Extend Your Own Correspondences: Unsupervised Distant Point Cloud Registration by Progressive Distance Extension

Quan Liu[1] Hongzi Zhu[1]* Zhenxi Wang[1] Yunsong Zhou[1] Shan Chang[2] Minyi Guo[1]
[1]Shanghai Jiao Tong University [2]Donghua University

https://github.com/liuQuan98/EYOC

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.01967

Abstract

Registration of point clouds collected from a pair of distant vehicles provides a comprehensive and accurate 3D view of the driving scenario, which is vital for driving safety related applications, yet existing literature suffers from the expensive pose label acquisition and the deficiency to generalize to new data distributions. In this paper, we propose EYOC, an unsupervised distant point cloud registration method that adapts to new point cloud distributions on the fly, requiring no global pose labels. The core idea of EYOC is to train a feature extractor in a progressive fashion, where in each round, the feature extractor, trained with near point cloud pairs, can label slightly farther point cloud pairs, enabling self-supervision on such far point cloud pairs. This process continues until the derived extractor can be used to register distant point clouds. Particularly, to enable high-fidelity correspondence label generation, we devise an effective spatial filtering scheme to select the most representative correspondences to register a point cloud pair, and then utilize the aligned point clouds to discover more correct correspondences. Experiments show that EYOC can achieve comparable performance with state-of-the-art supervised methods at a lower training cost. Moreover, it outwits supervised methods regarding generalization performance on new data distributions.

1. Introduction

Registering point clouds obtained on distant vehicles of 5 meters to 50 meters apart [28, 29] can greatly benefit a rich set of self-driving vision tasks, ranging from detection [55, 58, 61] and segmentation [42, 50] to birds' eye view (BEV) representation [27, 38] and SLAM [33, 34], and ultimately improve the overall driving safety. Traditional supervised registration methods not only heavily rely on *accurate* pose labels during training [6, 21, 43] but cannot attain expected performance on new data distributions as they do on existing datasets [9, 20], making them infeasible to use in real-world driving scenarios. In light of the ever-growing LiDAR-equipped vehicles and the tremendous amount of sequential unlabelled point cloud data, *can we finetune a registration network on a new point cloud distribution with no pose labels so that distant point clouds on the new distribution can be accurately registered on the fly?*

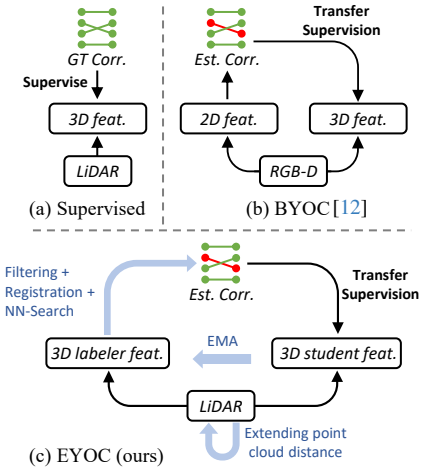

Figure 1. (a) Supervised registration require ground-truth (GT) pose, and (b) BYOC requires RGB-D images for supervision [12]. (c) In contrast, EYOC acquires supervision from LiDAR sequences directly, enabling single-modal unsupervised training.

In the literature, a rich set of supervised indoor [20, 24, 26, 37, 53, 56] or synthetic [1, 15, 48, 52] low-overlap registration methods have been proposed. Most of these methods simply fail on outdoor distant point clouds due to the patch-similarity assumption [31, 56] or structural prior such as optimal-transport [37, 54] no longer hold. While simpler networks (*e.g.*, CNNs) showed better robustness on distant point clouds [9, 20], they still need expensive ground-truth poses for training, as depicted in Fig. 1(a). As pointed out by Banani and Johnson [12], unsupervised registration is all about establishing correspondences. BYOC [12], Unsupervised R&R [13], and UDPReg [32] have bypassed correspondence acquisition in the indoor setting by borrowing

*Corresponding author

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.01967

correspondences from RGB channel or GMM matching, as depicted in Fig. 1(b), but they suffer from the discrepancy between dense surround images and a sparse point cloud in outdoor settings. As a result, there is no successful solution, to the best of our knowledge, to the unsupervised distant point cloud registration problem.

In this paper, we propose *Extend Your Own Correspondences* (EYOC), a fully unsupervised outdoor distant point cloud registration method *requiring neither pose labels nor any input of other modality*. As depicted in Fig. 1(c), our core idea is to adopt a progressive self-labeling scheme to train a feature extractor in multiple rounds. Specifically, in each round, a labeler model trained with near point cloud pairs can generate correspondence labels for farther-apart point clouds, which are used to train a student model. Particularly, the Siamese labeler-student models are synchronized using the exponential moving average (EMA). This process repeats until a full-fledged student model, capable of extracting effective features for distant point cloud registration, is obtained. Two main challenges are encountered in the design of EYOC as follows.

First, it is extremely challenging to prevent the self-labelling process from diverging, given the extreme low-overlap and density-variation of a distant point cloud pair, as witnessed even in supervised training [28]. To deal with this challenge, we take a gradual learning methodology by breaking the hard learning problem into a series of learning steps with increasing learning difficulties. Specifically, in the first step, considering the spatial locality of two consecutive frames in a LiDAR point cloud sequence, we assume that two consecutive frames approximately have no transformation, which can be used as supervision to train a basic model. After the model first converges to a decent set of weights, we enable the labeler-student self-labelling process and gradually extend the interval of training frames in each learning step. As a result, the student model can converge smoothly.

Second, it is nontrivial for the labeler in one learning step to generate sufficient correspondence labels of high quality for the next harder learning step. We observe *the near-far diversity phenomenon* of LiDAR point clouds, *i.e.*, when the observation distance changes, the point density variation of near objects is larger than that of far objects. This means that features extracted from low-density (far-from-LiDAR) regions are more stable along with distance changes. Inspired by this insight, we develop a spatial filtering technique to effectively discover a set of initial quality correspondences in low-density regions. Furthermore, to obtain more widespread correspondences, we perform a live registration using the initial correspondences followed by another round of nearest-neighbor search (NN-Search) to further dig out and amplify correct correspondences, readied for supervision of the student.

We evaluate EYOC design with trace-driven experiments on three major self-driving datasets, *i.e.*, KITTI [16], nuScenes [6], and WOD [43]. EYOC reaps comparable performance with state-of-the-art (SOTA) fully supervised registration methods while outwitting them by 17.4% mean registration recall in an out-of-domain unlabelled setting. To summarize, our contributions are listed as follows:

- We analyzed the *near-far diversity* of point clouds, where low-density regions of a point cloud produce consistent feature correspondences during a distance extension step.
- We propose an unsupervised distant point cloud registration method that can effectively adapt to new data distributions without pose labels or other input modalities.
- The performance and applicability of EYOC are validated with extensive experiments on three self-driving datasets.

2. Related Work

2.1. Supervised Registration

Recent registration techniques are highly monopolized by learning based methods [2–4, 9, 11, 19, 20, 24, 28, 29, 31, 35, 37, 53, 54, 57], due to both superior accuracy and faster inference speed compared with traditional extractors [22, 41, 46] or pose estimators such as RANSAC [14].

Local feature extractors. Correspondence-based local feature extractors have long been diverged into patch-based methods [2, 11, 19, 35, 57] and fully-convolutional methods [4, 9, 20, 28, 29]. 3DMatch [57] initiated the patch-based genre, while PointNet [36], smoothed density value and reconstruction were later introduced by PPF-Net [11], PerfectMatch [19], and DIP [35], respectively. The recent pinnacle SpinNet [2] and BUFFER [3] combine SO(2) equivalent cylindrical features with fully convolutional backbones. On the other hand, following FCGF [9], fully convolutional methods process the point cloud as a whole. KPConv [45] backbones are equipped with keypoint detection in D3Feat [4] and overlap attention in Predator [20]. APR [28] and GCL [29] further enhanced outdoor distant low-overlap registration with reconstruction and group-wise contrastive learning. We build our method upon fully convolutional methods because they are deemed most suitable for fast and generalizable outdoor registration.

Pose estimators. Pose estimators [5, 7, 10, 14, 25, 59, 60] take in feature maps and output the most probable pose estimation, where RANSAC [14] is a common time-consuming baseline. While DGR [10], PointDSC [5] and DHVR [25] opted for learned correspondence weight with FCNs, Non-local Module, and Hough Voting, respectively, non-parametric methods such as SC^2-PCR [7] and MAC [59] hit higher marks through the second order compatibility or maximal clique search.

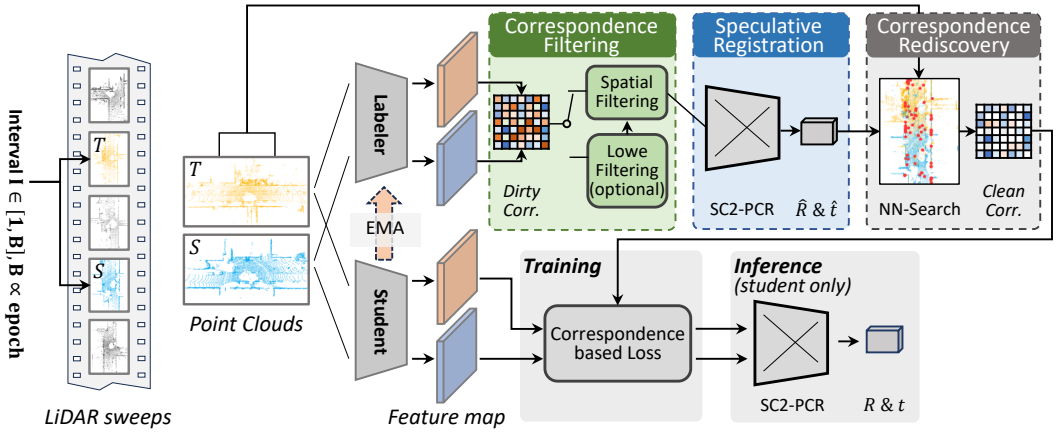

Figure 2. **Overview of Extend Your Own Correspondences (EYOC)**. It exhibits a two-branch student-labeler structure with periodic synchronization, where the labeler generates correspondences for the student. Point cloud pairs are selected at random frame interval, whose range extends with time. Labeler dirty correspondences are filtered before the speculative registration which outputs an estimated pose. Finally, correspondence rediscovery with NN-search on re-aligned input point clouds recovers clean correspondence labels.

Keypoint-free registration. Keypoint-free methods borrowed the idea of superpixels [17] from image matching to match heavily downsampled points (*i.e.*, superpoints), each representing a local patch [24, 31, 37, 53, 54, 56]. HRegNet [31] proposed to refine global pose with different stages of downsampling. CoFiNet [54], GeoTransformer [37], and PEAL [56] treat superpoints as seeds and match promising seed patches only. Another line of work, DeepPRO [24] and REGTR [53], regress correspondences directly without feature matching. However, their assumption that superpoint patches should share high overlap no longer holds considering extreme density-variation and low-overlap.

2.2. Unsupervised Registration

Compared with supervised methods, unsupervised registration is less explored especially for the outdoor scenario. BYOC [12] highlighted that random 2D CNNs could generate image correspondences good enough to supervise a 3D network, therefore indoor RGB-D images are used for self-supervision. UnsuperisedR&R [13] in turn seeked help from differentiable rendering of RGB-D images as mutual supervision after differentiable registration. UDPReg [32] enforced multiple losses on GMM matching to generate correspondences for indoor point clouds. However, outdoor unsupervised registration remain an exciting yet unexplored field of research, calling for more work on this area.

3. Problem Definition

Given two point clouds $\mathcal{S} \in \mathbb{R}^{n\times 3}, \mathcal{T} \in \mathbb{R}^{n\times 3}$, point cloud registration aims to uncover their relative transformation $R \in SO(3), t \in \mathbb{R}^3$ so that $\mathcal{S}R^T + t^T$ aligns with $\mathcal{T}$. When the LiDARs are placed on two distant vehicles separated at a distance of $d \in [5m, 50m]$, the sub-problem is referred to as *distant* point cloud registration [28]. Contrary to previous settings [20, 24, 53], distant point clouds share extreme low-overlap and density-variation leading to network divergence when directly applied to training. This is usually mitigated through a staged training strategy with pretraining on high-overlap pairs and finetuning on low-overlap pairs [28].

4. Method

The overview of EYOC is illustrated in Fig. 2, which composes of a siamese student-labeler network structure followed by correspondence filtering, speculative registration, and correspondence rediscovery. During training, two distant point clouds, $\mathcal{S}, \mathcal{T}$, are fed into the student and labeler networks to extract point-wise features $F^{stu}_{\mathcal{S}}, F^{lab}_{\mathcal{S}} \in \mathbb{R}^{n\times k}$ and $F^{stu}_{\mathcal{T}}, F^{lab}_{\mathcal{T}} \in \mathbb{R}^{m\times k}$. The labeler features are then processed by correspondence filtering to obtain a decent correspondence set $C^{lab} = \{(i,j)|p_i \in \mathcal{S}, q_j \in \mathcal{T}\}$. It is later fed into speculative registration to decide an optimal transformation $\hat{R} \in SO(3), \hat{t} \in \mathbb{R}^3$ between $\mathcal{S}$ and $\mathcal{T}$. The high-fidelity estimated transformation is used to re-align input point clouds, which allows us to rediscover correspondences using NN-Search for supervision of the student.

4.1. Extension of Point Cloud Distance

Unlike the supervised setting, it is impossible to calculate the accurate distance between LiDARs in the unsupervised setting. However, leveraging the spatial locality of LiDAR sequences, we can limit the translational upper bound by limiting the frame interval I between two frames in a sequence. Improving upon the staged training strategy [28], we propose to randomly select the frame interval $I \in \mathbb{N}^+, I \in [1, B]$ for every pair, where B grows from 1 to 30 during the course of training, forming 30 tiny steps. When $B = 1$, we assume identity transformation and apply supervised training. Our *progressive distance extension*

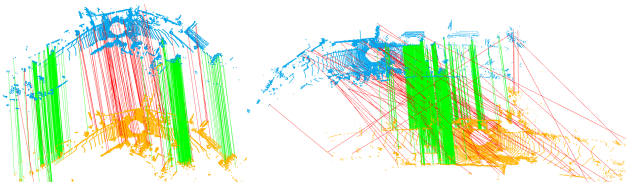

Figure 3. **The dirty correspondence labels generated by closer-range labeler (Left: $B = 1$; Right: $B = 10$) on farther-apart point clouds (Left: $d = 10m$; Right: $d = 30m$) in KITTI [16] before spatial filter.** Correct ones are colored green and false ones red. Close-to-LiDAR features are less generalizable to farther pairs than far-from-LiDAR features.

strategy increases the problem difficulty gradually to facilitate smooth convergence.

4.2. Labeler-Student Feature Extraction

Given a pair of distant point clouds, we pass them through two homogeneous 3D sparse convolutional backbones parameterized by W^{lab} and W^{stu}, to obtain point-wise feature maps. The student is periodically updated to the labeler in a gradual manner of exponential moving average (EMA), which keeps the labeler both stable and up-to-date, facilitating consistent label generation. Specifically, we update the labeler weights as in Eq. (1) after every epoch, where $\lambda \in [0, 1)$ is a decay factor:

$$W_{t+1}^{lab} \longleftarrow \lambda W_t^{lab} + (1-\lambda) W_t^{stu} \tag{1}$$

4.3. Correspondence Filtering

The correspondence filtering module aims to maximize the portion of correct correspondences produced by the labeler to enable unsupervised label generation. Different from BYOC, random 3D CNNs cast much worse correspondences than random 2D CNNs [12, 40, 47], so the dirty correspondences obtained by matching 3D labeler features $F_{\mathcal{S}}^{lab}$ and $F_{\mathcal{T}}^{lab}$ is likely to be rife with different fault patterns from RGB-D images. With that in mind, we investigate two types of filtering techniques on both feature space and Euclidean space based on data-centric observations.

Lowe filtering. Previous literature [12, 13] have found Lowe's Ratio [30] a good match for rating the most unique correspondences on indoor RGB-D point clouds. Specifically, given two corresponding features $f_{\mathcal{S}}^i \in F_{\mathcal{S}}^{lab}, f_{\mathcal{T}}^j \in F_{\mathcal{T}}^{lab}$, the significance is calculated according to Eq. (2), where $D(\cdot,\cdot)$ denotes the cosine similarity. Contrary to previous literature, we find Lowe filtering to deteriorate correspondence quality drastically as discussed in Sec. 5.3.

$$\omega_{i,j} = 1 - \frac{D(f_{\mathcal{S}}^i, f_{\mathcal{T}}^j)}{\min_{f_{\mathcal{T}}^k \in F_{\mathcal{T}}^{lab}, k \neq j} D(f_{\mathcal{S}}^i, f_{\mathcal{T}}^k)} \tag{2}$$

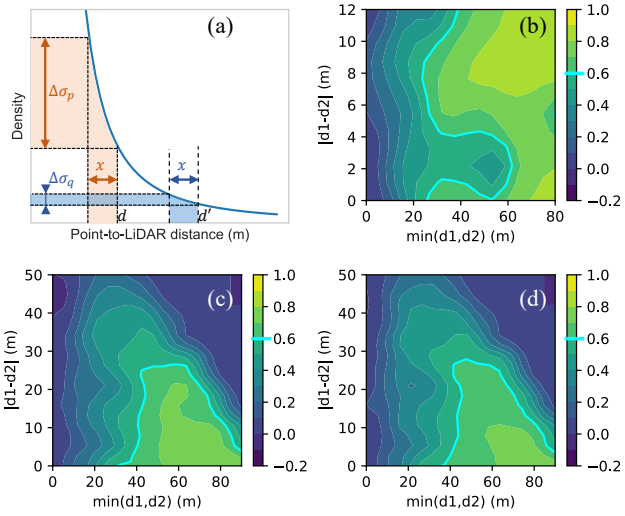

Figure 4. **Visual groundings** for our hypothesis on KITTI [16]. (a) Density of close-to-LiDAR points are more sensitive to movement than far-from-LiDAR points. (b-d) Cosine similarity of correspondences with its distance to two LiDARs, d_1, d_2, under (b) $I \in [1, 1]$, (c) $I \in [1, 15]$, and (d) $I \in [1, 30]$.

Spatial characteristic of labeler correspondences. In response to the failure of Lowe filtering, we conduct a label-driven investigation based on the *the near-far diversity phenomenon*, where far objects should have more consistent densities when the viewpoint undergoes displacements. We hereby examine the quality of raw feature correspondences for a labeler model on farther-apart point clouds than those in the labeler's training set, as depicted in Fig. 3, and propose the following hypothesis:

Hypothesis 4.1 *Correct correspondences are more likely to be clustered in low-density regions far from the LiDARs during the distance extension.*

Proof. We provide the rationale of a simplified case here based on the LiDAR sensor model [23]. A LiDAR can be modeled as a light source emitting light uniformly in all directions, and the probability density of a point being scanned is proportional to its energy absorption rate. Specifically, given two points in the world coordinate $p = (d, 0, 0)^T, q = (d', 0, 0)^T, 0 < d < d'$ and the current LiDAR center $O = (0, 0, 0)^T$, their respective densities are $\sigma_p = \frac{\alpha}{d^2}, \sigma_q = \frac{\alpha}{d'^2}$, where α is an unknown constant depending on the LiDAR resolution and incident angle, which we assume are the same for p and q. Suppose the LiDAR center now moves to $O' = (x, 0, 0)^T$ where $0 < x < d < d'$, the delta densities are $\Delta\sigma_p = \frac{\alpha}{(d-x)^2} - \frac{\alpha}{d^2}, \Delta\sigma_q = \frac{\alpha}{(d'-x)^2} - \frac{\alpha}{d'^2}$. It is easy to prove that $\Delta\sigma_p > \Delta\sigma_q$, as illustrated in Fig. 4(a). As widely acknowledged, CNN features are sensitive to density variation [20, 37, 44, 49, 51], therefore making close-range point features less robust under vehicle translation.

Spatial filtering design. Based on the findings, we quantitatively explore the relationship between distance from a correspondence to two LiDARs denoted by d_1, d_2, and the cosine similarity of that feature correspondence, as depicted in Fig. 4(b-d). We refer readers to Appendix Sec. 11.1 for similar results on other datasets. We confirm that close-to-LiDAR regions contain most correspondences, but are consistently under-performing and, therefore, could be purged to improve supervision quality. We hereby propose two sets of spatial filtering strategies:

- **Hard**: Discard points where $\min(d_1, d_2) < d_{thresh}$, regardless of training progression;
- **Adaptive**: Discard regions with $\leq s_{thresh}$ similarity in Fig. 4, where the decision boundary at $s_{thresh} = 0.6$ is highlighted in cyan. The similarities are exhaustively recorded from the pretraining dataset.

Emperically, both methods suffice to cut over 70% of the false correspondences while only 9% correct correspondences are discarded.

4.4. Speculative Registration

After the correspondence filtering, we adopt a SOTA registration algorithm SC^2-PCR [7] for accurate real-time registration to amplify the most promising set of correspondences. Although the correspondences have been heavily cleansed down to several hundred pairs, only 20% among which are correct on average, which is below the bar for direct supervision as discussed in Sec. 5.3; However, literature has shown that this ratio is high enough for a successful registration [7, 59]. Intuitively, if the input point clouds could be correctly registered, we could imitate fully-supervised training where correspondences are obtained directly from aligned input point clouds instead of matched features. Moreover, the searched nearest neighbors can vastly outnumber the heavily filtered labeler correspondences, making the training process more data-efficient. Therefore, we propose to obtain an estimated pose $\hat{R} \in SO(3), \hat{t} \in \mathbb{R}^3$ between the input point clouds $\mathcal{S}, \mathcal{T}$ on the fly with real-time registration algorithms.

4.5. Correspondence Rediscovery

With the input point clouds S, T and the estimated transformation $\hat{R} \in SO(3), \hat{t} \in \mathbb{R}^3$, we could simply follow supervised training to search correspondences for dense supervision. Specifically, we transform $\mathcal{S}' = R^T\mathcal{S} + t^T$, and obtain the nearest neighbors according to Eq. (3), where $\beta_{inlier} = 2m$ is a loosened match threshold to tolerate minor pose errors.

$$C_{ST} = \left\{ (i,j) \middle| p^i_{\mathcal{S}} \in \mathcal{S}', j = \arg\min_{p^j_{\mathcal{T}} \in \mathcal{T}} ||p^i_{\mathcal{S}} - p^j_{\mathcal{T}}||, \; s.t. \; ||p^i_{\mathcal{S}'} - p^j_{\mathcal{T}}|| < \beta_{inlier} \right\} \tag{3}$$

4.6. Loss Design

We adopt the widely-used Hardest-Contrastive Loss [9] as the training loss for the student. As nearest-neighbor search is not differentiable, we only back-propagate gradients to the student but not the labeler. Specifically, the loss is formulated as Eq. (4):

$$L = \frac{1}{|C_{\mathcal{ST}}|} \sum_{(i,j)\in C_{\mathcal{ST}}} \left[m + P(f^i_{\mathcal{S}}, f^j_{\mathcal{T}}) - \min_{j\neq k\in\mathcal{N}} P(f^i_{\mathcal{S}}, f^k_{\mathcal{T}}) \right]_+ + \frac{1}{|C_{\mathcal{TS}}|} \sum_{(j,i)\in C_{\mathcal{TS}}} \left[m + P(f^j_{\mathcal{T}}, f^i_{\mathcal{S}}) - \min_{i\neq k\in\mathcal{N}} P(f^j_{\mathcal{T}}, f^k_{\mathcal{S}}) \right]_+ \tag{4}$$

Where $\mathcal{N}$ is a subset of feature indices, m is the positive margin, $[\cdot]_+$ rounds negative values to 0, $P(\cdot,\cdot)$ denotes the squared distance between two vectors. $C_{\mathcal{TS}}$ follows Eq. (3) but is calculated in the reverse direction from $\mathcal{T}$ to $\mathcal{S}$.

5. Results

We demonstrate the superiority of EYOC against state-of-the-art methods on three major self-driving datasets, KITTI [16], nuScenes [6], and WOD [43]. We then provide an ablation study, finetuning strategies, and time analysis. Visualizations for the labeler are available in Fig. 6.

5.1. Experiment Setup

Datasets. Apart from our progressive dataset extension strategy, shorthanded as *progressive dataset*, we also follow existing literature [28, 29] to prepare the point cloud pairs based on the distance between two LiDARs, referred to as *traditional dataset*. The latter works under supervised settings, where the point cloud pairs have a random Euclidean distance between two LiDARs, denoted with $d \in [M, N]$ in meters. The traditional datasets are also used during all test sections. On the other hand, progressive datasets work for either supervised or unsupervised training, where point cloud pairs are selected with a random frame interval $I \in [1, B]$ due to the absence of pose labels. We set the initial value to $B = 1$ which grows linearly to $B = 30$ during 200 epochs. All datasets are cut into train-val-test splits by official recommendations.

Training. For supervised comparison methods, we follow common practice [28] to train the model on traditional datasets with $d \in [5, 20]$ and further finetune on $d \in [5, 50]$. The strategy applies to all baselines, while pretrained weights will be used for those whose training does not converge (denoted with *). On the other hand, EYOC needs only one course of training thanks to the progressive dataset. When a labelled pretraining dataset is available, the parameters of adaptive spatial filtering are acquired with the

Test Set	No.	Method	Pretrain Dataset	Finetune Dataset	Supervised	Progressive Dataset	mRR	RR @ $d \in$ [5,10]	[10,20]	[20,30]	[30,40]	[40,50]
KITTI	a	FCGF [9]	WOD	-	✓	-	71.8	98.0	92.5	85.0	52.6	30.7
		Predator [20]	WOD	-	✓	-	72.3	99.5	98.9	90.9	56.8	15.3
	b	FCGF [9]	-	KITTI	✓	-	77.4	98.4	95.3	86.8	69.7	36.9
		FCGF + C	-	KITTI	✓	✓	84.6	**100.0**	97.5	90.1	79.1	56.3
		Predator [20]	-	KITTI	✓	-	**87.9**	**100.0**	98.6	**97.1**	**80.6**	**63.1**
		SpinNet* [2]	-	KITTI	✓	-	39.1	99.1	82.5	13.7	0.0	0.0
		D3Feat* [4]	-	KITTI	✓	-	66.4	99.8	98.2	90.7	38.6	4.5
		CoFiNet [54]	-	KITTI	✓	-	82.1	99.9	**99.1**	94.1	78.6	38.7
		GeoTrans.* [37]	-	KITTI	✓	-	42.2	**100.0**	93.9	16.6	0.7	0.0
	c	EYOC (ours)	-	KITTI	-	✓	83.2	99.5	96.6	89.1	78.6	52.3
			WOD	KITTI	-	✓	80.6	99.5	95.6	89.1	75.1	43.7
WOD	d	FCGF [9]	KITTI	-	✓	-	69.9	97.1	87.9	61.8	59.0	43.9
		Predator [20]	KITTI	-	✓	-	70.7	98.1	97.6	81.2	53.2	23.6
	e	FCGF [9]	-	WOD	✓	-	**89.5**	**100.0**	98.6	91.2	**83.5**	**74.0**
		FCGF + C	-	WOD	✓	✓	77.2	98.1	89.9	75.8	64.7	57.7
		Predator [20]	-	WOD	✓	-	86.4	**100.0**	**100.0**	**95.3**	79.1	57.7
	f	EYOC (ours)	-	WOD	-	✓	78.4	97.6	91.3	78.2	65.5	59.3
			KITTI	WOD	-	✓	77.3	97.1	90.3	75.8	65.5	57.7
nuScenes	g	FCGF [9]	WOD	-	✓	-	67.1	98.9	**93.9**	**73.6**	42.6	26.3
		Predator [20]	WOD	-	✓	-	34.5	93.0	55.2	11.8	6.0	6.7
	h	FCGF [9]	-	nuScenes	✓	-	39.5	87.9	63.9	23.6	11.8	10.2
		FCGF + C	-	nuScenes	✓	✓	59.3	96.2	85.1	59.6	35.8	20.0
		Predator [20]	-	nuScenes	✓	-	51.0	**99.7**	72.2	52.8	16.2	14.3
	i	EYOC (ours)	-	nuScenes	-	✓	61.7	96.7	85.6	61.8	37.5	26.9
			WOD	nuScenes	-	✓	**68.4**	98.9	91.7	73.3	**44.3**	**33.7**

Table 1. **Comparison of mRR(%) and RR (%) between SOTA methods and EYOC over five test sets with $d \in [b_1, b_2]$ on KITTI [16], WOD [43], and nuScenes [6], respectively**, with increasing point cloud distance and registration difficulty. We group the tests denoted by letters *a-i*, **where *c,f,i* denotes EYOC, *a,d,g* are the fair generalization results of supervised methods and *b,e,h* mark the oracle supervised performance with labels on the new dataset**. EYOC is the only unsupervised method. We use 'FCGF + C' to denote FCGF trained with progressive datasets, which is a theoretical upper bound for EYOC. All features are registered using RANSAC.

help of pose labels, presumably from KITTI or WOD; Otherwise, we use hard spatial filtering. The complete training of EYOC consists of 200 epochs with 0.001 learning rate and 1×10^{-4} weight decay, same as FCGF, implemented with MinkowskiEngine [8] and Pytorch3D [39].

Inference. When conducting a comparison with previous methods, we apply RANSAC [14] on all methods including EYOC for fairness. Otherwise, we default EYOC inference to SC^2-PCR [7] for speed and performance.

Metrics. We report 5 metrics according to existing literature [9, 18, 28]: Registration Recall (RR), Relative Rotation Error (RRE), Relative Translation Error (RTE), Mean RR (mRR), and Inlier Ratio (IR), the formal definition of which can be found in Appendix Sec. 7.2. We apply IR *on the generated labeler correspondences* to indicate their quality during training.

5.2. Overall Performance

We compare both a generalization setting (*a,d,g*) and finetuning setting (*b,e,h*) for SOTA supervised methods, against the unsupervised EYOC (*c,f,i*) on three datasets, KITTI [16], WOD [43], and nuScenes [6], respectively in Tab. 1.

We first notice that supervised methods do fail to generalize to different datasets, according to *a-b,d-e,* and *g-h* in Tab. 1. Generalizing from WOD to KITTI, which are both 64-line datasets with small domain shift, supervised methods suffer 5.6% and 15.6% mRR drop respectively for FCGF [9] and Predator [20], when compared with models trained on KITTI directly (rows *a* and *b*). Similar results are seen generalizing from KITTI to WOD as well (rows *d* and *e*), with 19.6% and 15.7% mRR drop for FCGF and Predator, respectively. On the other hand, a harder dataset, nuScenes with only a 32-laser LiDAR, struggles to support supervised training. We witness worse supervised performance than generalization scores from WOD

No.	LF	SF-h	SF-a	SR+CR	PD	1st Epoch Labeler IR	mRR	[40, 50] RR	RRE	RTE
a	-	-	-	-	✓	5.1				
b	✓	-	-	-	✓	1.5				
c	✓	-	-	✓	✓	0.6				
d	✓	-	✓	-	✓	5.9		N/C		
e	✓	✓	-	-	✓	5.9				
f	-	-	✓	✓	-	0.0				
g	✓	-	✓	✓	✓	7.8	87.5	66.8	**1.3**	**29.7**
h	-	-	-	✓	✓	18.4	84.6	60.3	1.4	33.9
i	-	-	✓	✓	✓	43.3	**88.0**	**68.8**	**1.3**	31.8
j	-	✓	-	✓	✓	**53.2**	87.6	67.8	1.31	32.2

λ	[40,50]
0.0	71.9
0.1	70.4
0.2	**73.9**
0.3	71.4
0.4	69.3
0.5	71.9
0.6	69.8
0.7	72.9
0.8	67.3
0.85	58.8
0.9	N/C
0.99	N/C

d_{thresh}	1st Epoch Labeler IR
0	18.4
5	18.5
10	25.2
15	31.1
20	31.4
25	29.4
30	45.1
35	49.0
40	**53.2**
45	43.3

s_{thresh}	1st Epoch Labeler IR
0.0	18.4
0.1	25.4
0.2	30.7
0.3	31.5
0.4	31.1
0.5	34.9
0.6	**43.3**
0.7	N/C
0.8	N/C
0.9	N/C

Table 2. **Ablation study of EYOC.** Labeler IR (%), mRR (%), RR@$[40, 50]$ (%), RRE (°), and RTE (cm) on KITTI val set are presented. Lowe Filtering (LF), Spatial Filtering of *hard* (SF-h) or *adaptive* (SF-a) strategies, Speculative Registration and Correspondence Rediscovery (SR+CR), progressive Dataset (PD), EMA decay factor λ, and two parameters of Spatial Filtering, d_{thresh}, s_{thresh}, are ablated.

for FCGF when comparing the rows *g* and *h*. Nonetheless, their generalization scores are also subpar, hitting merely 67.1% and 34.5% mRR with FCGF and Predator, respectively on nuScenes in row *g*. Additionally, contrary to the common belief, Predator performs much worse than FCGF on an out-of-domain dataset, nuScenes, in row *g*.

Does unsupervised finetuning improve upon supervised methods on out-of-domain unlabelled data? By comparing *a-c*, *d-f*, and *g-i* in Tab. 1, we confirm that EYOC improves upon fixed supervised models by a considerable margin through unsupervised finetuning. On KITTI, EYOC surpasses raw FCGF, achieving $83.2\%(+11.4\%)$ and $80.6\%(+8.8\%)$ mRR by training from scratch and finetuning, respectively. On WOD and nuScenes, the respective figures are $78.4\%(+8.5\%)$ and $77.3\%(+7.4\%)$ on WOD, $61.7\%(-5.3\%)$ and $68.4\%(+1.3\%)$ on nuScenes compared to FCGF. We conclude that, given a pretrained model and an incoming unlabelled dataset, applying EYOC for unsupervised training/finetuning provides a considerable performance boost on the new dataset.

Is EYOC comparable to supervised methods on labelled data? Unsupervised methods have to perform similarly to supervised ones in order to be considered valuable. Through comparing *b-c*, *e-f*, and *h-i* in Tab. 1, we find that EYOC exhibits comparable performance with SOTA fully-supervised methods when trained on the same dataset. On KITTI, mRR of EYOC is only 4.7% and 1.4% lower than that of the best-performing Predator and FCGF+C, respectively. Other low-overlap registration methods, excluding CoFiNet [54], are less suitable for outdoor scenarios, as SpinNet [2], D3Feat [4], and Geotransformer [37] suffer from divergence. In the meantime, different results are reported on WOD where EYOC is 10.9% behind FCGF but 1.2% ahead of FCGF+C, indicating that FCGF+C is not always effective on all datasets. EYOC echibits stronger results on nuScenes, surpassing FCGF+C by 9.9% instead. We conclude that EYOC does perform similarly to fully supervised methods while requiring no pose labels at all.

5.3. Ablation

Structural components. We first ablate supporting structures of EYOC in Tab. 2, including Lowe Filtering (LF), Spatial Filtering with both *hard* (SF-h) and *adaptive* (SF-a) strategies, Speculative Registration + Correspondence Rediscovery (SR+CR), and the Progressive Dataset (PD). Judging from *a-b* and *g-i*, Lowe's filter deteriorates IR by 3.60% and 35.5%, respectively, contrary to previous findings on indoor RGB-D images. We keep Lowe filtering as an option in case of other datasets. Also, lone Spatial Filtering or speculative registration both fail to support training according to *c,d,e*. The best-performing setup (*i*) fails completely without Progressive Dataset (*f*) at 0.0% IR, indicating the importance of the Progressive Dataset strategy. On the other hand, converged setups reveal consistently higher IR up to 53.2%. SF-h (*i*) and SF-s (*j*) achieve 88.0% and 87.6% mRR, respectively, slightly better than not using Spatial Filtering (*h*) which achieves 84.6% mRR. Similar trends are observed with respective performance on long-range pairs as well, where *i* and *j* outperforms *h* by 8.5% and 7.5% RR, respectively. We default EYOC structure to SF-a, SR+CR, and PD (*i*) for the highest performance.

Parameter choices. Three parameter choices, λ, d_{thresh}, and s_{thresh}, are discussed in Tab. 2 as well. For the EMA decay factor λ, any value less than 0.7 achieves similar results averaging at 71.4%, while larger λ quickly drains the performance. On the other hand, similar to our previous findings Sec. 4.3, IR marks better scores with stricter thresholds of d_{thresh} and s_{thresh} (*i.e.*, using regions farther from the LiDAR), but the number of correspondences could shrink to the point of causing divergence under an extreme threshold. In light of this phenomenon, we choose $\lambda = 0.2$, $d_{thresh} = 40m$, and $s_{thresh} = 0.6$ are default parameters for the best performance just before the divergence line. Should a divergence occur on new datasets, these thresholds

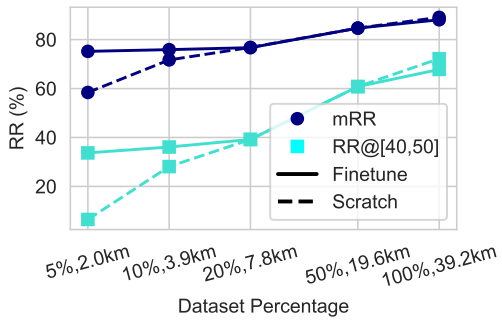

Figure 5. **Comparison between finetuning from WOD and training from scratch for EYOC**, with the first 5% to 100% of unlabelled KITTI, where both RR on $d \in [40, 50]$ and mRR are displayed. The horizontal axis is in log scale. Finetuning exhibits more stability before 20%, while training from scratch performs better after 50%.

	Training (one pass)						# Training
	Data	NN-S	Feat.	Label Gen.	Loss	Total	Required
FCGF [9]	692	-	128	-	356	1176	×2
FCGF*	17	33	152	-	301	503	×2
EYOC	18	-	170	381	296	865	×1

Table 3. Time analysis of EYOC, FCGF [9], and FCGF with GPU-accelerated NN-Search (denoted with *) in milliseconds. The number of complete training routines required for a network is listed in the last column.

could be lowered to cater to new data distributions.

5.4. Finetuning versus From Scratch

We further compare the finetuning and training-from-scratch strategies for EYOC with different portions of the new dataset KITTI, while assuming a pretrained model on WOD is available. Metrics including RR @ $d \in [40, 50]$, mRR, and driving distance (km) on KITTI are displayed with the first 5% to 100% of KITTI, as illustrated in Fig. 5. Overall, performance of both methods increase with the amount of training data; However, finetuning grants more stability by inheriting knowledge from the previous dataset, therefore performing better with smaller datasets then 20% $(7.8km)$ where the mRR stablizes around 75%. On the other hand, training from scratch achieves better results after 50% $(19.6km)$, peaking at the full dataset with 89.1% mRR and 72.2% RR @ $d \in [40, 50]$, respectively. We conclude that finetuning is better for datasets roughly shorter than 10km, while training from scratch would be a better choice for larger datasets.

5.5. Time Analysis

We break down the training time for FCGF [9] and EYOC in Tab. 3. Because the NN-search in Correspondence Rediscovery of EYOC is accelerated with GPU using Pytorch3D [39], it is necessary to apply the same trick to the baseline FCGF for fair comparison, which is termed 'FCGF*'. While EYOC needs an additional 381ms for label generation, it completes training once and for all, resulting in the lowest total training time. On the other hand, FCGF* is trained twice to prevent divergence [28] as detailed in Sec. 5.1. In comparison, vanilla FCGF ranks the slowest due to a prolonged data loading time of 692ms. We conclude that EYOC enjoys a lower training cost than its supervised counterpart.

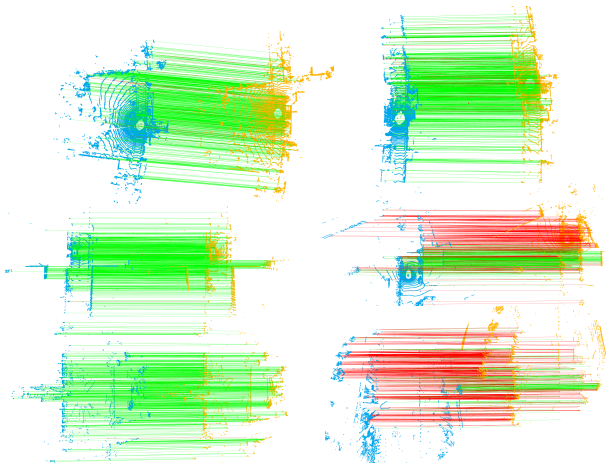

Figure 6. **Visualization of clean correspondence labels on KITTI (top row), nuScenes (middle row), and WOD (bottom row)**, where correspondences with $\leq 1m$ location error are colored green and otherwise red. Even when Speculative Registration fails, most of the false correspondences are in parallel to correct ones, they are just not precise but still informative.

6. Conclusion

We have proposed EYOC, an unsupervised distant point cloud registration technique that requires nothing more than consecutive LiDAR sweeps, which is easily acquired on-the-fly with self-driving vehicles. With the correspondence filtering pipeline built upon our investigations, EYOC allows a 3D feature extractor to generate labels for itself, enabling fully unsupervised training. Extensive experiments demonstrate that, while enjoying comparable performance to supervised methods, EYOC also has a lower training cost, thus being preferable compared to the traditional 'manual labelling + supervised training' paradigm. EYOC's unrivalled capability of finetuning on new data distributions marks a step towards the mass deployment of collaborative sensing on SDVs.

Acknowledgement. This work was supported in part by the Natural Science Foundation of Shanghai (Grant No.22ZR1400200) and the Fundamental Research Funds for the Central Universities (No. 2232023Y-01).

References

[1] Sk Aziz Ali, Kerem Kahraman, Gerd Reis, and Didier Stricker. RPSRNet: End-to-end trainable rigid point set registration network using Barnes-Hut 2D-Tree representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13100–13110, 2021. 1

[2] Sheng Ao, Qingyong Hu, Bo Yang, Andrew Markham, and Yulan Guo. SpinNet: Learning a general surface descriptor for 3D point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11753–11762, 2021. 2, 6, 7, 1

[3] Sheng Ao, Qingyong Hu, Hanyun Wang, Kai Xu, and Yulan Guo. BUFFER: Balancing accuracy, efficiency, and generalizability in point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1255–1264, 2023. 2

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6359–6367, 2020. 2, 6, 7, 1

[5] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. PointDSC: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15859–15869, 2021. 2

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 5, 6, 3

[7] Zhi Chen, Kun Sun, Fan Yang, and Wenbing Tao. SC2-PCR: A second order spatial compatibility for efficient and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13221–13231, 2022. 2, 5, 6, 1

[8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 6

[9] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1, 2, 5, 6, 8

[10] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2514–2523, 2020. 2

[11] Haowen Deng, Tolga Birdal, and Slobodan Ilic. PPFNet: Global context aware local features for robust 3D point matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–205, 2018. 2

[12] Mohamed El Banani and Justin Johnson. Bootstrap your own correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6433–6442, 2021. 1, 3, 4

[13] Mohamed El Banani, Luya Gao, and Justin Johnson. UnsupervisedR&R: Unsupervised point cloud registration via differentiable rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7129–7139, 2021. 1, 3, 4

[14] 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. 2, 6

[15] Kexue Fu, Shaolei Liu, Xiaoyuan Luo, and Manning Wang. Robust point cloud registration framework based on deep graph matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8893–8902, 2021. 1

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012. 2, 4, 5, 6

[17] Rémi Giraud, Vinh-Thong Ta, Aurélie Bugeau, Pierrick Coupé, and Nicolas Papadakis. SuperPatchMatch: An algorithm for robust correspondences using superpixel patches. *IEEE Transactions on Image Processing*, 26(8):4068–4078, 2017. 3

[18] Zan Gojcic, Caifa Zhou, and Andreas Wieser. Learned compact local feature descriptor for tls-based geodetic monitoring of natural outdoor scenes. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:113–120, 2018. 6

[19] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3D point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5545–5554, 2019. 2

[20] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. PREDATOR: Registration of 3D point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4267–4276, 2021. 1, 2, 3, 4, 6

[21] Behley Jens, Garbade Martin, Milioto Andres, Quenzel Jan, Behnke Sven, Stachniss Cyrill, and Gall Jurgen. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1

[22] Andrew E Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999. 2

[23] Felix Järemo Lawin, Martin Danelljan, Fahad Shahbaz Khan, Per-Erik Forssén, and Michael Felsberg. Density adaptive point set registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3829–3837, 2018. 4

[24] Donghoon Lee, Onur C Hamsici, Steven Feng, Prachee Sharma, and Thorsten Gernoth. DeepPRO: Deep partial

point cloud registration of objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5683–5692, 2021. 1, 2, 3

[25] Junha Lee, Seungwook Kim, Minsu Cho, and Jaesik Park. Deep hough voting for robust global registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15994–16003, 2021. 2

[26] Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5554–5564, 2022. 1

[27] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 1

[28] Quan Liu, Yunsong Zhou, Hongzi Zhu, Shan Chang, and Minyi Guo. APR: Online distant point cloud registration through aggregated point cloud reconstruction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1204–1212. International Joint Conferences on Artificial Intelligence Organization, 2023. Main Track. 1, 2, 3, 5, 6, 8

[29] Quan Liu, Hongzi Zhu, Yunsong Zhou, Hongyang Li, Shan Chang, and Minyi Guo. Density-invariant features for distant point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18215–18225, 2023. 1, 2, 5, 3

[30] H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293 (5828):133–135, 1981. 4

[31] Fan Lu, Guang Chen, Yinlong Liu, Lijun Zhang, Sanqing Qu, Shu Liu, and Rongqi Gu. HRegNet: A hierarchical network for large-scale outdoor LiDAR point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16014–16023, 2021. 1, 2, 3

[32] Guofeng Mei, Hao Tang, Xiaoshui Huang, Weijie Wang, Juan Liu, Jian Zhang, Luc Van Gool, and Qiang Wu. Unsupervised deep probabilistic approach for partial point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13611–13620, 2023. 1, 3

[33] Michael Montemerlo, Sebastian Thrun, Daphne Koller, Ben Wegbreit, et al. FastSLAM: A factored solution to the simultaneous localization and mapping problem. *Association for the Advancement of Artificial Intelligence / Innovative Applications of Artificial Intelligence Conference*, 593598, 2002. 1

[34] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 1

[35] Fabio Poiesi and Davide Boscaini. Distinctive 3D local deep descriptors. In *Proceedings of the International Conference on Pattern Recognition*, pages 5720–5727. IEEE, 2021. 2

[36] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 2

[37] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11143–11152, 2022. 1, 2, 3, 4, 6, 7

[38] Zequn Qin, Jingyu Chen, Chao Chen, Xiaozhi Chen, and Xi Li. UniFusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8690–8699, 2023. 1

[39] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv:2007.08501*, 2020. 6, 8

[40] Amir Rosenfeld and John K Tsotsos. Intriguing properties of randomly weighted networks: Generalizing while learning next to nothing. In *Proceedings of the Conference on Computer and Robot Vision*, pages 9–16. IEEE, 2019. 4

[41] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3212–3217. IEEE, 2009. 2

[42] Martin Simon, Karl Amende, Andrea Kraus, Jens Honer, Timo Samann, Hauke Kaulbersch, Stefan Milz, and Horst Michael Gross. Complexer-YOLO: Real-time 3D object detection and tracking on semantic point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[43] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5, 6, 3

[44] Gusi Te, Wei Hu, Amin Zheng, and Zongming Guo. RGCNN: Regularized graph cnn for point cloud segmentation. In *Proceedings of the ACM international conference on Multimedia*, pages 746–754, 2018. 4

[45] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. KPConv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6411–6420, 2019. 2

[46] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *Proceedings of the European Conference on Computer Vision*, pages 356–369. Springer, 2010. 2

[47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 4

[48] Yue Wang and Justin M Solomon. PRNet: Self-supervised learning for partial-to-partial registration. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[49] Wenxuan Wu, Zhongang Qi, and Li Fuxin. PointConv: Deep convolutional networks on 3D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019. 4

[50] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. SqueezeSegV3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 1–19. Springer, 2020. 1

[51] Zi Jian Yew and Gim Hee Lee. 3DFeat-Net: Weakly supervised local 3D features for point cloud registration. In *Proceedings of the European Conference on Computer Vision*, pages 607–623, 2018. 4

[52] Zi Jian Yew and Gim Hee Lee. RPM-Net: Robust point matching using learned features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11824–11833, 2020. 1

[53] Zi Jian Yew and Gim Hee Lee. REGTR: End-to-end point cloud correspondences with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6677–6686, 2022. 1, 2, 3

[54] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. CoFiNet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34:23872–23884, 2021. 1, 2, 3, 6, 7

[55] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022. 1

[56] Junle Yu, Luwei Ren, Yu Zhang, Wenhui Zhou, Lili Lin, and Guojun Dai. PEAL: Prior-embedded explicit attention learning for low-overlap point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17702–17711, 2023. 1, 3

[57] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1811, 2017. 2

[58] Xumiao Zhang, Anlan Zhang, Jiachen Sun, Xiao Zhu, Y Ethan Guo, Feng Qian, and Z Morley Mao. EMP: Edge-assisted multi-vehicle perception. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 545–558, 2021. 1

[59] Xiyu Zhang, Jiaqi Yang, Shikun Zhang, and Yanning Zhang. 3D registration with maximal cliques. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2023. 2, 5

[60] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *Proceedings of the European Conference on Computer Vision*, pages 766–782. Springer, 2016. 2

[61] Hanqi Zhu, Jiajun Deng, Yu Zhang, Jianmin Ji, Qiuyu Mao, Houqiang Li, and Yanyong Zhang. VPFNet: Improving 3D object detection with virtual point based LiDAR and stereo data fusion. *IEEE Transactions on Multimedia*, 2022. 1