Shooting Labels by Virtual Reality

Pierluigi Zama Ramirez , Claudio Paternesi, Daniele De Gregorio, Luigi Di Stefano
Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{pierluigi.zama, daniele.degregorio, luigi.distefano}@unibo.it, claudio.paternesi@studio.unibo.it

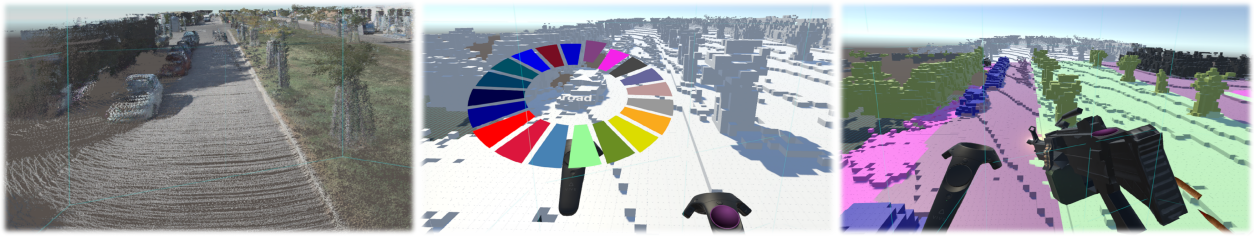

Figure 1: Left: coloured 3D point cloud obtained by a sequence from KITTI[8]. Centre: voxelized point cloud ready to be labeled together with the label (i.e. colour) selection tool. Right: labeled scene alongside with the *label shooting* tool.

Abstract

As availability of a few, large-size, annotated datasets, like ImageNet, Pascal VOC and COCO, spawned the deep learning revolution that has disrupted computer vision research so dramatically, we argue that new tools to facilitate generation of many more may finally popularize data-driven AI throughout applications and domains. In this work we propose a new tool based on Virtual Reality (VR) which makes semantic annotation of 3D data as easy and fun as a video game. Besides, our framework allows for projecting the 3D annotations into 2D images, thereby speeding up a notoriously slow and expensive task such as pixel-wise semantic labeling.

1. Introduction

Two major leitmotifs in nowadays computer vision read like *Convolutional Neural Networks have surpassed human performance in classification tasks"* [13] and *"The success of the modern Deep Neural Networks (DNNs) is ascribable to the availability of large datasets"* [5]. As for the latter, one might just consider the dramatic advances brought in by large annotated datasets like ImageNet [20] and Pascal VOC [7] in the fields of image classification and object detection, as well as by KITTI [8] and Cityscapes [5] in the realm of dense scene understanding. Indeed, the key issue in modern computer vision deals more and more with how to speed-up and facilitate acquisition of large annotated datasets. Innovative start-ups, like Scale.ai (https://scale.ai/), Superannotate.ai (https://superannotate.ai/) and many others, have received hundreds of millions of dollars in funding to develop advanced image labeling tools. This suggests data generation qualifying itself as a business as relevant as the development of data-driven AI techniques.

The annotation processes is notoriously tedious and expensive. Moreover, the more complex the perception task, the slower and more costly becomes the annotation work. If we consider, e.g., 2D Semantic Segmentation [4], among the most complex annotation tasks together with Instance Segmentation [12], labeling a single image may take several minutes and cost several dollars. Thus, as proposed in [14, 6, 17, 3], directly annotating a 3D reconstruction of the scene in order to then be able to project the 3D labels into 2D images may facilitate the data generation process.

Based on these considerations, in this work we propose a novel tool based on Virtual Reality (VR) to facilitate and speed-up dense 3D semantic labeling. This enables to obtain both 3D and 2D data endowed with semantic annotations. To the best of our knowledge, ours is the first system which allows for handling efficiently large-scale semantic labeling, such as, e.g. labeling whole city blocks. Moreover, our approach is inspired by VR games, which paves the way for full-fledged gamification of this type of activities. Our open source framework, based on Unity, Blender, and open3D [23] provides an immersive VR experience within large environments represented as 3D meshes, wherein the user can "color" sur-

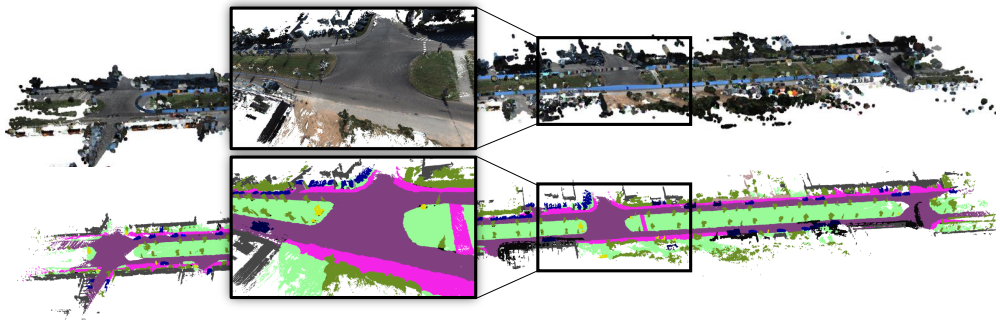

Figure 2: 3D Labeling of a Kitti sequence. Top: RGB point cloud. Bottom: labeled point cloud obtained using our tool.

faces semantically in a highly engaging way by *shooting labels*. Figure 1 depicts some in-game visualizations. Project page at https://github.com/pierlui92/Shooting-Labels.git.

2. Related Works

Semantic Segmentation Datasets Several datasets featuring 2D images annotated with semantic labels are available. The best known are KITTI [8] and Cityscapes [5], which, yet, contain a relatively small number of images semantically annotated by hand. They were two of the first datasets proposed in this area (urban outdoor), so the focus was more on the data than on how to generate them. On the other hand, the Mapillary dataset [19] includes much more images, though the labeling was still performed manually image per image. The same is true for some indoor datasets, such as [21] and [22], in which, although intelligent graphic tools are used to produce 2D frame-by-frame annotations, the order of magnitude of the available images is only slightly higher. Conversely, the authors of [14] proposed an efficient pipeline for indoor environments, which allows for scanning a room, reconstruct it in 3D and then label easily the gathered 3D data rather than each single image. In [6] the authors formally extend the procedure with the a re-projection module which brings back in 2D the 3D labels based on known camera poses. The label re-projection approach was then exploited in other datasets such as [1] and [3]. It can be observed that, leveraging on 3D reconstruction and camera tracking to facilitate labeling, may be thought of as shifting the cost of labeling each individual image toward the complexity of the requirements necessary to obtain a suitable dataset (tracked camera). This benefit is even more evident in similar synthetic datasets, such as [11, 17, 16], where obviously both camera tracking and 3D reconstruction are no longer external elements but inherent to the rendering engine. Recently [15, 10, 2] proposed large urban outdoor 2D and 3D datasets where labeling was carried out on point clouds and then 2D semantic labeled images were attained by the re-projection.

User Experience As for "how to generate the data", and specifically to the case of the real datasets, the user experience that takes place during labeling, which almost all authors define as 3D Dense Annotation, is rarely addressed in literature as it basically concerns a 3D modeling experience or the likes (e.g. think of the interaction needed to "color" a table inside a 3D model of an entire room). Some authors have expressly addressed this by proposing valid smart solutions. In [9] the authors have proposed an interactive procedure by means of which the user can physically touch the scene within a classical 3D reconstruction pipeline, so as to "color" large parts of the scene by exploiting region growing techniques. In [18], instead, the authors build a physical device able to reproduce the pipeline while the user navigates the environment in Augmented Reality, using a laser pointer to identify the homogeneous areas of the scene and assign them a correct label. Our proposed method takes advantage of the reconstruction pipeline but introduces a Virtual Reality framework to navigate within the reconstructed environments, providing the user with a series of tools, oriented to gamification, to "color" the world in a fast and intuitive manner. To the best of our knowledge, our is the first method that allows for labeling very-large-scale scenes in a short time by a VR approach.

3. Tool

In this section we will briefly describe the most important features of our VR labeling tool, which can work with the most popular 3D representations, such as point cloud and meshes, obtained by any kind of 3D reconstruction technique. Moreover, with our tool we can also load a 3D labeling already obtained by any technique (e.g. a CNN for 3D semantic segmentation) in order to refine it. Our pipeline can be summarized in 3 main steps: 1) Pre-Processing of 3D Data; 2) Virtual Reality Labeling; 3) Post-Processing of 3D Labeled Data.

Pre-Processing of 3D Data When dealing with point clouds we need to obtain a suitable visualization in terms of both efficiency and user experience. The tool proposes two

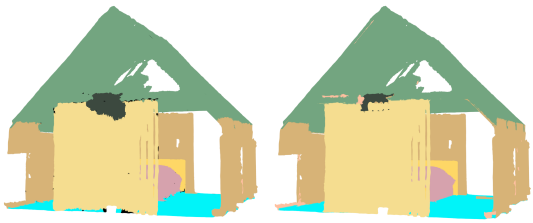

Figure 3: Qualitative results on a scene of Matterport 3D dataset [3]. Left: our labeling; Right: Matterport labels.

different visualization experience depending on the kind of 3D data at hand. If we deal with meshes we can directly visualize and interact with them in the tool, whilst point clouds need to be converted into a lighter format to enable efficient and responsive interaction with the game rendering engine. Thus, we voxelize an input point cloud and during labeling the player will navigate in a voxelized world. Moreover, when dealing with large scale scenarios we need to optimize the run time rendering. We employ a Level-Of-Detail strategy where objects closer to the player are loaded at a higher resolution than those farther away. We split our 3D data into several chunks and we save several versions at different resolution. Depending on the distance between the player and the object we load the version at the most suitable resolution.

Virtual Reality Labeling In this step the user can explore the reconstruction immersively through Virtual Reality. The player can teleport or physically move around the scene to reach each portion of the scene. Several features are implemented to ease the user experience provided by the tool. The user can pick the label directly from a color palette and choose between different *label shooting* weapons which feature different action ranges, thereby enabling either a more precise or faster labeling. Two different visualization are implemented: play and visualization mode. In play mode (Figure 1 middle and right images) the user can shoot labels and visualize the progress by seeing only those surfaces not yet labeled (i.e. not colored). On the other hand, in visualization mode (Figure 1 left image) the tool visualizes the RGB version of the 3D data (if available) so as to provide the user with a better understanding of the semantics of the scene and thus facilitate the process.

Post-Processing of 3D Labeled Data The labeled scene can be exported as either a 3D mesh or a point cloud, depending on the input type. We also employ the Blender render engine to project the 3D labels into 2D images. Thus, if camera parameters are available, we can easily obtain labeled 2D images by positioning the camera in Blender, setting the parameters and rendering it. As we are interested in rendering only the labels even a computer graphic render engine is perfect for this purpose. In those 3D points where we do not have any information we assume a void label.

Figure 4: 2D Semantic labels of the Kitti dataset [8] obtained through re-projection technique. On the left the RGB image, on the right the semantic labels.

4. Experiments

To evaluate the efficiency and performance of our tool we tested it with both indoor and large outdoor scenarios.

Indoor Labeling We labeled a few scenes from Matterport 3D [3] and we qualitatively compared our results with their labeling. In Figure 3 we show on the left the labeling obtained by our tool and on the right that provided by Matterport 3D. We could obtain almost the same 3D labeling in only a few minutes while walking immersively within the reconstructed room.

Large Scale Outdoor Labeling We evaluated the effectiveness of our tool in a challenging outdoor scenario: the Kitti Odometry dataset [8]. We used the provided 3D Lidar data of the static sequence [1] consisting of more than 1000 images equipped with ground truth camera poses. We reconstructed the point cloud, then voxelized and labeled it by our tool. Thus, we obtained both point cloud and 2D images exploiting the re-projection technique. In Figure 2 we can see the 3D reconstructed sequence in both RGB and Labeled version. In Figure 4 we can see some qualitative examples of 2D semantic images obtained through re-projection. We were able to label the whole sequence in approximately 8 hours, a very shorter time with respect to other non-VR tool such as [2] which needed about 51 hours for each sequence.

5. Conclusions and Future Works

We have proposed the first 3D semantic labeling tool based on Virtual Reality (VR). Our tool exploits VR alongside with gamification to ameliorate and expedite semantic labeling of large scale scenarios. The tool works with the most popular 3D data structures, such as meshes and point clouds. We will release both the 3D and corresponding 2D semantic labels for a whole outdoor sequence from the Kitti dataset which accounts for more than 1000 images. We hope that our contribution will help in vastly simplifying and accelerating the tedious and time-consuming data annotation process required by state-of-the-art deep learning architectures for computer vision.

[1] `Kitti Sequence 2011_09_30_drive_0020_sync`

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. A dataset for semantic segmentation of point cloud sequences. *arXiv preprint arXiv:1904.01416*, 2019. 2, 3

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 1, 2, 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 1

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1, 2

[7] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2, 3

[9] Stuart Golodetz, Michael Sapienza, Julien PC Valentin, Vibhav Vineet, Ming-Ming Cheng, Victor A Prisacariu, Olaf Kähler, Carl Yuheng Ren, Anurag Arnab, Stephen L Hicks, et al. Semanticpaint: interactive segmentation and learning of 3d worlds. In *ACM SIGGRAPH 2015 Emerging Technologies*, page 22. ACM, 2015. 2

[10] Timo Hackel, N. Savinov, L. Ladicky, Jan D. Wegner, K. Schindler, and M. Pollefeys. SEMANTIC3D.NET: A new large-scale point cloud classification benchmark. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-1-W1, pages 91–98, 2017. 2

[11] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4077–4085, 2016. 2

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

[14] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 92–101. IEEE, 2016. 1, 2

[15] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2

[16] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *British Machine Vision Conference (BMVC)*, 2018. 2

[17] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2678–2687, 2017. 1, 2

[18] Ondrej Miksik, Vibhav Vineet, Morten Lidegaard, Ram Prasaath, Matthias Nießner, Stuart Golodetz, Stephen L Hicks, Patrick Pérez, Shahram Izadi, and Philip HS Torr. The semantic paintbrush: Interactive 3d mapping and recognition in large outdoor spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3317–3326. ACM, 2015. 2

[19] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017. 2

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[21] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 2

[22] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 2

[23] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 1