DPCN++: Differentiable Phase Correlation Network for Versatile Pose Registration

Zexi Chen, Yiyi Liao, Haozhe Du, Haodong Zhang, Xuecheng Xu, Haojian Lu, Member, IEEE, Ronxiong and Yue Wang

Abstract—Pose registration is critical in vision and robotics. This article focuses on the challenging task of initialization-free pose registration up to 7DoF for homogeneous and heterogeneous measurements. While recent learning-based methods show promise using differentiable solvers, they either rely on heuristically defined correspondences or require initialization. Phase correlation seeks solutions in the spectral domain and is correspondence-free and initialization-free. Following this, we propose a differentiable solver and combine it with simple feature extraction networks, namely DPCN++. It can perform registration for homo/hetero inputs and generalizes well on unseen objects. Specifically, the feature extraction networks first learn dense feature grids from a pair of homogeneous/heterogeneous measurements. These feature grids are then transformed into a translation and scale invariant spectrum representation based on Fourier transform and spherical radial aggregation, decoupling translation and scale from rotation. Next, the rotation, scale, and translation are independently and efficiently estimated in the spectrum step-by-step. The entire pipeline is differentiable and trained end-to-end. We evaluate DPCN++ on a wide range of tasks taking different input modalities, including 2D bird’s-eye view images, 3D object and scene measurements, and medical images. Experimental results demonstrate that DPCN++ outperforms both classical and learning-based baselines, especially on partially observed and heterogeneous measurements.

Index Terms—Pose registration, differentiable solver, end-to-end learning.

I. INTRODUCTION

Pose registration aims to estimate the relative pose given a pair of measurements. It stems as a core competence for numerous applications, including object pose estimation [58], scene reconstruction [43], localization [44], and medical imaging [74]. Pose registration remains a challenging task in the initialization-free setting, in particular for partially observed or heterogeneous measurements.

Recently, learning-based pose registration methods have shown promise in addressing these challenges. The pioneering work directly learns pose regression [39], however, the lack of interpretability hinders generalization. More recent approaches incorporate a differentiable and explicit solver into deep neural networks. This allows for enhancing the interpretability by splitting the task into two stages, trainable feature extraction and classical pose solving, while being able to be trained end-to-end. Despite usually being parameter-free, the pose solver is a key component as it provides gradients for the feature extraction network. In this work, we investigate a key question, what is an ideal solver for learning-based pose registration?

Existing learning-based methods leverage two types of solvers depending on whether they rely on explicit correspondences. Correspondence-based solvers, e.g., an SVD solver, infer the pose based on a set of matched points [59], [71], [72]. This line of approaches can obtain a global optimum given good correspondences. However, it requires the network to explicitly learn intermediate (typically redundant) keypoints and their matching relationships, thereby forcing the network to address a higher-dimensional problem than the pose registration itself. Moreover, it heavily relies on outlier elimination to obtain robust pose registration, and the network can be misguided when the solver fails given bad correspondences.

Another line of work avoids intensive feature matching by leveraging correspondence-free solvers. Given a pair of extracted features instead of explicit correspondences, this type of solver iteratively updates the pose to maximize the feature similarity along the gradients [1]. This gradient-based solver requires a good initialization and easily gets stuck in the local minimum. The suboptimal solution can also mislead the feature extraction network.

Analyzing these two lines of work, we identify two key properties of an ideal solver for learning-based pose registration: i) It should avoid explicit correspondence, thus relieving the burden of learning heuristically defined features and preventing outlier elimination. ii) It should avoid initializations and achieve global registration [69] in a non-iterative manner, hence providing direct supervision to the feature extraction network towards the best outcome. Classical phase correlation [5], [55] who decouple rotation, translation, and seek registration solutions in the spectral domain satisfies both criteria, i.e., being correspondence-free and initialization-free. Following this phase correlation, we propose a differentiable solver and combine it with simple feature extraction networks, namely Deep Phase Correlation Network (DPCN++). DPCN++ can perform registration for homo/hetero inputs and generalizes well on unseen objects, achieving superior performance compared to existing learned-based methods.
In contrast to the classical phase correlation, our learning-based approach not only achieves better performance but can also estimate relative poses from limited overlaps as well as heterogeneous measurements, e.g., CT and MRI. Moreover, we provide a general pose registration framework for both bird’s eye view (BEV) 2D images and 3D measurements.

Fig. 1 illustrates our general pose registration framework by leveraging a simple deep feature extraction network and a differentiable correspondence-free global solver. Specifically, we reformulate the pose registration problem as feature grid matching. Given a pair of homogeneous/heterogeneous measurements, we adopt feature extraction networks to learn the dense feature grids, respectively. At the core of our method is a correspondence-free, differentiable solver that decouples the estimation of rotation, translation, and scale. Based on the feature grids, we build the translation and scale-invariant spectrum representation to decouple the translation and scale from rotation, using Fourier transform and spherical radial aggregation. Next, the rotation, scale and translation are efficiently estimated in the spectrum step-by-step independently. In each step, the estimator is built upon the whole sub-solution space (translation, scale or rotation) correlation, thus is initialization-agnostic. By modeling the estimator in a probabilistic manner, our solver is fully differentiable and allows for updating the feature extraction network from pose loss end to end. After employing deep neural networks, we do not claim that the whole DPCN++ is globally convergent but rather employ a global solver for end-to-end learning. Please refer to Appendix B.3 for further discussion, available online. We summarize our contributions as follows:

- Our core contribution is to incorporate a globally convergent, correspondence-free, up to 7 ° of freedom (7DoF) phase correlation solver into learning-based pose registration. This allows for learning a simple feature matching network guided by pose estimation error instead of manually designed correspondences.
- Our framework is generally applicable to both 3D measurements and gravity aligned BEV images. Moreover, our method can serve inputs across different modalities by simply leveraging two different feature extraction networks.
- We validate the effectiveness of our method via extensive experiments. Our method achieves superior performance compared to both classical and learning-based baselines, especially on partially observed or heterogeneous measurements.
- We release a multi-modal Aero-Ground Dataset, intending to facilitate cooperative localization between ground mobile robots, micro aerial vehicles (MAVs) and satellites.

Our work is an extension of a conference article [10], where we introduce the Differentiable Phase Correlation Network (DPCN) that solves the pose registration problem between two gravity-aligned BEV images. This article extends our conference article in the following ways: 1) Scope: We consider the registration dimension of both 2D to 3D measurements with degrees of freedom up to 7 DoF. Our conference article only considers the 2D registration task with 4 DoF which is treated as a degenerated version of the 7 DoF case in this article. 2) Methodology: We introduce a general algorithm leveraging spherical $S^2$ and $SO(3)$ Fourier transforms to address 3D representations in the spectral domain. This leads to a new and comprehensive framework capable of solving pose registration for both homogeneous and heterogeneous measurements in 2D or 3D. 3) Experiments: We conduct extensive experiments on versatile 3D-3D pose registration tasks on both synthetic and real-world datasets, including object-level and scene-level point clouds and meshes as well as medical images. We further conduct case studies on symmetric objects and show that our method can identify all plausible solutions.

II. RELATED WORKS

A. 3D-3D Homogeneous Registration

3D-3D pose registration based on homogeneous measurements is one of the most studied scenarios in pose registration. We categorize previous works on this problem into direct
regression, correspondence-based and correspondence-free methods.

Direct Regression Methods. To guide the feature learning by a differentiable pose regression solver in an end-to-end manner, some methods learn to directly regress the pose by a fully connected network. AlignNet [19] utilizes a MLP layer for alignment estimation from the global feature. PCRNNet [52] introduces a canonical pose as an intermediate stage to reduce the space for direct pose regression. 3DRegNet [40] follows a similar way but utilizes correspondence as the input. These methods reveal the trend of end-to-end learning. One drawback is that the output pose is directly regressed by neural networks instead of being estimated by an explicit solver, so the lack of interpretability usually leads to weak generalization.

Correspondence-Based Methods. The correspondence-based registration method originates from the well-known iterative closest point (ICP) [4]. Many follow-up variants are proposed [11], [18], [49], [50]. The solver of the point-to-point ICP is global convergent. The main limitation is the nearest neighbor based correspondence strategy, which highly depends on the initialization. Therefore, learning is employed to improve the feature extraction and feature matching. DCP [59] incorporates DGCNN [60] for point cloud embedding and an attention-based module for feature matching, followed by a differentiable SVD solver for an end-to-end pose estimation architecture. DGR [12] uses fully convolutional geometric features [13] for feature extraction and applies 6-dimensional segmentation for correspondence prediction. DeepGMR [71] learns to find pose-invariant correspondences between Gaussian mixture models that approximate the shape, and compute the transformation based on the model parameters, so that the noise in the raw point cloud can be suppressed. The main limitation of correspondence-based learning methods is the robustness of the SVD solver. It enforces feature matching to be almost outlier free, which is unfortunately hard to achieve for the feature network.

Concerning the problem mentioned above, another line of efforts has been made for solver improvement. Go-ICP [70] utilizes a Branch-and-Bound (BnB) method to search the whole SE(3) space to get a globally optimal solution. FGR [75] uses second-order optimization and applies a Geman-McClure cost function to reach global registration of high accuracy. DGR [12] proposes an outlier robust registration method as post-processing to finetune the result yielded by the learning stage. TEASER [69] is the first certifiable global registration algorithm dealing with a large percentage of outliers, of which the main idea is the decoupling of scale, translation and rotation. However, these solvers are usually not differentiable, making it hard to guide the features and matching in an end-to-end manner.

Correspondence-Free Methods. The core idea of correspondence-free methods is to estimate the pose based on the similarity of features between two measurements. According to the convergence of solver, we further categorize this line of methods into locally convergent and globally convergent ones.

The locally convergent correspondence-free pose registration is inspired by the optical flow in 2D image registration. A pioneering work, PointNetLK [1], extracts global features using PointNet [46] and extends Lucas-Kanade algorithm [35] to feature space for iterative estimation of the pose. Li et al. [31] follow the idea of PointNetLK and propose an analytical form of Jacobian to improve the performance in mismatched conditions. Huang et al. [23] further add a feature-metric loss to the PointNetLK framework as a side task, which directly enforces the feature similarity between the aligned measurements. One of the limitations of this line of methods is the iterative solver, which is sensitive to the initialization, and the local minima may mislead the feature learning.

The globally convergent correspondence-free method mainly employs the idea of phase correlation. Bülow et al. [5] and PHASER [3] both utilize spherical and spatial Fourier transforms to estimate the relative pose using correlation [57] in the spectrum. The global convergence lies in the correlation, which is an intrinsically exhaustive search, but can be evaluated efficiently via decoupling in the spectrum. Inspired by these methods, we introduce a differentiable version of phase correlation to enable end-to-end learning based on a globally convergent solver. In contrast to [3], [5], our framework achieves better registration performance and is applicable to versatile pose registration tasks by learning from data. Recently, with the progress of geometric deep learning, Zhu et al. [77] apply SO(3)-equivariance embedding for feature learning. The rotation is globally estimated in the feature space, skipping the stage of making correspondence. However, this method may fail when the relative pose has both relative rotation and translation.

B. 3D-3D Heterogeneous Registration

Multi-modal data may vary in the data structure, physical and anatomical principles, intensity and noise [51], which makes the registration task much more challenging. Heterogeneous pose registration tasks mainly come from medical image analysis. In order to register heterogeneous measurements, one line of works [8], [9], [53] employs handcrafted mutual information as a similarity measure [61] and utilizes bio-inspired optimization algorithm for minimization, like Particle Swarm Optimization (PSO). Kisaki et al. [28] accelerate the registration by the Levenberg-Marquardt optimization algorithm, which, however, depends on the initial value. Several algorithms also use the Branch-and-Bound framework to overcome the randomness of the bio-inspired optimization algorithms [17], [41]. Yang et al. [68] propose a general framework for certifiable robust geometric perception, which exploits the registration between the 3D mesh model and point clouds. As the efforts for solver improvement in homogeneous pose registration, these solvers show good performance when the feature is correctly designed, but they are not differentiable, thus cannot be applied to guide the automatic feature learning.

Deep learning is also introduced to solve the heterogeneous registration, mainly for learning the similarity metric or estimating transformation directly. Lee et al. [30] reformulate the problem of learning similarity measures to a binary classification of aligned and misaligned patches, tested on CT-MRI and PET-MRI volumes. CNN/RNN based methods [20], [54], [65] are proposed to learn local similarity metrics for better registration. Reinforcement learning is utilized in [32], [36] to learn the
pose regression directly, which can be trained in an end-to-end manner, but may have a weaker generalization due to the lack of the explicit solver. Cao et al. [7] employ the network to warp the image in a non-rigid way for medical image registration. These methods show the advantages of the learning-based methods, especially in learning the feature space across heterogeneous models. One of their limitations is the dependency on the initialization, which constrains the convergence basin of the learned similarity and warping strategy.

C. 2D-2D Registration

The trend in 2D-2D registration is similar to 3D-3D registration. Early methods employ hand-crafted image feature point matching to generate correspondences, like SIFT [33], which is then fed to the solver for pose registration. To deal with large appearance changes in two images, learning-based methods are applied for feature detection and matching, like SuperPoint [15], R2D2 [47] and D2Net [16] which shows good performance in visual localization [73]. To avoid the ground truth correspondences, direct regression is connected to the feature network in an end-to-end way in DAM [45]. However, as in 3D-3D registration, such solvers may have weaker generalization performance due to the lack of explicit constraints. Currently, a common practice in image registration is to employ a robust solver as a post-processing step [24].

General image registration shown above deals with 6DoF pose. In this article, we focus on the 4DoF registration problem for BEV images. It can be solved by the correspondence-based methods above, but also correspondence-free methods. To build robust similarity under appearance change for images, Kaslin et al. [25] apply normalized cross-correlation (NCC) on geometric measurements, and WNCC [67] recursively rotates the source in a small range to match the rotation. Kummerle et al. [29] and Ruchti et al. [48] utilize hand-craft features to localize LiDAR against satellite maps. To learn the feature instead of hand-crafted ones, Kim et al. [27] utilize feature maps trained from other tasks to measure the similarity. However, it is unclear how to design tasks to improve feature learning. More recently, Lu et al. [34] and Tang et al. [56] propose to learn the embedding with a differentiable exhaustive search solver in the discretized solution space, and show good results. But due to the expensive search, the efficiency is low, and the discretization is only applied in the local range. To improve the search efficiency, Barnes et al. [2] propose to evaluate the similarity in the frequency spectrum. To further eliminate the constraint of local range, phase correlation (PC) [55], which searches the global optimal pose in the whole solution space by reformulating the problem into the spectrum, is investigated in the context of end-to-end deep learning for homogeneous [63] and heterogeneous pose registration [10]. These works show promises of introducing a global convergent and a differentiable solver, which guides feature learning towards the best outcome. However, both methods can only be applied to 2D. In contrast, we leverage the classic phase correlation [5], [55], combine them with deep learning, and propose a general framework for versatile pose registration that is applicable to both 3D-3D and 2D-2D pose registration.

III. OVERVIEW

Given two measurements with homogeneous or heterogeneous modality, e.g., point cloud and mesh, our goal is to estimate their relative pose up to 7DoF, including rotation, scale and translation, without referring to an initial value. In this work, we first derive the method for 7DoF 3D-3D pose registration and consider the 4DoF orthogonal 2D-2D pose registration as its degenerated version. We now formally state the problem in sequel.

A. Problem Statement

Let \( v_1 \) and \( v_2 \) denote two measurements linked by an unknown relative pose \( T = \{ t, r, \mu \} \) with translation \( t \in \mathbb{R}^3 \), rotation \( r \) in \( SO(3) \), and isotropic scale \( \mu \). A general end-to-end learning-based pose registration problem can be stated as

\[
\min_{\theta_1, \theta_2} \| T^* - \arg \max_T \mathcal{C}(Q_{\theta_1}(v_1), Q_{\theta_2}(\Omega TV_2)) \|,
\]

where \( Q_{\theta_1} \) and \( Q_{\theta_2} \) are two deep networks extracting features from the inputs, such as point clouds, meshes and signed distant fields (SDF), \( \Omega_T \) is the pose transformation applied to the input with proper operation generated by \( T \), \( T^* \) is the ground truth, \( \mathcal{C} \) is a scoring function measure the similarity. The key challenge is the design of a differentiable and initialization-agnostic \( \arg \max \mathcal{C}(\cdot, \cdot) \), so that the gradient of \( Q_{\theta_1} \) and \( Q_{\theta_2} \) can be correctly derived for learning.

In correspondence-based methods, \( \arg \max \mathcal{C}(\cdot, \cdot) \) is substituted with euclidean distance between correspondences and closed-form SVD solver, thus differentiable and initialization-agnostic. But these methods assume that the correspondence yielded by \( Q_{\theta_1} \) and \( Q_{\theta_2} \) is perfect, which may be difficult in practice, especially for heterogeneous inputs. In correspondence-free methods, \( \arg \max \mathcal{C}(\cdot, \cdot) \) is approximated by a numerical iterative process, which is differentiable, but the gradient-based iteration may mislead the learning of \( Q_{\theta_1} \) and \( Q_{\theta_2} \) due to the local minimums, thus highly depends on an initial value in both training and testing stage. Our proposed approach aims to solve the challenge of initialization dependency in corresponding-free registration methods by decoupling and seeking correlations between measurements in the spectral domain. We train the deep feature extractors to predict features of the inputs, regardless of their completeness and modality, to obtain the highest correlation when the measurements are aligned with the correct pose. This allows the method to be effective in a variety of scenarios, including partial-to-partial and heterogeneous registration.

B. Method Overview

Our method, DPCN++, also consists of a trainable feature extraction network and a differentiable pose registration solver. To address the challenge, we follow the classic idea of correspondence-free phase correlation in [5], and present a differentiable phase correlation solver as \( \arg \max \mathcal{C}(\cdot, \cdot) \).
Fig. 2. Overall pipeline of the DPC on grids. Rotation: The translation and scale are decoupled from the rotation once we focus on the magnitudes of the corresponding Fourier transform and project them to the sphere. Then the SO(3) rotation on the sphere is estimated with spherical phase correlation. Scale: The source is rotation-compensated with the target considering the voxelization process is given in Appendix B.2, available online. The translation and scale from the rotation [5], and estimate the rotation, scale and translation step-by-step inversely.

Translation Invariance. Following the common practice that lowercase letters indicate the time domain variables and the upperscases indicate the frequency domain, we have the corresponding 3D Discrete Fourier Transform of the grid $g(k)$ as $G(j)$

$$G(j) = \frac{1}{(2B)^3} \sum_{k} g(k) e^{-i2\pi j^{T}k},$$

where $j$ is the sampled frequency with the axes of $\{j_1, j_2, j_3\}$. With (6), the relationship between $g_1(k)$ and $g_2(k)$ as in (5) can be given in the frequency domain as

$$G_1(j) = \mu^{-3}G_2(r_{\alpha,\beta,\gamma}j\mu^{-1})e^{i2\pi(r_{\alpha,\beta,\gamma}j\mu^{-1})^{T}t}.$$  

By taking the magnitude of the frequency spectrum, we have

$$|G_1(j)| = \mu^{-3}|G_2(r_{\alpha,\beta,\gamma}j\mu^{-1})|,$$

Note that the two magnitude spectrum is invariant to translation $t$, and only depend on the relative rotation $r_{\alpha,\beta,\gamma}$ and scale $\mu$. For brevity, we further denote the magnitude as $\hat{G}(j)$

$$\hat{G}(j) = |G(j)| = |\hat{F}(g(k))|,$$

where $\hat{F}$ is the Fourier transform.

Scale Invariance. We further eliminate the scale dependency in $G(j)$. Specifically, we formulate a spherical function $s(\lambda)$ defined on the unit sphere $\lambda \in S^2$ by projecting each cell of $G(j)$ to sphere coordinates following

$$s(\lambda) = \sum_{j \in \Gamma_\lambda} \hat{G}(j)$$

$$\Gamma_\lambda = \left\{ j | \lambda = \left[ \arctan \left( \frac{j_2}{j_1} \right), \arccos \left( \frac{j_3}{\sqrt{j_1^2 + j_2^2 + j_3^2}} \right) \right] \right\},$$
where \( j = [j_1, j_2, j_3] \). In spherical coordinates, \( \lambda \) actually encodes a ray direction, so all cells passed by the ray are summed as the value of \( s(\lambda) \).

With radial aggregation, we note that \( s(\lambda) \) is invariant to scale change, and only depends on the rotation \( r_{\alpha, \beta, \gamma} \) as
\[
s_1(\lambda) = \mu^{-3} s_2(\lambda(r_{\alpha, \beta, \gamma}j)).
\]

(12)

By further normalizing the grids, \( \mu^{-3} \) in (12) is eliminated with \( s_1(\lambda) = s_2(\lambda(r_{\alpha, \beta, \gamma}j)) \). The projected sphere is further resampled following the “DH-Grid” sampling theorem by Driscoll and Healy [21] and create a resampled grid on the sphere with the size of \( 2B \times 2B \). Such sampling of \( s(\lambda) \) is implemented with interpolation, yielding \( \hat{s}(\hat{\lambda}) \) with the spherical coordinate \( \hat{\lambda} \) mapped from \( \lambda \).

A. Rotation Registration

Based on the analysis above, we arrive at two spherical functions derived from the raw measurements, which is translation and scale invariant. Now we propose global and differentiable rotation registration solver to align \( \hat{s}_1(\hat{\lambda}) \) and \( \hat{s}_2(\hat{\lambda}) \).

Spherical Phase Correlation. Given two spherical functions with relative rotation \( r_{\alpha, \beta, \gamma} \), we define the correlation on unit sphere as
\[
f(r_{\alpha, \beta, \gamma}) = \int_{\lambda \in S^2} \hat{s}_1(\hat{\lambda})\hat{s}_2(r_{\alpha, \beta, \gamma}\hat{\lambda})d\hat{\lambda},
\]

(13)

As shown in [37], the \( SO(3) \) Fourier transform of \( f(r_{\alpha, \beta, \gamma}) \), denoted as \( F_{mn}^l \), can be efficiently evaluated by point-wise multiplication, \( \odot \), of the spherical Fourier transform of two spheres as
\[
F_{mn}^l = \tilde{F}_{m,1}^l \odot \tilde{F}_{n,2}^l,
\]

(14)

where \( \tilde{F}_{m,1}^l \) and \( \tilde{F}_{n,2}^l \) are the spherical Fourier transform of \( \hat{s}_1(\hat{\lambda}) \) and \( \hat{s}_2(\hat{\lambda}) \), and \( \tilde{F}_{n,2}^l \) is the conjugate. In our case, \( m = n \). The spherical Fourier transform are defined as
\[
\tilde{S}_{m}^l = \frac{\sqrt{2\pi}}{2B} \sum_{\lambda} a_m s(\lambda) Y_m^l(\hat{\lambda}),
\]

(15)

where \( Y_m^l \) is the \((2l + 1)\) spherical harmonics of degree \( l \) and order \( m \), \( a_m \) is the weight to compensate for an oversampling at the pole, and empirically given by [21]. The complete derivation of the spherical Fourier transform can be found in [37]. Note that in our case, \( m = n \) and \( l = B \) in (14). Considering (13), (14), and (15), we have
\[
\tilde{F}_{SO(3)}(f(r_{\alpha, \beta, \gamma})) = \tilde{F}_{SO(3)}(\hat{s}_1(\hat{\lambda})) \odot \tilde{F}_{SO(3)}(\hat{s}_2(r_{\alpha, \beta, \gamma}\hat{\lambda})),
\]

(16)

where \( \tilde{F}_{SO(3)} \) and \( \tilde{F}_{SO(3)} \) denote for the \( SO(3) \) Fourier transform and spherical Fourier transform respectively.

Therefore, by taking the inverse \( SO(3) \) Fourier transform, denoted as \( iS_{SO(3)} \) of (16), we have the correlation map defined on the \( Z - Y - Z \) Euler angle space. The rotation registration can then be solved by searching the index with the highest correlation
\[
[\hat{\alpha}, \hat{\beta}, \hat{\gamma}] = \arg \max f(r_{\alpha, \beta, \gamma})
\]
\[
[\arg \max i\tilde{F}_{SO(3)}(\tilde{F}_{SO(3)}(f(r_{\alpha, \beta, \gamma})))
\]
\[
[\arg \max i\tilde{F}_{SO(3)}(\tilde{F}_{SO(3)}(\hat{s}_1(\hat{\lambda})))
\]
\[
\odot \tilde{F}_{SO(3)}(\hat{s}_2(r_{\alpha, \beta, \gamma}\hat{\lambda}))).
\]

(17)

The rotation registration solver actually evaluates the correlation for all possible Euler angles with the discretization resolution of \( \pm \frac{\pi}{2B} \) for \( \alpha \) and \( \gamma \), and \( \pm \frac{\pi}{2B} \) for \( \beta \) in an efficient way. Obviously, this solver is agnostic to the initial value and guarantees a global solution of rotation at the error level determined by the resolution.

Probabilistic Approximation. To make the solver differentiable, we approximate \( \arg \max \) by probabilistic modeling. We map the resultant correlation \( f(r_{\alpha, \beta, \gamma}) \) in (17) to a discrete probability density function \( p(r_{\alpha, \beta, \gamma}) \) by softmax function
\[
p(f(r_i)) = \frac{e^{\xi_r f(r_i)}}{\sum_j e^{\xi_r f(r_j)}},
\]

(18)

where \( \xi_r \) is the solver temperature controlling the peak property of the density, and can be learned by data. With an abuse of notation, we use the subscript \( i, j \) to indicate the enumeration of the discretized 3D \( Z - Y - Z \) Euler angle space. We take the expectation of the rotation as the estimation
\[
\hat{r} = \sum_r r_i p(f(r_i)).
\]

(19)

We design two types of losses for supervision. The L1 loss between \( \hat{r} \) and the ground truth \( r^* \)
\[
\mathcal{L}_{r, L1} = \|r^* - \hat{r}\|_1,
\]

(20)

and the KL-Divergence between the rotation density and a gaussian density \( \delta_r \) peaking at the ground truth \( r^* \)
\[
\mathcal{L}_{r, kld} = \text{KLD}(p(f(r)), \delta_r),
\]

(21)

where the standard deviation of the gaussian is a hand-crafted parameter. By probabilistic estimation, we approximate the global solver (17) using a differentiable process.

B. Scale Registration

Recall the magnitude spectrum in (8), we can eliminate the effect of rotation by rotating \( \hat{G}_2^l(j) \) with the estimated \( \hat{r}_{\hat{\alpha}, \hat{\beta}, \hat{\gamma}} \) in (17) and form \( \hat{G}_2^l(j) \)
\[
\hat{G}_2^l(j) = \hat{G}_2^l(\hat{r}_{\hat{\alpha}, \hat{\beta}, \hat{\gamma}}j),
\]

(22)

which is differed to \( \hat{G}_1(j) \) only by scale \( \mu \)
\[
\hat{G}_1(j) = \mu^{-3} \hat{G}_2^l(j \mu^{-1}).
\]

(23)

When dealing with such isotropic scale, we simplify the problem into 2D by accumulating the cubic grids along one axis (e.g. \( \hat{J}_1 \)) and form a 2D square grid
\[
\hat{G}_{(2d)}(j_{(2d)}) = \sum_{j_1 = -B}^{B - 1} \hat{G}_1(j),
\]

(24)
where $j_{(2d)} \in [-B, B - 1]^2$ stands for axes in 2D grid $G_{(2d)}$ (e.g. $\{j_1, j_2\}$). Therefore, (23) becomes

$$G_{1(2d)}(j_{(2d)}) = \mu^{-3} \tilde{G}_{2(2d)}(j_{(2d)}) \mu^{-1}. \quad (25)$$

By representing $G_{1(2d)}(j_{(2d)})$ in the log-polar coordinate, we have

$$\rho_{(\log |j_{(2d)}|, \angle j_{(2d)})} = \hat{G}_{1(2d)}(j_{(2d)}). \quad (26)$$

We then aggregate the second dimension of the log-polar representation by summation, yielding

$$\hat{\rho}_{(\log |j_{(2d)}|)} = \sum_{j_{(2d)}} \rho_{(\log |j_{(2d)}|, \angle j_{(2d)})}, \quad (27)$$

which leads to the representation reflecting the scale as shift

$$\hat{\rho}_1(\log |j_{(2d)}|) = \mu^{-3} \hat{\rho}_2(\log |j_{(2d)}|) + \log \mu^{-1}, \quad (28)$$

where $\log \mu^{-1}$ and $\log |j_{(2d)}|$ are further denoted as $\hat{\mu}$ and $\hat{j}$ for brevity.

**Cartesian Phase Correlation.** To solve the scale registration in (28), we apply the similar phase correlation as (13) for rotation, but in Cartesian coordinates. Specifically, the correlation is defined as

$$\hat{f} (\hat{\mu}) = \int \hat{\rho}_1(\hat{j}) \hat{\rho}_2(\hat{j} - \hat{\mu}) d\hat{j}. \quad (29)$$

We again apply Fourier transform to (28) to find the solution. Based on (7), we have

$$P_1(\hat{j}) = \mu^{-3} P_2(\hat{j}) e^{i2\pi \hat{j}^T \hat{\mu}}, \quad (30)$$

where $P_1$ and $P_2$ are the Fourier spectrum of $\hat{\rho}_1$ and $\hat{\rho}_2$, $\hat{j}$ is the sampled frequency. To evaluate the correlation in frequency domain, we calculate the cross-power spectrum $F_{\mu}$ as

$$\hat{F}_\mu(P_1, P_2) = P_1 \cdot P_2 = \mu^{-3} |P_2|^2 e^{i2\pi \hat{j}^T \hat{\mu}}. \quad (31)$$

Note that the phase of the cross-power spectrum is equivalent to the phase difference between grids. We then estimate the scale by taking the inverse Fourier transform ($i\hat{\mathcal{F}}$) of $\frac{\hat{F}_\mu}{|P_2|^2}$, yielding a normalized and impulsed correlation map $f(\hat{\mu})$, which ideally peaks at the real scale

$$f(\hat{\mu}) = i\mathcal{F} \left( \frac{\hat{F}_\mu}{|P_2|^2} \right) = i\mathcal{F} (\mu^{-3} e^{i2\pi \hat{j}^T \hat{\mu}}). \quad (32)$$

As (19), the estimation is built as

$$\hat{\mu} = \sum_{\hat{\mu}} \hat{\mu} p(f(\hat{\mu})), \quad (33)$$

where $p(f(\hat{\mu}))$ is given by softmax function

$$p(f(\hat{\mu})) = \frac{e^{\mu \hat{\mu} f(\hat{\mu})}}{\sum_{\hat{\mu}} e^{\mu \hat{\mu} f(\hat{\mu})}}, \quad (34)$$

following the probabilistic approximation in (18). After estimating $\hat{\mu}$ and $p(f(\hat{\mu}))$, we can easily recover $\hat{\mu}$ and $p(f(\mu))$ with $\mu = \log \mu^{-1}$. Then we define both L1 loss nad the KL-Divergence for scale registration

$$\mathcal{L}_{\mu,1} = \|\hat{\mu} - \hat{\mu}\|_1 \quad \text{(35)}$$
$$\mathcal{L}_{\mu,kld} = KLD(p(\mu), \delta_{\mu'}). \quad \text{(36)}$$

**C. Translation Registration**

With the estimated $\hat{\mu}$, we can further eliminate the effect of scale in frequency spectrum in (5) as

$$g_1(k) = g_2^2 \hat{\mu}(k - t). \quad (37)$$

Note that the form is similar to (28) for the scale registration, but with the translation in 3D, so we apply 3D Cartesian phase correlation for the correlation of the translation as

$$f(t) = i\mathcal{F} \left( \frac{\hat{F}}{|G|^2} \right) = i\mathcal{F} (e^{i2\pi \hat{\mu}^T t}), \quad (38)$$

and build the expectation for translation registration

$$\hat{t} = \sum_t t_i p(f(t_i)), \quad (39)$$

where $p(f(t_i))$ is derived by

$$p(f(t_i)) = e^{\xi \hat{f}(t_i)} \sum_j e^{\xi \hat{f}(t_i)} \hat{f}(t_i). \quad (40)$$

We can also define both L1 loss and the KL-Divergence for translation registration

$$\mathcal{L}_{t,1} = \|\hat{t} - \hat{t}\|_1 \quad \text{(41)}$$
$$\mathcal{L}_{t,kld} = KLD(p(\hat{t}), \delta_{\hat{t}}). \quad \text{(42)}$$

Finally, we finish the three-stage differentiable phase correlation solver, $Q_{DPc}$, for rotation, scale and translation registration which is able to back-propagate the error from losses to the input feature grids $g_1(k)$ and $g_2(k)$, as well as the three solver temperature parameters $\{\xi_r, \xi_s, \xi_t\}$.

**D. Extension to 2D**

In some applications, the pitch and roll, as well as the height can be aligned with onboard sensor e.g. air-ground BEV image registration, leaving the 7DoF pose to 4DoF. So we also extend the phase correlation to 4DoF pose registration with the input of 2D grids.

Given two 2D grids $g_{1(2d)}$ and $g_{2(2d)}$ linked by the relative translation $t$, rotation $r$ and scale $\mu$

$$g_{1(2d)}(k_{(2d)}) = g_{2(2d)} (r_k k_{(2d)} t), \quad (43)$$

where $k_{(2d)} \in [-B, B - 1]^2$ denotes the index of 2D grid.

**Decoupling of Translation, Scale and Rotation.** We also follow the decoupling idea in 7DoF by first eliminating the translation. Replace (8) and (9) with 2D Fourier transform, we decouple the translation from pose registration

$$\hat{G}_{1(2d)}(j_{(2d)}) = \hat{G}_{2(2d)}(\mu r_j j_{(2d)}), \quad (44)$$

where $\hat{G}_{(2d)}$ is the magnitude of Fourier spectrum of $g_{(2d)}$, and $j_{(2d)}$ is the corresponding frequency coordinate in 2D. We apply
Cartesian phase correlation for scale and rotation registration at the same time.

We further represent \( \hat{G}_{1,2d} \) and \( \hat{G}_{2,2d} \) in the log-polar coordinate

\[
\rho_1(\log|\mathbf{j}_{(2d)}|, \angle \mathbf{j}_{(2d)}) = \rho_2(\log|\mathbf{j}_{(2d)}| + \log \mu, \angle \mathbf{j}_{(2d)} + \alpha).
\]

Note that for 2D grids, the rotation is only determined by one angle, so it is encoded linearly in one axis.

Phase Correlation for Rotation, Scale and Translation. Note that (45) has the similar form to (37), where the rotation and scale are in the two axes, thus we similarly apply 2D Cartesian phase correlation like (38), (39) and (40) in 3D for estimation, and employ (42) as losses. Once the rotation and scale are estimated, they are applied to compensate \( g_{2,2d} \) to yield \( g_{r,\mu}^{*}\mu \), which leads to

\[
g_{1,2d}(k_{(2d)}) = \hat{r}_{(2d)}(k_{(2d)} - t).
\]

To estimate \( t \), we also apply 2D Cartesian phase correlation as that in rotation and scale stage, which finally completes the differentiable phase correlation solver for 4DoF case.

V. DEEP PHASE CORRELATION NETWORK

With the differentiable phase correlation \( Q_{DPC} \), we have error backpropagated to the feature grids both in 3D and 2D. The remaining problem is the design of \( Q_\theta \), and \( Q_\theta \) for building the feature grid. The function consists of two parts, a voxelization part, which converts data in different modals to a unified grid based representation, and a deep network part, which extracts the dense feature from the grid Fig. 3.

A. Data Voxelization

In the voxelization part, the process is simple. We first create a grid \( V(k) \) with \( k \in [-B, B - 1]^3 \), and fill all voxels with zeros. Then, for non-grid data e.g., point cloud, we label the voxel with 1 if there is a point occupying this voxel. Therefore, the resultant voxel is binary for pure point cloud, \( V(k) \in \{0, 1\} \). We can also assign voxels with the hand-crafted feature of points occupying e.g. density, resulting in a real valued grid, \( V(k) \in \mathbb{R} \).

For data with additional features e.g., colored point cloud, a multi-channel grid can be built \( V(k) \in \mathbb{R}^C \), where \( C \) is the size of the feature. In this article, we focus on the first two types of grid. For grid data like SDF, we directly regard it as \( V(k) \). For 2D cases, since the input images are already represented in the grid, we only need to reshape the images to the shape of \( 2B \times 2B \).

B. Deep Feature Extraction

In conventional phase correlation, low pass filter is applied to suppress high-frequency noise in the two inputs, which can be seen as a special form of feature extractor [5], [55]. For more general forms of inputs, the hand-crafted low pass filter is far from sufficient. Considering that there is no common feature to directly supervise the feature extractor, an end-to-end learning of feature extractor leveraged by the differentiable phase correlation is able to address the problem without assuming feature types like points, lines or edges.

Given a pair of inputs \( v_1 \) and \( v_2 \), after voxelization to data grids \( V(k) \), features grids are then extracted using UNet \( U(V(k)) \). We apply different feature extraction networks for rotation/scale registration, and translation registration

\[
g_{r,\mu,1}(k) = Q_{\theta,1,\nu}(v_1) = U_{\theta,1,\nu}(V_1(k)) \quad (47)
g_{r,\mu,1}(k) = Q_{\theta,1,\nu}(v_1) = U_{\theta,1,\nu}(V_1(k)) \quad (48)
g_{r,\mu,2}(k) = Q_{\theta,2,\nu}(v_2) = U_{\theta,2,\nu}(V_2(k)) \quad (49)
g_{r,\mu,2}(k) = Q_{\theta,2,\nu}(v_2) = U_{\theta,2,\nu}(V_2^{\mu}(k)), \quad (50)
\]

where \( v_2^{\mu} \) and \( V_2^{\mu}(k) \) are the rotation and scale compensated \( v_2 \) and the corresponding voxelized data grid respectively. Note that in the training stage, \( \{r, \mu\} \) is given by the ground truth \( \{r^*, \mu^*\} \) while in the inference stage, they are given by the estimated results \( \{\hat{r}, \hat{\mu}\} \). The whole process of the DPCN++ forward inference is summarized in Algorithm 1.

Architecture in detail. The UNet3D is adopted from [64], with each UNet constructed with 4 down-sampling encoder layers and 4 up-sampling decoder layers to extract features. We choose the LeakyReLU as activation [66]. In the training stage, the parameters of the UNets as well as the temperature \( \xi \) for softmax are tuned.

C. The Chain Rule of DPC++

We present the chain rule to derive the back-propagation evaluation to train DPC++. We first show the gradient of \( Q_{DPC} \), by which the gradient of loss with respect to the network parameter \( \{\theta_{1,r,\mu}, \theta_{2,\mu,\nu}, \theta_{1,t}, \theta_{2,t}\} \) is derived.

Gradient of DPC With Respect to Input. We set rotation as an example to derive the gradient of the loss with respect to the first input \( g_1(k) \) as

\[
\frac{\partial L_r}{\partial g_1(k)} = \frac{\partial}{\partial \hat{r}} \sum_{r_i} \hat{r} f(r_i) \frac{\partial f(r_i)}{\partial g_1(k)}, \quad (51)
\]

where the gradient consists of three terms. The first term is common. The second term, according to (18), we have

\[
\frac{\partial}{\partial f(r_i)} = \sum_j \frac{\xi_\nu(r_i - r_j) e^{\xi_\nu(f(r_i) + f(r_j))}}{(\sum_k e^{\xi_\nu(f(r_k))})^2}. \quad (52)
\]

For the third term, it is the gradient of the Fourier and spherical transforms. Note that these transforms are actually linear, so the derivation is not difficult. We leave the derivation in Appendix B.1, available online.

Gradient of DPC With Respect to Temperature. For the temperature parameter \( \xi_\nu \), the gradient is

\[
\frac{\partial L_r}{\partial \xi_\nu} = \frac{\partial}{\partial \xi_\nu} \frac{\partial f(r_i)}{\partial g_1(k)} \frac{\partial}{\partial \xi_\nu} f(r_i), \quad (53)
\]

where the first term is the same as that in (51), and the second term is given as

\[
\frac{\partial}{\partial \xi_\nu} = \sum_i \sum_j \frac{r_i e^{\xi_\nu(f(r_i) + f(r_j))}(r_i - r_j)}{(\sum_k e^{\xi_\nu(f(r_k))})^2}. \quad (54)
\]
Algorithm 1: Deep Phase Correlation Network (DPCN++).

**Input:** heterogeneous representation \( \{ v_1, v_2 \} \), bandwidth \( B \), ground truth of relative pose \( \{ t^*, r^*, \mu^* \} \).

**Output:** \( t, r, \mu \)

1: \( \triangleright \) Feature Grids in Rotation(Scaling) Stage with \( Q_{\theta, \mu} \)
2: \( \{ g_{\mu,1}, g_{\mu,2} \} \leftarrow \{ Q_{\theta_1}(v_1), Q_{\theta_2}(v_2) \} \)
3: \( \triangleright \) Decouple Translation
4: \( \{ G_1, G_2 \} \leftarrow \{ \text{FourierTransform}(\{ g_{\mu,1}, g_{\mu,2} \}) \} \)
5: \( \triangleright \) Decouple Scale
6: if dimension of \( g_{\mu,1} = 2 \) then
7: \( \{ s_1, s_2 \} \leftarrow \text{LogPolar}(\{ G_1, G_2 \}) \)
8: else if dimension of \( g_{\mu,1} = 3 \) then
9: \( \{ s_1, s_2 \} \leftarrow \text{Spherical}(\{ G_1, G_2 \}) \)
10: end if
11: \( \triangleright \) Estimate \( \hat{r}, \hat{\mu} \)
12: if dimension of \( g_{\mu,1} = 2 \) then
13: \( \{ \hat{r}, \hat{\mu} \} \leftarrow \text{CartesianPhaseCorrelation}(s_1, s_2) \)
14: else if dimension of \( g_{\mu,1} = 3 \) then
15: \( \hat{r} \leftarrow \text{SphericalPhaseCorrelation}(s_1, s_2) \)
16: \( \triangleright \) Decouple Rotation
17: \( g_{\mu,2}^\hat{r} \leftarrow \text{Transform}(g_{\mu,2}, \hat{r}) \)
18: \( \{ g_{1,2D}, g_{2,2D}^\hat{r} \} \leftarrow \{ \sum_j g_{\mu,1}, \sum_j g_{\mu,2}^\hat{r} \} \)
19: \( \hat{\mu} \leftarrow \text{CartesianPhaseCorrelation}(\text{LogPolar}(\{ \text{FourierTransform}(\{ g_{1,2D}, g_{2,2D}^\hat{r} \}) \})) \)
20: end if
21: \( \triangleright \) Estimate \( t \)
22: \( v_2^\hat{r}, \hat{\mu} \leftarrow \text{Transform}(v_2, \{ \hat{r}, \hat{\mu} \}) \)
23: \( \triangleright \) Feature Grids in Translation Stage with \( Q_{\theta, t} \)
25: \( \{ g_{t,1}, g_{t,2} \} \leftarrow \{ Q_{\theta_1}(v_1), Q_{\theta_2}(v_2^\hat{r}, \hat{\mu}) \} \)
26: \( t \leftarrow \text{CartesianPhaseCorrelation}(g_{t,1}, g_{t,2}) \)
27: \( \triangleright \) Train All Feature Extractors
28: \( \text{BackPropagate}(L(\{ t, \hat{r}, \hat{\mu}, \{ t^*, r^*, \mu^* \}))) \)
29: return \( t, \hat{r}, \hat{\mu} \)

**Gradient of DPCN++**. We finally show the full chain rule of the rotation supervised loss with respect to the feature extraction network parameters \( \theta_{1, r_s} \)

\[
\frac{\partial L_r}{\partial \theta_{1, r_s}} = \sum_k \frac{\partial L_r}{\partial g_1(k)} \frac{\partial g_1(k)}{\partial \theta_{1, r_s}},
\]

where the first term of the gradient is given in (51), the second term is given by the gradient of the UNet, which we can calculate by auto-gradient in public available deep learning package.

For the back-propagation pathway from loss to translation and scale, as well as the gradient with respect to the second input, a similar process can be followed. Note that the derivation above does not depend on any specific feature network, it is a general framework for implementing the back-propagation of versatile pose registration using DPCN++.

**D. Training Details**

Following the forward and backward process introduced above, we can implement the whole DPCN++ with gradients except for several sampling processes including: spherical transform, DH-Grid transform and log-polar transform. We apply sampling by linear interpolation, through which the re-sampled pixels are a continuous function of pre-sampled pixels, keeping valid gradients.

**Loss and Learning Setting.** We conclude the DPCN++ by presenting the total loss of DPCN++, which is a combination of rotation loss \( L_r \), translation loss \( L_t \), and scale loss \( L_\mu \)

\[
L = L_r + L_t + L_\mu
\]

where \( w \) are the weights for each loss. In the experiments, these weights are \{1, 3, 3, 1, 1, 3\}. This total loss guides the learning of both feature extractor parameters and the optimizer temperature parameters.

The code is implemented under PyTorch, with the hardware settings of CPU i9 12900 k, GPU RTX3090 × 2, and RAM 128 GB. The learning rate is set to 3e−4 for heterogeneous and 5e−5 for homogeneous with epoch decay. Due to the memory consumption of 3D UNets, the training batch size is set to 1, which also means there is no batch level operation.

**VI. EXPERIMENTAL RESULTS**

We first conduct experiments on 2D-2D BEV image registration, then, we show extensive comparisons of 3D-3D heterogeneous and heterogeneous pose registration. Finally, we conduct case studies to show the ability of registering symmetric object.

Additional quantitative and qualitative results, case studies and ablation studies can be found in the Appendix, available online.

**A. 2D-2D Pose Registration**

**Datasets.** We evaluate our approach on two 2D datasets. First, we construct a simulation dataset consisting of random 2D primitives, enabling a comparison of all methods using noise-free ground truth. As shown in Fig. 4 (left), We further apply Gaussian filtering to the target images to additionally simulate challenging heterogeneous image pairs. This dataset contains 2,000 randomly generated images pairs for training and 1,000 pairs for evaluation.

To further verify our performance in real-world BEV image pose registration, we collect a multi-modal Aero-Ground (AG) Dataset. This dataset allows for evaluating cooperative localization between ground mobile robots, micro aerial vehicles (MAVs) and satellite. Specifically, it contains several different image pairs as follows:

- “LiDAR Local Map” to “Drone’s Birds-eye Camera”;
- “LiDAR Local Map” to “Satellite Map”;
- “Stereo Local Map” to “Drone’s Birds-eye Camera”;
- “Stereo Local Map” to “Satellite Map”.

The 4DoF ground truth of the dataset is manually labeled. This dataset contains three scenes, where we split the training and testing regions without spatial overlapping, see Fig. 4 (right). Both the simulation dataset and the AG dataset have image
Fig. 4. Simulation dataset (left) containing “Homogeneous”, “Heterogeneous” and “Heterogeneous w/ Outlier” sets. Aero-Ground Dataset (right) containing “drone’s view”, “LiDAR intensity”, “stereo” and “satellite”. The experiments are carried out on location (a) and (b) separately in which the model is trained on images pairs generated inside red areas and validated on images pairs generated inside blue area. The generalization is carried out with estimating poses of images inside location (c) with models trained on (a) and (b).

resolutions of 256 × 256. For each input pair, we constrain both horizontal (x) and vertical (y) translation changes in the range of [−50, 50] pixels, together with rotation and scale change within [0, π] and [0.8, 1.2], respectively.

**Baselines.** For 2D image 4DoF registration, we first compare the performance with Phase Correlation (PC) [55], a learning-free version of our approach by eliminating the feature extraction network. We further compare to learning-based methods, Relative Pose Regression (RPR) [26], DAM [45], R2D2 [47], and Dense Search (DS) [2]. RPR [26] directly regresses the pose without integrating an explicit pose solver. DAM is similar to RPR but introduces feature correlations as an intermediate representation. As DAM does not estimate scale, we provide the ground truth scale to DAM in both training and testing. We retrain both RPR and DAM on our datasets. R2D2 is a feature point based pose registration method, which, estimates a similarity transform using feature correspondences together with an SVD solver. This is recently popular for visual localization [24], [73]. Finally, DS is similar to our work but uses a different solver which exhaustively rotates the source image among a set of candidate angles to find the best match wrt. the target. Due to the exhaustive search, DS is computationally expensive, thus we also relax it to 3DoF with ground truth scale as DAM, and train it in a small range, [0◦, 15◦].

**Evaluation Metrics.** We evaluate the percentage of estimation with an error lower than a given threshold, Accuracy in Units (Acc)

\[
Acc_{r1} = \frac{1}{n} \sum_{i=1}^{n} \#\{ |r_i^* - \hat{r}_i| \leq 1^\circ \} \times 100\% \tag{57}
\]

\[
Acc_{t0} = \frac{1}{n} \sum_{i=1}^{n} \#\{ |t_i^* - \hat{t}_i| \leq 10 \text{pixels} \} \times 100\% \tag{58}
\]

\[
Acc_{\mu0.2} = \frac{1}{n} \sum_{i=1}^{n} \#\{ |\mu_i^* - \hat{\mu}_i| \leq 0.2 \} \times 100\% , \tag{59}
\]

where \# is the count of the set \{·\}, and n is the total amount of image pairs. Note that for the AG dataset, each ground image is generated at a scale of 0.1 m per pixel. Thus, the threshold error of 10 pixels for translation indicates for 1 m in the real world.

1) Heterogeneous Registration: We now compare with the baselines in the challenging heterogeneous registration. The comparisons given homogeneous inputs are shown in Appendix C.3, available online.

**Simulation Dataset.** Experimental results on the simulation dataset are shown in Table I (sim). It can be seen that our

| Method     | Exp | Acc_{r1} | Acc_{t0} | Acc_{r1} | Acc_{t0} | Acc_{\mu0.2} | Runtime |
|------------|-----|----------|----------|----------|----------|--------------|---------|
| PC [55]    | sim | 69.1     | 45.7     | 72.3     | 97.6     |             | 18.3    |
| R2D2 [47]  | sim | 31.4     | 21.4     | 40.2     | 75.4     | 244.2        |         |
| DS [2]     | sim | 87.1     | 91.9     | 23.9     |          | 301.3        |         |
| DAM [45]   | sim | 99.6     | 99.2     | 80.8     |          | 114.2        |         |
| RPR [26]   | sim | 62.7     | 49.1     | 78.3     | 96.7     | 6.47         |         |
| DPCN++     | sim | 100      | 100      | 100      | 100      |             | 22.1    |
| R2D2 [47]  | l2a(a) | 32.7     | 41.1     | 37.6     | 71.2     | 244.1        |         |
| R2D2 [47]  | l2b(a) | 39.1     | 45.7     | 30.1     | 73.5     | 244.1        |         |
| DAM [45]   | l2a(a) | 47.7     | 41.9     | 32.9     | 75.9     | 244.2        |         |
| DAM [45]   | l2b(a) | 38.4     | 70.8     | 37.8     |          | 110.6        |         |
| RPR [26]   | l2a(a) | 39.4     | 66.8     | 22.5     |          | 117.3        |         |
| RPR [26]   | l2b(a) | 35.2     | 53.6     | 24.1     |          | 111.4        |         |
| RPR [26]   | l2a(a) | 51.5     | 63.9     | 33.9     |          | 114.2        |         |
| DPCN++     | l2a(a) | 96.9     | 98.0     | 99.2     | 95.5     | 24.75        |         |
| DPCN++     | l2b(a) | 98.2     | 94.0     | 99.2     | 94.2     | 26.37        |         |
| DPCN++     | l2a(b) | 90.9     | 97.8     | 97.4     | 93.7     | 23.61        |         |
| DPCN++     | l2b(b) | 91.3     | 92.6     | 99.3     | 93.5     | 24.72        |         |
| R2D2 [47]  | l2a(b) | 20.6     | 39.4     | 22.4     | 64.3     | 244.2        |         |
| R2D2 [47]  | l2b(b) | 21.8     | 28.9     | 25.1     | 66.8     | 244.0        |         |
| DAM [45]   | l2a(b) | 30.1     | 42.2     | 35.1     |          | 113.9        |         |
| DAM [45]   | l2b(b) | 30.9     | 49.6     | 27.3     |          | 116.5        |         |
| DPCN++     | l2a(b) | 96.2     | 89.2     | 99.7     | 99.7     | 24.51        |         |
| DPCN++     | l2b(b) | 91.6     | 90.6     | 99.4     | 95.0     | 25.63        |         |

Note: “l2a”, “l2d”, “l2a”, “l2d” are the abbreviation for “LiDAR local map” to “satellite map”, “LiDAR local map” to “drone’s birds-eye camera”, “stereo local map” to “satellite map”, “stereo local map” to “drone’s birds-eye Camera”, respectively. The runtime are measured in milliseconds. The best numbers are in bold and the secondaries are underlined.
approach reaches 100% accuracy rate on the simulation dataset, outperforming all baselines. Though PC and R2D2 perform well in the homogeneous case as shown in Appendix C.3, available online, they fail in heterogeneous cases, demonstrating the effectiveness of our combination of learned feature extractors and PC. Note that DPCN++ only requires ground truth pose for supervision, while R2D2 relies on supervision in the form of dense correspondences that are harder to obtain. The better performance of DAM against RPR validates the effectiveness of an explicit intermediate representation to bring inductive bias. But the direct regression of DAM causes the lower accuracy than that of DPCN++. For DS, it is mainly limited by the search range due to the expensive cost. It is worth noting that DPCN++ can also be considered as an exhaustive solver, but the solution space is significantly reduced by decoupling rotation, scale and translation. In terms of efficiency, our method is slower than and RPR, yet still allows for pose registration at 50 FPS. The performance is also evaluated by Mean Square Error (MSE) and more thresholds of *Acc* in the Appendix C.3, available online.

**AG Dataset.** In this experiment, we select R2D2 and DAM as representative baselines for solver-based and solver-free methods, respectively. Table I shows evaluation results on scene (a) and (b) of the AG dataset. As can be seen, when estimating 4DoF poses across real-world sensors, our approach achieves accuracy rate above 89.2% at a threshold of 1 m, typically sufficient for many applications. Though DAM is relaxed to 3DoF registration given ground truth scale, our approach still outperforms it by a large margin. The result of R2D2 indicates that, when the style of inputs changes drastically, matching local feature tends to be difficult. The Mean Square Error (MSE) and more thresholds for *Acc* are evaluated in the Appendix C.3 C.4, available online. The additional examples of the experiments comparing with the classical PC is shown in the Appendix C.1 and C.2, available online.

To evaluate the generalization of DPCN++, we conduct experiments on scene (c) with DPCN++ trained on scene (a) and (b) and DAM model trained on partial data of scene (c). The results shown in the Appendix C.5, available online, indicate that our approach is capable of estimating poses with a similar accuracy regardless of scene changes, and still outperforms DAM which is specifically trained on partial data of scene (c). Therefore, the generalization of employing an explicit interpretable solver is verified.

We also conduct a case study for the 2D registration by visualizing and analyzing the intermediate results of our pipeline during the 2D heterogeneous registration on the AG dataset in Appendix F, available online.

### B. 3D-3D Pose Registration

**Datasets.** For homogeneous registration, we evaluate DPCN++ and the baselines on two different datasets, an object-level dataset MVP Benchmark [42], and a scene-level dataset 3DMatch [72]. The MVP Benchmark contains 6,400 training pairs and 1,200 testing pairs. The 3DMatch dataset consists of 54 scenes for training and 8 scenes for testing. For both datasets, we apply random 7DoF relative transformation between inputs pairs. Based on the completeness of measurement, three scenarios are tested including complete-to-complete, partial-to-complete, and partial-to-partial.

For heterogeneous registration, we first construct a heterogeneous dataset based on Linemod [22] as a proof of concept. By extracting point cloud and building SDF from the mesh provided by Linemod, we evaluate versatile pose registration across these three modalities. We follow the settings of homogeneous registration in relative pose generation and scenarios. More detailed settings are elaborated in the Appendix A.1, available online.

We further evaluate heterogeneous registration using medical imaging dataset, including CT-MRI pairs of RIRE [62], and CT-Ultrasound pairs on “USCT” [38]. RIRE provides registered CT and MRI images of the human and “USCT” provides CT and Ultrasound images of humans spines, canine spines, and lamb spines. For all datasets, we randomly initialize relative transformations for each inputs pairs following the rules in Section VI-B1. More details regarding the settings are provided in Appendix A.2, available online.

**Baselines.** We compare homogeneous registration performance of DPCN++ to three lines of methods: 1) Learning-free methods: PC, the learning-free version of our approach by eliminating the feature extraction network but for 7DoF pose registration. TEASER [69], a state-of-the-art 7DoF robust point cloud registration method with superior performance in handling large outlier rates. We also compare TEASER with certification (CERT) [69], an improved version of TEASER at the cost of increasing the inference time. For correspondences of TEASER, we utilize the FPFH introduced in [69] following the original implementation. 2) Correspondence-based learning methods: DCP [59] and DeepGMR [71], both methods overcome the local convergence of ICP by learning the feature correspondences. 3) Correspondence-free learning methods: PointNetLK [1], a method employs a gradient based solver to minimize the similarity between learned features. Note that DCP, DeepGMR and PointNetLK are designed for 6DoF pose registration, thus we provide them the ground truth scale, and only evaluate rotation and translation. For all methods, we do not apply post-processing like ICP.

**Evaluation Metrics.** We follow the metrics for evaluating 3D poses in [59], [69], [71], with the error for translation $E_t$, rotation $E_r$ and scale $E_\mu$ defined as

$$E_t = \| \hat{t} - t^* \|$$

$$E_r = \arccos \frac{tr(\hat{r}^T r^*) - 1}{2}$$

$$E_\mu = \| \hat{\mu} - \mu^* \|,$$

where $\hat{}$ and $^*$ are the estimated result and the ground truth of $\cdot$, respectively.

1) **Homogeneous Registration: Object-Level Registration.** The MVP Benchmark contains complete point clouds and the corresponding partial version, we thus consider three scenarios for object-level point cloud registration: i) complete-to-complete, ii) partial-to-complete, and iii) partial-to-partial. The
target relative pose is generated by randomly sampling from \(|t| \leq 1\text{m}, r \in SO(3), \text{ and } \mu \in [0.8, 1.2]|. To evaluate the robustness against noisy input, we also add random outliers to one of the point clouds. The outlier rate ranges from 0% to 50% with an increase of 10% at a time.

Fig. 5 shows the translation, rotation and scale error wrt. varying outlier rate for three scenarios. DPCN++ is evaluated at three different bandwidths \(\{64, 100, 128\}\), named DPCN++64, DPCN++100 and DPCN++128. Fig. 5(a) shows that when registering two complete point clouds, DPCN++ is competitive with the state-of-the-art method TEASER given various outlier rates. The learning-free PC at bandwidth 128 (PC128) also shows similar performance throughout different rates of outliers. These results demonstrate the advantage of the global solver in the initialization-free setting when the data is completed.

However, when the source or the target is partially observed as shown in Fig. 5(b) and (c), the result is different. Both TEASER and PC have degenerated performance, indicating their sensitivity to completeness. The missing part of the source makes the corresponding part of the target become a significant outlier. In contrast, DPCN++ retains a similar performance compared to the complete-to-complete setting and achieves the best performance, validating the effectiveness of the deep feature extraction network. By training end-to-end guided by the differentiable solver, the network is capable of filtering the outliers and enhancing crucial features for registration. For other methods with deep feature extractor including DCP, DeepGMR and PNKL, their performances also degenerate less than TEASER and PC when the point cloud is not complete. However, these methods struggle to achieve low error based on their own solvers, reflecting the advantage of the initialization-agnostic and correspondence-free phase correlation solver used in DCPN++. The qualitative results in Fig. 6 lead to consistent conclusions.

Scene-Level Registration. We follow the same way as in object-level registration to generate the target relative pose between a pair of point clouds, which is also the same settings in [69]. We do not add additional random outliers as the point clouds are collected in the real world and contains sensor noises. We employ 500 pairs in the Kitchen scene for training and leave other scenes for testing. We follow the same metrics as in [69] that the registration is success when 1) rotation error smaller than \(10^\circ\), and 2) translation error less than 30 cm.
Fig. 6. Qualitative Comparison on Homogeneous Object-level Registration on the MVP dataset. We show point cloud-wise registration results with 50% outliers. Note: Blue and green point the source and target and the red are the outliers.

Table II
EVALUATION RESULTS ON 3DMatch Dataset

| Baselines  | Home 1 | Home 2 | Hotel 1 | Hotel 2 | Hotel 3 | Study | MIT Lab | Time (ms) |
|------------|--------|--------|---------|---------|---------|-------|---------|-----------|
| TEASER     | 92.9   | 86.5   | 97.6    | 89.4    | 94.4    | 91.1  | 83.1    | 39        |
| TEASER (CERT) | 94.1   | 88.7   | 98.2    | 91.9    | 94.4    | 94.3  | 88.6    | > 1000    |
| DCP       | 85.1   | 84.9   | 86.2    | 89.3    | 80.5    | 79.6  | 85.4    | 73        |
| DeepGMR   | 83.5   | 88.1   | 82.2    | 85.6    | 83.1    | 81.3  | 81.7    | 63        |
| PNLK [61] | 81.2   | 82.9   | 81.5    | 84.3    | 79.1    | 77.4  | 80.5    | > 200     |
| DPCN++64  | 93.5   | 91.2   | 94.7    | 93.1    | 95.7    | 92.3  | 92.1    | 17        |
| DPCN++100 | 94.4   | 92.2   | 95.6    | 95.3    | 96.2    | 94.1  | 92.6    | 23        |
| DPCN++128 | 96.3   | 94.1   | 96.1    | 95.9    | 97.3    | 94.7  | 93.5    | 37        |

The best numbers are in bold and the secondaries are underlined.

Table II shows the success rate of all methods, upon which the results of TEASER and TEASER (CERT) are taken from [69]. It can be seen that our method performs the best on most scenes. TEASER is also highly competitive on this dataset as the overlap between partial observations is larger compared to the object-level registration, thus TEASER is less affected by the completeness of the point cloud. TEASER (CERT) raises a higher accuracy with certification, but at the cost of much more computation time. Compared with other learning-based methods, our method still achieves higher accuracy owing to the solver. Fig. 7 shows the cases for qualitative evaluation. Considering that only one threshold may not fully reflect the registration error, we additionally report results evaluated with varying thresholds in Appendix D.2, available online.

2) Heterogeneous Registration: Point Cloud-Mesh-SDF. To verify the heterogeneous $7$DoF registration, we carry out experiments across 3D representations including i) point cloud, ii) mesh and iii) Signed Distance Field (SDF). The relative pose between the two measurements is generated following the settings in homogeneous registration. We take the best performing baseline among all sub-second methods, TEASER, for comparison. As TEASER is not directly applicable for heterogeneous registration, we convert mesh and SDF to point cloud to evaluate it in the homogeneous setting. We denote this baseline as TEASER*. Moreover, we also consider a baseline DPCN++ which also takes a pair of point clouds as input. Note that both TEASER* and DPCN++ assume the conversion between different types of representations are known, while the conversion is non-trivial in many applications, e.g., medical imaging.

Figs. 8, 10, and 12 show the translation, rotation and scale error cloud to mesh, point cloud to SDF, and mesh to SDF registration respectively. Note that in all scenarios, DPCN++ achieves similar performance to that of DPCN++* without applying the conversion in prior. Instead, the common features from heterogeneous input can be built by the extraction network. This result verifies the usefulness of heterogeneous registration in practice. In addition, as the result of homogeneous registration, DPCN++ in complete-to-complete scenario has smaller errors than those of the partial-to-complete scenario by a small margin, while TEASER* has a large degeneration. The qualitative cases of the versatile registration are shown in Figs. 9, 11 and 13.

CT-MRI-3D Ultrasound. We further evaluate the heterogeneous 3D registration with real medical images captured by CT, MRI and 3D ultrasound. In this task, the different modality captures the different part of the body, e.g., CT images usually reflect the structure of the bone while MRI, muscle and tissue. As the conversion between different modalities is unknown, we only evaluate DPCN++ in this problem. We focus on the rigid part, i.e. bone, registration for medical imaging. Figs. 15 and 16 show the quantitative and qualitative results. Note that even with large noise in 3D ultrasound for the human spine, a successful registration between 3D ultrasound and CT is achieved. These results again show that by training DPCN++ end-to-end, the feature extractors are able to learn common features from data to enable challenging heterogeneous registration, demonstrating the value of DPCN++ in real applications.
Fig. 7. Qualitative Comparison on Homogeneous Scene-level Registration on 3DMatch. We show one successful case for all baselines. The flaws of several baselines are captured in red boxes.

Fig. 8. Quantitative Comparison on Heterogeneous Registration of point cloud to mesh. The result of “Partial to Partial” is demonstrated in Appendix D.3, available online.

C. Case Study on Symmetric Object

We conduct 4 more experiments to demonstrate the performance of DPCN++ when handling the symmetric objects in the Linemod dataset: i) a complete-to-complete point cloud-to-mesh registration of the central symmetric object Bowl, ii) a partial-to-complete point cloud-to-mesh registration of the central symmetric object Bowl, iii) a complete-to-complete point cloud-to-mesh registration of the axial symmetric object Eggbox, and iv) a RGB-D registration of the central symmetric object Bowl.
An intriguing observation is that, despite all possible rotations selected from the “red curve” in the correlation cube leading to the rotation of the bowl about the axis of symmetry, DPCN++ is able to locate the most accurate match in proximity to the ground truth. This can be attributed to the imperfections in the symmetry of the model, which are apparent upon close examination of the mesh model and the presence of flaws and noise brought by the 3D scanner.

**Bowl, Partial Point Cloud to Complete/Partial Mesh.** We also conduct the experiment on partially overlapping objects. For this case, DPCN++ presents similar results to case i), showing that it can handle the symmetric object registration well even with fewer overlaps. The qualitative demonstration is shown in Fig. 14 (row 2 and 3).

**Eggbox, Complete Point Cloud to Complete Mesh.** We conduct experiments on the axial symmetric object Eggbox. The demonstration in Fig. 14 (row 3) indicates that the DPCN++
serves correlation result with two peaks, the transformation from which differs from each other by 180°. The axial symmetry of the Eggbox results in the multi-solution for rotation registration. However, different from the bowl, the solution number is reduced to 2 which corresponds to the correlation. We also demonstrate the rotated result with respect to “ZYZ” angles indicated by the secondary peak, it is less perfect due to the same reason: imperfect symmetry.

**Bowl, RGB-D Registration.** We first paint the point cloud red, and then randomly paint an area green. The voxel grids for such point clouds are extended in channels (for RGB) resulting in the retraining of the network with more input channels for the feature extractor. Demonstration in Fig. 14 (row 4) shows that when the symmetric model Bowl is texturized, the curve-shape peak in the correlation cube as in Fig. 14 (row 1) disappears, and leaves a unique and sharp peak. This is accounted for by the colored bowls are no longer symmetric when we consider the textures (colors).

By providing these cases, we hope to better illustrate the mechanism behind decoupling and phase correlation. This can also demonstrate the advantages of this solver, which can handle multiple solutions, avoid initial value interference, and promote learning more effectively. More detailed demonstration can be found in Appendix D.4, available online.

**D. Limitation**

Though DPCN++ is shown to be versatile, robust, and effective in experiments, the main bottleneck of the method is the intensive memory consumption. In this article, we choose the UNet as the deep feature extractor, which limits the bandwidth of the DPCN++ up to 128 while taking two RTX3090 for training. We believe that with more lightweight feature extractors, and with the advancement of the hardware, the limitation can be relieved in the future. We also report a representative failure case in Appendix D.1, available online, which reveals the fact DPCN++ might be degraded to the inputs with relatively small overlaps. The authors believe that with attention involves, the degeneration brought by small overlaps can be further reduced. Also, there are discretization approximations in the decoupling process that will accumulate errors in the final result. However, in practical applications, the larger source of error is the
noise in the data and non-overlapping features which is much greater compared to the approximation error. Therefore, we aim to reduce this type of error by introducing a network to our approach.

VII. CONCLUSION

We present a general end-to-end feature grid registration framework, DPCN++, for versatile pose registration problems. In DPCN++, the core component is a differentiable, initialization-free, correspondence-free phase correlation solver that enables back-propagating the pose error to the learnable feature extractors, which addresses the outlier sensitivity of correspondence-based methods, and the local convergence of previous correspondence-free methods. With the DPC solver, it is unnecessary to define the specific form of feature, which is learned fully driven by data. We evaluate DPCN++ on both 2D and 3D registration tasks including BEV images, object and scene level 3D measurements, as well as medical images. All results show that our method is able to register both homo and hetero sensor measurements with competitive performance, verifying the strong inductive bias brought by the DPC solver. In the future, we would like to investigate the 2D-3D pose registration to address the needs for cross-dimension registration as well as investigate the theoretic global convergence of a neural network-based pose solver.

REFERENCES

[1] Y. Aoki, H. Goforth, R. A. Srivatsan, and S. Lucey, “PointNetLK: Robust & efficient point cloud registration using PointNet,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7156–7165.
[2] D. Barnes, R. Weston, and J. Posner, “Masking by moving: Learning distraction-free radar odometry from pose information,” in Proc. 3rd Conf. Robot Learn., 2019, pp. 303–316.
[3] L. Bernreiter, L. Ott, J. Nieto, R. Siegwart, and C. Cadena, “PHASER: A robust and correspondence-free global pointcloud registration,” IEEE Robot. Automat. Lett., vol. 6, no. 2, pp. 855–862, Apr. 2021.
[4] P. J. Besl and N. D. McKay, “Method for registration of 3-D shapes,” in Proc. Sensor Fusion IV: Control Paradigms Data Structures, 1992, pp. 586–606.
[5] H. Bülow and A. Birk, “Scale-free registrations in 3D: 7 degrees of freedom with Fourier mellin soft transforms,” Int. J. Comput. Vis., vol. 126, no. 7, pp. 731–750, 2018.
[6] 3D Slicer community, 3D-Slicer (Version 4.11.20210226) [Software], 2021.
[7] X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, and D. Shen, “Deep learning based inter-modality image registration supervised by intra-modality similarity,” in Proc. Int. Workshop Mach. Learn. Med. Imag., 2018, pp. 55–63.
[8] Y.-W. Chen, C.-L. Lin, and A. Mimori, “Multimodal medical image registration using particle swarm optimization,” in Proc. 8th Int. Conf. Intell. Syst. Des. Appl., 2008, pp. 127–131.
[9] Y.-W. Chen, A. Mimori, and C.-L. Lin, “Hybrid particle swarm optimization for 3-D image registration,” in Proc. IEEE 16th Int. Conf. Image Process., 2009, pp. 1753–1756.
[10] Z. Chen, X. Xu, Y. Wang, and R. Xiong, “Deep phase correlation for end-to-end heterogeneous sensor measurements matching,” in Proc. Conf. Robot Learn., 2020, pp. 2359–2375.
[11] D. Chetverikov, D. Stepanov, and P. Krsek, “Robust Euclidean alignment of 3D point sets: The trimmed iterative closest point algorithm,” Image Vis. Comput., vol. 23, no. 3, pp. 299–309, 2005.
[12] C. Choy, W. Dong, and V. Koltun, “Deep global registration,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2511–2520.
[13] C. Choy, J. Park, and V. Koltun, “Fully convolutional geometric features,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 8958–8966.
[14] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, “Spherical CNNs,” 2018, arXiv:1801.10130.
[15] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superpoint: Self-supervised interest point detection and description,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2018, pp. 337–33712.
[16] M. Dusmanu et al., “D2-Net: A trainable CNN for joint detection and description of local features,” 2019, arXiv:1905.03561.
[17] O. Enqvist, K. Josephson, and F. Kahl, “Optimal correspondences from pairwise constraints,” in Proc. IEEE 12th Int. Conf. Comput. Vis., 2009, pp. 1295–1302.
[18] S. Granger and X. Pennec, “Multi-scale EM-ICP: A fast and robust approach for surface registration,” in Proc. Eur. Conf. Comput. Vis., 2002, pp. 418–432.
[19] J. Groff, A. Olep, and B. Leibe, “AlignNet-3D: Fast point cloud registration of partially observed objects,” in Proc. Int. Conf. 3D Vis., 2019, pp. 623–632.

[20] G. Haskins et al., “Learning deep similarity metric for 3D MR–TRUS image registration,” Int. J. Comput. Assist. Radiol. Surg., vol. 14, no. 3, pp. 417–425, 2019.

[21] D. M. Healy, D. N. Rockmore, P. J. Kostelec, and S. Moore, “FFTs for the 2-sphere-improvements and variations,” J. Fourier Anal. Appl., vol. 9, no. 4, pp. 341–385, 2003.

[22] S. Hinterstoisser et al., “Model based training, detection and pose estimation of textured and texture-less 3D objects in heavily cluttered scenes,” in Proc. Asian Conf. Comput. Vision, 2012, pp. 548–562.

[23] X. Huang, G. Mei, and J. Zhang, “Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 11363–11371.

[24] Y. Jiao, Y. Wang, X. Ding, B. Fu, S. Huang, and R. Xiong, “2-entropy random sample consensus for robust visual localization: Framework, methods, and verifications,” IEEE Trans. Ind. Electron., vol. 68, no. 5, pp. 4519–4528, May 2021.

[25] R. Kaslin et al., “Collaborative localization of aerial and ground robots through elevation maps,” in Proc. IEEE Int. Symp. Saf. Secur. Rescue Robot., 2016, pp. 284–290.

[26] A. Kendall and R. Cipolla, “Geometric loss functions for camera pose estimation learning,” in Proc. 3rd Int. Conf. 3-D Imag. Imag. Process., 2001, pp. 145–152.

[27] E. Sattat and T. Theoharis, “An application independent review of multimodal 3D registration methods,” Comput. Graph., vol. 91, pp. 153–178, 2020.

[28] V. Sarode et al., “PCRNet: Point cloud registration network using PointNet encoding,” 2019, arXiv:1908.07906.

[29] L. Schwab, M. Schmitt, and R. Wanka, “Multimodal medical image registration using particle swarm optimization with influence of the data’s initial orientation,” in Proc. IEEE Conf. Comput. Intell. Biol. Comput., 2015, pp. 1–8.

[30] A. Sedghi et al., “Semi-supervised deep metrics for image registration,” 2018, arXiv:1804.01565.

[31] B. S. Reddy and B. N. Chatterji, “An FFT-based technique for translation, rotation, and scale-invariant image registration,” IEEE Trans. Image Process., vol. 5, no. 8, pp. 1266–1271, Aug. 1996.

[32] Y. T. Yang, D. De Martini, D. Barnes, and P. Newman, “RSL-Net: Localising in satellite images from a radar on the ground,” IEEE Robot. Autom. Lett., vol. 5, no. 2, pp. 1087–1094, Apr. 2020.

[33] C. Wang, X. Jing, and C. Zhao, “Local upsampling Fourier transform for accurate 2D/3D image registration,” Comput. Electr. Eng., vol. 38, no. 5, pp. 1346–1357, 2012.

[34] C. Wang et al., “DenseFusion: 6D object pose estimation by iterative dense fusion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3338–3347.

[35] Y. Wang and J. M. Solomon, “Deep closest point: Learning representations for point cloud registration,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3522–3531.

[36] Y. Wang, Y. Sun, Z. Liu, S. E. Arzuma, M. M. Bronstein, and J. M. Solomon, “Dynamic graph CNN for learning on point clouds,” ACM Trans. Graph., vol. 38, no. 5, pp. 1–12, 2019.

[37] W. M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, “Multimodal volume registration by maximization of mutual information,” Med. Image Anal., vol. 1, no. 1, pp. 35–51, 1996.

[38] J. West et al., “Comparison and evaluation of retrospective intermodality brain image registration techniques,” J. Comput. Assist. Tomogr., vol. 21, no. 4, pp. 554–568, 1997.

[39] R. Weston, M. Gadd, D. De Martini, P. Newman, and I. Posner, “FastMbyM: Leveraging translational invariance of the Fourier transform for efficient and accurate radial orometry,” 2022, arXiv:2203.04059.

[40] A. Wolny et al., “Accurate and versatile 3D segmentation of plant tissues at cellular resolution,” Elife, vol. 9, Jul. 2020, Art. no. e57613.

[41] R. Wright et al., “LSTM spatial co-transformer networks for registration of 3D fetal US and MR brain images,” in Proc. Int. Workshop Data Driven Treat. Response Assessment Preterm Perinatal Paediatric Image Anal., 2018, pp. 35–42.

[42] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” 2015, arXiv:1505.00853.

[43] X. Xu, Z. Chen, J. Guo, Y. Wang, Y. Wang, and R. Xiong, “Collaborative localization of aerial and ground mobile robots through orthomosaic map,” in Proc. IEEE Int. Conf. Real-Time Comput. Robot., 2020, pp. 122–127.

[44] H. Yang and L. Carluccio, “Certifiable outlier-robust geometric perception: Exact semidefinite relaxations and scalable global optimization,” 2021, arXiv:2109.03349.
[69] H. Yang, J. Shi, and L. Carlone, “TEASER: Fast and certifiable point cloud registration,” *IEEE Trans. Robot.*, vol. 37, no. 2, pp. 314–333, Apr. 2021.

[70] J. Yang, H. Li, D. Campbell, and Y. Jia, “Go-ICP: A globally optimal solution to 3D ICP point-set registration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2241–2254, Nov. 2016.

[71] W. Yuan, B. Eckart, K. Kim, V. Jampani, D. Fox, and J. Kautz, “DeepGMR: Learning latent Gaussian mixture models for registration,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 733–750.

[72] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, “3DMatch: Learning local geometric descriptors from RGB-D reconstructions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 199–208.

[73] Z. Zhang, T. Sattler, and D. Scaramuzza, “Reference pose generation for long-term visual localization via learned features and view synthesis,” *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 821–844, 2021.

[74] H. Zhong, N. Wen, J. J. Gordon, M. A. Elshaikh, B. Movas, and I. J. Chetty, “An adaptive MR-CT registration method for MRI-guided prostate cancer radiotherapy,” *Phys. Med. Biol.*, vol. 60, no. 7, pp. 2837–2851, 2015.

[75] Q.-Y. Zhou, J. Park, and V. Koltun, “Fast global registration,” in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 766–782.

[76] Q.-Y. Zhou, J. Park, and V. Koltun, “Open3D: A modern library for 3D data processing,” 2018, arXiv:1801.09847.

[77] M. Zhu, M. Ghaffari, and H. Peng, “Correspondence-free point cloud registration with SO(3)-equivariant implicit shape representations,” in *Proc. Conf. Robot Learn.*, PMLR, 2022, pp. 1412–1422.

Zexi Chen received the BEng degree in mechatronical engineering from Zhejiang University, Hangzhou, China, in 2019. He is currently working toward the PhD degree with the State Key Laboratory of Industrial Control and Technology, and Institute of Cyber-Systems and Control, Zhejiang University. His research interests include representation registration, 3D vision and scene understanding.

Yiyi Liao received the PhD degree from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2018. From 2018 to 2021, she was a postdoctoral researcher with the Autonomous Vision Group, University of Tübingen and Max Planck Institute for Intelligent Systems, Germany. She is currently an assistant professor with Zhejiang University. His research interests include 3D vision and scene understanding.

Haozhe Du received the BEng degree in control science and engineering and a minor in mechatronical engineering from Zhejiang University, Hangzhou, China, in 2022, where he is currently working toward the MS degree. His latest research interests include representation registration and 3D vision.

Haodong Zhang received the BS degree from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2020, where he is currently working toward the MS degree. His latest research interests include imitation learning and transfer learning.

Xuecheng Xu received the BS degree from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2019, where he is currently working toward the PhD degree. His latest research interests include SLAM and multi-robot systems.

Haojian Lu (Member, IEEE) received the BEng degree in mechatronical engineering from BIT, and the PhD degree in robotics from CityU Hong Kong. He is currently a professor with the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China. His research interests include micro/nanorobotics, bioinspired robotics, medical robotics, and micro aerial vehicle and soft robotics.

Rong Xiong received the PhD degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China, in 2009. She is currently a professor with the Department of Control Science and Engineering, Zhejiang University. Her latest research interests include motion planning and SLAM.

Yue Wang received the PhD degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, P.R. China, in 2016. He is currently an associate professor with the Department of Control Science and Engineering, Zhejiang University. His latest research interests include mobile robotics and robot perception.