End-to-end globally consistent registration of multiple point clouds

Zan Gojcic*§ Caifa Zhou*§ Jan D. Wegner§ Leonidas J. Guibas† Tolga Birdal†

§ETH Zurich †Stanford University

Abstract

We present a novel, end-to-end learnable, multiview 3D point cloud registration algorithm. Registration of multiple scans typically follows a two-stage pipeline: the initial pairwise alignment and the globally consistent refinement. The former is often ambiguous due to the low overlap of neighboring point clouds, symmetries and repetitive scene parts. Therefore, the latter global refinement aims at establishing the cyclic consistency across multiple scans and helps in resolving the ambiguous cases. In this paper we propose the first end-to-end algorithm for joint learning of both parts of this two-stage problem. Experimental evaluation on benchmark datasets shows that our approach outperforms state-of-the-art by a significant margin, while being end-to-end trainable and computationally less costly. A more detailed description of the method, further analysis, and ablation studies are provided in the original CVPR 2020 paper [11].

1. Introduction

The capability of aligning and fusing multiple scans is essential for the tasks of structure from motion and 3D reconstruction. As such, it has several use cases in augmented reality and robotics. While for pairwise registration well accepted methods do exist [23, 9, 12, 7, 10], registering multiple scans globally remains a challenge, because i) the global registration methods are typically dependent on a good pairwise initialization, and ii) it is unclear how to best make use of the quadratic pairwise relationships. Most methods for global alignment of multiview data aim at synchronizing the pairwise transformation parameters with good initializations [14, 19, 3, 2, 6], or incorporate pairwise keypoint correspondences in a joint optimization [23, 20, 4]. In general, these methods work well on data with low synthetic noise, but struggle on real scenes with high levels of clutter and occlusion [5]. A general drawback of this hierarchical procedure is that the global noise distribution over all nodes in the pose graph ends up being far from random, i.e. significant biases persist due to the highly correlated initial pairwise alignments.

*First two authors contributed equally to this work.

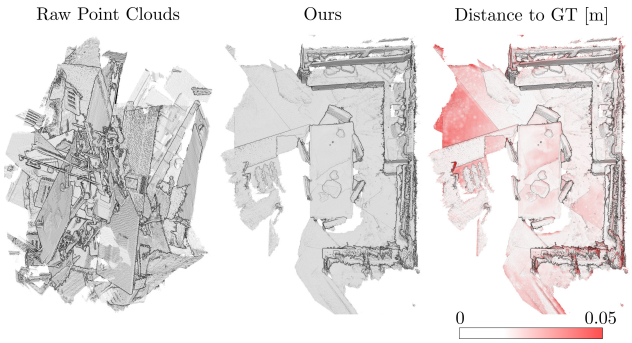

Figure 1. Result of our end-to-end reconstruction on the 60 scans of *Kitchen* scene from *3DMatch benchmark* [21].

Contributions The main contributions of our work are:

1. We reformulate the traditional two-stage approach as an end-to-end differentiable, declarative neural network that solves two differentiable optimization problems during the forward pass: (i) the Procrustes problem for the estimation of the pairwise transformations and (ii) the spectral relaxation of the transformation synchronization.
2. We propose a confidence estimation block that uses a novel *overlap pooling* layer to predict the confidence in the estimated pairwise transformation parameters.
3. We formulate the registration of multiple 3D point clouds as an IRLS problem and iteratively refine both the pairwise and absolute transformation estimates.
4. We integrate all these into, to the best of our knowledge, the first end-to-end data driven multiview point cloud registration algorithm.

Resulting from the aforementioned contributions, the proposed multiview registration algorithm (i) is very efficient to compute, (ii) achieves more accurate scan alignments because the residuals are being fed back to the pairwise network in an iterative manner, (iii) outperforms current state-of-the art on pairwise point cloud registration as well as transformation synchronization.

2. End-to-End Multiview 3D Registration

Consider a set of potentially overlapping point clouds $\mathcal{S} = \{\mathbf{S}_i \in \mathbb{R}^{N\times 3}, 1 \leq i \leq N_{\mathcal{S}}\}$ capturing a 3D scene from different viewpoints (i.e. poses). The task of *multiview registration* is to recover the rigid, absolute poses

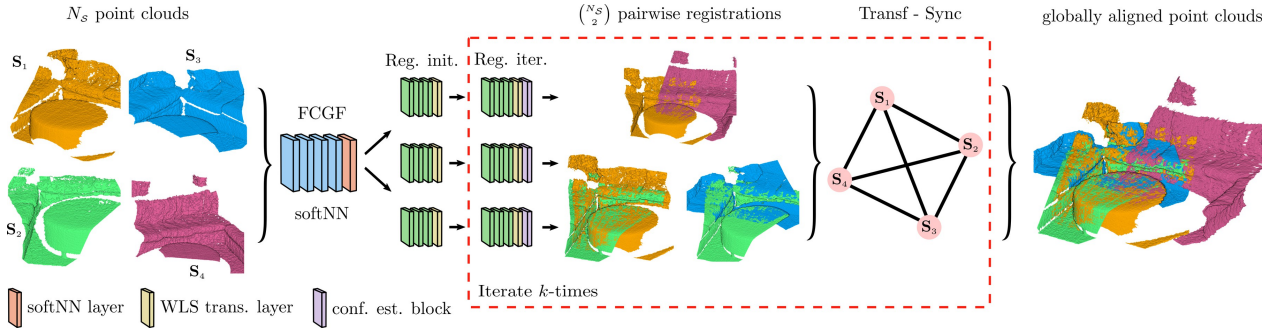

Figure 2. Proposed pipeline for end-to-end multiview 3D point cloud registration.

$\{\mathbf{M}_i^* \in SE(3)\}_i$ given the scan collection, where

$$SE(3) = \left\{ \mathbf{M} \in \mathbb{R}^{4\times4} \colon \mathbf{M} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \right\}. \quad (1)$$

$\mathbf{R}_i \in SO(3)$ and $\mathbf{t_i} \in \mathbb{R}^3$. $\mathcal{S}$ can be augmented by connectivity information resulting in a finite graph $\mathcal{G} = (\mathcal{S}, \mathcal{E})$, where each vertex represents a single point set and the edges $(i, j) \in \mathcal{E}$ encode the information about the relative rotation $\mathbf{R}_{ij}$ and translation $\mathbf{t}_{ij}$ between the vertices.

Our end-to-end multiview 3D registration approach consists of three modules: i) learned correspondence module, ii) pairwise registration module, and iii) iterative transformation synchronization module. For each of the input point clouds $\mathbf{S}_i$ we extract FCGF [7] features that are fed into the *softNN* layer to compute the stochastic correspondences for $\binom{N_s}{2}$ pairs. These correspondences are used as input to the two registration blocks (*Reg. init.* and *Reg. iter.*), whose outputs are used to build the graph. After each iteration of *Transf-Sync* layer that solves the spectral relaxation of the transformation synchronization problem [2], the estimated transformation parameters are used to pre-align the correspondences that are concatenated with the weights from the previous iteration and the residuals and feed anew to *Reg. iter.* block.

SoftNN layer In order to establish the pointwise correspondences a nearest neighbor (NN) search in the FCGF feature space has to be carried out. However, the selection rule of such hard assignments is not differentiable. We therefore relax the NN-assignment in a probabilistic manner by computing a probability (weight) vector $\mathbf{s}$ of the categorical distribution [17], which is used to hallucinate the corresponding points as a weighted average of the coordinates. These soft correspondences enable the flow of the gradients to the initial layers during training.

Differentiable pairwise registration The pairwise registration block is formulated as an IRLS problem for which the per-correspondence weights are obtained by combining the 3D outlier filtering network [13] with the order-aware blocks proposed in [22]. Specifically, our initial pairwise registration block (*Reg. init.*) takes the coordinates of the putative correspondences as input and outputs weights, indicating if the established putative correspondence is an outlier or an inlier. The inferred weights together with the coordinates of the putative correspondences are then fed into the weighted Procrustes problem

$$\hat{\mathbf{R}}_{ij}, \hat{\mathbf{t}}_{ij} = \underset{\mathbf{R}_{ij}, \mathbf{t}_{ij}}{\arg\min} \sum_{l=1}^{N} w_l ||\mathbf{R}_{ij}\mathbf{p}_l + \mathbf{t}_{ij} - \mathbf{q}_l)||^2. \quad (2)$$

Differentiable closed-form solution based on the Kabsch algorithm is then defined as follows:

$$\overline{\mathbf{p}} := \sum_{l=1}^{N} w_l \mathbf{p}_l / |\mathbf{w}|_1, \quad \overline{\mathbf{q}} := \sum_{l=1}^{N} w_l \mathbf{q}_l / |\mathbf{w}|_1 \quad (3)$$

where $\overline{\mathbf{p}}$ and $\overline{\mathbf{q}}$ denote the weighted centroids of point clouds $\mathbf{P}_i \in \mathbb{R}^{N\times3}$ and $\mathbf{Q}_j \in \mathbb{R}^{N\times3}$, where $\mathbf{P}_i \sim \mathbf{Q}_j$ are under correspondence, respectively. The centered point coordinates can then be computed as $\widetilde{\mathbf{p}}_l := \mathbf{p}_l - \overline{\mathbf{p}}$, $\widetilde{\mathbf{q}}_l := \mathbf{q}_l - \overline{\mathbf{q}}$, $l = 1, \ldots, N$. Arranging the centered points back to the matrix forms $\widetilde{\mathbf{P}} \in \mathbb{R}^{N\times3}$ and $\widetilde{\mathbf{Q}} \in \mathbb{R}^{N\times3}$, a weighted covariance matrix $\mathbf{S} \in \mathbb{R}^{3\times3}$ can be computed as $\mathbf{S} = \widetilde{\mathbf{P}}^T \mathbf{W} \widetilde{\mathbf{Q}}$, where $\mathbf{W} = \text{diag}(w_1, \ldots, w_N)$. The solution to the registration problem in Eq. 2 is then given by the projection of $\mathbf{S}$ onto the $SO(3)$ manifold as: $\hat{\mathbf{R}}_{ij} = \mathbf{V} \cdot \text{diag}([1, 1, \det(\mathbf{V}\mathbf{U}^T)]) \cdot \mathbf{U}^T$ where $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the singular value decomposition of $\mathbf{S}$. $\det(\cdot)$ denotes computing the determinant and is used here to avoid creating a reflection matrix. Finally, $\hat{\mathbf{t}}_{ij}$ is computed as $\hat{\mathbf{t}}_{ij} = \overline{\mathbf{q}} - \hat{\mathbf{R}}_{ij}\overline{\mathbf{p}}$. Motivated by the results in [18, 22] we add another registration block (i. e. *Reg. iter.*) to our network. This block is identical to *Reg. init.* except the fact that it uses the weights as well as the pointwise residuals along with the original input to further refine the registration result.

Confidence estimation The unknown overlap ratio between point clouds makes it difficult to i) determine the quality of the estimated relative pose parameters, ii) build the graph connections between point clouds for the transformation synchronization. Hence, we introduce the confidence estimation network whose output c_{ij} indicates the confidence in the estimated pairwise transformations $\hat{\mathbf{M}}_{ij}$.

Methods		Rotation Error						Translation Error (m)					
		3°	5°	10°	30°	45°	Mean/Med.	0.05	0.1	0.25	0.5	0.75	Mean/Med.
Pairwise (All)	*FGR* [23]	9.9	16.8	23.5	31.9	38.4	76.3°/-	5.5	13.3	22.0	29.0	36.3	1.67/-
	Ours (1^{st} iter.)	32.6	37.2	41.0	46.5	49.4	65.9°/48.8°	25.1	34.1	40.0	43.4	46.8	1.37/0.94
FGR (Good)	*FastGR* [23]	12.4	21.4	29.5	38.6	45.1	68.8°/-	7.7	17.6	28.2	36.2	43.4	1.43/-
	EIGSE3 (*FGR*) [2]	1.5	4.3	12.1	34.5	47.7	68.1°/-	1.2	4.1	14.7	32.6	46.0	1.29/-
	L2Sync (*FGR*) [16]	34.4	41.1	49.0	58.9	62.3	42.9°/-	2.0	7.3	22.3	36.9	48.1	1.16/-
Ours (Good)	*EIGSE3* [2]	63.3	70.2	75.6	80.5	81.6	23.0°/1.7°	42.2	58.5	69.8	76.9	79.7	0.45/0.06
	Ours (1^{st} iter.)	57.7	65.5	71.3	76.5	78.1	28.3°/1.9°	44.8	60.3	69.6	73.1	75.5	0.57/0.06
	Ours (4^{th} iter.)	60.6	68.3	73.7	78.9	81.0	24.2°/1.8°	47.1	63.3	72.2	76.2	78.7	0.50/**0.05**
	Ours (After Sync)	**65.8**	**72.8**	**77.6**	**81.9**	**83.2**	**20.3°/1.6°**	**48.4**	**67.2**	**76.5**	**79.7**	**82.0**	**0.42/0.05**

Table 1. Multiview registration evaluation on *ScanNet* [8] dataset. We report the ECDF values for rotation and translation errors. Best results are shown in bold.

Differentiable transformation synchronization The global transformation parameters can be estimated either jointly (*transformation synchronization*) [14, 3, 2, 6] or by dividing the problem into *rotation synchronization* [1]

$$\mathbf{R}_i^* = \underset{\mathbf{R_i} \in SO(3)}{\arg\min} \sum_{(i,j)\in\mathcal{E}} c_{ij} ||\hat{\mathbf{R}}_{ij} - \mathbf{R}_i \mathbf{R}_j^T||_F^2 \qquad (4)$$

and *translation synchronization* [15]

$$\mathbf{t}_i^* = \underset{\mathbf{t_i}}{\arg\min} \sum_{(i,j)\in\mathcal{E}} c_{ij} ||\hat{\mathbf{R}}_{ij}\mathbf{t}_i + \hat{\mathbf{t}}_{ij} - \mathbf{t}_j||^2 \qquad (5)$$

where the weights c_{ij} represent the estimated confidence in the relative transformation parameters. Both of these optimization problems admit differentiable closed form solutions under spectral relaxation as follows [1, 15]. Consider a symmetric matrix $\mathbf{L} \in \mathbb{R}^{3N_\mathcal{S} \times 3N_\mathcal{S}}$ resembling a block Laplacian matrix, defined as

$$\mathbf{L} = \begin{bmatrix} \mathbf{I}_3 \sum_i c_{i1} & -c_{12}\hat{\mathbf{R}}_{12} & \cdots & -c_{1N_\mathcal{S}}\hat{\mathbf{R}}_{1N_\mathcal{S}} \\ -c_{21}\hat{\mathbf{R}}_{21} & \mathbf{I}_3 \sum_i c_{i2} & \cdots & -c_{2N_\mathcal{S}}\hat{\mathbf{R}}_{2N_\mathcal{S}} \\ \vdots & & \ddots & \vdots \\ -c_{N_\mathcal{S}1}\hat{\mathbf{R}}_{N_\mathcal{S}1} & -c_{N_\mathcal{S}2}\hat{\mathbf{R}}_{N_\mathcal{S}2} & \cdots & \mathbf{I}_3 \sum_i c_{iN_\mathcal{S}} \end{bmatrix}$$

where $N_\mathcal{S}$ denotes the number of nodes in the graph. The least squares estimates of the global rotation matrices $\mathbf{R}_i^*$ are then given, under relaxed orthonormality and determinant constraints, by the three eigenvectors $\mathbf{v}_i \in \mathbb{R}^{3N_\mathcal{S}}$ corresponding to the smallest eigenvalues of $\mathbf{L}$. Consequently, the nearest rotation matrices under Frobenius norm can be obtained by a projection of the 3×3 submatrices of $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \in \mathbb{R}^{3N_\mathcal{S} \times 3}$ onto $SO(3)$ analogous to the Kabsch algorithm in pairwise registration.

Similarly, the closed-form solution to the least squares formulation of the translation synchronization can be written as [16] $\mathbf{t}^* = \mathbf{L}^+\mathbf{b}$ where $\mathbf{t}^* = [\mathbf{t}_1^{*^T}, \ldots, \mathbf{t}_{N_\mathcal{S}}^{*^T}]^T \in \mathbb{R}^{3N_\mathcal{S}}$ and $\mathbf{b} = [\mathbf{b}_1^{*^T}, \ldots, \mathbf{b}_{N_\mathcal{S}}^{*^T}]^T \in \mathbb{R}^{3N_\mathcal{S}}$ with $\mathbf{b}_i := -\sum_{j\in\mathcal{N}(i)} c_{ij}\hat{\mathbf{R}}_{ij}^T\hat{\mathbf{t}}_{ij}$. where $\mathcal{N}(i)$ denotes the neighboring vertices of $\mathbf{S}_i$ in $\mathcal{G}$ and $^+$ shows the pseudoinverse.

3. Experimental evaluation

Training The individual parts of the network are connected into an end-to-end multiview 3D registration algorithm as shown in Fig. 2. We pre-train the individual subnetworks (training details available in [11]) before fine-tuning the whole model in an end-to-end manner on the 3DMatch dataset [21] using the official train/test data split.

Multiview registration We evaluate the performance of our approach on the task of multiview registration using the *ScanNet* [8] dataset. To ensure a fair comparison, we follow [16] and use the same 32 randomly sampled scenes for evaluation. For each scene we randomly sample 30 RGBD images that are 20 frames apart. The temporal sequence of the frames is discarded. We compare our method to a SOTA transformation synchronization [2], as well as the recent learned approach [16]. We evaluate [2] with both, [23] and our pairwise transformation estimates as input. "Good" in Tab. 1 denotes that the edges were punned according to [16] before the transformation synchronization.

As shown in Tab. 1 our approach can achieve a large improvement on the multiview registration tasks compared to the baselines. Not only our estimation of the initial pairwise relative transformations estimated using are more accurate than the ones of FGR [23], but they can also be further improved in the subsequent iterations. This clearly confirms the benefit of the feed-back loop of our algorithm. In Fig. 3, a qualitative comparison of the global registration of scene **Hotel 1** from *3DMatch* dataset is presented.

4. Conclusions

We have introduced an end-to-end learnable, global, multiview point cloud registration algorithm. Our method departs from the common two-stage approach and directly learns to register all views in a globally consistent manner. Experimental evaluation on benchmark datasets show that our method outperforms SOTA by more than 25 percentage points on average, considering the rotation error.

Acknowledgements. This work is partially supported by Stanford-Ford Alliance, NSF grant IIS-1763268, Vannevar Bush Faculty Fellowship, Samsung GRO program and the Stanford SAIL Toyota Research Center. We thank NVIDIA Corp. for providing the GPUs used in this work.

Figure 3. Registration of multiple scans of the **Hotel 1** scene from *3DMatch* dataset. The superiority of our method over the state of the art is also apparent when the outcome are qualitatively inspected.

References

[1] F. Arrigoni, L. Magri, B. Rossi, P. Fragneto, and A. Fusiello. Robust absolute rotation estimation via low-rank and sparse matrix decomposition. In *IEEE 3DV*, 2014. 3

[2] F. Arrigoni, B. Rossi, and A. Fusiello. Spectral synchronization of multiple views in se(3). *SIAM Journal on Imaging Sciences*, 2016. 1, 2, 3

[3] F. Bernard, J. Thunberg, P. Gemmar, F. Hertel, A. Husch, and J. Goncalves. A solution for multi-alignment by transformation synchronisation. In *IEEE CVPR*, 2015. 1, 3

[4] U. Bhattacharya and V. M. Govindu. Efficient and robust registration on the 3d special euclidean group. In *IEEE ICCV*, 2019. 1

[5] T. Birdal and S. Ilic. Cad priors for accurate and flexible instance reconstruction. In *IEEE ICCV*, 2017. 1

[6] T. Birdal, U. Simsekli, M. O. Eken, and S. Ilic. Bayesian pose graph optimization via bingham distributions and tempered geodesic mcmc. In *NIPS*, 2018. 1, 3

[7] C. Choy, J. Park, and V. Koltun. Fully convolutional geometric features. In *IEEE ICCV*, 2019. 1, 2

[8] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE CVPR*, 2017. 3

[9] H. Deng, T. Birdal, and S. Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *ECCV*, 2018. 1

[10] H. Deng, T. Birdal, and S. Ilic. 3d local features for direct pairwise registration. In *IEEE CVPR*, 2019. 1

[11] Z. Gojcic, C. Zhou, J. Wegner, L. Guibas, and T. Birdal. Learning multiview 3d point cloud registration. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[12] Z. Gojcic, C. Zhou, J. D. Wegner, and A. Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *IEEE CVPR*, 2019. 1

[13] Z. Gojcic, C. Zhou, and A. Wieser. Robust pointwise correspondences for point cloud based deformation monitoring of natural scenes. In *4th JISDM*, 2019. 2

[14] V. M. Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *IEEE CVPR*, 2004. 1, 3

[15] X. Huang, Z. Liang, C. Bajaj, and Q. Huang. Translation synchronization via truncated least squares. In *NIPS*, 2017. 3

[16] X. Huang, Z. Liang, X. Zhou, Y. Xie, L. J. Guibas, and Q. Huang. Learning transformation synchronization. In *CVPR*, 2019. 3

[17] T. Plötz and S. Roth. Neural nearest neighbors networks. In *NIPS*, 2018. 2

[18] R. Ranftl and V. Koltun. Deep fundamental matrix estimation. In *ECCV*, 2018. 2

[19] P. Theiler, J. D. Wegner, and K. Schindler. Markerless point cloud registration with keypoint-based 4-points congruent stes. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume II-5/W2, 2013. 1

[20] P. Theiler, J. D. Wegner, and K. Schindler. Globally consistent registration of terrestrial laser scans via graph optimization. *ISPRS Journal of Photogrammetry and Remote Sensing*, 109:126–136, 2015. 1

[21] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser. 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In *IEEE CVPR*, 2017. 1, 3

[22] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, L. Quan, and H. Liao. Learning two-view correspondences and geometry using order-aware network. In *ICCV*, 2019. 2

[23] Q.-Y. Zhou, J. Park, and V. Koltun. Fast global registration. In *ECCV*, 2016. 1, 3