DRS-VPT: Directly Relocalizing in a Scan with Vision Point Transformers

Lanke Frank Tarimo Fu and Maurice Fallon

University of Oxford, Oxford OX2 6NN, United Kingdom
{fu,mfallon}@robots.ox.ac.uk

Abstract. We present **DRS-VPT**, a feed-forward transformer architecture for foundational image-to-scan registration. Given query images and a reference 3D point cloud, the model predicts the scan pose and point map alongside the poses and point maps of each camera, all expressed in the first camera's frame. It additionally predicts a coarse-to-fine pyramid of per-point and per-pixel features for direct reprojective alignment of the scan to the first image. This formulation unifies downstream tasks such as camera–LiDAR calibration in autonomous driving and indoor camera-to-map relocalization. A single DRS-VPT model achieves state-of-the-art performance for image-to-LiDAR registration in autonomous driving, competitive indoor relocalization without training map-specific weights, and strong zero-shot transfer to unseen environments. We also show qualitatively that the model learns complex scan-to-image projection properties such as occlusion of back-facing points.

1 Introduction

Recent progress in feed-forward visual geometry has shown that large transformer architectures can infer dense scene structure and camera poses directly from images, without requiring explicit calibration or iterative optimization. Models such as DUSt3R [38], MASt3R [16], VGGT [36], and MapAnything [15] demonstrate that generalizable multi-view geometry can be learned at scale. However, these approaches operate purely in the image domain. When geometry priors are available [12, 15], they are assumed to be co-registered with an image view. In many applications—including robotics, autonomous driving, and digital surveying—scene geometry is instead captured independently by sensors such as LiDAR or structured-light scanners, and is not aligned to any camera. This is the problem of *image-to-scan registration*: aligning a query image to a reference 3D point cloud (see Figure 1).

Image-to-scan registration takes different forms. In autonomous driving, it arises as camera–LiDAR calibration. Classical methods [6, 17, 50] require controlled targets, known intrinsics, or accurate initialization; learned approaches [7, 13] assume small pose offsets and train on a single sensor configuration, limiting generalization. Similarly, in indoor and outdoor camera-to-map relocalization, a

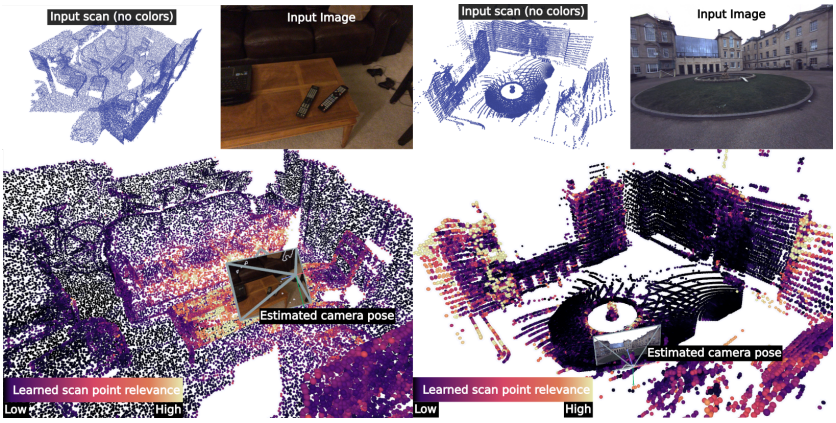

Fig. 1: As input, our model ingests raw point clouds (without color information), alongside images taken from different poses. The model accurately registers the pose of the camera in the point cloud. (Left) camera to dense map localization on a 12scenes (unseen during training). (Right) camera to sparse scan localization on the Oxford Spires dataset.

query image must be registered against a 3D scan of the environment. Existing methods [26, 29, 47] perform image-to-scan alignment by either lifting image features into SFM reconstructions, or rendering image views from dense XYZRGB maps where an expensive TLS sensor provides per-point color. Both approaches rely on an external source of pre-aligned per-point color baked into the 3D map.

In this work, we introduce **DRS-VPT**, a feed-forward transformer architecture for image-to-scan registration. By training at scale across diverse sensors and environments, the model learns true cross-modal alignment between raw point clouds and images without baking image features into the input 3D points, or prior knowledge of the scan's pose. For each camera, the model predicts a dense point map and camera pose. For scan alignment, the model predicts a coarse scan pose and dense cross-modal feature pyramids for both the scan and first image, which together drive differentiable direct image-to-scan alignment, achieving state-of-the-art camera–LiDAR registration and competitive relocalization against raw point cloud maps.

Our contributions are as follows:

- We introduce **DRS-VPT**, a transformer architecture that unifies feed forward geometry prediction with differentiable direct image-to-scan alignment, enabling image-to-scan registration without pose initialization or per-point image features in the reference map.
- Through foundational training across diverse sensors and environments, a single DRS-VPT model achieves state-of-the-art camera–LiDAR registration and competitive relocalization against raw point cloud maps, generalizing across domains where existing single-dataset methods fail catastrophically.
- As we show in an ablation, reprojective feature-metric alignment significantly outperforms geometric point-alignment alternatives such as ICP and Kabsch, reducing mean translation error by 5× on camera–LiDAR registration.

2 Related Work

Learned Wide Baseline Image–Scan Alignment. Motivated by online camera-LiDAR calibration fault recovery for autonomous driving, DeepI2P [18] introduced image–scan alignment from wide offsets ($\approx 5\,\mathrm{m}$) without pose initialization. DeepI2P uses a decoupled two-stage design: a cross-modal classifier trained with binary cross-entropy labels each scan point inside or outside the camera frustum, then a separate inverse camera projection solver recovers the pose, with no gradient flowing back to the classifier. Later, CorrI2P [25], RelaI2P [10], and VP2P [49] also employ the two-stage design but instead supervise point-to-pixel correspondences end-to-end, by regressing pose via a differentiable PnP solver. More recently, TrafficLoc [43] additionally applies attention supervision to improve cross-modal feature alignment. Trained independently on KITTI and nuScenes, these methods generalize poorly across datasets—even between the two—a failure mode we confirm in our experiments.

Learned Image–Dense Scan Alignment. A related line of work targets image-to-dense-point-cloud alignment under high overlap. P2-Net [35] jointly detects and describes keypoints in a shared cross-modal descriptor space, without feature fusion between modalities. 2D3D-MATR [19] removes the detection step, matching coarse-to-fine patch descriptors with a detection-free transformer that fuses image and point cloud features. UniCorrn [9] unifies 2D–2D, 2D–3D, and 3D–3D correspondence under a single shared-weight transformer. These methods address a considerably more constrained problem than the camera–LiDAR setting above. The reference point cloud is back-projected from a dense depth map, and image–scan overlap exceeds 50%. Neither of these conditions holds when a narrow-FOV camera registers against a sparse 360° LiDAR sweep.

Image–Scan Alignment in Camera Relocalization. Motivated by camera relocalization against pre-built 3D maps, these methods register query images to 3D scans by embedding image appearance into the map. PixLoc [26] lifts appearance features from mapping images into an SfM reconstruction and retrieves query poses via feature-metric direct alignment. InLoc [29] and Zhang *et al.* [47] instead render synthetic image views—from a posed RGBD database and a TLS XYZRGB map respectively—and match against the query image. All require per-point image appearance baked into the map, precluding localization against geometry-only scans.

Fine Camera–LiDAR Alignment. Accurate camera–LiDAR alignment is fundamental to sensor fusion between these modalities. Where bespoke calibration targets are available, classical methods optimize the joint alignment of plane and line observations across both sensors [50], or the alignment of planar geometry alongside the intensity patterns of the target [6]. In unstructured environments, calibration instead relies on mutual-information maximization [23], intensity–edge matching [17], or online alignment of LiDAR and camera intensity signals [22, 24]. Learning-based approaches [11, 21, 45] directly regress the

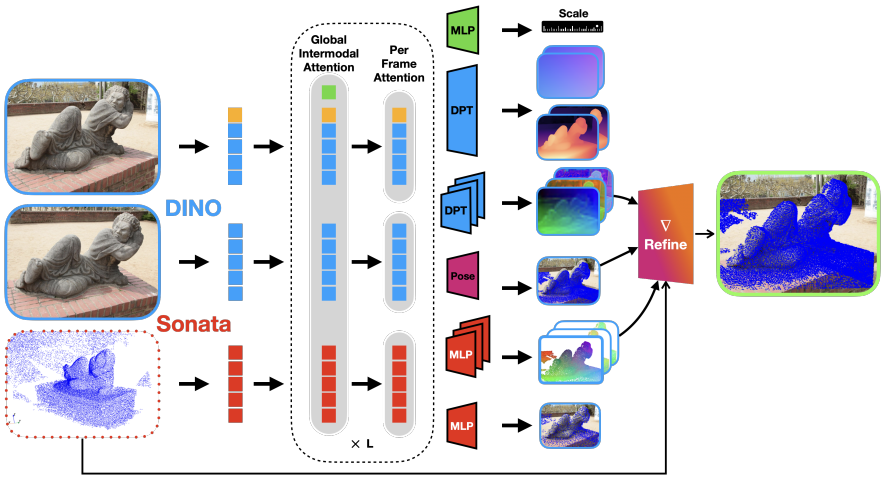

Fig. 2: DRS-VPT architecture. Query images and a reference 3D point cloud are encoded using DINOv2 and a Sonata point transformer respectively. Their tokens are fused using a transformer with alternating global cross-modal attention and modality-specific local attention. From the fused representation the model predicts camera poses, dense point maps, and alignment features for both modalities. The predicted scan pose provides an initial alignment which is refined using differentiable direct image–scan alignment, producing the final cloud pose in camera 1 frame.

camera–LiDAR transformation from paired observations, but exhibit poor generalization across sensor configurations [7]. DXQNet [13] and VoxCal [7] improve on this by coupling differentiable pose optimization into the training loop, allowing pose supervision to shape the learned features. All these methods assume a relatively accurate initial guess pose—they perform local refinement, not recovery from wide offsets.

Feed-Forward Visual Geometry Models. Early convolutional models for joint depth and motion estimation [33, 48] showed limited generalisation, prompting methods to embed geometry-grounded differentiable optimization into training [31,32,37,51]. DUSt3R [38] then demonstrated that dense two-view geometry can be solved end-to-end with a large transformer trained at scale, motivating MASt3R [16] and VGGT [36], which extended this to many views. Pow3R [12] and MapAnything [15] further extend these architectures to ingest ground-truth geometry alongside images, but require the geometry to be frame-aligned to camera views.

3 DRS-VPT

Our end-to-end model (see Figure 2) ingests a set of N RGB images together with a reference 3D point cloud. We denote the image set as $\hat{\mathcal{I}} = \{\hat{I}_i\}_{i=1}^{N}$ and the reference point cloud as $\hat{P}$. All predicted geometric quantities are expressed in the coordinate frame of the first image $\hat{I}_1$, which serves as the reference frame for

the scene. During both training and inference the model operates on a small set of image observations (typically up to four views) jointly with the reference point cloud. Crucially, the reference point cloud is not assumed to be aligned with any of the input images. As a result, the model must infer the alignment between the visual observations and the 3D geometry of the scan while estimating the camera poses of the remaining images relative to $\hat{I}_1$.

Formally, the network implements

$$f_{\text{DRS-VPT}}(\hat{\mathcal{I}}, \hat{P}) = \Big\{ s,\, (Q_i, \tilde{\mathbf{t}}_i, R_i, \tilde{D}_i, W_i^X)_{i=1}^N, \\ (\tilde{X}^P, Q^P, \tilde{\mathbf{t}}^P, W^P), \\ (F_I^{(l)}, F_P^{(l)}, W_I^{(l)}, W_P^{(l)})_{l=1}^3 \Big\}, \tag{1}$$

Here $s \in \mathbb{R}^+$ denotes the predicted global metric scale. For each image i, the model predicts a unit quaternion $Q_i \in \mathbb{H}$ representing camera rotation and an up-to-scale translation $\tilde{\mathbf{t}}_i \in \mathbb{R}^3$ expressed in the coordinate frame of the first image $\hat{I}_1$. The tilde notation $\tilde{(\cdot)}$ indicates quantities defined up to scale.

In addition, the model predicts per-view ray directions $R_i \in \mathbb{R}^{3\times H\times W}$ and up-to-scale ray depths $\tilde{D}_i \in \mathbb{R}^{1\times H\times W}$, from which local point maps may be recovered. For each reconstructed image point, the model also predicts confidence weights $W_i^X \in \mathbb{R}^{H\times W}$ that estimate the reliability of the predicted geometry.

For the reference scan, the model predicts transformed scan points $\tilde{X}^P \in \mathbb{R}^{3\times K}$ together with a quaternion rotation $Q^P \in \mathbb{H}$ and an up-to-scale translation $\tilde{\mathbf{t}}^P \in \mathbb{R}^3$ describing the pose of the scan relative to the first camera $\hat{I}_1$. The model additionally predicts per-point confidence weights $W^P \in \mathbb{R}^K$ that capture the reliability of each transformed scan point. The predicted scan pose provides an initialization for subsequent direct image–scan alignment.

To refine the scan's pose, the network also predicts dense features for both modalities. We produce a three-level pyramid of coarse-to-fine alignment features $\{F_I^{(l)}\}_{l=1}^3$ and $\{F_P^{(l)}\}_{l=1}^3$ for the first image and point cloud branches respectively, alongside corresponding feature confidence maps $\{W_I^{(l)}\}_{l=1}^3$ and $\{W_P^{(l)}\}_{l=1}^3$. These weighted feature pyramids are used for differentiable direct alignment between the reference image and the 3D scan.

3.1 Encoding Images and Point Clouds

Given a set of N images and a reference point cloud, DRS-VPT encodes both modalities into a shared latent space of dimension 1024. For image inputs, we use the same frozen DINOv2 encoder as MapAnything [15]. Specifically, we extract the final-layer patch features from DINOv2 ViT-L, producing

$$F_I \in \mathbb{R}^{1024\times H/14\times W/14}, \tag{2}$$

which serve as the per-patch image tokens used throughout the model.

For the reference point cloud, we employ the Sonata variant of PointTransformerV3 [40, 41]. We use adaptive voxel downsampling with grid sizes ranging from 5 mm to 10 cm depending on the dataset and scene extent. For non-metric datasets such as MegaDepth, the grid size is determined using the median nearest-neighbour distance of the point cloud. To accommodate scenes of varying scale, the downsampled point cloud $\hat{P} \in \mathbb{R}^{3\times K}$ is normalized by the scene scale (logarithm of the mean point range). Sonata is adapted to produce features in the same latent dimension (1024). The resulting encoder produces a set of per-point features

$$F_P = \{f_k^{\text{pt}}\}_{k=1}^{K}, \quad f_k^{\text{pt}} \in \mathbb{R}^{1024}. \tag{3}$$

To help regress the original scene scale, the normalization term is encoded and added to the scan tokens via a learned gate initialised to zero.

3.2 Cross-modal Attention

The encoded tokens are processed by a transformer composed of alternating global and local attention layers. An extra learnable scale token participates in global attention to propagate metric information across the image and point cloud representations. A single *base-frame embedding* is added to the tokens of the first image $\hat{I}_1$, defining the reference coordinate frame and anchoring the representation for subsequent geometric prediction. We adopt a 24-layer alternating-attention architecture similar to MapAnything [15], consisting of 12 global and 12 local attention blocks with a latent dimension of 768 (projected down from 1024 at input), 12 attention heads, and an MLP ratio of 4. The transformer weights are initialized from the corresponding MapAnything model and subsequently fine-tuned for the cross-modal registration task.

To maintain tractable attention complexity, the encoded point cloud tokens are uniformly downsampled to 5,000 tokens in Sonata's space-filling curve order before entering the transformer. During the global attention blocks, all tokens from both modalities and the scale token attend jointly. In the alternating local attention blocks, the modalities are processed independently. Point cloud tokens perform self-attention within the cloud branch and image tokens perform self-attention within each image view independently.

After the transformer layers, we fuse the output cloud tokens with their respective original high-resolution point features using a learned gating mechanism. For each point, the final feature representation is computed as

$$f_k^{\text{out}} = \sigma(g_k)\, f_k^{\text{ctx}} + (1 - \sigma(g_k))\, f_k^{\text{local}}, \tag{4}$$

where f_k^{ctx} denotes the contextual feature obtained from the transformer tokens, f_k^{local} is the original point encoder feature, and $\sigma(g_k)$ is a learned gating weight predicted from the concatenated features. The gated fusion is specifically valuable because it recovers high-resolution per-point features that the transformer would otherwise lose due to the 5,000-token downsampling, with performance degrading when fewer points are available (see Tab. 6).

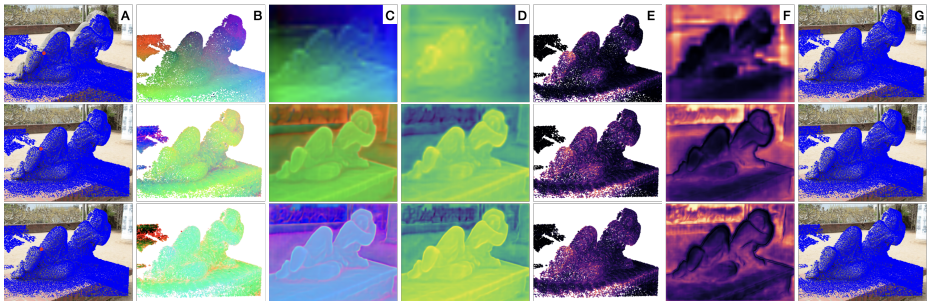

Fig. 3: Coarse-to-fine Direct Alignment. The cloud pose estimate (A) is refined by aligning the cloud dense features (B) to the image features (C). (A) in the first row shows the pose head output for the cloud pose in camera 1 frame. Notice that it is slightly misaligned. The heatmap in (D) shows the cosine similarity of the queried cloud point marked in red in (A) to the feature at each pixel in (C). Note that this heatmap evolves from global saliency at the coarsest level (first row), to sharper, more local saliency at finer levels. (E) and (F) show the feature weights of the cloud and image features respectively. The post-alignment overlay for each pyramid level is shown in (G). See also Figure 4 for a more detailed example on ETH3D.

3.3 Output Reconstruction

For each image view i, the model predicts ray directions R_i, up-to-scale ray depths $\tilde{D}_i$, camera rotation Q_i, and up-to-scale translation $\tilde{\mathbf{t}}_i$, all expressed in the coordinate frame of the first image $\hat{I}_1$. The image branch uses the same DPT-based depth and convolutional pose heads as MapAnything [15]. For the scan output, a lightweight linear decoder maps each upsampled cloud token (see Equation 4) to transformed coordinates $\tilde{X}^P$ in the frame of $\hat{I}_1$, alongside per-point confidence weights. A global pose head additionally predicts scan rotation Q^P and up-to-scale translation $\tilde{\mathbf{t}}^P$, providing a coarse initialization for the direct image–scan alignment stage. The global metric scale s is predicted using the scale token via a two-layer MLP with exponential output to ensure positivity.

From the first image and point cloud tokens, we decode three-level coarse-to-fine feature pyramids for cross-modal alignment. Image features (dimension 24 at each level) are decoded using DPT heads with upsampling skip connections; point cloud features are decoded using a lightweight MLP with skip connections. Both branches also output per-element confidence weights. These pyramids drive the differentiable direct image–scan alignment described next.

3.4 Unrolled Direct Image–Scan Alignment Refinement

While the network predicts a decent initial pose for the reference scan, accurate cross-modal localization benefits from fine direct alignment as illustrated by the coarse-to-fine optimization process in Fig. 3.

Let $\hat{\mathbf{P}}_j$ denote the input 3D measurements from the reference scan. These points are not assumed to be aligned with the camera coordinate frame and

remain in their original sensor coordinates. Instead, the predicted global scale s is used to rescale the translation component of the predicted scan pose $(Q^P, \tilde{\mathbf{t}}^P)$, yielding the initialization $\mathbf{t}^P = s\,\tilde{\mathbf{t}}^P$ while the rotation is given by $R(Q^P) \in SO(3)$. This scaled pose provides the initial alignment of the scan relative to the first camera view.

Since the predicted scan pose is expressed in camera 1 coordinates, each scan point $\hat{\mathbf{P}}_j$ is projected into the image of camera 1 via the standard projection $\mathbf{u}_{ij} = \Pi\Big(R(Q^P)\,\hat{\mathbf{P}}_j + \mathbf{t}^P\Big)$, given known intrinsics. Following PixLoc [26], at pyramid level p we compute a feature residual between the point-cloud feature and the image feature bilinearly sampled at the projected pixel location, $\mathbf{r}_j^p = \mathbf{F}_{P,j}^p - \mathbf{F}_I^p(\mathbf{u}_{ij}) \in \mathbb{R}^{D_p}$, where $\mathbf{F}_{P,j}^p$ is the feature of point j in the point-cloud branch and $\mathbf{F}_I^p(\cdot)$ is the image feature map evaluated at the projected location. The alignment energy at level p is then $E_p = \sum_j w_j^P\, w_j^I\, \rho\big(\|\mathbf{r}_j^p\|^2\big)$, where w_j^P and w_j^I are learned confidence weights for the point cloud and image features respectively, and $\rho(\cdot)$ is a robust penalty to suppress outliers. The product of confidence weights allows the optimizer to jointly downweight unreliable geometric observations and ambiguous image regions. Remarkably, the point-cloud weights w_j^P often learn to downweight geometrically occluded points, such as scan measurements lying behind visible surfaces from the current camera viewpoint (see Figure 5).

To optimize the pose, we linearize the residuals with respect to an incremental pose update $\boldsymbol{\xi} \in \mathbb{R}^6$ in the Lie algebra of SE(3). Stacking the residuals across all points and feature channels yields the Jacobian $J_{k+(j-1)D_p} = \frac{\partial r_{j,k}^p}{\partial \boldsymbol{\xi}}$, where $r_{j,k}^p$ denotes the k-th feature channel of the residual $\mathbf{r}_j^p$. Let W denote a diagonal weight matrix containing the combined confidence weights and robust kernel derivatives. The Gauss–Newton approximation to the Hessian is $H = J^\top W J$, and the pose increment is obtained by solving the normal equations,

$$H\,\boldsymbol{\xi} = -J^\top W \mathbf{r}, \tag{5}$$

and applying the update via the SE(3) exponential map,

$$\begin{bmatrix} R_{i+1} & \mathbf{t}_{i+1} \\ 0 & 1 \end{bmatrix} = \exp(\widehat{\boldsymbol{\xi}})^\top \begin{bmatrix} R_i & \mathbf{t}_i \\ 0 & 1 \end{bmatrix}. \tag{6}$$

We unroll a fixed number of Gauss–Newton iterations (typically five) at each pyramid level and proceed from coarse to fine across the feature hierarchy. The final supervision is applied using a reprojection loss between the predicted and ground-truth poses,

$$\mathcal{L}_{\text{reproj}} = \sum_p \sum_j \rho\left(\left\| \Pi^p\Big(\bar{R}\,\hat{\mathbf{P}}_j + \bar{\mathbf{t}}\Big) - \Pi^p\Big(R_M^p\,\hat{\mathbf{P}}_j + \mathbf{t}_M^p\Big)\right\|_2^2\right), \tag{7}$$

where $(\bar{R}, \bar{\mathbf{t}})$ denotes the ground-truth pose and $(R_M^p, \mathbf{t}_M^p)$ denotes the pose after the final unrolled optimization step at pyramid level p. This coarse-to-fine optimization allows the model to refine the initial scan alignment and achieve accurate image-to-scan localization.

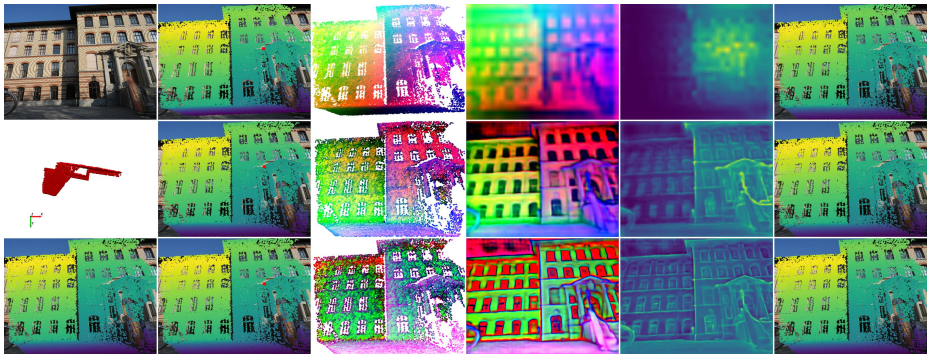

Fig. 4: Coarse-to-fine direct alignment on ETH3D Facade. The first column shows the model's input image, the input misaligned cloud, and the final aligned overlay. Each subsequent column corresponds to one pyramid level, showing (top to bottom): the cloud overlay at the current estimate with a query point in red, point cloud features, image features, the query point's cross-modal similarity heatmap, and the overlay after refinement. The starting pose at the first level (top row) is produced by the transformer's scan pose head.

3.5 Training Loss Formulation

We supervise the predicted ray directions R_i, rotations Q_i, translations $\tilde{\mathbf{t}}_i$, depths $\tilde{D}_i$, local point maps $\tilde{L}_i$, world-frame point maps $\tilde{X}_i$, and confidence maps C_i against their ground-truth counterparts $\hat{R}_i$, $\hat{Q}_i$, $\hat{\mathbf{t}}_i$, $\hat{D}_i$, $\hat{L}_i$, and $\hat{X}_i$.

The ray and rotation losses are computed using ℓ_2 distances,

$\mathcal{L}_{\text{rays}} = \sum_i \|\hat{R}_i - R_i\|_2$ and $\mathcal{L}_{\text{rot}} = \sum_i \min(\|\hat{Q}_i - Q_i\|_2, \| - \hat{Q}_i - Q_i\|_2)$, which accounts for the two-to-one symmetry of unit quaternions. Translations are optimized in a scale-invariant manner, $\mathcal{L}_{\text{translation}} = \sum_i \left\| \frac{\hat{\mathbf{t}}_i}{\hat{z}} - \frac{\tilde{\mathbf{t}}_i}{\tilde{z}} \right\|_2$, where $\hat{z}$ and $\tilde{z}$ denote the ground-truth and predicted scene scales.

We train depth and point-map predictions in log-space using,
$f_{\log}(x) = \frac{x}{\|x\|} \log(1 + \|x\|)$. The corresponding losses are
$\mathcal{L}_{\text{depth}} = \sum_i \|f_{\log}(\hat{D}_i/\hat{z}) - f_{\log}(\tilde{D}_i/\tilde{z})\|_2$, $\mathcal{L}_{\text{lpm}} = \sum_i \|f_{\log}(\hat{L}_i/\hat{z}) - f_{\log}(\tilde{L}_i/\tilde{z})\|_2$, and the confidence-weighted point-map loss,
$\mathcal{L}_{\text{pointmap}} = \sum_i C_i \|f_{\log}(\hat{X}_i/\hat{z}) - f_{\log}(\tilde{X}_i/\tilde{z})\|_2 - \alpha \log C_i$.

The global metric scale s is optimized using $\mathcal{L}_{\text{scale}} = \|f_{\log}(\hat{z}) - f_{\log}(z_{\text{metric}})\|_2$, where $z_{\text{metric}} = s \cdot \text{sg}(\tilde{z})$, combines the predicted scale s with a detached normalization term $\tilde{z}$ to prevent gradient coupling through scale.

Confidence-weighted point-map supervision is applied to both image and point cloud outputs. Scale supervision is included only for datasets with metric ground truth.

The final training objective combines geometric supervision with the reprojection loss from the differentiable alignment stage,

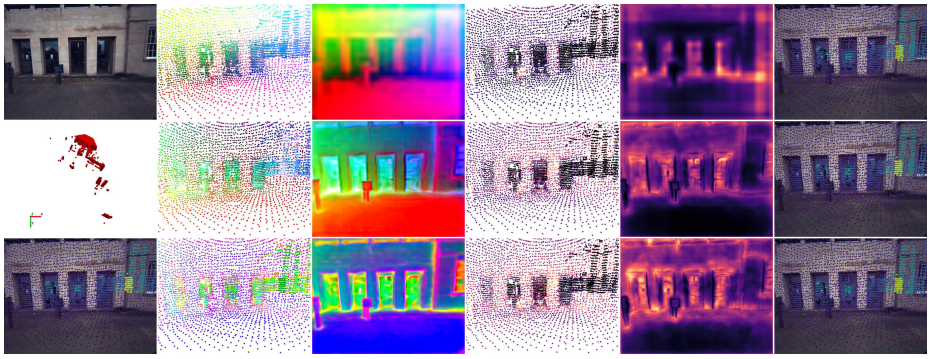

Fig. 5: Learned feature saliency and confidence weighting on Oxford Spires. The model assigns high saliency to geometrically distinctive structures while the learned confidence weights (columns 4) suppress scan points that are occluded from the current camera viewpoint (pillars behind the facade of the building shown as dark streaks).

$$\begin{aligned}\mathcal{L}_{\text{total}} =& 10\,\mathcal{L}_{\text{pointmap}} + \mathcal{L}_{\text{rays}} + \mathcal{L}_{\text{rot}} + \mathcal{L}_{\text{translation}} \\ &+ \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{lpm}} + \mathcal{L}_{\text{scale}} + \mathcal{L}_{\text{reproj}}.\end{aligned} \tag{8}$$

3.6 Datasets and Multi-View Sampling

We train DRS-VPT on a diverse collection of multi-modal datasets spanning indoor, outdoor, and large-scale driving environments: MegaDepth [20], OxfordSpires [30], KITTI [8], ETH3D [27], Argoverse [39], nuScenes [3], PandaSet [44], 7-Scenes [28], ScanNet++ [46], WildRGBD [42], and GrandTour [4, 5]. These datasets provide a wide range of sensing modalities including structure-from-motion reconstructions, LiDAR scans, and dense laser-scanner depth maps. Training across this heterogeneous collection enables the model to generalize across both metric and up-to-scale geometric regimes as well as across different sensor configurations.

Datasets. For **MegaDepth**, we use the publicly available structure-from-motion reconstructions, treating the recovered 3D points as up-to-scale supervision for the point cloud branch. **OxfordSpires** provides synchronized camera–LiDAR data collected in outdoor environments; we use sparse LiDAR returns within the camera frustums for geometric supervision. In **KITTI**, **Argoverse**, **nuScenes**, and **PandaSet**, the model ingests full LiDAR scans captured by autonomous driving platforms, providing the low-overlap alignment challenge between narrow-FOV cameras and full 360° vehicle-mounted LiDAR sweeps. **ETH3D**, **ScanNet++**, and **7-Scenes**, **WildRGBD** provide dense indoor reconstructions obtained from structured-light, depth camera or laser scanning systems, which we treat as depth-derived point clouds associated with the corresponding camera views. Finally, the **GrandTour** dataset contributes camera–LiDAR sequences

Table 1: Camera–LiDAR alignment under wide initialization. For each dataset we report translation t (in meters) and rotation r (in degrees) errors, shown as mean followed by median. Accuracy is the percentage of predictions within 2 m translation and 5° rotation error.

	KITTI			**nuScenes**		
Method	t [m]	r [°]	Acc	t [m]	r [°]	Acc
CorrI2P [25]	3.78 / –	5.89 / –	72.42	3.04 / –	3.73 / –	49.00
VP2P [49]	1.91 / 1.23	5.17 / 2.69	64.13	0.89 / –	2.15 / –	88.33
CoFiI2P [14]	0.32 / 0.27	1.27 / 1.06	99.41	1.21 / 0.75	2.54 / 1.70	90.87
TrafficLoc [43]	0.17 / 0.14	0.82 / 0.55	99.18	2.26 / 0.23	5.68 / 1.63	90.06
DRS-VPT (Ours)	**0.09** / **0.05**	**0.34** / **0.25**	**99.73**	**0.33** / **0.13**	**0.67** / **0.33**	**97.76**

from a robot-mounted platform across deployment-relevant in-the-wild settings such as forests and warehouses, providing the model with registration experience in the unstructured environments where robotic systems must operate.

Multi-View Sampling. To ensure consistent geometric overlap during training, we adopt a covisibility-based sampling strategy similar to MapAnything [15]. For each dataset, we precompute pairwise frame overlaps using reprojection-based visibility checks derived from ground-truth poses or depth maps. Training samples are constructed such that all selected image views share at least 25% geometric overlap with the reference scan.

During training, we sample connected sets of N covisible images together with a reference point cloud. The reference scan may correspond either to a LiDAR frame or to a reconstructed point cloud associated with one of the views. To encourage generalizability, we apply random rotation and translation augmentations to the input scan during training.

For datasets with smaller-scale scenes (e.g., WildRGBD, and ScanNet++), we occasionally aggregate several nearby point clouds into a single reference frame before presenting them to the model. This produces larger map-like scans while preserving the single-scan input formulation of the network. For 7-Scenes, all posed depths are aggregated into a map frame and downsampled to train relocalization. Consequently, the model is consistently trained in an image-to-scan relocalization setting rather than in a local reconstruction configuration.

4 Experiments

We evaluate DRS-VPT on the task of image-to-scan relocalization across both outdoor autonomous driving and indoor RGB-D environments. Unlike most prior relocalization approaches, which train scene-specific models tied to a fixed map representation, DRS-VPT directly ingests a 3D scan and image at inference time and predicts camera poses relative to that scan.

We evaluate the method in three complementary settings. First, we measure wide-baseline camera–LiDAR alignment accuracy on two autonomous driving

datasets — KITTI and nuScenes. Second, we evaluate indoor camera relocalization on the 7Scenes benchmark, where DRS-VPT dynamically ingests the scene map instead of learning a dedicated scene model. Finally, we test cross-dataset generalization by localizing images in completely unseen environments without fine-tuning.

Together, these experiments evaluate in-domain performance as well as the ability of a unified multi-modal transformer to generalize across different sensors, scene scales, and environmental statistics.

4.1 Autonomous Driving Alignment

We first evaluate wide-baseline camera–LiDAR alignment on KITTI and nuScenes. Following prior work, we report translation and rotation errors together with the percentage of predictions within 2 m translation and 5° rotation thresholds.

For this experiment, we use KITTI odometry sequences 9 and 10 for the wide-baseline localization task. None of these sequences are included in our training set, making this a genuine held-out evaluation.

In the wide-baseline setting, we perturb the LiDAR point cloud with a random rigid transformation of up to ± 10 m in the x–z plane and an arbitrary rotation about the vertical (y) axis, following the standard evaluation protocol of prior camera–LiDAR registration methods [25, 49]. This yaw-and-translation protocol reflects the dominant degrees of freedom in ground-vehicle operation; the model receives no initial pose and must recover the full alignment globally from the perturbed input.

Table 1 shows that DRS-VPT achieves state-of-the-art performance on both datasets. Despite operating with a unified architecture that directly processes point clouds and images, our method outperforms specialized correspondence-based methods such as CorrI2P, VP2P, CoFiI2P, and TrafficLoc. Notably, the model achieves the lowest translation and rotation errors on KITTI while maintaining strong performance on nuScenes, demonstrating that the learned geometric representation transfers effectively across different autonomous driving sensor configurations.

4.2 Indoor Camera–Map Localization

We evaluate indoor relocalization on the 7Scenes dataset, a standard benchmark for camera pose estimation in RGB-D environments. Unlike most existing methods, which train scene-specific neural models that encode a map representation during training, DRS-VPT directly ingests a 3D geometric map at inference time and predicts the camera pose relative to that map.

For this experiment, 3D maps are constructed from the RGB-D sequences provided with the dataset. 3D maps are constructed from the training sequences of each scene; at evaluation time, query images are drawn from the held-out test split and localized against these training-sequence maps.

Table 2: Pose estimation results on the 7Scenes dataset (indoor). Values are translation error (cm) / rotation error (deg). The last column reports overall recall (%).

Method	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Recall↑
InLoc [29]	3/1.05	3/1.07	2/1.16	3/1.05	5/1.55	4/1.31	9/2.47	66.3
DSAC* [2]	**2**/1.10	**2**/1.24	**1**/1.82	3/1.15	**4**/1.34	4/1.68	**3**/**1.16**	**85.2**
PixLoc [26]	**2**/0.80	**2**/**0.73**	**1**/0.82	3/**0.82**	**4**/1.21	**3**/**1.20**	5/1.30	75.7
ACE [1]	**2**/**0.68**	**2**/0.87	**1**/**0.67**	3/0.84	**4**/**1.15**	4/1.34	5/1.25	75.6
DRS-VPT (ours)	3/0.84	**2**/1.01	2/1.00	3/0.88	7/1.67	5/1.30	27/3.26	66.4

Table 2 compares our approach to several established relocalization systems. Methods such as DSAC* and ACE learn scene-specific neural scene representations, while PixLoc performs feature-based alignment between query images and reference images stored in a map. In all cases, these baselines operate entirely within the image domain, matching pixels between query and reference views captured by the same camera and relying on scene-specific training to encode the map structure.

In contrast, DRS-VPT localizes query images directly against a 3D map constructed from depth observations, introducing a cross-modal alignment problem between RGB images and point clouds. Despite this additional modality gap and the absence of scene-specific training, our method remains competitive with existing approaches, demonstrating that the learned cross-modal features provide sufficient geometric signal for reliable localization.

Performance on the *Stairs* scene is noticeably lower than on other scenes. This environment contains highly repetitive staircase structures that produce strong geometric aliasing, making pose disambiguation difficult even for specialized relocalization systems.

These results suggest that dynamic map ingestion may provide a viable alternative to scene-specific relocalization models. We further validate this hypothesis in the next experiment, where we localize images in a completely unseen indoor dataset using only the provided 3D map, demonstrating strong cross-environment generalization.

4.3 Zero-Shot Indoor Relocalization

To further evaluate cross-environment generalization, we test DRS-VPT on the 12Scenes [34] dataset, which contains indoor environments similar to 7Scenes but was never used during training. As in the previous experiment, the model localizes query images against 3D maps constructed from depth-camera observations of each scene.

Table 3 reports aggregated results across subscenes, showing median translation and rotation errors together with recall under the standard 5 cm and 5° accuracy threshold. Despite never observing these environments during training,

Table 3: DRS-VPT results on Twelve Scenes (scene-level aggregation).

	apt1	apt2	office1	office2
Trans (cm)	4	5	4	7
Rot (deg)	1.92	2.23	1.82	2.78
Recall (%)	56.8	48.9	62.0	33.7

Table 4: KITTI → nuScenes generalization for image-to-point alignment.

Method	t [m]	r [°]	Recall (%)
TrafficLoc	16.55	95.55	1.9
CoFiI2P	17.46	93.40	2.9
DRS-VPT	**0.538**	**1.14**	**66.74**

DRS-VPT achieves consistent localization across all scenes with an overall recall of 53.6%.

Although recall is lower than on 7Scenes due to the domain shift and the absence of scene-specific training, the results demonstrate that the model can successfully localize images in completely unseen environments using only a provided geometric map. This experiment highlights the ability of DRS-VPT to generalize beyond the training distribution and operate in a true zero-shot relocalization setting.

4.4 Cross-Dataset Generalization

Finally, we evaluate cross-dataset generalization in the autonomous driving setting. For this experiment, baseline methods are trained on KITTI and evaluated directly on nuScenes, following the protocol used in prior work. For DRS-VPT, we train a single model on our full multi-dataset training corpus while explicitly excluding nuScenes.

Table 4 reports localization performance on nuScenes. Methods trained only on KITTI fail to generalize under the significant domain shift between the two datasets, resulting in extremely large pose errors and near-zero recall. In contrast, DRS-VPT maintains accurate alignment with a median translation error of 0.538 m and rotation error of 1.14°, achieving a recall of 66.74%.

These results demonstrate that the geometric representations learned by DRS-VPT transfer effectively across different autonomous driving platforms and sensor configurations. More broadly, the ability to localize images in previously unseen environments using only a provided 3D map suggests that dynamic map ingestion can serve as a practical alternative to scene-specific localization pipelines.

4.5 Ablation: Refinement Strategy and Point Density

We also explored geometric alignment on predicted point maps as an alternative refinement strategy. Tab. 5 compares Iterative Closest Point (ICP, aligning the camera-view point map to the input cloud) and Kabsch (aligning the input cloud to the cloud-head point map). While these methods improve rotation, neither significantly reduces translation error. We observe that point-based alignment demands globally accurate metric depth from the model—a fundamentally harder regression target than learning cross-modal feature alignment. Feature-metric direct alignment instead exploits discontinuities in learned cross-modal

features at surface boundaries, a signal the end-to-end training naturally produces, reducing mean translation error by 5×. Tab. 6 further shows performance improving with point density, demonstrating the benefits of the gated fusion mechanism (see Equation 4) in recovering high-resolution per-point detail.

Table 5: Comparison of refinement strategies on KITTI.

Method	t [m]	r [°]	Acc
Raw pose head	0.50 / 0.44	1.45 / 1.38	99.64
+ ICP refinement	0.52 / 0.45	1.25 / 1.12	99.57
+ Kabsch alignment	0.61 / 0.52	0.94 / 0.83	98.82
+ **Feature-metric**	**0.09** / **0.05**	**0.34** / **0.25**	**99.73**

Table 6: Robustness of GN refinement to point cloud density on KITTI.

N Points	t [m]	r [°]	Acc
5,000	0.14 / 0.09	0.46 / 0.35	99.64
10,000	0.10 / 0.06	0.35 / 0.29	99.71
20,000	0.09 / 0.06	0.34 / 0.25	99.71
50,000	**0.08** / **0.05**	**0.32** / **0.23**	**99.71**

5 Discussion and Future Work

We presented DRS-VPT, a feed-forward transformer for image-to-scan relocalization that unifies geometric prediction and differentiable direct alignment in a single architecture. Unlike prior methods that explicitly classify frustum overlap before solving for correspondences, DRS-VPT learns to identify relevant scan regions implicitly through the reprojective inductive bias of the alignment objective (see Figure 1). Strikingly, the learned confidence weights suppress back-facing, occluded scan points without any explicit occlusion supervision (see Figure 3), evidence that the reprojective objective instils physically meaningful geometric understanding. Across autonomous driving and indoor benchmarks, DRS-VPT achieves state-of-the-art wide-baseline camera–LiDAR alignment, remains competitive with specialized indoor relocalization systems, and generalizes well to unseen environments.

Several open directions remain. The current formulation assumes a single rigid reference scan, therefore, extending to multi-scan or hierarchical map representations would broaden applicability to large-scale environments. Handling articulated and dynamic objects would open applications in robot state tracking. Additionally, generalizing to allowing multiple output instances per point cloud, unlocks object-level instance detection as a downstream task.

Acknowledgements

This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200 — EUROfusion) and from the EPSRC [grant number EP/W006839/1. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. The authors thank Matias Mattamala, Edgar Sucar, Eldar Insafutdinov and Benjamin Ramtoula for valuable discussions on earlier versions of this work, and Emil Jonasson and Yifu Tao for their careful review of the manuscript.

References

1. Brachmann, E., Cavallari, T., Prisacariu, V.A.: Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5044–5053 (2023)
2. Brachmann, E., Rother, C.: Visual Camera Re-Localization From RGB and RGB-D Images Using DSAC . IEEE Transactions on Pattern Analysis & Machine Intelligence **44**(09), 5847–5865 (2022). https://doi.org/10.1109/TPAMI.2021.3070754
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
4. Frey, J., Tuna, T., Fu, F., Patterson, K., Xu, T., Fallon, M., Cadena, C., Hutter, M.: Grandtour: A legged robotics dataset in the wild for multi-modal perception and state estimation (2026), https://arxiv.org/abs/2602.18164, *Equal contribution (Turcan Tuna and Jonas Frey).
5. Frey, J., Tuna, T., Fu, L.F.T., Weibel, C., Patterson, K., Krummenacher, B., Müller, M., Nubert, J., Fallon, M., Cadena, C., Hutter, M.: Boxi: Design Decisions in the Context of Algorithmic Performance for Robotics. In: Proceedings of Robotics: Science and Systems. Los Angeles, United States (2025), *Equal contribution (Jonas Frey and Turcan Tuna and Frank Fu).
6. Fu, L.F.T., Chebrolu, N., Fallon, M.: Extrinsic calibration of camera to lidar using a differentiable checkerboard model. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1825–1831 (2023). https://doi.org/10.1109/IROS55552.2023.10341781
7. Fu, L.F.T., Fallon, M.: Batch differentiable pose refinement for in-the-wild camera/lidar extrinsic calibration. In: Tan, J., Toussaint, M., Darvish, K. (eds.) Proceedings of The 7th Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 229, pp. 1362–1377. PMLR (2023)
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012). https://doi.org/10.1109/CVPR.2012.6248074
9. Goswami, P., Ding, T., Liu, F., Jiang, H.: Unicorrn: Unified correspondence transformer across 2d and 3d. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9943–9954 (2026)
10. Hou, M., Wang, G., Wang, Z., Ma, B.: Relai2p: Relational learning for image-to-point cloud registration. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10887630
11. Iyer, G., Ram, R.K., Jatavallabhula, K.M., Krishna, K.M.: 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)
12. Jang, W., Weinzaepfel, P., Leroy, V., Agapito, L., Revaud, J.: Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1071–1081 (2025)

13. Jing, X., Ding, X., Xiong, R., Deng, H., Wang, Y.: Dxq-net: Differentiable lidar-camera extrinsic calibration using quality-aware flow. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6235–6241 (2022). https://doi.org/10.1109/IROS47612.2022.9981418
14. Kang, S., Liao, Y., Li, J., Liang, F., Li, Y., Zou, X., Li, F., Chen, X., Dong, Z., Yang, B.: CoFiI2P: Coarse-to-fine correspondences-based image to point cloud registration. IEEE Robotics and Automation Letters **9**(11), 10264–10271 (2024). https://doi.org/10.1109/LRA.2024.3466068
15. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. In: 2026 International Conference on 3D Vision (3DV). pp. 499–509. IEEE (2026)
16. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXII. pp. 71–91. Springer-Verlag (2024). https://doi.org/10.1007/978-3-031-73220-1_5
17. Levinson, J., Thrun, S.: Automatic online calibration of cameras and lasers. In: Robotics: science and systems (2013). https://doi.org/10.15607/RSS.2013.IX.029
18. Li, J., Lee, G.H.: Deepi2p: Image-to-point cloud registration via deep classification. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 15955–15964 (2021), https://api.semanticscholar.org/CorpusID:233181563
19. Li, M., Qin, Z., Gao, Z., Yi, R., Zhu, C., Guo, Y., Xu, K.: 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14128–14138 (2023)
20. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
21. Lv, X., Wang, B., Dou, Z., Ye, D., Wang, S.: Lccnet: Lidar and camera self-calibration using cost volume network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 2894–2901 (2021)
22. Napier, A., Corke, P., Newman, P.: Cross-calibration of push-broom 2d lidars and cameras in natural scenes. In: 2013 IEEE International Conference on Robotics and Automation. pp. 3679–3684 (2013). https://doi.org/10.1109/ICRA.2013.6631094
23. Pandey, G., McBride, J.R., Savarese, S., Eustice, R.M.: Automatic targetless extrinsic calibration of a 3d lidar and camera by maximizing mutual information. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. p. 2053–2059. AAAI'12 (2012). https://doi.org/10.1609/aaai.v26i1.8379
24. Pascoe, G., Maddern, W.P., Newman, P.: Direct visual localisation and calibration for road vehicles in changing city environments. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW) pp. 98–105 (2015), https://api.semanticscholar.org/CorpusID:14372564
25. Ren, S., Zeng, Y., Hou, J., Chen, X.: Corri2p: Deep image-to-point cloud registration via dense correspondence. IEEE Transactions on Circuits and Systems for Video Technology **33**(3), 1198–1208 (2023). https://doi.org/10.1109/TCSVT.2022.3208859

26. Sarlin, P.E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Hammarstrand, L., Kahl, F., et al.: Back to the feature: Learning robust camera localization from pixels to pose. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3247–3257 (2021)
27. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
28. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in rgb-d images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2930–2937 (2013). https://doi.org/10.1109/CVPR.2013.377
29. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: InLoc: Indoor Visual Localization with Dense Matching and View Synthesis . IEEE Transactions on Pattern Analysis & Machine Intelligence **43**(04), 1293–1307 (2021). https://doi.org/10.1109/TPAMI.2019.2952114
30. Tao, Y., Muñoz Bañón, M.A., Zhang, L., Wang, J., Fu, L.F.T., Fallon, M.: The oxford spires dataset: Benchmarking large-scale lidar-visual localisation, reconstruction and radiance field methods. Int. J. Rob. Res. **45**(6), 839–857 (2026). https://doi.org/10.1177/02783649251369905
31. Teed, Z., Deng, J.: Deepv2d: Video to depth with differentiable structure from motion (2020), https://arxiv.org/abs/1812.04605
32. Teed, Z., Deng, J.: Droid-slam: deep visual slam for monocular, stereo, and rgb-d cameras. In: Proceedings of the 35th International Conference on Neural Information Processing Systems (2021)
33. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5622–5631 (2017). https://doi.org/10.1109/CVPR.2017.596
34. Valentin, J., Dai, A., Niessner, M., Kohli, P., Torr, P., Izadi, S., Keskin, C.: Learning to navigate the energy landscape. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 323–332 (2016). https://doi.org/10.1109/3DV.2016.41
35. Wang, B., Chen, C., Cui, Z., Qin, J., Lu, C.X., Yu, Z., Zhao, P., Dong, Z., Zhu, F., Trigoni, N., Markham, A.: P2-net: Joint description and detection of local features for pixel and point matching. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15984–15993 (2021). https://doi.org/10.1109/ICCV48922.2021.01570
36. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025) — *Best Paper Award* (2025). https://doi.org/10.48550/arXiv.2503.11651, https://arxiv.org/abs/2503.11651
37. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21686–21697 (2024)
38. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20697–20709 (2024)
39. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Carr, P., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In:

Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021) (2021)
40. Wu, X., DeTone, D., Frost, D., Shen, T., Xie, C., Yang, N., Engel, J., Newcombe, R., Zhao, H., Straub, J.: Sonata: Self-supervised learning of reliable point representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22193–22204 (2025)
41. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler faster stronger. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4840–4851 (2024)
42. Xia, H., Fu, Y., Liu, S., Wang, X.: Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22378–22389 (2024)
43. Xia, Y., Lu, Y., Song, R., Dhaouadi, O., Henriques, J.F., Cremers, D.: Trafficloc: Localizing traffic surveillance cameras in 3d scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 28685–28695 (2025)
44. Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., Li, Z., Wu, J., Sun, K., Jiang, K., Wang, Y., Yang, D.: Pandaset: Advanced sensor suite dataset for autonomous driving. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). p. 3095–3101. IEEE Press (2021). https://doi.org/10.1109/ITSC48978.2021.9565009
45. Ye, C., Pan, H., Gao, H.: Keypoint-based lidar-camera online calibration with robust geometric network. IEEE Transactions on Instrumentation and Measurement **71**, 1–11 (2022). https://doi.org/10.1109/TIM.2021.3129882
46. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
47. Zhang, L., Tao, Y., Lin, J., Zhang, F., Fallon, M.: Visual localization in 3d maps: comparing point cloud, mesh, and nerf representations. Auton. Robots **50**(1) (2026). https://doi.org/10.1007/s10514-025-10232-5
48. Zhou, H., Ummenhofer, B., Brox, T.: Deeptam: Deep tracking and mapping. In: Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI. p. 851–868. Springer-Verlag, Berlin, Heidelberg (2018). https://doi.org/10.1007/978-3-030-01270-0_50
49. Zhou, J., Ma, B., Zhang, W., Fang, Y., Liu, Y.S., Han, Z.: Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
50. Zhou, L., Li, Z., Kaess, M.: Automatic extrinsic calibration of a camera and a 3d lidar using line and plane correspondences. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5562–5569 (2018). https://doi.org/10.1109/IROS.2018.8593660
51. Zhou, T., Brown, M., Snavely, N., Lowe, D.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6612–6619 (2017). https://doi.org/10.1109/CVPR.2017.700