Towards a Unified Latent Space for Metric Correspondence Learning

Lukas Beer, Anton Backhaus, Thorsten Luettel, Mirko Maehlisch *

Abstract: Learning metric correspondences across heterogeneous sensor observations is a key requirement for robust Bird's-Eye View (BEV) perception in autonomous driving. This paper studies uni-, cross-, and multi-modal keypoint matching as a unified latent-space problem, where camera and LiDAR observations are projected into a common BEV representation and matched through modality-agnostic keypoints and descriptors. We instantiate this formulation through BEVMatch, a self-supervised framework that learns metric keypoints and descriptors directly in BEV space.

The framework employs modality-specific encoders, a cross-modal fusion module, and lightweight prediction heads operating on a shared BEV latent space. Training uses relative odometry as geometric supervision and modality dropout to encourage sensor-agnostic representations without requiring manual keypoint annotations.

Experiments on nuScenes demonstrate that a single model can estimate correspondences across camera-only, LiDAR-only, cross-modal, and multi-modal settings, including scenarios with larger relative translation and rotation. To assess practical applicability, we further integrate the learned correspondences into a keypoint-based SLAM pipeline deployed on a research vehicle. While the resulting system is evaluated as a feasibility study rather than a production-ready localization solution, the results indicate that unified metric BEV correspondences can support downstream localization and mapping tasks across varying sensor configurations.

Keywords: keypoints, localization, SLAM, self-supervised learning

1 Introduction

Robust autonomous driving requires representations that combine semantic scene understanding with geometrically consistent localization. Modern vehicles use complementary sensors such as cameras, which provide dense appearance cues, and Light Detection and Ranging (LiDAR), which provides accurate metric structure. The key challenge is to learn a representation in which heterogeneous observations can be compared, fused, and matched despite changing sensor availability.

Bird's-Eye View (BEV) representations offer a natural coordinate system for this goal. By projecting camera and LiDAR observations into a shared top-down metric space, recent methods achieve strong results in detection, segmentation, and online mapping [1, 2, 3, 4]. However, BEV is typically used as a dense prediction space, while its potential for metric correspondence learning remains underexplored. Such correspondences are essential for ego-motion estimation, relocalization, loop closure, and map alignment.

*All authors are with the Institute for Autonomous Driving, University of the Bundeswehr Munich, 85579 Neubiberg, Germany. Contact author email: lukas.beer@unibw.de.

Classical and learned image keypoint methods [5, 6, 7, 8, 9, 10, 11, 12, 13] provide robust features within images, and 3D methods estimate correspondences directly in point clouds [14, 15, 16, 17, 18]. Cross-modal registration further addresses image-to-point-cloud alignment [19, 20, 21, 22, 23]. Yet these approaches generally target specific modality pairs rather than a single metric space that supports temporal uni-, cross-, and multi-modal matching.

We formulate metric correspondence learning as a unified latent-space problem for BEV perception. The goal is to learn keypoints and descriptors that are repeatable across time and matchable across camera-only, LiDAR-only, camera–LiDAR, and fused inputs. Instead of treating each modality pair separately, we view them as different observations of the same metric BEV correspondence space.

We instantiate this idea with **BEVMatch**, a self-supervised framework for BEV keypoint detection and description. Modality-specific encoders project camera and LiDAR inputs into a shared BEV coordinate system, cross-modal attention fuses their features, and a common head predicts keypoint probabilities and local descriptors. Relative odometry provides geometric supervision, eliminating manual keypoint annotations, while modality dropout promotes robustness to varying sensor configurations.

This design offers three advantages. First, keypoints have a direct metric interpretation in BEV, enabling SE(2) pose estimation without scale recovery. Second, descriptors are learned in a modality-agnostic latent space, supporting direct uni-, cross-, and multi-modal matching. Third, correspondence estimation can reuse features from a BEV perception backbone, reducing redundancy between perception and localization.

Our contributions are:

- We formulate uni-, cross-, and multi-modal keypoint matching as metric correspondence learning in a shared BEV latent space.
- We introduce BEVMatch, a self-supervised framework for learning repeatable BEV keypoints and descriptors from odometry supervision.
- We use modality dropout to encourage sensor-agnostic descriptors across camera-only, LiDAR-only, cross-modal, and fused inputs.
- We evaluate BEVMatch on nuScenes [24], demonstrating metric correspondence estimation across multiple sensor configurations.
- We conduct a feasibility study of the resulting keypoints in a SLAM system.

2 Method

We formulate uni-, cross-, and multi-modal keypoint matching as metric correspondence learning in a shared BEV latent space. BEVMatch maps available camera and LiDAR observations into BEV, predicts keypoint probabilities and descriptors, and learns self-supervised from relative odometry between temporal frames. The framework is backbone-agnostic and can be attached to existing BEV perception networks. In our implementation, multi-view images are encoded with ResNet-50 [25] and lifted using Lift-Splat-Shoot (LSS) [1], while LiDAR point clouds are encoded with SECOND [26].

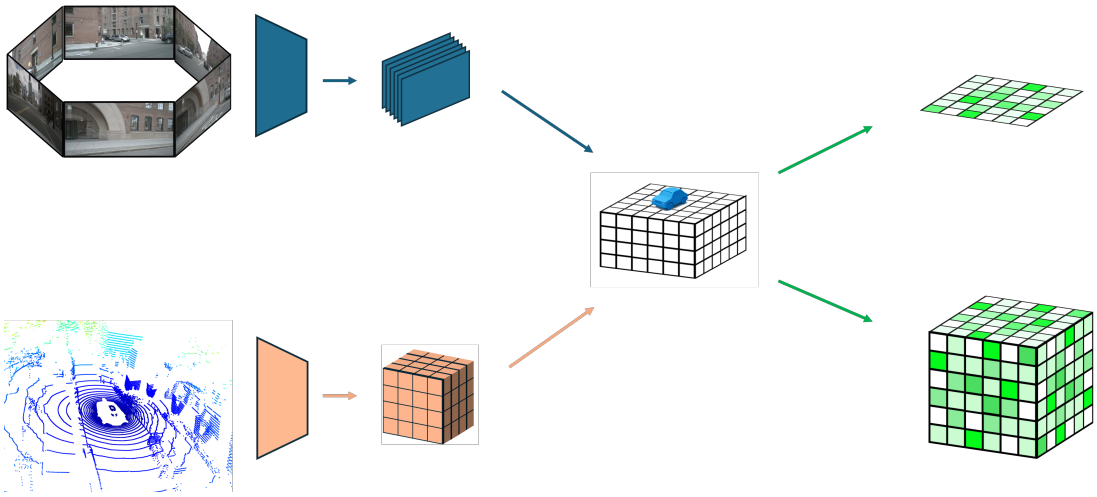

Figure 1: BEVMatch architecture. Camera and LiDAR observations are encoded into modality-specific BEV feature maps, fused in a unified latent space, and processed by lightweight heads that predict keypoint probabilities and local descriptors in metric BEV coordinates.

2.1 Shared BEV Representation

Camera and LiDAR inputs are first projected into separate BEV feature maps. We apply bidirectional cross-attention between downsampled camera and LiDAR features at 32×32 resolution, upsample the attention-enhanced features, concatenate them with the original modality features, and fuse them by a convolutional layer. The resulting latent space is defined on a 180×180 grid covering $108\,\mathrm{m} \times 108\,\mathrm{m}$, corresponding to $0.6\,\mathrm{m}$ per cell.

Two compact convolutional heads predict keypoint probabilities and 128-dimensional descriptors at each BEV grid location. The design follows learned keypoint pipelines such as SiLK [12], but operates directly in metric BEV space.

2.2 Self-Supervised Correspondence Generation

Training uses temporally separated frame pairs with known relative SE(2) motion. For source and target BEV feature maps $\mathcal{F}_0$ and $\mathcal{F}_1$, relative odometry defines dense correspondences by warping source coordinate p into the target frame. Correspondences outside the valid BEV region are discarded, and a round-trip consistency check removes geometrically inconsistent matches. These odometry-derived correspondences supervise descriptor learning and keypoint repeatability without manual annotations.

2.3 Descriptor Loss

The descriptor loss pulls corresponding BEV locations together and pushes non-corresponding locations apart. For each sampled correspondence i, descriptor d_0^i is paired with its warped counterpart $d_1^{c(i)}$, while all other sampled descriptors act as negatives. With ℓ_2-normalized descriptors, temperature τ, and N_D sampled correspondences, we use the symmetric In-

foNCE loss:

$$\mathcal{L}_D = -\frac{1}{2N_D}\sum_{i=1}^{N_D}\left[\log\frac{\exp(d_0^i \cdot d_1^{c(i)}/\tau)}{\sum_{j=1}^{N_D}\exp(d_0^i \cdot d_1^j/\tau)} + \log\frac{\exp(d_1^{c(i)} \cdot d_0^i/\tau)}{\sum_{j=1}^{N_D}\exp(d_1^{c(i)} \cdot d_0^j/\tau)}\right]. \tag{1}$$

This yields descriptors that are discriminative across BEV cells and consistent across time and sensor configurations.

2.4 Keypoint Loss

The detector is trained to assign high probability to repeatable and matchable locations. For each ground-truth correspondence (p_0, p_1), we obtain p_{pred} by selecting the target descriptor with maximum similarity. A prediction is correct if it lies within radius r of the warped target. This tolerance accounts for grid discretization and odometry noise.

The resulting binary keypoint maps g_0 and g_1 are optimized with focal loss [27]. For keypoint probability s_i^b and label g_i^b at location i in frame $b \in \{0, 1\}$,

$$\mathcal{L}_k = -\frac{1}{2N_K}\sum_{b=0}^{1}\sum_{i=1}^{N_K}\left[\alpha(1-s_i^b)^\gamma g_i^b \log(s_i^b) + (1-\alpha)(s_i^b)^\gamma(1-g_i^b)\log(1-s_i^b)\right]. \tag{2}$$

Focal weighting mitigates the imbalance between sparse keypoints and many non-keypoint cells.

2.5 Repulsion Loss

To avoid clustered detections, we add a repulsion loss over the top-k keypoints in each frame. For selected keypoints i and j with BEV grid distance d_{ij}, distances below $d_{\min}$ are penalized:

$$\mathcal{P}_{ij} = \text{ReLU}(d_{\min} - d_{ij}). \tag{3}$$

The full loss is

$$\mathcal{L}_{rep} = \frac{1}{2k(k-1)}\sum_{b=0}^{1}\sum_{i\neq j}\mathcal{P}_{ij}^b s_i^b s_j^b. \tag{4}$$

Thus, nearby high-confidence keypoints are discouraged, improving spatial coverage for SE(2) estimation.

2.6 Auxiliary Segmentation Loss

An auxiliary HD-map segmentation head regularizes the BEV representation during training. It encourages semantic structure such as drivable areas and lane markings and stabilizes image-to-BEV lifting. The branch is optimized with per-class focal loss:

$$\mathcal{L}_{seg} = \frac{1}{C}\sum_{c=1}^{C}\mathcal{L}_c, \tag{5}$$

where C is the number of semantic classes. Segmentation is used only for representation-level regularization and does not filter or weight keypoints.

Table 1: Keypoint matching results on nuScenes [24] for different modality configurations and relative motion ranges. We report mean translation error RTE (↓), mean rotation error RRE (↓), and recall (↑) at multiple SE(2) thresholds. Relative motion is grouped into close (0–5 m and $\leq 10°$), mid (5–10 m and $\leq 20°$), and far (10–15 m and $\leq 40°$). Best results are shown in **bold**.

Method	Mod.	RTE / RRE (m / °) ↓			Recall (%) ↑								
		close	mid	far	0.5 m / 5°			2 m / 10°			5 m / 20°		
					close	mid	far	close	mid	far	close	mid	far
SIFT [5]	SC2SC	1.02/7.61	2.60/16.83	4.87/24.78	30.0	6.5	2.3	72.3	26.3	11.1	86.4	62.8	34.0
ORB [7]	SC2SC	0.54/3.99	2.03/12.13	4.63/20.60	59.3	21.4	6.2	83.3	43.2	19.0	95.5	72.9	41.3
SuperPoint [9]	SC2SC	0.23/2.46	1.13/6.92	3.35/12.21	77.7	45.0	18.7	89.3	67.4	39.9	98.3	81.8	59.0
LoFTR [28]	SC2SC	**0.15**/2.38	0.87/6.48	2.92/10.80	82.0	54.2	27.1	90.7	71.8	46.9	98.6	83.3	61.7
GICP [29]	L2L	0.28/**0.13**	3.52/1.68	10.07/6.56	91.8	35.6	7.8	93.0	37.9	8.3	99.8	47.3	10.8
RANSAC-FPFH [30]	L2L	0.26/0.29	0.67/**0.53**	4.06/3.23	**94.6**	**80.7**	**38.2**	98.0	90.8	50.7	**100.0**	91.4	51.6
HRegNet [31]	L2L	0.67/1.04	3.22/4.85	8.44/8.17	69.6	10.6	0.3	81.0	32.3	3.2	97.7	70.9	16.5
BEVMatch (Ours)	SC2SC	2.40/1.76	5.17/4.08	9.54/9.52	12.2	0.8	0.1	39.1	6.6	2.8	96.2	29.4	11.1
BEVMatch (Ours)	C2C	1.10/0.90	2.31/1.77	4.17/2.91	37.4	11.4	4.5	82.4	52.8	34.1	99.4	81.7	59.7
BEVMatch (Ours)	C2L	1.38/1.73	1.99/2.20	3.17/3.22	11.4	8.3	5.3	60.5	54.9	39.3	88.3	84.0	66.3
BEVMatch (Ours)	L2L	0.37/0.54	0.58/0.97	1.29/2.05	77.1	55.0	32.0	**99.9**	97.7	83.7	**100.0**	99.7	89.7
BEVMatch (Ours)	M2M	0.36/0.43	**0.57**/0.73	**1.19**/**1.39**	77.4	55.5	34.3	**99.9**	**98.4**	**85.6**	**100.0**	**99.8**	**90.4**

2.7 Overall Objective and Modality Dropout

The final objective is

$$\mathcal{L}_{total} = \lambda_D \mathcal{L}_D + \lambda_k \mathcal{L}_k + \lambda_{seg} \mathcal{L}_{seg} + \lambda_{rep} \mathcal{L}_{rep}. \tag{6}$$

Here, $\mathcal{L}_D$ learns the correspondence space, $\mathcal{L}_k$ selects repeatable locations, $\mathcal{L}_{seg}$ regularizes BEV semantics, and $\mathcal{L}_{rep}$ prevents redundant detections. After an initial descriptor warm-up, λ_k is progressively increased to emphasize sparse keypoint detection.

To support varying sensor configurations, we use progressive modality dropout. After a warm-up phase with all modalities active, camera and LiDAR inputs are independently dropped with probability p, increased up to $p_{\max} = 0.5$. If all inputs of a frame are dropped, one modality is retained. This exposes the model to camera-only, LiDAR-only, cross-modal, and fused inputs during training, enabling uni-, cross-, and multi-modal matching at inference.

3 Experiments

We evaluate BEVMatch on nuScenes [24] for metric correspondence learning in a unified BEV latent space. A single model is used for all settings; only the active input modalities change at inference. We consider single-camera matching (SC2SC), multi-camera matching (C2C), camera-to-LiDAR matching (C2L), LiDAR-to-LiDAR matching (L2L), and multi-modal matching (M2M).

3.1 Evaluation Protocol

For each temporal frame pair, we match predicted BEV keypoints and estimate the relative SE(2) transformation with Random Sample Consensus (RANSAC) [32]. We report

relative translation error (RTE), relative rotation error (RRE), and recall at 0.5 m / 5°, 2.0 m / 10°, and 5.0 m / 20°. Mean RTE and RRE are computed over successful estimates only. Results are grouped by relative motion into close, mid, and far ranges as defined in Table 1.

3.2 Baselines

We compare against image-based methods SIFT [5], ORB [7], SuperPoint [9], and LoFTR [28], evaluated on front-camera pairs. The SC2SC variant of BEVMatch provides the corresponding single-camera baseline. Since image-plane methods do not recover metric scale directly, their scale is resolved using ground-truth motion.

For LiDAR matching, we compare against GICP [29], RANSAC-FPFH [30], and HRegNet [31]. Unlike these modality-specific methods, BEVMatch uses one shared BEV latent space for L2L, C2L, C2C, and M2M inference.

3.3 Main Results

Table 1 shows that BEVMatch supports metric correspondence estimation across all modality configurations. The best results are obtained for L2L and M2M, where LiDAR provides stable geometric structure. For larger displacements, BEVMatch remains robust, while classical point-cloud registration degrades strongly in the far-motion regime.

M2M achieves the best overall robustness, especially for mid and far motion. Compared with L2L, adding cameras gives moderate accuracy gains but increases redundancy, since the same model can operate with fused, camera-only, LiDAR-only, or cross-modal inputs.

Camera-only matching benefits from the multi-view BEV formulation. C2C clearly improves over SC2SC, indicating that surround-view aggregation compensates for information loss during image-to-BEV lifting. While single-camera image matchers remain strong for small motion, BEVMatch estimates metric correspondences directly in top-down coordinates without scale recovery.

C2L results show that the shared latent space can bridge heterogeneous sensor observations. Although C2L is harder than L2L and M2M, it remains competitive with C2C, supporting the use of one metric correspondence space for uni-, cross-, and multi-modal matching.

4 Deployment

We deploy BEVMatch for keypoint-based SLAM to assess practical use of the learned BEV correspondence representation. The system estimates a planar trajectory from learned BEV keypoints without wheel odometry, IMU measurements, or external localization at runtime. We did not remove any sensors during runtime, therefore it performs purely multi-modal and not cross-modal. For deployment, we trained BEVMatch on our own dataset using in-vehicle odometry as supervision. Figure 2 shows our test vehicle, MuCAR-3, which is equipped with four surround-view cameras and a Velodyne VLS-128 LiDAR. We evaluate the proposed architecture on this platform.

Figure 2: The testing vehicle, MuCAR-3.

4.1 System Pipeline

The trained BEV keypoint network is executed with TensorRT. Its dense heatmap and descriptor field are converted into sparse keypoint frames using confidence thresholding, local non-maximum suppression, sub-cell peak refinement, and descriptor extraction. Only keypoints, scores, and descriptors are passed to the SLAM backend.

The backend uses factor-graph optimization to estimate the current SE(2) pose relative to a local landmark map. Stable detections are inserted or updated as landmarks, keyframes are selected by accumulated motion, and local pose-graph optimization is run periodically. Landmarks are associated using descriptor similarity, and SE(2) alignment is estimated with RANSAC. If too few landmarks are available, the system falls back to frame-to-frame matching.

4.2 Accuracy and Limitations

Table 2: Translation error (t_{err}) vs. track length L.

driven length (m)	100	200	300	400	500	600	700	800
t_{err} (%)	5.07	6.00	6.79	7.39	7.45	7.38	7.05	6.28

The deployed system produces a continuous SE(2) trajectory on a 2.7 km suburban campus test drive. Table 2 shows the relative translation accuracy, in the style of the KITTI odometry benchmark [33]. The results are not yet comparable to a tuned production localization stack and should be interpreted as a feasibility study. Furthermore, we only evaluated multi-modal to multi-modal. In this setup, no sensor dropouts were simulated.

The main limitation is the current BEV grid resolution, which improves efficiency but limits localization precision through discretization and feature downsampling. Overall, the deployment shows that unified metric BEV correspondences can be connected to a practical SLAM backend through a simple sparse keypoint interface, while leaving room for improved resolution, dynamic-scene handling, and backend robustness.

5 Conclusion

In this work, we presented BEVMatch, a novel keypoint detection and matching framework designed for perception in the BEV space. By leveraging a shared latent space, BEVMatch enables metric SE(2) estimation without scale ambiguity and supports seamless modality interchange. Our evaluation showed strong performance in L2L and M2M settings and competitive results in C2C. We demonstrated that while adding cameras to LiDAR yields marginal accuracy gains, it provides critical redundancy.

Nevertheless, there remain some limitations. As a BEV-based method, it is limited to planar motions and cannot handle a full 6-DoF transformation. As a method relying on geometric warping for self-supervision, BEVMatch assumes a static scene during the correspondence generation. We do not explicitly model dynamic objects or occlusions; consequently, keypoints on moving vehicles may result in incorrect correspondences. Future work will address this by incorporating flow estimation or semantic masking to filter dynamic regions. In addition to that, down- and upsampling operations in the BEV encoder and fusion module lead to a loss of fine-grained details, affecting the precision of keypoint localization and description. Other attention strategies, such as deformable attention [34] or flash attention [35, 36] could be explored to reduce the computational overhead of the cross-attention-based fusion module.

We believe metric BEV keypoints offer a promising way for unified perception and localization systems in multi-sensor autonomous driving.

Acknowledgment

This research paper is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr [project MORE] which we gratefully acknowledge. dtec.bw is funded by the European Union – NextGenerationEU.

References

[1] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210. Springer, 2020.

[2] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *ECCV*, 2024.

[3] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE Int. Conf. Robot. Autom.*, pages 2774–2781. IEEE, 2023.

[4] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. MapTR: Structured modeling and learning for online vectorized hd map construction. In *ICLR*, 2023.

[5] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *ECCV*, pages 404–417. Springer, 2006.

[7] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB an efficient alternative to SIFT or SURF. In *ICCV*, pages 2564–2571. Ieee, 2011.

[8] Pablo F Alcantarilla, Jesus Nuevo, and Adrien Bartoli. Fast explicit diffusion for accelerated features in nonlinear scale spaces. In *BMVC*. BMVA Press, 2013.

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *CVPRW*, pages 224–236, 2018.

[10] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint detection and description of local features. In *CVPR*, pages 8092–8101, 2019.

[11] Jerome Revaud, Philippe Weinzaepfel, César De Souza, and Martin Humenberger. R2D2: Repeatable and reliable detector and descriptor. In *NeurIPS*, pages 12414–12424, 2019.

[12] Pierre Gleize, Weiyao Wang, and Matt Feiszli. SiLK – simple learned keypoints. In *CVPR*, pages 22499–22508, 2023.

[13] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020.

[14] Zi Jian Yew and Gim Hee Lee. 3DFeat-Net: Weakly supervised local 3d features for point cloud registration. In *ECCV*, pages 607–623, 2018.

[15] Jiaxin Li, Benjamin Chen, and Gim Hee Lee. USIP: Unsupervised stable interest point detection from 3d point clouds. In *ICCV*, pages 361–370, 2019.

[16] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3Feat: Joint learning of dense detection and description of 3d local features. In *CVPR*, pages 6359–6368, 2020.

[17] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. PREDATOR: Registration of 3d point clouds with low overlap. In *CVPR*, pages 4267–4276, 2021.

[18] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Junhui Li, Haibin Wang, and Chiew-Lan Tai. GeoTransformer: Fast and robust point cloud registration with geometric transformer. In *CVPR*, pages 11115–11125, 2022.

[19] Jiaxin Li and Gim Hee Lee. Deepi2p: Image-to-point cloud registration via deep classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15960–15969, 2021.

[20] Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen. Corri2p: Deep image-to-point cloud registration via dense correspondence. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1198–1208, 2022.

[21] Lin Bie, Shouan Pan, Siqi Li, Yining Zhao, and Yue Gao. Graphi2p: Image-to-point cloud registration with exploring pattern of correspondence via graph learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22161–22171, 2025.

[22] Xinjun Li, Wenfei Yang, Jiacheng Deng, Zhixin Cheng, Xu Zhou, and Tianzhu Zhang. Implicit correspondence learning for image-to-point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16931, 2025.

[23] Filippo D'Addeo, Lorenzo Cipelli, Adriano Cardace, Emanuele Ghelfi, Andrea Zinelli, and Massimo Bertozzi. Calibbev: Lidar-camera calibration via bev alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4345–4354, 2026.

[24] 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*, pages 11621–11631, 2020.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[26] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[28] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 2021.

[29] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-ICP. In *Robotics: Science and Systems*, volume 2, pages 435–442, 2009.

[30] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3d registration. In *IEEE Int. Conf. Robot. Autom.*, pages 3212–3217, 2009.

[31] Fan Lu, Guang Chen, Yinlong Liu, Zhang Lijun, Sanqing Qu, Shu Liu, and Rongqi Gu. HRegNet: A hierarchical network for large-scale outdoor lidar point cloud registration. In *ICCV*, 2021.

[32] 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.

[33] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

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

[35] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[36] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024.