Virtual Multi-view Fusion for 3D Semantic Segmentation

Abhijit Kundu[1] **Xiaoqi Yin**[2] **Alireza Fathi**[3] **David Ross**[4]
Brian Brewington[5] **Thomas Funkhouser**[6] **Caroline Pantofaru**[7]

Google Research

Abstract

Semantic segmentation of 3D meshes is an important problem for 3D scene understanding. In this paper we revisit the classic multiview representation of 3D meshes and study several techniques that make them effective for 3D semantic segmentation of meshes. Given a 3D mesh reconstructed from RGBD sensors, our method effectively chooses different virtual views of the 3D mesh and renders multiple 2D channels for training an effective 2D semantic segmentation model. Features from multiple per view predictions are finally fused on 3D mesh vertices to predict mesh semantic segmentation labels. Using the large scale indoor 3D semantic segmentation benchmark of ScanNet, we show that our virtual views enable more effective training of 2D semantic segmentation networks than previous multiview approaches. When the 2D per pixel predictions are aggregated on 3D surfaces, our virtual multiview fusion method is able to achieve significantly better 3D semantic segmentation results compared to all prior multiview approaches and competitive with recent 3D convolution approaches.

1. Introduction

emantic segmentation of 3D scenes is a fundamental problem in computer vision. Given a 3D representation of a scene (e.g., a textured mesh of an indoor environment), the goal is to output a semantic label for every surface point. The output could be used for semantic mapping, site monitoring, training autonomous navigation, and several other applications. State-of-the-art (SOTA) methods for 3D semantic segmentation currently use 3D sparse voxel convolution operators for processing input data. For example, MinkowskiNet [7] and SparseConvNet [11] each load the input data into a sparse 3D voxel grid and extract features with sparse 3D convolutions. These "placecentric" methods are designed to recognize 3D patterns and thus work well for types of objects with distinctive 3D shapes (e.g., chairs), and not so well for others (e.g., wall pictures). They also take a considerable amount of memory, which limits spatial resolutions and/or batch sizes. Alternatively, when posed RGB-D images are available, several researchers have tried using 2D networks designed for processing photographic RGB images arXiv:2007.13138v1 [cs.CV] 26 Jul 20202 A. Kundu et al. to predict dense features and/or semantic labels and then aggregate them on visible 3D surfaces [15,41], and others project features onto visible surfaces and convolve them further in 3D [10,40,18,19]. Although these "view-centric" methods utilize massive image processing networks pretrained on large RGB image datasets, they do not achieve SOTA performance on standard 3D segmentation benchmarks due to the difficulties of occlusion, lighting variation, and camera pose misalignment in RGB-D scanning datasets. None of the view-based methods is currently in the top half of the current leaderboard for the 3D Semantic Label Challenge of the ScanNet benchmark. In this paper, we propose a new view-based approach to 3D semantic segmentation that overcomes the problems with previous methods. The key idea is to use synthetic images rendered from "virtual views" of the 3D scene rather than restricting processing to the original photographic images acquired by a physical camera. This approach has several advantages that address the key problems encountered by previous view-centric method [3,21]. First, we select camera intrinsics for virtual views with unnaturally wide field-of-view to increase the context observed in each rendered image. Second, we select virtual viewpoints at locations with small variation in distances/angles to scene surfaces, relatively few occlusions between objects, and large surface coverage redundancy. Third, we render non-photorealistic images without view-dependent lighting efffects and occlusions by backfacing surfaces – i.e., virtual views can look into a scene from behind the walls, floors, and ceilings to provide views with relatively large context and little occlusion. Fourth, we aggregate pixel-wise predictions onto 3D surfaces according to exactly known camera parameters of virtual views, and thus do not encounter "bleeding" of semantic labels across occluding contours. Fifth, virtual views during training and inference can mimic multi-scale training and testing and avoid scale in-variance issues of

2D CNNs. We can generate as many virtual views as we want during both training and testing. During training, more virtual views provides robustness due to data augmentation. During testing, more views provides robustness due to vote redundancy. Finally, the 2D segmentation model in our multiview fusion approach can benefit from large image pre-training data like ImageNet and COCO, which are unavailable for pure 3D convolution approaches. We have investigated the idea of using virtual views for semantic segmentation of 3D surfaces using a variety of ablation studies. We find that the broader design space of view selection enabled by virtual cameras can significantly boost the performance of multiview fusion as it allows us to include physically impossible but useful views (e.g., behind walls). For example, using virtual views with original camera parameters improves 3D mIoU by 3.1original photograpic images, using additional normal and coordinates channels and higher field of view can further boost mIoU by 5.7gain of 2.1capture the 3D information in the scenes and optimize for training 2D CNNs. Overall, our simple system is able to achieve state-of-the-art results on both 2D and 3D semantic labeling tasks in ScanNet Benchmark [9], and is significantlyVirtual Multiview Fusion 3 better than the best performing previous multi-view methods and very competitive with recent 3D methods based on convolutions of 3D point sets and meshes. In addition, we show that our proposed approach consistently outperforms 3D convolution and real multi-view fusion approaches when there are fewer scenes for training. Finally, we show that similar performance can be obtained with significantly fewer views in the inference stage. For example, multi-view fusion with 12 virtual views per scene will outperform that with all 1700 original views per scene. The rest of the paper is organized as follows. We introduce the research landscape and related work in §2. We describe the proposed virtual multiview fusion approach in detail in §3-§5. Experiment results and ablation studies of our proposed approach are presented in §6. Finally we conclude the paper with discussions of future directions in §7.

2. Related Work

There has been a large amount of previous work on semantic segmentation of 3D scenes. The following reviews only the most related work. Multi-view labeling. Motivated by the success of view-based methods for object classification [35], early work on semantic segmentation of RGB-D surface reconstructions relied on 2D networks trained to predict dense semantic labels for RGB images. Pixel-wise semantic labels were backprojected and aggregated onto 3D reconstructed surfaces via weighted averaging [15,41], CRFs [25], Bayesian fusion [24,41,46], or 3D convolutions [10,18,19]. These methods performed multiview aggregation only for the originally captured RGB-D photographic images, which suffer from limited fields-of-view, restricted viewpoint ranges, view-dependent lighting effects, and misalignments with reconstructed surface geometry, all of which reduce semantic segmentation performance. To overcome these problems, some recent work has proposed using synthetic images of real data in a multiview labeling pipeline [3,21,12], but they still use camera parameters typical of real images (e.g., small field of view), propose methods suitable only for outdoor environments (lidar point clouds of cities), and do not currently achieve state-of-the-art results. 3D convolution. Recent work on 3D semantic segmentation has focused on methods that extract and classify features directly with 3D convolutions. Network architectures have been proposed to extract features from 3D point clouds [29,30,31,33,38,16], surface meshes [14,17], voxel grids [34], and octrees [32]. Current state-of-the-art methods are based on sparse 3D voxel convolutions [7,8,11],

where submanifold sparse convolution operations are used to compute features on sparse voxel grids. These methods utilize memory more efficiently than dense voxel grids, but are still limited in spatial resolution in comparison to 2D images4 A. Kundu et al. and can train with supervision only on 3D datasets, which generally are very small in comparison to 2D image datasets. Synthetic data. Other work has investigated training 2D semantic segmentation networks using computer graphics renderings of 3D synthetic data [47]. The main advantage of this approach is that image datasets can be created with unlimited size by rendering novel views of a 3D scene [22,26]. However, the challenge is generally domain adaptation – networks trained on synthetic data and

tested on real data usually do not perform well. Our method avoids this problem by training and testing on synthetic images rendered with the same process.

The guidelines below will be enforced for initial submissions and camera-ready copies. Figure 1 is a brief summary.

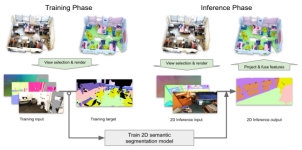

Figure 1. Virtual multi-view fusion system overview.

3. Method Overview

The proposed multiview fusion approach is illustrated in Figure 1. At a high level, it consists of the following steps. Training stage. During the training stage, we first select virtual views for each 3D scene, where for each virtual view we select camera intrinsics, camera extrinsics, which channels to render, and rendering parameters (e.g., depth

range, backface culling). We then generate training data by rendering the selected virtual views for the selected channels and ground truth semantic labels. We train 2D semantic segmentation models using the rendered training data and use the model in the inference stage. Inference stage. At inference stage, we select and render virtual views using a similar approach as in the training stage, but without the ground truth semantic labels. We conduct 2D semantic segmentation on the rendered virtual views using the trained model, project the 2D semantic features to 3D, then derive the semantic category in 3D by fusing multiple projected 2D semantic features.

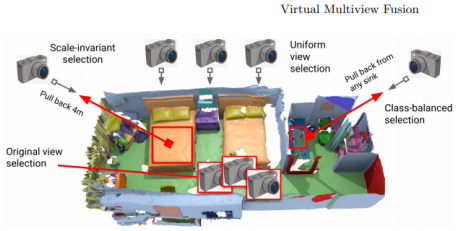

Figure 2. Virtual multi-view fusion system overview.

4. Virtual view selection

Virtual view selection is central to the proposed multiview fusion approach as it brings key advantages over multiview fusion with original image views. First, it allows us to freely select camera parameters that work best for 2D semantic segmentation tasks, and with any set of 2D data augmentation approaches. Second, it significantly broadens the set of views to choose from by relaxing the physical constraints of real cameras and allowing views from unrealistic but useful camera positions that significantly boost model performance, e.g. behind a wall. Third, it allows 2D views to capture additional channels that are difficult to capture with real cameras, e.g., normals and coordinates. Finally, by selecting and rendering virtual views, we have essentially eliminated any errors in the camera calibration and pose estimation, which are common in the 3D reconstruction process. Lastly, sampling views consistently at different scales resolves scale in-variance issues of traditional 2D CNNs.