A Diverse, Real-World Data Set for Training and Evaluation on 3D Point Clouds

Florian Gawrilowicz
Department of Applied Mathematics and Computer Science
Technical University of Denmark
flgw@dtu.dk

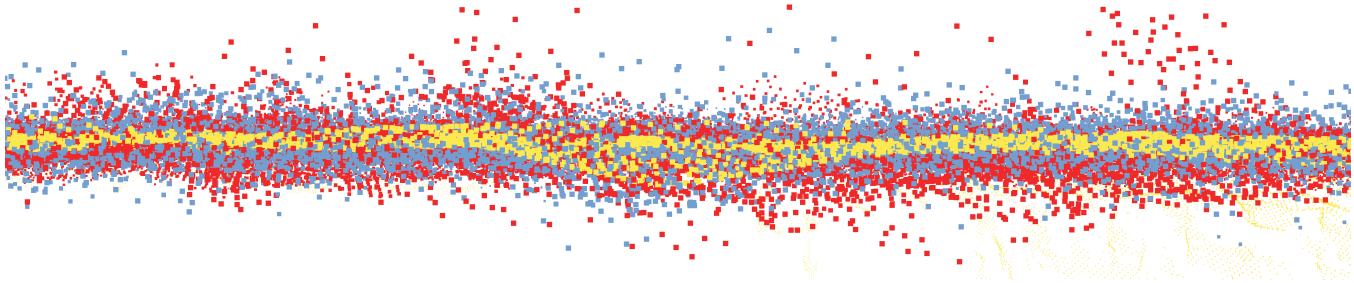

Figure 1: Closeup on superimposed point clouds, in blue color the stuctured light scan, in red the MVS approach by Furukawa and Ponce [9] and in yellow the result of the combined robust implicit moving least squares [13] (RIMLS) processing. While the structured light results exhibits less outliers than the MVS approach, it still is noisy, whereas the combined processing produces results of much higher fidelity.

ABSTRACT

We present a unique data set for training and evaluation on real-world point clouds. More specifically for post-processing of multi-view stereo results, e.g. denoising and consolidation[2].

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**; **Matching**; *Scene understanding*; *Shape representations*; *Shape inference*; *Supervised learning by regression*; *Neural networks*; 3D imaging; Shape analysis.

1 INTRODUCTION

There has always been a synergy between algorithmic advances and new data modalities. But with deep learning data has become not only an inspiration but a necessity.

While deep learning on image processing tasks has been hugely successful, extending these techniques to the three-dimensional domain is still an active and open research topic. Much of the challenges stem from the very different data representations needed in 3D processing. While the single modality for images commonly used samples on a regular grid, this does not scale well to three dimensions - it is simply not feasible to use a voxel grid for today's high-resolution scans, and this situation will only continue to worsen in the future. Using point clouds, where each data point is stored with its Euclidean coordinates, is therefore a more efficient alternative widely used in practice.

But this very different representation requires new methods to be developed to foster a comparable surge in development as we have seen on image data. Not only the methods have to be adapted and developed, but it also requires an adequate amount of data in order to become successful. Currently, much of the research done in deep learning on point clouds uses synthetic data. Often they are point samples from a surface model or CAD models. While this allows generating arbitrary amounts of training data by simply re-sampling the surface, the applicability and transferability to real-world data is not guaranteed.

We address this shortcoming by approximating ground truth for real-world data. Of course, it is virtually impossible to attain the actual accurate positions for scenes of reasonable complexity. This is because any measurement and algorithm introduces noise and outliers. To this end, we propose a combination of recording modalities, i.e. multi-view stereo algorithms (MVS) and structured light (STL) scanner. While MVS does not require special hardware and produces densely sampled point clouds, STL needs a calibrated projector in conjunction with one or more cameras. Methods trained on this data would allow achieving the quality of STL scans with the acquisition simplicity of MVS methods. This would also befit downstream task like surface reconstruction as they are usually sensitive to noise in the input.

2 RELATED WORK

2.1 Base Dataset

We leverage a previously published [1, 12] and openly available[1] data set. The data consist of 124 different scenes, seen from 49 or 64 calibrated camera positions. Images from each position are taken under seven different lighting conditions, some of them are depicted in Figure 2. Results of three different MVS algorithms [5, 9, 17] in conjunction with structured light scans for each camera position are provided in the data set.

[1] Available here: http://roboimagedata.compute.dtu.dk/

Figure 2: Photos of 40 of the 124 scenes in our data set.

2.2 Recent Deep Learning Publications

A short list of recent deep learning methods on point clouds, in conjunction with their application and used data set is given in Table 1. It is by no means a comprehensive list, but rather exemplifies the current landscape in this field. A much wider range of publications on classification, and semantic segmentation exists. This might be the result of the availability of huge data set like the ones listed below, which all target classification an/or segmentation tasks:

- **Shapenet** [7]: Triangular meshes of 220,000 CAD models classified into 3,135 categories for classification.
- **ModelNet40** [20]: Triangular meshes of CAD models of 40 categories (mostly man-made, e.g. furniture), 9,843 shapes for training and 2,468 for testing of classification and shape retrieval.
- **SHREC15** [14]: Triangular meshes of 1200 shapes from 50 categories. Each category contains 24 models, such as horses, cats, etc. in various poses. The original purpose is classification into the 50 categories.
- **ScanNet** [8]: 1513 scanned and reconstructed indoor scenes. 1201 scenes for training, 312 scenes for test. The data set was designed for 3D scene understanding tasks, such as 3D object classification, semantic labeling, and shape retrieval.
- **Matterport** [6]: RGB-D data set containing 10,800 panoramic views with surface reconstructions and camera poses. The task is also semantic segmentation.

As these data sets are not applicable to tasks like normal estimation from noisy point clouds, it is common to generate training data by sampling the surface and adding Gaussian noise to the samples, e.g. [4], [10], or [11]. Although this allows for practically infinite data it also reduces the capturing process to a very idealized noise-model. Much of the characteristics of real-world data is not faithfully captured in this, e.g. viewing direction depended effects, and systematic or correlated noise like they are obvious in [9], see Figure 3 third column. In [19] a scanning process is emulated, but still only in a very simplified fashion of rendering depth images of the models and recombining these into a point clouds.

3 CONTRIBUTION

We extend the base data set [1, 12] to make it readily usable for training of deep learning algorithms. While the original data set comes already with a method for evaluation, this error measure is not well suited for training as it does not give a one-to-one correspondence of the MVS points to the STL reference, and also exhibits some noise in the STL reference. We therefore estimate a reference by consolidating the MVS points via robust implicit moving least squares [13] (RIMLS) based on the STL reference, i.e. the implicit surface is defined by the STL points and the MVS points are relocated onto this surface. This not only results in a drastic reduction in noise, as can be seen in Figure 3, but also gives a good estimate of the normals as the displacement induced by RIMLS approximates mean curvature motion, which is perpendicular to the local tangent plane. With this processing we are able to facilitate the advantages of STL scans and transfere them to the MVS points.

4 EXAMPLE TRAINING RESULTS

As an example use case, we trained the PCPNet [10] implementation[2] published by the authors on our data set. The error function is defined by

$$\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N}(1 - |\cos(n_i, \hat{n}_i)|)^2 \ , \tag{1}$$

where n is the target normal vector and $\hat{n}$ the prediction. It penalizes the angle between the ground truth normal vectors and the prediction without considering orientation. In Figure 4 the loss for training and test split is plotted. It reaches a value of approximately 0.2 on the test data, which corresponds to a average difference of 16 degrees.

5 CONCLUSION AND FUTURE WORK

We make use of the unique and rich data set provided in [1, 12] and extend its usability further to the realms of deep learning. None of the published data sets discussed above was designed for normal estimation or consolidation of point clouds. We hope to close this gap and enable another application of deep learning.

We will release the addition to the data set on our website (http://roboimagedata.compute.dtu.dk/), as well as the source code to align MVS results with the STL reference. This will then allow facilitating other MVS methods, beyond the three already provided, and will help to further diversify the data set.

Besides that, there are two major topics to address in the future. One is working on the fidelity of our ground truth estimate. The other is evaluating more published work on this data set and assessing its benefit.

Our next step towards a more faithful data set will be formulating the consolidation process in a Bayesian framework. This will allow us to not only have a point estimate of each 3D position, but also assess the uncertainty of this estimate. Furthermore with a Bayesian approach we obtain a fitted, generative model. This will allow us to draw new samples from the estimated distribution, which in turn can be used to augment the data set and increase the robustness of the algorithms trained on our data.

A broader evaluation of algorithms is foremost a computationally bounded endeavor, as for example the training of PCPNet on our data took almost 10 days. But as this data set will become more widely known new publications might include an evaluation on it right away.

REFERENCES

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. 2016. Large-Scale Data for Multiple-View Stereopsis. *International Journal of Computer Vision* (2016), 1–16.

[2] Available here: https://github.com/paulguerrero/pcpnet

Table 1: Recent deep learning approaches on point clouds and the data sets used. No entry in test data means the same data set as for training was used and split into training & test data.

Algorithm	Application	Training data	Test data
PointNet [15]	Classification & Segmentation	Shapenet [7]	
PointNet++ [16]	Classification & Segmentation	Shapenet [7]	
PCPNet [10]	Normal & Curvature estimation	8 standard meshes, e.g. Armadillo	16 similar meshes
EC-Net [19]	Consolidation	24 CAD models and 12 everyday objects	Shapenet [7]
NormalNet [4]	Normal estimation	Simple synthetic shapes like cubes	Stanford Dragon
MC-Net [11]	Classification, Segmentation, Normal estim.	*ModelNet40* [20] (for normals)	
PCNN [3]	Classification, Segmentation, Normal estim.	*ModelNet40* [20] (for normals)	
DGCNN [18]	Classification, Segmentation, Normal estim.	*ModelNet40* [20] (for normals)	

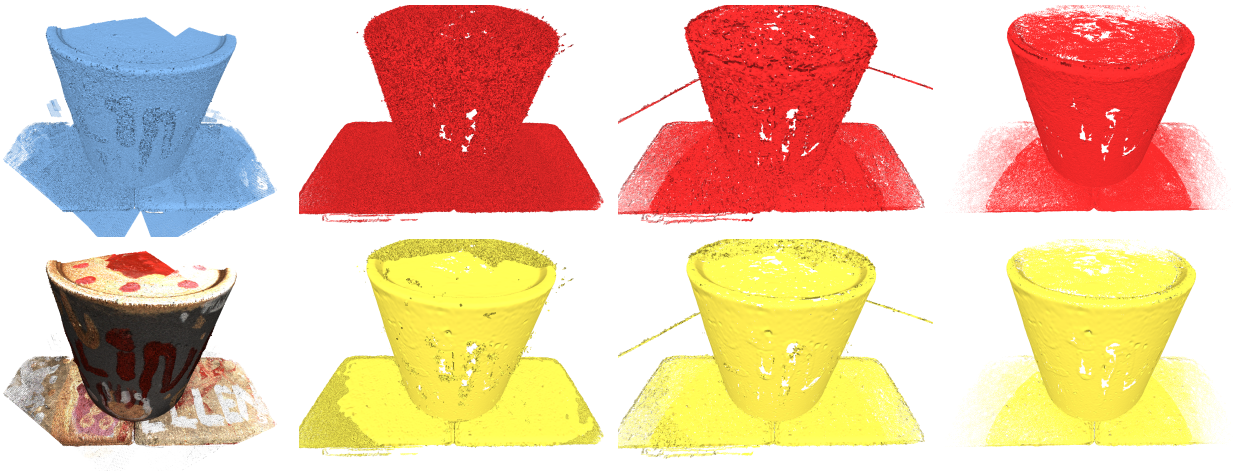

Figure 3: Structured light scan (blue) and same scan with recorded color below, Multi-view stereo results (red) from left to right: Campbell et al. [5], Furukawa and Ponce [9], Tola et al. [17], below estimated ground truth corresponding to the stereo results above.

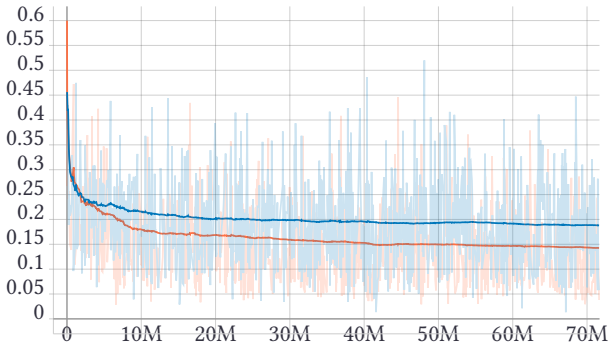

Figure 4: Progress of the loss (see eq. (1)) on training (orange) and test (blue) data for PCPNet [10].

[2] Marc Alexa, Johannes Behr, Daniel Cohen-Or, Shachar Fleishman, David Levin, and Claudio T. Silva. 2003. Computing and rendering point set surfaces. *IEEE Transactions on visualization and computer graphics* 9, 1 (2003), 3–15.

[3] Matan Atzmon, Haggai Maron, and Yaron Lipman. 2018. Point Convolutional Neural Networks by Extension Operators. *ACM Trans. Graph.* 37, 4, Article 71 (July 2018), 12 pages. https://doi.org/10.1145/3197517.3201301

[4] Alexandre Boulch and Renaud Marlet. 2016. Deep Learning for Robust Normal Estimation in Unstructured Point Clouds. *Computer Graphics Forum* 35, 5 (2016), 281–290. https://doi.org/10.1111/cgf.12983 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12983

[5] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. 2008. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*. Springer, 766–779.

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)* (2017).

[7] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015. *ShapeNet: An Information-Rich 3D Model Repository.* Technical Report arXiv:1512.03012 [cs.GR]. Stanford University — Princeton University — Toyota Technological Institute at Chicago.

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE.*

[9] Yasutaka Furukawa and Jean Ponce. 2010. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence* 32, 8 (2010), 1362–1376.

[10] Paul Guerrero, Yanir Kleiman, Maks Ovsjanikov, and Niloy J. Mitra. 2018. PCPNet: Learning Local Shape Properties from Raw Point Clouds. *Computer Graphics Forum* 37, 2 (2018), 75–85. https://doi.org/10.1111/cgf.13343

[11] Pedro Hermosilla, Tobias Ritschel, Pere-Pau Vázquez, Àlvar Vinacua, and Timo Ropinski. 2018. Monte Carlo convolution for learning on non-uniformly sampled point clouds. In *SIGGRAPH Asia 2018 Technical Papers.* ACM, 235.

[12] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. 2014. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 406–413.

[13] A Cengiz Öztireli, Gael Guennebaud, and Markus Gross. 2009. Feature preserving point set surfaces based on non-linear kernel regression. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 493–501.

[14] D. Pickup, X. Sun, P. L. Rosin, R. R. Martin, Z. Cheng, Z. Lian, M. Aono, A. Ben Hamza, A. Bronstein, M. Bronstein, S. Bu, U. Castellani, S. Cheng, V. Garro, A. Giachetti, A. Godil, J. Han, H. Johan, L. Lai, B. Li, C. Li, H. Li, R. Litman, X. Liu, Z. Liu, Y. Lu, A. Tatsuma, and J. Ye. 2014. SHREC'14 track: Shape Retrieval of Non-Rigid 3D Human Models. In *Proceedings of the 7th Eurographics workshop on 3D Object Retrieval (EG 3DOR'14)*. Eurographics Association, 10.

[15] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2016. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *arXiv preprint arXiv:1612.00593* (2016).

[16] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *arXiv preprint arXiv:1706.02413* (2017).

[17] Engin Tola, Christoph Strecha, and Pascal Fua. 2012. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications* 23, 5 (2012), 903–920.

[18] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. 2019. Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphics (TOG)* (2019).

[19] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. 2018. EC-Net: an Edge-aware Point set Consolidation Network. In *ECCV*.

[20] Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and J. Xiao. 2015. 3D ShapeNets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1912–1920. https://doi.org/10.1109/CVPR.2015.7298801