The Perfect Match: 3D Point Cloud Matching with Smoothed Densities Supplementary material

Zan Gojcic Caifa Zhou Jan D. Wegner Andreas Wieser

ETH Zurich

{firstname.lastname@geod.baug.ethz.ch}

In this supplementary material we provide additional information about the evaluation experiments (Sec. 1, 2 and 3) along with the detailed per-scene results (Sec. 4) and some further visualizations (Fig. 1 and 2). The source code and all the data needed for comparison are publicly available at https://github.com/zgojcic/3DSmoothNet.

1. Evaluation metric

This section provides a detailed explanation of the evaluation metric adopted from [2] and used for all evaluation experiments throughout the paper.

Consider two point cloud fragments $\mathcal{P}$ and $\mathcal{Q}$, which have more than 30% overlap under ground-truth alignment. Furthermore, let all such pairs form a set of fragment pairs $\mathcal{F} = \{(\mathcal{P}, \mathcal{Q})\}$. For each fragment pair the set of correspondences obtained in the feature space is then defined as

$$\mathcal{C} = \{\{\mathbf{p}_i \in \mathcal{P}, \mathbf{q}_j \in \mathcal{Q}\}, f(\mathbf{p}_i) = \text{nn}(f(\mathbf{q}_j), f(\mathcal{P})) \wedge f(\mathbf{q}_j) = \text{nn}(f(\mathbf{p}_i), f(\mathcal{Q}))\} \quad (1)$$

where $f(\mathbf{p})$ denotes a non-linear function that maps the feature point $\mathbf{p}$ to its local feature descriptor and nn() denotes the nearest neighbor search based on the $l2$ distance. Finally, the quality of the correspondences in terms of average recall R per scene is computed as

$$R = \frac{1}{|\mathcal{F}|} \sum_{f=1}^{|\mathcal{F}|} \mathbb{1}\Big(\big[\frac{1}{|\mathcal{C}_f|} \sum_{i,j \in \mathcal{C}_s} \mathbb{1}\big(||\mathbf{p}_i - T_f(\mathbf{q}_j)||_2 < \tau_1 \big) \big] > \tau_2 \Big) \quad (2)$$

where T_f denotes the ground-truth transformation alignment of the fragment pair $f \in \mathcal{F}$. τ_1 is the threshold on the Euclidean distance between the correspondence pair (i, j) found in the feature space and τ_2 is a threshold on the inlier ratio of the correspondences [2]. Following [2] we set $\tau_1 = 0.1$m and $\tau_2 = 0.05$ for both, the *3DMatch* [8] as well as the *ETH* [5] data set. The evaluation metric is based on the theoretical analysis of the number of iterations k needed by RANSAC [3] to find at least $n = 3$ corresponding points with the probability of success $p = 99.9\%$. Considering, $\tau_2 = 0.05$ and the relation

$$k = \frac{\log(1-p)}{\log(1-\tau_2^n)}, \quad (3)$$

the number of iterations equals $k \approx 55000$ and can be greatly reduced if the number of inliers τ_2 can be increased (e.g. $k = 860$ if $\tau_2 = 0.2$).

Method	Parameter	*3Dmatch* data set	*ETH* data set
FPFH [6]	r_f[m]	0.093	0.310
	r_n[m]	0.093	0.310
SHOT [7]	r_f[m]	0.186	0.620
	r_n[m]	0.093	0.310
3DMatch [8]	W[m]	0.300	1.500[1]
	n_{voxels}	30^3	30^3
CGF [4]	r_f[m]	0.186	0.620
	r_n[m]	0.093	0.310
	r_{min}[2][m]	0.015	0.05
PPFNet [2]	k_n[points]	17	/
	r_f[m]	0.300	/
PPF-FoldNet [1]	k_n[points]	17	/
	r_f[m]	0.300	/

Table 1: Parameters used for the state-of-the-art methods in the evaluation experiments.

2. Baseline Parameters

In order to perform the comparison with the state-of-the-art methods, several parameters have to be set. To ensure a fair comparison we set all the parameters relative to our voxel grid width W which we set as $W_{3DMatch} = 0.3$m and $W_{ETH} = 1$m for *3DMatch* and *ETH* data sets respectively. More specific, for the descriptors based on the spherical support we use a feature radius $r_f = \sqrt[3]{\frac{3}{4\pi}} W$ that yields a sphere with the same volume as our voxel grid and

[1] Larger voxel grid width used due to the memory restrictions.

[2] Used to avoid the excessive binning near the center, see [4]

	FPFH [6] (33 dim)	SHOT [7] (352 dim)	3DMatch [8] (512 dim)	CGF [4] (32 dim)	PPFNet [2] (64 dim)	PPF-FoldNet [1] (512 dim)	Ours (16 dim)	Ours (32 dim)
Kitchen	43.1	74.3	58.3	60.3	89.7	78.7	93.1	**97.0**
Home 1	66.7	80.1	72.4	71.1	55.8	76.3	93.6	**95.5**
Home 2	56.3	70.7	61.5	56.7	59.1	61.5	86.5	**89.4**
Hotel 1	60.6	77.4	54.9	57.1	58.0	68.1	95.6	**96.5**
Hotel 2	56.7	72.1	48.1	53.8	57.7	71.2	90.4	**93.3**
Hotel 3	70.4	85.2	61.1	83.3	61.1	94.4	**98.2**	**98.2**
Study	39.4	64.0	51.7	37.7	53.4	62.0	92.8	**94.5**
MIT Lab	41.6	62.3	50.7	45.5	63.6	62.3	92.2	**93.5**
Average	54.3	73.3	57.3	58.2	62.3	71.8	92.8	**94.7**
STD	11.8	7.7	7.8	14.2	11.5	9.9	3.4	2.7

Table 2: **Detailed quantitative results on the *3DMatch* dataset.** For each scene we report the average recall in percent over all overlapping fragment pairs. Best performance is shown in bold.

for all voxel-based descriptors we use the same voxel grid width W. For descriptors that require, along with the coordinates also the normal vectors, we use the point cloud library (PCL) built-in function for normal vector computation, using all the points in the spherical support with the radius $r_n = \frac{r_f}{2}$. Tab. 1 provides all the parameters that were used for the evaluation. If some parameters are not listed in Tab 1 we use the original values set by the authors. For the handcrafted descriptors, FPFH [6] and SHOT [7] we use the implementation provided by the original authors as a part of the PCL[3]. We use the PCL version 1.8.1 x64 on Windows 10 and use the parallel programming implementations (omp) of both descriptors. For 3DMatch [8] we use the implementation provided by the authors[4] on Ubuntu 16.04 in combination with the CUDA 8.0 and cuDNN 5.1. Finally, for CGF [4] we use the implementation provided by the authors[5] on a PC running Windows 10. Note that we report the results of PPFNet [2] and PPF-FoldNet [1] as reported by the authors in the original papers, because the source code is not publicly available. Nevertheless, for the sake of completeness we report the feature radius r_f and the k-nearest neighbors k_n used for the normal vector computation, which were used by the authors in the original works. For the *3DRotatedMatch* and *3DSparseMatch* data sets we use the same parameters as for the*3DMatch* data set.

Performance of the 3DMatch descriptor The authors of the 3DMatch descriptor provide along with the source code and the trained model also the precomputed truncated distance function (TDF) representation and inferred descriptors for the *3DMatch* data set. We use this descriptors directly for all evaluations on the original *3DMatch* data set. For the evaluations on the *3DRotatedMatch*, *3DSparseMatch* and *ETH* data sets we use their source code in combination with the pretrained model to infer the descriptors.

[3]https://github.com/PointCloudLibrary/pcl
[4]https://github.com/andyzeng/3dmatch-toolbox
[5]https://github.com/marckhoury/CGF

When analyzing the *3DSparseMatch* data set results, we noticed a discrepancy. The descriptors inferred by us achieve better performance than the provided ones. We analyzed this further and determined that the TDF representation (i.e. the input to the CNN) is identical and the difference stems from the inference using their provided weights. In the paper this is marked by a footnote in the results section. For the sake of consistency, we report in this Supplementary material all results for *3DMatch* data set using the precoumpted descriptors and the results on all other data set using the descriptors inferred by us.

3. Preprocessing of the benchmark data sets

***3DMatch* data set** The authors of *3DMatch* data set provide along with the point cloud fragments and the ground-truth transformation parameters also the indices of the interest points and the ground-truth overlap for all fragments. To make the results comparable to previous works, we use these indices and overlap information for all descriptors and perform no preprocessing of the data except for the random rotations (*3DRotatedMatch*) and random subsampling (*3DSparseMatch*).

***3DSparseMatch* data set** In order to test the robustness of our approach to variations in point density we create a new data set, denoted as *3DSparseMatch*, using the point cloud fragments from the *3DMatch* data set. Specifically, we first extract the indices of the interest points provided by the authors of the *3DMatch* data set and then randomly downsample the remaining points, keeping 50%, 25% and 12.5% of the points. We consider two scenarios in the evaluation. In the first scenario we use one of the fragments to be registered with the full and the other one with the reduced point cloud density (*Mixed*), while in the second scenario we evaluate the descriptors on the fragments with the same level of sparsity (*Both*).

	FPFH [6] (33 dim)	SHOT [7] (352 dim)	3DMatch [8] (512 dim)	CGF [4] (32 dim)	PPFNet [2] (64 dim)	PPF-FoldNet [1] (512 dim)	Ours (16 dim)	Ours (32 dim)
Kitchen	43.5	74.1	2.4	60.5	0.2	78.9	93.3	**97.2**
Home 1	66.7	80.1	3.8	71.2	0.0	78.2	93.6	**96.2**
Home 2	56.3	70.2	5.3	57.2	1.4	64.4	87.0	**90.9**
Hotel 1	62.4	77.0	1.8	57.2	0.4	67.7	95.6	**96.5**
Hotel 2	56.7	72.1	6.7	53.8	0.0	69.2	91.4	**92.3**
Hotel 3	72.2	85.2	1.9	83.3	0.0	96.3	**98.2**	**98.2**
Study	39.7	65.1	2.7	38.7	0.0	62.7	93.2	**94.5**
MIT Lab	41.6	62.3	3.9	45.5	0.0	67.5	92.2	**93.5**
Average	54.9	73.3	3.6	58.5	0.3	73.1	93.0	**94.9**
STD	12.2	7.6	1.7	14.0	0.5	11.1	3.2	2.5

Table 3: **Detailed quantitative results on the *3DRotatedMatch* data set.** For each scene we report the average recall in percent over all overlapping fragment pairs. Best performance is shown in bold.

	FPFH (33 dim)	SHOT (352 dim)	3DMatch (512 dim)	CGF (32 dim)	Ours (16 dim)	Ours (32 dim)
Kitchen	89	154	103	125	200	**274**
Home 1	142	206	134	156	252	**324**
Home 2	125	182	125	142	247	**318**
Hotel 1	86	131	73	90	192	**272**
Hotel 2	94	124	64	94	178	**238**
Hotel 3	119	159	64	130	210	**276**
Study	56	84	64	55	130	**171**
MIT Lab	74	121	84	78	194	**246**
Average	98	145	88	108	200	**264**

Table 4: **Average number of correct correspondences on *3DMatch* data set.** We report the average number of correct correspondences over all overlapping fragments of individual scenes.

***ETH* data set** For the *ETH* data set we use the point clouds and the ground-truth transformation parameters provided by the authors of the data set. We start by downsampling the point clouds using a voxel grid filter with the voxel size equal to 0.02m. The authors of the data set also provide the ground-truth overlap information, but due to the downsampling step we opt to compute the overlap on our own as follows. Let $\mathbf{p}_i \in \mathcal{P}$ and $\mathbf{q}_i \in \mathcal{Q}$ denote points in the point clouds $\mathcal{P}$ and $\mathcal{Q}$, which are part of the same scene of *ETH* data set, respectively. Given the ground-truth transformation $T_{\mathcal{P}}^{\mathcal{Q}}$ that aligns the point cloud $\mathcal{Q}$ with the point cloud $\mathcal{P}$, we compute the overlap $\psi_{\mathcal{P},\mathcal{Q}}$ relative to point cloud $\mathcal{P}$ as

$$\psi_{\mathcal{P},\mathcal{Q}} = \frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \mathbb{1}\big(||\mathbf{p}_i - \text{nn}(\mathbf{p}_i, T_{\mathcal{P},\mathcal{Q}}(\mathcal{Q})||_2 < \tau_\psi\big) \quad (4)$$

where nn denotes the nearest neighbor search based on the $l2$ distance in the Euclidean space and τ_ψ thresholds the distance between the nearest neighbors. In our evaluation experiments, we select $\tau_\psi = 0.06$m, which equals three times the resolution of the point clouds after the voxel grid downsampling, and consider only the point cloud pairs for which both $\psi_{\mathcal{P},\mathcal{Q}}$ and $\psi_{\mathcal{Q},\mathcal{P}}$ are bigger than 0.3. Because no indices of the interest points are provided we randomly sample 5000 interest points that have more than 10 neighbor points in a sphere with a radius $r = 0.5$m in every point cloud. The condition of minimum ten neighbors close to the interest point is enforced in order to avoid the problems with the normal vector computation.

4. Detailed results

***3DMatch* data set** Detailed per scene results on the *3DMatch* data set are reported in Tab. 2. *Ours* (32 dim) consistently outperforms all state-of-the-art by a significant margin and achieves a recall higher than 89% on all of the scenes. However, the difference between the performance of individual descriptors is somewhat masked by the selected low value of τ_2, e.g. same average recall on Hotel 3 scene achieved by *Ours* (16 dim) and *Ours* (32 dim). Therefore, we additionally perform a more direct evaluation of the quality of found correspondences, by computing the average number of correct correspondences established by each individual descriptor (Tab 4). Where the term correct correspondences, denotes the correspondences for which the distance between the points in the coordinate space after the ground-truth alignment is smaller than 0.1m. Results in Tab. 4 again show the dominant performance of the 3DSmoothNet compared to the other state-of-the-art but also highlight the difference between *Ours* (32 dim) and *Ours* (16 dim). Remarkably, *Ours* (32 dim) can establish almost two times more correspondences than the closest competitor.

***3DRotatedMatch* data set** We additionally report the detailed results on the *3DRotatedMatch* data set in Tab 3. Again, 3DSmoothNet outperforms all other descriptor on all the scenes and maintains a similar performance as on the *3DMatch* data set. As expected the performance of the rotational invariant descriptors [6, 7, 4, 1] is not affected by the rotations of the fragments, whereas the performance of the

	3DSparseMatch data set					
	Mixed			*Both*		
	50%	25%	12.5%	50%	25%	12.5%
FPFH [6]	54.4	52.0	48.3	52.2	49.7	41.5
SHOT [7]	71.1	69.8	69.8	70.8	68.4	66.4
3DMatch [8]	73.0	72.7	70.2	73.8	72.8	72.8
CGF [4]	54.2	49.0	37.5	50.3	38.3	24.4
Ours (16 dim)	92.5	92.3	91.3	92.7	91.7	90.5
Ours (32 dim)	**95.0**	**94.5**	**94.1**	**95.0**	**94.5**	**93.7**

Table 5: **Results on the *3DSparseMatch* data set**. 'Mixed' denotes Scenario 1 in which only one of the fragments was downsampled and 'Both' denotes that both fragments were downsampled. We report average recall in percent over all scenes. Best performance is shown in bold.

descriptors, which are not rotational invariant [8, 2] drops to almost zero. This greatly reduces the applicability of such descriptors for general use, where one considers the point cloud, which are not represented in their canonical representation.

***3DSparseMatch* data set** Tab 5 shows the results on the three different density levels (50%, 25% and 12, 5%) of the *3DSparseMatch* data set. Generally, all descriptors perform better when the point density of only one fragments is reduced, compared to when both fragments are downsampled. In both scenarios, the recall of our approach drops marginally by max 1 percent point and remains more than 20 percent points above any other competing method. Therefore, 3DSmoothNet can be labeled as invariant to point density changes.

References

[1] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *European conference on computer vision (ECCV)*, 2018. 1, 2, 3

[2] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 4

[3] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 1

[4] Marc Khoury, Qian-Yi Zhou, and Vladlen Koltun. Learning compact geometric features. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3, 4

[5] François Pomerleau, M. Liu, Francis Colas, and Roland Siegwart. Challenging data sets for point cloud registration algorithms. *The International Journal of Robotics Research*, 31(14):1705–1711, 2012. 1

[6] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2009. 1, 2, 3, 4

[7] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *European conference on computer vision (ECCV)*, 2010. 1, 2, 3, 4

[8] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 4

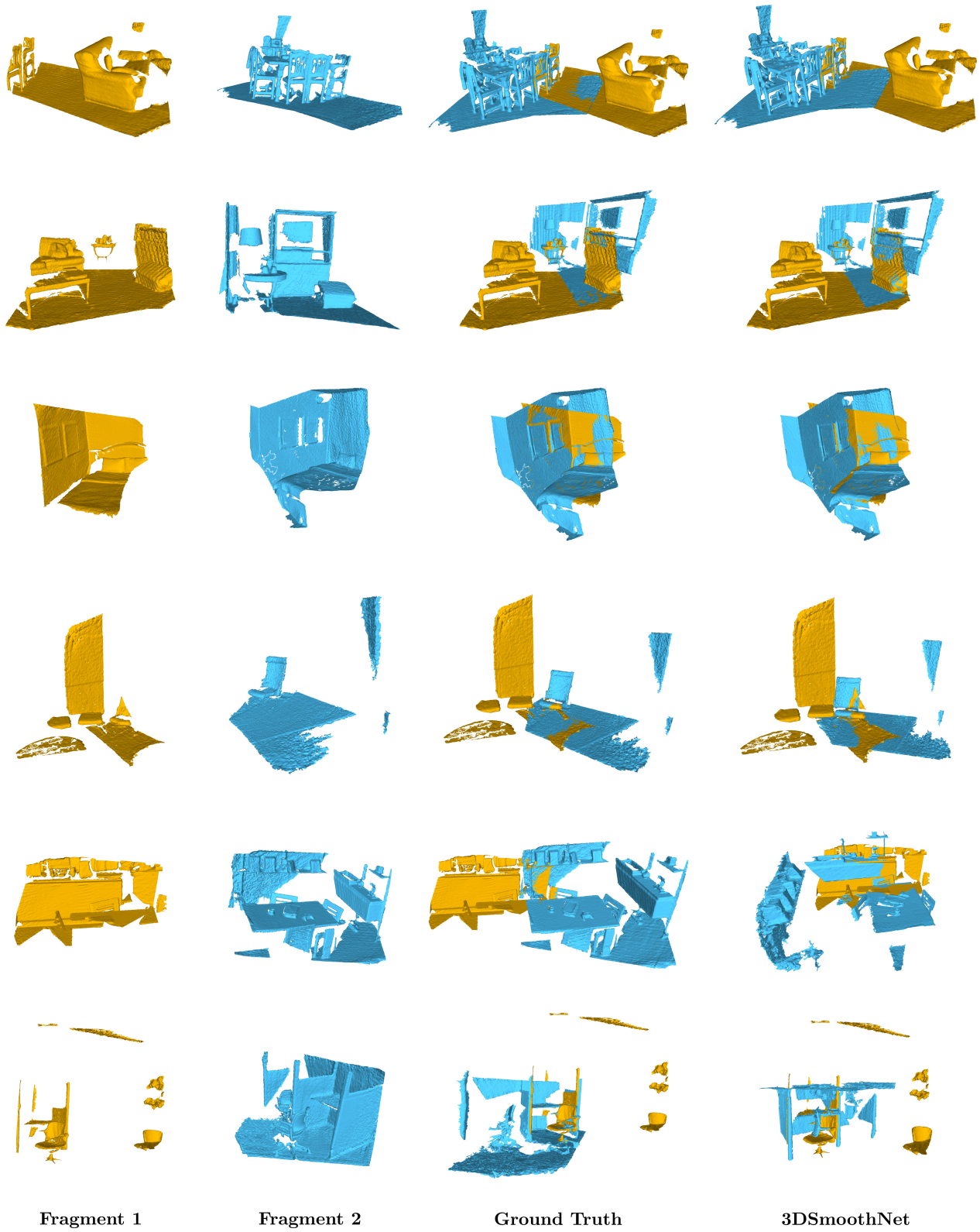

Figure 1: **Additional qualitative results of 3DSmoothNet on the *3DMatch* data set**. First three rows show hard examples for which the 3DSmoothNet succeeds, whereas the last three rows show some of the failure cases. 3DMatch and CGF fail for all these examples.

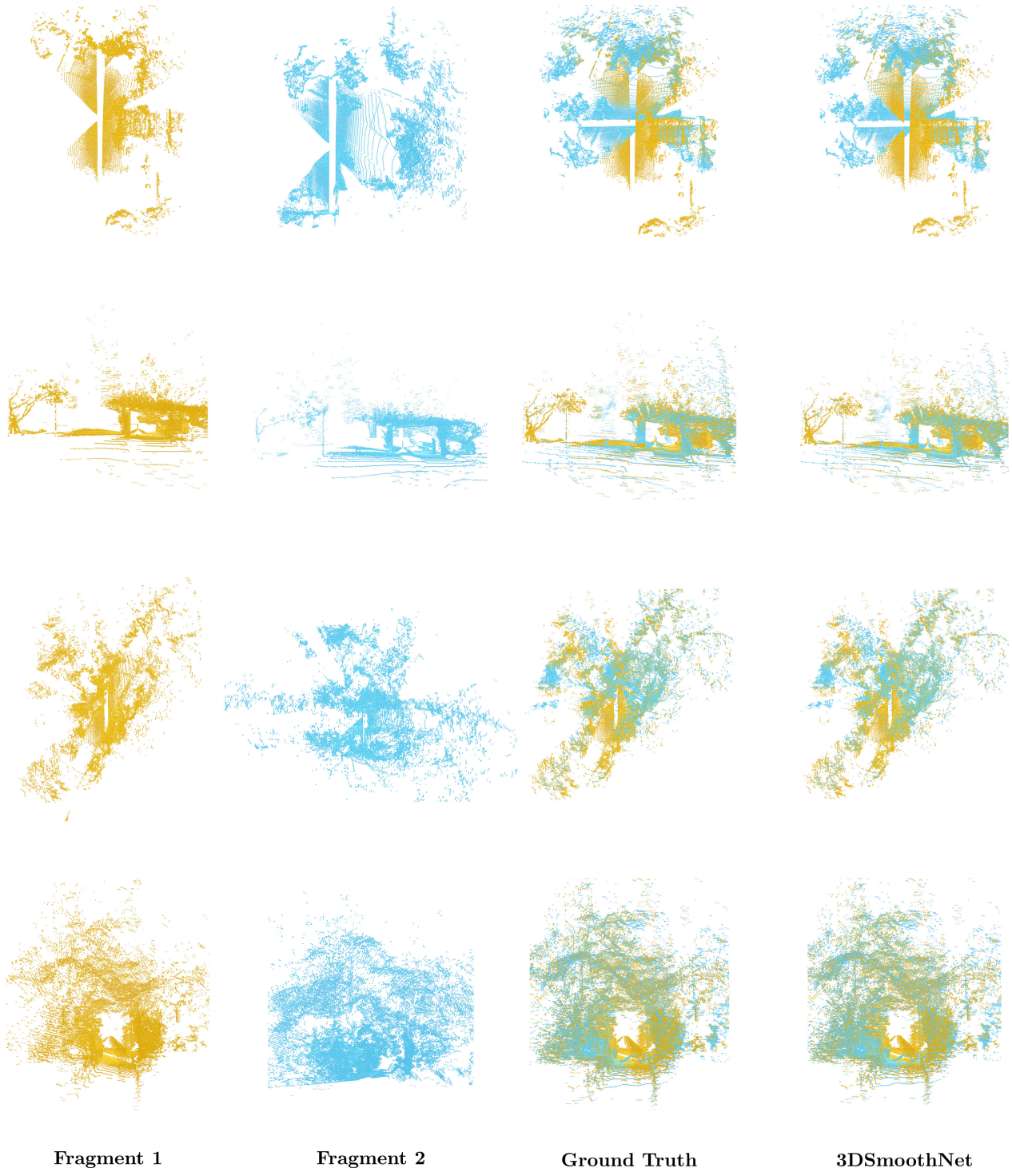

Figure 2: **Qualitative results of the 3DSmoothNet on the *ETH* data set**. 3DSmoothNet trainined only on the indoor reconstructions from RGB-D images can generalize to outdoor natural scenes, which consist of high level of noise and predominantly unstructured vegetation. The data set is made even harder by the introduced dynamic between the epochs (e.g. walking persons)