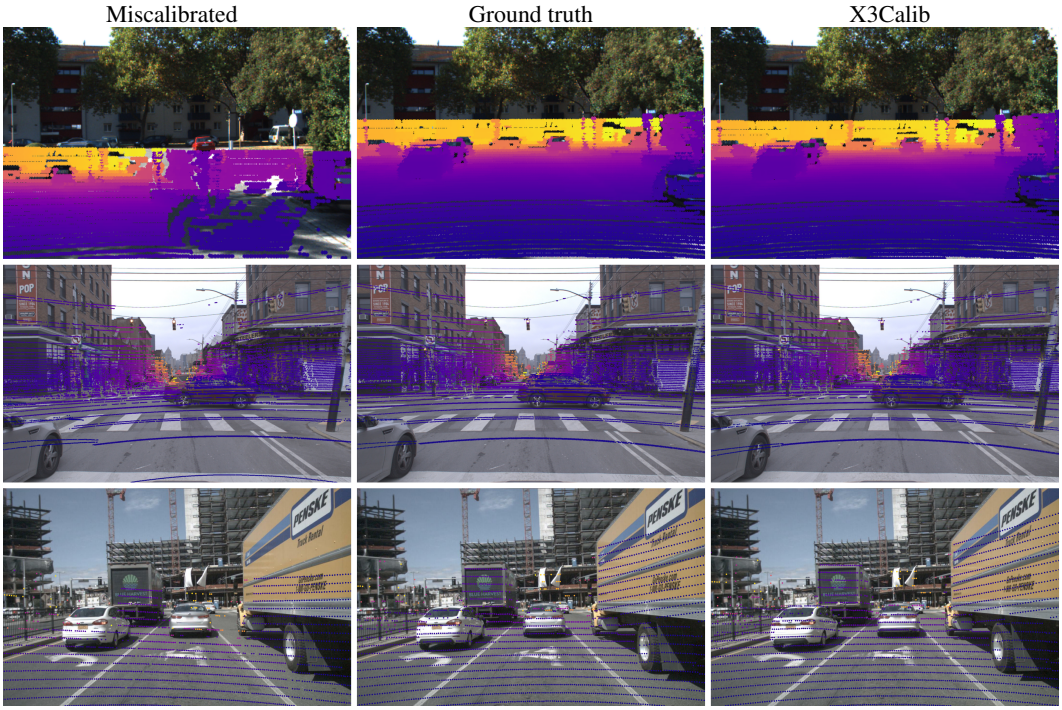

Figure 3: **LiDAR points projected into the image** with the miscalibrated extrinsics (left), the ground truth (centre), and the X3Calib estimate (right); colour encodes depth. *Top*: KITTI, in distribution. *Middle*: Argoverse, zero-shot, PnP solver; the rotation error falls from $2.49°$ to $0.48°$ and the translation error from $0.54\,\mathrm{m}$ to $0.36\,\mathrm{m}$ on this frame. *Bottom*: nuScenes, zero-shot. Table 2 gives the statistics over all frames of each of the three unseen datasets.

exceeds the $0.87\,\mathrm{m}$ any perturbation can produce. Its recall at $(5°, 0.5\,\mathrm{m})$ and $(10°, 1\,\mathrm{m})$, 1.1% and 15.4%, is therefore below that of Identity (9.2%, 47.9%) and far below that of X3Calib with the 3D-3D solver (15.1%, 51.7%), trained on KITTI alone.

X3Calib is the only method that improves on Identity zero-shot. Averaged over the three datasets, X3Calib with the PnP solver lowers the median rotation error from $10.26°$ to $4.71°$ and raises recall from 12.9% to 29.9% at $(5°, 0.5\,\mathrm{m})$ and from 48.6% to 72.9% at $(10°, 1\,\mathrm{m})$; the 3D-3D solver reaches $5.48°$, 25.4%, and 67.3%. The gain is largest on Argoverse and PandaSet (90.5% and 80.0% at $(10°, 1\,\mathrm{m})$ with PnP, against 51.1% and 46.9%) and modest on nuScenes (51.7% with 3D-3D, against 47.9%). X3Calib never scores below Identity on any dataset or metric, whereas LCCNet and CalibNet fall below it on almost every one and CMRNext on three of the five on nuScenes. This supports the premise of the reformulation: correspondences learned on local 3D geometry transfer to new sensors, whereas neither a regressed correction nor image-to-projection matching trained on three datasets does. Figure 3 shows a KITTI, an Argoverse, and a nuScenes frame.

4.4 Ablations

Fitting strategy: consensus search instead of iterated least squares. Iterated least squares is a local method. The standard alternative is a global consensus search: we feed the same correspondences to SupeRANSAC (Barath, 2025), with a three-point closed-form 3D-3D solver or a minimal PnP solver, and score hypotheses with the same residuals (Table 3). It needs no initial estimate. In distribution the two strategies are close: with the 3D-3D solver, the consensus search is better at the tightest threshold (62.0% against 57.7%) and iterated least squares at the two looser ones (94.2% and 99.7% against 93.3% and 99.5%). Zero-shot the consensus search fails with either solver. Its recall falls below Identity at $(5°, 0.5\,\mathrm{m})$ and $(10°, 1\,\mathrm{m})$ on every dataset, down to 4.0% on Argoverse against 51.1%, and its median translation error is 2.5 to $6.7\,\mathrm{m}$ with 3D-3D and 17.6 to $64.3\,\mathrm{m}$ with PnP, far beyond the $0.87\,\mathrm{m}$ the perturbation can produce. It finds a large set of mutually con-

Table 2: **Zero-shot results on three unseen datasets**, X3Calib, LCCNet, and CalibNet trained only on KITTI. *Identity* always returns the identity transform, whatever the input. X3Calib rows use iterated least squares with the named solver. *AVG*: unweighted mean of the three per-dataset values. Medians are reported. †CMRNext (Cattaneo & Valada, 2025) is held out only on nuScenes; it was trained on the Argoverse sensor rig and on our PandaSet test scenes (Section 4.1), so its gray rows are not a comparison and are excluded from the ranking. **Bold**: best and underlined: second-best value per block and column, lower for errors and higher for recall.

	Method	Rot. med. (°) ↓	Trans. med. (m) ↓	Recall (%) ↑		
				1°, 0.25 m	5°, 0.5 m	10°, 1 m
nuScenes	CMRNext (Cattaneo & Valada, 2025)†	**2.59**	1.61	0.0	1.1	15.4
	LCCNet (Lv et al., 2021)	12.86	3.73	0.0	0.0	0.0
	CalibNet (Iyer et al., 2018)	10.57	**0.50**	0.0	5.6	46.6
	Identity	10.43	0.51	0.0	9.2	47.9
	X3Calib (PnP)	10.41	0.51	0.0	11.0	48.1
	X3Calib (3D-3D)	9.43	0.51	0.0	**15.1**	**51.7**
Argoverse	CMRNext (Cattaneo & Valada, 2025)†	0.18	0.05	99.8	100.0	100.0
	LCCNet (Lv et al., 2021)	14.68	4.25	0.0	0.0	0.0
	CalibNet (Iyer et al., 2018)	10.21	0.52	0.0	5.3	48.5
	Identity	9.91	**0.48**	0.4	15.9	51.1
	X3Calib (PnP)	**1.49**	**0.48**	**2.5**	**42.2**	**90.5**
	X3Calib (3D-3D)	3.02	**0.48**	0.6	33.1	79.7
PandaSet	CMRNext (Cattaneo & Valada, 2025)†	0.06	0.02	99.1	99.8	100.0
	LCCNet (Lv et al., 2021)	14.04	3.25	0.0	0.0	0.0
	CalibNet (Iyer et al., 2018)	14.19	0.58	0.0	0.6	21.7
	Identity	10.44	**0.49**	0.2	13.6	46.9
	X3Calib (PnP)	**2.22**	**0.49**	**1.2**	**36.6**	**80.0**
	X3Calib (3D-3D)	3.99	**0.49**	0.6	28.1	70.5
AVG	LCCNet (Lv et al., 2021)	13.86	3.74	0.0	0.0	0.0
	CalibNet (Iyer et al., 2018)	11.66	0.53	0.0	3.8	38.9
	Identity	10.26	**0.49**	0.2	12.9	48.6
	X3Calib (PnP)	**4.71**	**0.49**	**1.2**	**29.9**	**72.9**
	X3Calib (3D-3D)	5.48	**0.49**	0.4	25.4	67.3

0

0000

sistent but wrong correspondences and returns their pose with confidence. In our experiments, a certified-optimal branch-and-bound search on the same correspondences is no more accurate on a small KITTI subset, so the correspondence set, not the search, is the bottleneck.

Solver: PnP versus 3D-3D. Scoring the same correspondences with two residuals separates the effect of the lifted depth from that of the correspondences. In distribution, the 3D-3D residual gives higher recall with both fitting strategies, by a wide margin under the consensus search (93.3% against 74.5% at $(5°, 0.5\,\mathrm{m})$) and at every threshold under iterated least squares (57.7%, 94.2%, and 99.7% against 55.4%, 86.1%, and 96.7%), although PnP has the lower median rotation error there ($0.42°$ against $0.47°$). We expected the opposite, since the PnP residual ignores the lifted depth. We suspect that on KITTI the lifted depth is accurate enough for a metric residual to be a better-conditioned inlier criterion than a pixel residual, which weights near and far points differently. Zero-shot, the order depends on the fitting strategy: under the consensus search, 3D-3D is better on Argoverse and PandaSet and on translation everywhere, and both fail; under iterated least squares, PnP is better on Argoverse and PandaSet and 3D-3D on nuScenes. We therefore recommend the 3D-3D solver in distribution and the PnP solver on unseen rigs, where it is better on average.

Runtime. X3Calib takes $83.5\,\mathrm{ms}$ per frame (12 FPS) on one RTX 4090, 71.7% of it in the depth prediction network (Table 4), so faster depth networks speed it up. Recalibration need not run on every frame, so X3Calib can run in the background.

Limitations. X3Calib matches geometry alone: it ignores the LiDAR intensity and the image colour, so scenes with little 3D structure but rich texture, such as open roads, constrain it poorly. Feeding both to the correspondence network as point features is future work. X3Calib also inherits the errors of the depth network; a wrong metric scale, for instance, biases the 3D-3D solve. The same dependence is a strength: X3Calib will improve as metric depth networks improve.

Table 3: **Ablation: consensus search in place of iterated least squares.** The same correspondences are fed to SupeRANSAC (Barath, 2025) with a three-point 3D-3D solver or a PnP solver, on KITTI (in distribution) and on nuScenes, Argoverse, and PandaSet (zero-shot). For comparison, the *Iterated LS* rows give X3Calib with each solver, as in Tables 1 and 2. Medians as in Table 2. **Bold**: best value per dataset and column.

	Variant	Rot. med. (°) ↓	Trans. med. (m) ↓	Recall (%) ↑		
				1°, 0.25 m	5°, 0.5 m	10°, 1 m
KITTI	RANSAC (PnP)	0.64	0.32	32.8	74.5	94.4
	RANSAC (3D-3D)	0.50	**0.20**	**62.0**	93.3	99.5
	Iterated LS (PnP)	**0.42**	0.22	55.4	86.1	96.7
	Iterated LS (3D-3D)	0.47	0.21	57.7	**94.2**	**99.7**
nuScenes	RANSAC (PnP)	**8.79**	24.60	**1.0**	5.3	11.5
	RANSAC (3D-3D)	8.93	2.49	0.0	6.2	20.8
	Iterated LS (PnP)	10.41	**0.51**	0.0	11.0	48.1
	Iterated LS (3D-3D)	9.43	**0.51**	0.0	**15.1**	**51.7**
Argoverse	RANSAC (PnP)	8.14	17.60	0.0	0.6	3.2
	RANSAC (3D-3D)	7.07	6.70	0.0	1.3	4.0
	Iterated LS (PnP)	**1.49**	**0.48**	**2.5**	**42.2**	**90.5**
	Iterated LS (3D-3D)	3.02	**0.48**	0.6	33.1	79.7
PandaSet	RANSAC (PnP)	17.18	64.31	0.0	0.5	2.8
	RANSAC (3D-3D)	7.89	2.55	0.9	8.4	22.0
	Iterated LS (PnP)	**2.22**	**0.49**	**1.2**	**36.6**	**80.0**
	Iterated LS (3D-3D)	3.99	**0.49**	0.6	28.1	70.5

Table 4: **Per-frame runtime on KITTI** on one RTX 4090 (batch size 1, FP16 depth network). Matching runs on the GPU and iterated least squares (3D-3D) on the CPU.

Stage	Mean (ms)	Med. (ms)	Std. (ms)	% of total
Depth estimation	59.9	59.8	0.4	71.7
Correspondence network	17.4	17.1	1.1	20.9
Matching + iterated least squares	6.2	6.3	0.4	7.4
Total	83.5	83.2	1.3	100.0

5 Conclusion

We reformulated single-frame LiDAR-camera calibration as cross-modal 3D registration. Lifting the image into a metric point cloud lets an unmodified registration network learn correspondences between the two sensors, and iterated least squares with a coarse-to-fine inlier threshold turns them into extrinsics. In distribution, the correspondences are accurate enough to exceed public learned calibrators at the tightest threshold, and using the lifted depth in the solve beats discarding it. Zero-shot, no published calibrator improves on the recall of the Identity baseline, not even CMRNext on nuScenes, although it was trained on three datasets. X3Calib improves on it on every unseen dataset, strongly on Argoverse and PandaSet and modestly on nuScenes. Future work should recover translation off-distribution, for instance by estimating the depth scale jointly with the pose, and should test the method on real sensor rigs whose miscalibration is not bounded in advance. The code and trained model will be made public.

AI use statement

In this work, we used generative AI tools to polish the text, that is, to edit the paper for readability. We have reviewed all AI-assisted text and checked that it matches our results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

This work uses only publicly available autonomous-driving benchmark datasets (KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2020), Argoverse (Chang et al., 2019), and Pandaset (Xiao et al., 2021)) that were collected and released by their respective authors; it does not involve new data collection, human subjects, or personally identifiable information. We do not foresee harmful applications specific to this work beyond those inherent to autonomous-driving perception systems in general.

Reproducibility statement

We describe the model architecture, training objectives, and optimization setup in Section 3, and the datasets, baselines, miscalibration protocol, and evaluation metrics in Section 4.1. The code will be made public upon acceptance.

Acknowledgments

References

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In *CVPR*, pp. 1080–1089, 2022. doi: 10.1109/CVPR52688.2022.00116.

Daniel Barath. SupeRANSAC: One RANSAC to rule them all. *arXiv preprint arXiv:2506.04803*, 2025.

Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, pp. 11618–11628, 2020. doi: 10.1109/CVPR42600.2020.01164.

Daniele Cattaneo and Abhinav Valada. CMRNext: Camera to LiDAR matching in the wild for localization and extrinsic calibration. *IEEE Transactions on Robotics*, 41:1995–2013, 2025. doi: 10.1109/TRO.2025.3546784.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3D tracking and forecasting with rich maps. In *CVPR*, pp. 8740–8749, 2019. doi: 10.1109/CVPR.2019.00895.

Dmitry Chetverikov, Dmitry Svirko, Dmitry Stepanov, and Pavel Krsek. The trimmed iterative closest point algorithm. In *ICPR*, volume 3, pp. 545–548, 2002. doi: 10.1109/ICPR.2002.1047997.

Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, pp. 8957–8965, 2019. doi: 10.1109/ICCV.2019.00905.

Mathieu Cocheteux, Aaron Low, and Marius Bruehlmeier. UniCal: A single-branch transformer-based model for camera-to-LiDAR calibration and validation. *arXiv preprint arXiv:2304.09715*, 2023a.

Mathieu Cocheteux, Julien Moreau, and Franck Davoine. PseudoCal: Towards initialisation-free deep learning-based camera-LiDAR self-calibration. In *BMVC*, 2023b. arXiv:2309.09855.

Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3D object detection to common corruptions in autonomous driving. In *CVPR*, pp. 1022–1032, 2023. doi: 10.1109/CVPR52729.2023.00105.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981. doi: 10.1145/358669.358692.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, pp. 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE TPAMI*, 46(12):10579–10596, 2024. doi: 10.1109/TPAMI.2024.3444912.

Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. PREDATOR: Registration of 3D point clouds with low overlap. In *CVPR*, pp. 4265–4274, 2021. doi: 10.1109/CVPR46437.2021.00425.

Ganesh Iyer, R. Karnik Ram, J. Krishna Murthy, and K. Madhava Krishna. CalibNet: Geometrically supervised extrinsic calibration using 3D spatial transformer networks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1110–1117, 2018. doi: 10.1109/IROS.2018.8593693.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Viktor Larsson and contributors. PoseLib – minimal solvers for camera pose estimation. `https://github.com/vlarsson/PoseLib`, 2020.

Jesse Levinson and Sebastian Thrun. Automatic online calibration of cameras and lasers. In *Robotics: Science and Systems (RSS)*, 2013. doi: 10.15607/RSS.2013.IX.029.

Jiarong Lin and Fu Zhang. R^3LIVE: A robust, real-time, RGB-colored, LiDAR-inertial-visual tightly-coupled state estimation and mapping package. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10672–10678, 2022. doi: 10.1109/ICRA46639.2022.9811935.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2774–2781, 2023. doi: 10.1109/ICRA48891.2023.10160968.

Zhaotong Luo, Guohang Yan, Xinyu Cai, and Botian Shi. Zero-training LiDAR-camera extrinsic calibration method using segment anything model. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14472–14478, 2024. doi: 10.1109/ICRA57147.2024.10610983.

Xudong Lv, Boya Wang, Ziwen Dou, Dong Ye, and Shuo Wang. LCCNet: LiDAR and camera self-calibration using cost volume network. In *CVPRW*, pp. 2888–2895, 2021. doi: 10.1109/CVPRW53098.2021.00324.

Gaurav Pandey, James R. McBride, Silvio Savarese, and Ryan M. Eustice. Automatic extrinsic calibration of vision and Lidar by maximizing mutual information. *Journal of Field Robotics*, 32 (5):696–722, 2015. doi: 10.1002/rob.21542.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR*, pp. 10106–10116, 2024. doi: 10.1109/CVPR52733.2024.00963.

Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *CVPR*, pp. 11133–11142, 2022. doi: 10.1109/CVPR52688.2022.01086.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pp. 12159–12168, 2021. doi: 10.1109/ICCV48922.2021.01196.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2022. doi: 10.1109/TPAMI.2020.3019967.

Nick Schneider, Florian Piewak, Christoph Stiller, and Uwe Franke. RegNet: Multimodal sensor registration using deep neural networks. In *IEEE Intelligent Vehicles Symposium (IV)*, pp. 1803–1810, 2017. doi: 10.1109/IVS.2017.7995968.

Zachary Taylor and Juan Nieto. Automatic calibration of Lidar and camera images using normalized mutual information. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE TPAMI*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. PointPainting: Sequential fusion for 3D object detection. In *CVPR*, pp. 4603–4611, 2020. doi: 10.1109/CVPR42600.2020.00466.

Yue Wang and Justin M. Solomon. Deep closest point: Learning representations for point cloud registration. In *ICCV*, pp. 3522–3531, 2019. doi: 10.1109/ICCV.2019.00362.

Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, Yunlong Wang, and Diange Yang. PandaSet: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 3095–3101, 2021. doi: 10.1109/ITSC48978.2021.9565009.

Heng Yang, Pasquale Antonante, Vasileios Tzoumas, and Luca Carlone. Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection. *IEEE Robotics and Automation Letters*, 5(2):1127–1134, 2020. doi: 10.1109/LRA.2020.2965893.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pp. 10371–10381, 2024a. doi: 10.1109/CVPR52733.2024.00987.

Ze Yang, George Chen, Haowei Zhang, Kevin Ta, Ioan Andrei Bârsan, Daniel Murphy, Sivabalan Manivasagam, and Raquel Urtasun. UniCal: Unified neural sensor calibration. In *ECCV*, pp. 327–345, 2024b. doi: 10.1007/978-3-031-72764-1_19.

Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In *CVPR*, pp. 199–208, 2017. doi: 10.1109/CVPR.2017.29.

Haoxin Zhang, Shuaixin Li, Xiaozhou Zhu, Hongbo Chen, and Wen Yao. 3-D LiDAR and monocular camera calibration: A review. *IEEE Sensors Journal*, 25(7):10530–10555, 2025a. doi: 10.1109/JSEN.2025.3530925.

Ji Zhang and Sanjiv Singh. Visual-lidar odometry and mapping: Low-drift, robust, and fast. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2174–2181, 2015. doi: 10.1109/ICRA.2015.7139486.

Jiuling Zhang, Yurong Wu, and Huilong Jiang. Survey on monocular metric depth estimation. *Computers*, 14(11), 2025b. ISSN 2073-431X. doi: 10.3390/computers14110502. URL `https://www.mdpi.com/2073-431X/14/11/502`.

Qilong Zhang and Robert Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pp. 2301–2306, 2004. doi: 10.1109/IROS.2004.1389752.