Diffuser: Multi-View 2D-to-3D Label Diffusion for Semantic Scene Segmentation

Author(s):
Mascaro, Ruben; Teixeira, Lucas; Chli, Margarita

Publication date:
2021

Permanent link:
https://doi.org/10.3929/ethz-b-000484229

Originally published in:
https://doi.org/10.1109/ICRA48506.2021.9561801

Diffuser: Multi-View 2D-to-3D Label Diffusion for Semantic Scene Segmentation

Ruben Mascaro, Lucas Teixeira and Margarita Chli
Vision For Robotics Lab, ETH Zurich, Switzerland

***Abstract*— Semantic 3D scene understanding is a fundamental problem in computer vision and robotics. Despite recent advances in deep learning, its application to multi-domain 3D semantic segmentation typically suffers from the lack of extensive enough annotated 3D datasets. On the contrary, 2D neural networks benefit from existing large amounts of training data and can be applied to a wider variety of environments, sometimes even without need for retraining. In this paper, we present 'Diffuser', a novel and efficient multi-view fusion framework that leverages 2D semantic segmentation of multiple image views of a scene to produce a consistent and refined 3D segmentation. We formulate the 3D segmentation task as a transductive label diffusion problem on a graph, where multi-view and 3D geometric properties are used to propagate semantic labels from the 2D image space to the 3D map. Experiments conducted on indoor and outdoor challenging datasets demonstrate the versatility of our approach, as well as its effectiveness for both global 3D scene labeling and single RGB-D frame segmentation. Furthermore, we show a significant increase in 3D segmentation accuracy compared to probabilistic fusion methods employed in several state-of-the-art multi-view approaches, with little computational overhead.**

I. Introduction

Image-based semantic segmentation has long been studied in computer vision, and its extension to scene segmentation in 3D has become particularly relevant in robotics. By augmenting 3D scene maps with semantic information, this task can support a wide variety of applications requiring somewhat high-level reasoning, such as obstacle avoidance and mission planning for autonomous robotic navigation or physical interaction of a robot with its workspace.

Current state-of-the-art methods in 3D semantic segmentation mostly rely on end-to-end trainable neural networks that use 3D convolution operators for extracting features directly from 3D data [3], [4], [5]. These networks achieve unprecedented performance in scenes containing elements with distinct shapes, such as human-made objects, but their generalisation to a wide variety of environments becomes hindered by the lack of sufficiently large, labeled 3D datasets, which are usually hard to produce.

As an alternative, when images from calibrated cameras are available, per-pixel semantic labels can be extracted in the image space from multiple viewpoints and aggregated on visible 3D surfaces by exploiting the camera projection principles. Methods based on this multi-view fusion approach [6], [7], [8] benefit from image processing networks that can be trained on readily available, massive datasets, facilitating their adaptation to new environments and making them more suitable for robotic applications. However, label fusion schemes typically suffer from difficulties caused by occlusion and imprecise 2D segmentation boundaries, resulting in a downgraded 3D segmentation quality.

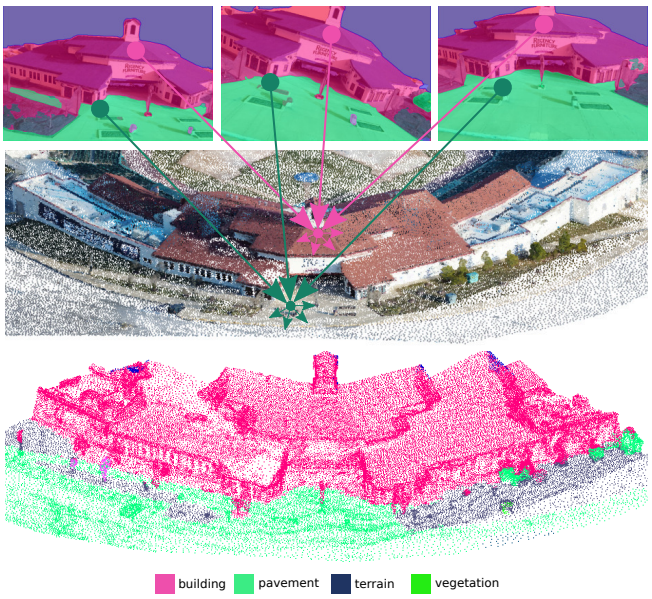

Fig. 1: Taking as input a 3D point cloud and a set of localized images processed by a 2D semantic segmentation network, 'Diffuser' uses an efficient graphical model that leverages geometry to propagate class labels from the 2D image space to the 3D map. The bottom image shows qualitative 3D segmentation results obtained in one sequence of the Aerial Depth Dataset [1]. A subset of the input images segmented by the MSeg [2] framework and the input 3D point cloud are shown in the top and middle rows, respectively. The pink and green arrows represent the proposed graphical model to transfer semantic labels from 2D pixels to 3D points.

In this paper, we address these issues by proposing a novel and efficient method for fusing multi-view 2D semantic predictions into a consistent and refined segmentation in 3D. We formulate the 3D segmentation task as a transductive label diffusion problem on a graph, where geometric context is leveraged to propagate class labels from 2D pixels to 3D points. Our method can be used locally, e.g. for single RGB-D frame segmentation, and for global 3D scene labeling. Furthermore, it can be integrated with any 2D semantic segmentation framework, it does not require 3D training data and it generalizes well to different types of environments.

This work was supported by the Swiss National Science Foundation (SNSF, Agreement no. PP00P2183720), the National Centre of Competence in Research Digital Fabrication (NCCR DFAB), and the Amazon Research Awards 2019 program.

The video is available at https://youtu.be/WWqaFLgK5Kk.

II. Related Work

Deep Learning in 3D. The success of deep neural networks in solving several scene understanding tasks on images has inspired their application in the context of 3D semantic segmentation. In recent years, various network architectures have been proposed in order to extract semantic features directly from 3D point clouds [9], [10], [11], [12], [13], [14], [15]. Current state-of-the-art methods process input data with sparse convolutions [3], [4], [5], which utilize memory more efficiently than approaches based on dense voxel grids. However, these methods are still limited in spatial resolution and can be trained with supervision only on 3D datasets, which are generally too small and more difficult to produce in comparison to 2D image datasets. Moreover, due to the morphological gap and variation in point density between 3D representations of different scenarios, these algorithms may generalize poorly to unseen environments.

Multi-view Approaches. Instead of directly processing 3D data, other approaches have focused on running semantic segmentation on images captured from known poses and making use of multi-view geometric relationships to project the predicted semantic labels onto the 3D space. Early works on online semantic mapping typically aggregated pixel-wise semantic features onto 3D reconstructed surfaces via Bayesian fusion [6], [16] or weighted averaging [17], and used computationally expensive Conditional Random Field (CRF) models to regularise the resulting 3D segmentation. More recently, offline end-to-end solutions that extract features from 2D images and convolve them further in 3D have been proposed [18], [19]. Although these learning-based methods generally achieve better results, they require labeled 3D data for training, which is a limiting factor in certain types of environments. To overcome problems originating from using RGB or RGB-D images, such as limited field of view and misalignment with reconstructed surface geometry, or in cases where aligned images are not available, other recent approaches have explored using synthetic images of real 3D data in multi-view labeling pipelines [20], [21], [22], [23]. These methods are able to sample good viewpoints, use artificial camera parameters and render multiple channels, generally improving segmentation in 2D, but still use simple probabilistic fusion methods that do not leverage geometric context to deal with inconsistent label predictions.

In contrast to these approaches, here we formulate an efficient label propagation scheme that simultaneously handles multi-view label prediction inconsistencies and exploits the geometry of the scene to better refine segmentation boundaries in 3D. Our approach does not require 3D training data, it is not restricted to any particular type of scene and could be used with any of the aforementioned frameworks to improve the multi-view fusion step.

Graphical Models for 2D-3D Fusion. Graph-based methods are well established in machine learning and have successfully been applied to point-cloud segmentation from aligned images in the past. For example, Koppula *et al.* [24] created nodes in a graph from segments of a complete 3D map and used hand-crafted geometric and visual features as edge potentials to infer the final semantic labeling. Wang *et al.* [25] proposed a semantic segmentation approach for image-aligned 3D point clouds that retrieves referenced labeled images of similar appearances and uses a graphical structure to propagate their labels to the 3D points. More recently, dense CRF models formed by unary and pairwise terms [6], sometimes enhanced with higher-order potentials [7], [8], have been applied as a post-processing step to address noise in online semantic scene segmentation pipelines.

Our approach is mostly inspired by [26], which introduces a graph-based label propagation method for instance segmentation of 3D LiDAR scans given a set of detection masks predicted with a 2D convolutional neural network on aligned RGB images. In this work, however, we propose an extended graphical model that is able to seamlessly fuse pixel-wise semantic labels predicted from multiple views and better exploits the observed surface geometry to regularise the resulting 3D segmentation.

III. Method

In this section, we introduce our 2D-to-3D label diffusion algorithm for semantic segmentation of 3D point clouds using multiple views. Approaching the task from a semi-supervised labeling perspective, we formulate a novel graph structure that leverages the output of a 2D semantic segmentation framework and uses geometric context to propagate class labels through the point cloud, resulting in a consistent and refined 3D semantic map.

A. Problem Statement and Notation

Our approach to semantic 3D segmentation through label diffusion from multiple views takes as input a 3D point cloud and a set of N_f localized images. We assume that the set of input images, $\mathcal{I} = \{\mathcal{I}_k\}_{k=1}^{N_f}$, has already been processed by a 2D semantic segmentation framework which, for each pixel coordinate $\boldsymbol{p}_j^{\mathcal{I}_k} \in \mathbb{N}^2$, $j \in \{1, \dots, N_p^{\mathcal{I}_k}\}$, predicts a class label $y_j^{\mathcal{I}_k} \in \{1, \dots, C\}$. The number of pixels in each image, $N_p^{\mathcal{I}_k}$, as well as the number of classes that the network is able to predict, C, are arbitrary. Let the point cloud be denoted as $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^3\}_{i=1}^{N_x}$, with N_x being the number of 3D points. The goal of the algorithm is to assign each 3D point $\boldsymbol{x}_i$ a class label y_i.

B. Graphical Model

Similarly to [26], our graphical model for 2D-to-3D label diffusion, $\boldsymbol{G}$, is formed by nodes that represent both 2D pixels and 3D points, although we specifically design it to handle multiple views. Assuming that the 2D pixels are labeled by means of any image semantic segmentation framework, this graph is then used to propagate class labels through the 3D points, which are all initially unlabeled. To guide the label diffusion process, edges between 2D pixels and 3D points as well as among 3D points are generated as follows:

1) Pixel-to-Point Edges: To allow the flow of information from 2D to 3D in our graph, for each frame $\mathcal{I}_k$ we construct a subgraph $\boldsymbol{G}^{\mathcal{I}_k \to \mathcal{X}}$ that can be represented as a $N_x \times N_p^{\mathcal{I}_k}$ adjacency matrix of the form:

$$\boldsymbol{G}_{ij}^{\mathcal{I}_k \to \mathcal{X}} = \begin{cases} \lambda & \text{if } \boldsymbol{p}_j^{\mathcal{I}_k} = \pi^{\mathcal{I}_k}(\boldsymbol{x}_i) \\ 0 & \text{otherwise} \end{cases} , \quad (1)$$

where λ is a hyperparameter that controls the amount of information being propagated from a pixel to a connected point, and $\pi^{\mathcal{I}_k}(\boldsymbol{x}_i)$ denotes the projected 2D location of 3D point $\boldsymbol{x}_i$ in frame $\mathcal{I}_k$, which is computed as follows:

$$\pi^{\mathcal{I}_k}(\boldsymbol{x}_i) = \boldsymbol{K}^{\mathcal{I}_k}\left(\boldsymbol{R}^{\mathcal{I}_k}\boldsymbol{x}_i + \boldsymbol{t}^{\mathcal{I}_k}\right) . \quad (2)$$

Here, $\boldsymbol{K}^{\mathcal{I}_k}$ is the intrinsics matrix, while $\boldsymbol{R}^{\mathcal{I}_k}$ and $\boldsymbol{t}^{\mathcal{I}_k}$ represent the rotation and translation in the known extrinsic parameters corresponding to image $\mathcal{I}_k$. To deal with occlusion in point clouds, we assume that a mapping from pixel coordinates to depth is available (e.g. provided by an RGB-D sensor or a 3D reconstruction pipeline) and perform a depth consistency check to keep only the points that are visible in each frame.

In our implementation, we set $\lambda = 10^{-4}$ to reduce the influence of any pixel on the final 3D labeling. Although we experimented with connecting each 3D point to multiple pixels within a neighboring region of its projected 2D location, as in [26], we empirically found that simply considering the projected pixel location leads to similar results while being computationally more efficient.

2) Point-to-Point Edges: Edges among 3D points are created by connecting each point $\boldsymbol{x}_i$ to its K nearest neighbors, $KNN(\boldsymbol{x}_i)$, according to Euclidean distance. The K nearest neighbor search is efficiently performed by using a KD-tree. The subgraph of 3D point connections, which we denote as $\boldsymbol{G}^{\mathcal{X} \to \mathcal{X}}$, is then defined as a $N_x \times N_x$ adjacency matrix of the form:

$$\boldsymbol{G}_{ii'}^{\mathcal{X} \to \mathcal{X}} = \begin{cases} \omega_{ii'} & \text{if } \boldsymbol{x}_{i'} \in KNN(\boldsymbol{x}_i) \\ 0 & \text{otherwise} \end{cases} , \quad (3)$$

$$\omega_{ii'} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|_2^2}{2\sigma_d^2} - \frac{\|\boldsymbol{n}_i - \boldsymbol{n}_{i'}\|_2^2}{2\sigma_s^2}\right) . \quad (4)$$

In our formulation, $\boldsymbol{n}_i$ represents the normal vector at point $\boldsymbol{x}_i$, computed based on a local 3D point neighborhood, while σ_d and σ_s are bandwidth hyperparameters for the Gaussian edge potential $\omega_{ii'}$. Specifically, we set $K = 24$, $\sigma_s = 0.2$ and $\sigma_d = 2\rho$, where ρ is an estimate of the mean point cloud density.

By connecting each point to the set of neighboring points in space and weighting the resulting edges based both on the distance between them and the difference between its corresponding normals, the subgraph $\boldsymbol{G}^{\mathcal{X} \to \mathcal{X}}$ effectively encodes the 3D geometry of the scene point cloud . Contrary to [26], where edges between points are simply weighted based on Euclidean distance, our formulation also takes into account the smoothness of the underlying surface to propagate information through the point cloud. The reasoning behind our approach is that labels should be more easily propagated between neighboring points if the underlying surface is smooth, i.e. if the difference between their corresponding normals is small. This allows us to better respect 3D boundaries in the graph, as shown in the experiments.

3) Label Diffusion Graph: The complete graph for label diffusion, including both the pixel-to-point and the point-to-point edges, is then defined as:

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}^{\mathcal{X} \to \mathcal{X}} & \boldsymbol{G}^{\mathcal{I}_1 \to \mathcal{X}} & \cdots & \boldsymbol{G}^{\mathcal{I}_{N_f} \to \mathcal{X}} \\ \boldsymbol{0} & \boldsymbol{I}^{\mathcal{I}_1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I}^{\mathcal{I}_{N_f}} \end{bmatrix} , \quad (5)$$

where each $\boldsymbol{I}^{\mathcal{I}_k}$ is an identity matrix of size $N_p^{\mathcal{I}_k} \times N_p^{\mathcal{I}_k}$. It is worth noting that, in contrast to approaches based on fully-connected CRFs [6], [7], our method builds upon a sparse graphical model, allowing for efficient computation during label propagation, and unifies the label fusion and map regularisation steps.

C. Graph-based Label Diffusion Principles

After constructing $\boldsymbol{G}$, a probabilistic transition matrix $\boldsymbol{P}$ can be computed by normalizing each row of the adjacency matrix to sum up to 1:

$$\boldsymbol{P}_{mn} = \frac{\boldsymbol{G}_{mn}}{\sum_{n'} \boldsymbol{G}_{mn'}}, \quad \forall m, n \in \{1, \ldots, N\} , \quad (6)$$

with $N = N_x + \sum_k N_p^{\mathcal{I}_k}$ being the total number of nodes in the label difusion graph. The element $\boldsymbol{P}_{mn}$ of matrix $\boldsymbol{P}$ can be interpreted as the probability of transition from node m to n. By indicating edges, along which information on class labels should be propagated, this matrix $\boldsymbol{P}$ will guide the diffusion process for 3D point labeling later on.

Assuming that a known number of classes C is predicted by the 2D semantic segmentation framework, we additionally define a label matrix $\boldsymbol{Z} \in \mathbb{R}^{N \times C}$ as:

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}^{\mathcal{X}} & \boldsymbol{Z}^{\mathcal{I}_1} & \cdots & \boldsymbol{Z}^{\mathcal{I}_{N_f}} \end{bmatrix}^T . \quad (7)$$

For each node in graph $\boldsymbol{G}$, $\boldsymbol{Z}$ contains C entries that will accumulate the likelihood of this node belonging to each of the candidate classes. In our case, entries corresponding to 3D points, $\boldsymbol{Z}^{\mathcal{X}}$, are all initialized to zero, while entries corresponding to 2D pixels, $\boldsymbol{Z}^{\mathcal{I}_k}$, are defined according to the 2D segmentation masks:

$$\boldsymbol{Z}_{jc}^{\mathcal{I}_k} = \delta^{\mathcal{I}_k}(j, c) , \quad (8)$$

with $\delta^{\mathcal{I}_k}(j, c)$ being a function that returns 1 if pixel $\boldsymbol{p}_j^{\mathcal{I}_k}$ is in the segmentation mask of class c and 0 otherwise.

Label diffusion can then be applied by iteratively performing the following matrix multiplication:

$$\boldsymbol{Z} \leftarrow \boldsymbol{P} \cdot \boldsymbol{Z} . \quad (9)$$

According to [27], since matrix $\boldsymbol{P}$ is row-normalized and ensures the labels of the source nodes (i.e. pixels in our case) to remain unchanged by the propagation step, the algorithm is guaranteed to converge.

D. Dimensionality Reduction

The formulation presented above leads to a probabilistic transition matrix $\boldsymbol{P}$ and a label matrix $\boldsymbol{Z}$ that grow rapidly with an increasing number of frames. However, since entries associated with 2D pixels in matrix $\boldsymbol{Z}$ remain fixed during label diffusion and we are solely interested in computing $\boldsymbol{Z}^{\mathcal{X}}$, i.e. the labels of the 3D points, the propagation step can be reduced to:

$$\boldsymbol{Z}^{\mathcal{X}} \leftarrow \boldsymbol{P}^{\mathcal{X}\to\mathcal{X}}\boldsymbol{Z}^{\mathcal{X}} + \boldsymbol{Z}^{\mathcal{I}\to\mathcal{X}} \ , \tag{10}$$

where $\boldsymbol{Z}^{\mathcal{I}\to\mathcal{X}}$ is a $N_x \times C$ constant matrix that encodes the amount of information being pushed to each 3D point from all corresponding 2D image pixels:

$$\boldsymbol{Z}^{\mathcal{I}\to\mathcal{X}} = \sum_k \boldsymbol{P}^{\mathcal{I}_k\to\mathcal{X}}\boldsymbol{Z}^{\mathcal{I}_k} \ . \tag{11}$$

By solving the problem this way, we eliminate the need for storing the full label matrix $\boldsymbol{Z}$, therefore reducing the memory footprint of the algorithm, and make the complexity of the diffusion process mainly depend on the size of the point cloud, not on the number of images used as sources to propagate the semantic labels.

In addition, the structure of the proposed graphical model allows building the matrix $\boldsymbol{Z}^{\mathcal{I}\to\mathcal{X}}$ in an incremental and parallel fashion. For each frame, we can compute a matrix $\boldsymbol{Z}^{\mathcal{I}_k\to\mathcal{X}}$ by first creating the subgraph of pixel-to-point connections and multiplying its adjacency matrix $\boldsymbol{G}^{\mathcal{I}_k\to\mathcal{X}}$ by the label matrix $\boldsymbol{Z}^{\mathcal{I}_k}$ generated from the predicted 2D segmentation masks:

$$\boldsymbol{Z}^{\mathcal{I}_k\to\mathcal{X}} = \boldsymbol{G}^{\mathcal{I}_k\to\mathcal{X}}\boldsymbol{Z}^{\mathcal{I}_k} \ . \tag{12}$$

Since, by definition, $\boldsymbol{Z}^{\mathcal{I}_k}$ is a matrix, where all elements in each row are set to zero except from one, whose value is set to 1, it can be demonstrated that:

$$\sum_j \boldsymbol{G}_{ij}^{\mathcal{I}_k\to\mathcal{X}} = \sum_c \left[\boldsymbol{G}^{\mathcal{I}_k\to\mathcal{X}}\boldsymbol{Z}^{\mathcal{I}_k}\right]_{ic} = \sum_c \boldsymbol{Z}_{ic}^{\mathcal{I}_k\to\mathcal{X}} \ . \tag{13}$$

Therefore, to ensure that the full probabilistic transition matrix of the underlying graph is row-normalized, $\boldsymbol{P}^{\mathcal{X}\to\mathcal{X}}$ and $\boldsymbol{Z}^{\mathcal{I}\to\mathcal{X}}$ in Eq. (10) can be computed as follows:

$$\boldsymbol{P}_{ii'}^{\mathcal{X}\to\mathcal{X}} = \frac{\boldsymbol{G}_{ii'}^{\mathcal{X}\to\mathcal{X}}}{\sum_{i''}\boldsymbol{G}_{ii''}^{\mathcal{X}\to\mathcal{X}} + \sum_c\sum_k \boldsymbol{Z}_{ic}^{\mathcal{I}_k\to\mathcal{X}}} \ , \tag{14}$$

$$\boldsymbol{Z}_{ic}^{\mathcal{I}\to\mathcal{X}} = \frac{\sum_k \boldsymbol{Z}_{ic}^{\mathcal{I}_k\to\mathcal{X}}}{\sum_{i'}\boldsymbol{G}_{ii'}^{\mathcal{X}\to\mathcal{X}} + \sum_{c'}\sum_k \boldsymbol{Z}_{ic'}^{\mathcal{I}_k\to\mathcal{X}}} \ . \tag{15}$$

After the matrices $\boldsymbol{P}^{\mathcal{X}\to\mathcal{X}}$ and $\boldsymbol{Z}^{\mathcal{I}\to\mathcal{X}}$ are constructed, label diffusion is iteratively applied according to Eq. (10) until convergence of $\boldsymbol{Z}^{\mathcal{X}}$ or until a maximum number of iterations is reached. Finally, the likelihood values are converted to 3D point labels according to:

$$y_i = \underset{c\in\{1,\dots,C\}}{\operatorname{argmax}} \boldsymbol{Z}_{ic}^{\mathcal{X}} \ , \tag{16}$$

meaning that each point gets assigned the most likely label. The label diffusion algorithm proposed in this paper is outlined in Algorithm 1.

Algorithm 1 Diffuser

Require: A 3D point cloud and a set of aligned images processed by a 2D semantic segmentation framework.

1: **for each** frame $\mathcal{I}_k$ **do**
2: Project 3D points and compute $\boldsymbol{G}^{\mathcal{I}_k\to\mathcal{X}}$ (Eq. 1).
3: Define matrix $\boldsymbol{Z}^{\mathcal{I}_k}$ from 2D segmentation (Eq. 8).
4: Compute $\boldsymbol{Z}^{\mathcal{I}_k\to\mathcal{X}} = \boldsymbol{G}^{\mathcal{I}_k\to\mathcal{X}}\boldsymbol{Z}^{\mathcal{I}_k}$.
5: Update $\sum_k \boldsymbol{Z}^{\mathcal{I}_k\to\mathcal{X}}$.
6: **end for**
7: Construct the subgraph matrix $\boldsymbol{G}^{\mathcal{X}\to\mathcal{X}}$ (Eq. 3).
8: Define label matrix $\boldsymbol{Z}^{\mathcal{X}}$ and initialize all entries to zero.
9: Compute matrices $\boldsymbol{P}^{\mathcal{X}\to\mathcal{X}}$ and $\boldsymbol{Z}^{\mathcal{I}\to\mathcal{X}}$ (Eqs. 14, 15).
10: Perform label diffusion until convergence or max. number of iterations is reached (Eq. 10).
11: Determine class labels for all 3D points (Eq. 16).

IV. Experiments

Aiming at showing that our approach generalizes well to various types of environments, we run a series of experiments on three completely different datasets for 3D semantic segmentation containing image sequences: ScanNet [28], a RGB-D video dataset established as one of the standard 3D segmentation benchmarks; the 3DRMS Challenge Dataset [29], a challenging outdoor dataset, featuring image sequences captured from virtual and real drives through semantically-rich garden scenes; and the Aerial Depth Dataset [1], a photorealistic aerial dataset that exhibits more challenging scenarios than the established indoor and car driving datasets. In all of them, ground truth 2D semantic annotations and camera poses, as well as semantically annotated 3D point clouds depicting the area of each sequence are provided for evaluation.

We also experiment with different network architectures and pre-trained models to run semantic segmentation on the input images. For both the indoor and outdoor datasets, we use a network based on the HRNet-W48 [30] architecture and trained on MSeg [2], a composite, multi-domain semantic segmentation dataset designed for training models, in order to achieve consistently high performance across domains and to generalize well to previously unseen datasets. In our experiments on the ScanNet dataset, we additionally use a variant of this model specifically trained on the ScanNet train split. We refer to this model as 'Oracle', as coined in [2]. In the outdoor experiments, besides the MSeg model, we use a PSPNet [31] backbone trained on Cityscapes [32].

For evaluation purposes, the classes predicted by the aforementioned models are mapped to the specific classes considered by each of our test datasets. Semantic segmentation accuracy is evaluated in terms of mean Intersection-over-Union (mIoU) over the considered classes, which is a common metric in semantic segmentation benchmarks [28].

A. Evaluation on the ScanNet Dataset

We use a subset of 25 scans in the validation split of the ScanNet dataset to test our multi-view fusion method

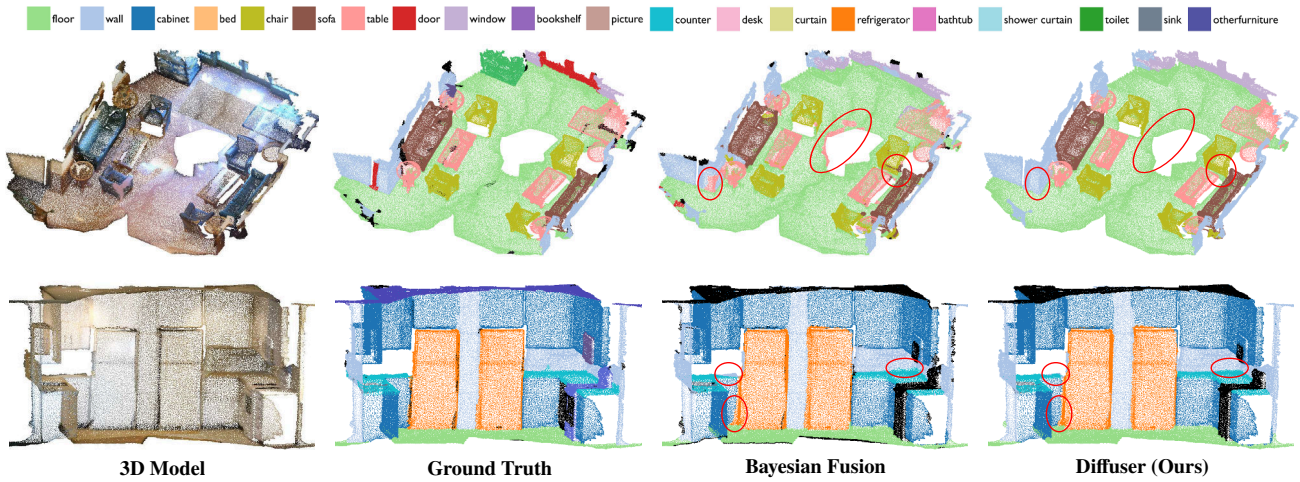

Fig. 2: Qualitative 3D point-cloud semantic segmentation results on the ScanNet dataset [28]. By exploiting 3D geometry to propagate labels, our approach is able to better handle semantic prediction inconsistencies and refine segmentation boundaries with respect to the Bayesian fusion approach used in state-of-the-art multi-view methods [21], [22]. Main differences are highlighted within red circles.

and compare it against alternative approaches for transferring 2D semantic labels to aligned 3D point-clouds. To show the flexibility of our approach, we evaluate it for both the task of global 3D point-cloud and single RGB-D frame semantic segmentation.

1) Global 3D point-cloud semantic segmentation: For each scan, we extract every 10th frame from the full RGB-D sequence and run semantic segmentation on the corresponding RGB images. We then use our multi-view fusion approach to propagate the predicted labels from all processed images to the vertices of the full 3D scene mesh. To increase the efficiency of the label propagation step, the scene point clouds are initially downsampled using a voxel grid of 2 cm. After convergence of the algorithm, each 3D point in the original point cloud is assigned the label of the nearest point in the downsampled point cloud.

As the focus of this work is on the multi-view fusion algorithm, independently of the 2D semantic segmentation strategy being used, we compare our approach against the Bayesian fusion scheme employed by several and recent multi-view methods [6], [8], [21], [22], [33], which consists in directly aggregating the semantic labeling likelihoods of all the corresponding pixels at each 3D point and picking the label with the highest score. As visible in the results in Table I, with both the MSeg and the Oracle models, the proposed Diffuser algorithm significantly improves the final 3D segmentation accuracy. Qualitative results in Figure 2 show that, by exploiting the 3D geometry of the point cloud, our approach better handles noisy semantic predictions and is able to refine 3D segmentation boundaries, thus removing the need for further, potentially more expensive regularisation strategies such as CRFs.

2) Single RGB-D frame semantic segmentation: These experiments are designed to show that our method can also be used to improve the segmentation of single RGB-D frames in online applications by taking into account semantic labels predicted at neighboring views. In this case, using the same subsampled sequences as in the global segmentation experiments, we back-project each depth image in 3D and then apply our multi-view fusion approach to label the resulting point cloud using the 2D semantic predictions from a subset of 5 neighboring frames. For evaluation purposes, the final 3D labels are projected back to the original frame and compared against the ground truth 2D labels.

We compare our multi-view fusion approach against single-view and multi-view 2D-to-3D label transferring alternatives. As shown in Table II, for the single-view case we consider direct projection labeling, where each 3D point is naively labeled based on the class assigned to its corresponding pixel, the recent LDLS [26] graph-based label

TABLE I: Global 3D point-cloud semantic segmentation results on 25 scenes of the ScanNet [28] validation split. The accuracy of the semantic segmentation models used in our experiments for predicting labels in 2D is also measured, as it greatly influences the performance of the multi-view fusion algorithms.

Model + Fusion Method	3D mIoU	2D mIoU
MSeg (HRNet-W48)	–	35.3
Oracle (HRNet-W48)	–	46.1
MSeg + Bayesian [21], [22]	45.8	–
MSeg + Diffuser (Ours)	**53.8**	–
Oracle + Bayesian [21], [22]	51.2	–
Oracle + Diffuser (Ours)	**61.2**	–

TABLE II: Single RGB-D frame semantic segmentation results on 25 sequences of the ScanNet [28] validation split. We compare our multi-view approach to both single-view and multi-view alternatives for transferring labels from 2D to 3D.

Method	Num. views	2D mIoU
Oracle + Direct labeling	1	53.7
Oracle + LDLS [26]	1	56.2
Oracle + Bayesian [21], [22]	5	57.0
Oracle + Diffuser (Ours)	1	58.4
Oracle + Diffuser (Ours)	5	**59.8**

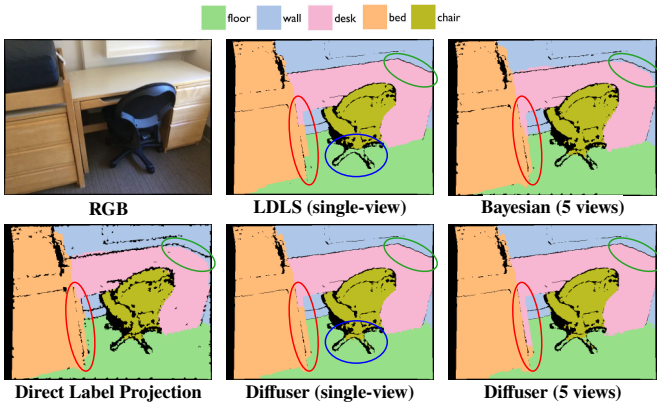

Fig. 3: Qualitative single-RGB-D-frame semantic segmentation results on ScanNet [28]. Using label propagation algorithms, both LDLS [26] and our approach are able to refine segmentation boundaries when compared to direct projection strategies (green circles). However, by considering surface normals, our approach better guides the diffusion process and achieves a higher level of detail (see the lower part of the chair within the blue circles). In addition, by using information from neighboring views, our approach is able to correct for potential single-frame prediction errors, such as the left side of the desk being confused with the bed in this case (red circles).

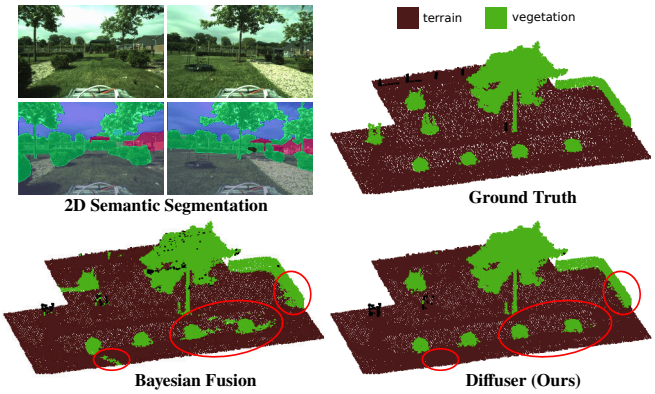

Fig. 4: Qualitative segmentation results on the real-world validation sequence of the 3DRMS Challenge dataset [29]. For evaluation purposes, the classes considered by the original dataset are associated with the MSeg universal classes [2] as follows: *grass* and *ground* are mapped to *terrain*, while *hedge*, *topiary*, *rose* and *tree* are mapped to *vegetation*. Main differences are highlighted within red circles.

propagation approach and the proposed Diffuser algorithm applied to the single RGB-D frame. For the multi-view case, where semantic labels predicted at 5 neighboring frames are fused, we also provide the results obtained with the Bayesian fusion approach.

Among the single-view alternatives, we observe that our method is able to improve segmentation accuracy when compared to the LDLS graphical model, which does not consider surface smoothness to propagate labels. In addition, when using information from multiple views along with geometric context, our framework is able to better correct for single-frame prediction errors, showing consistent improvement with respect to both single-view approaches and the multi-view Bayesian fusion strategy, which does not take 3D geometry into account (see Figure 3).

It is worth mentioning that, in our current GPU-accelerated Python implementation, building the graph of point-to-point edges and performing 100 label diffusion iterations on a point cloud containing 30,000 points takes 0.3 seconds on average on an Nvidia Quadro P2000, thus rendering our approach suitable for online applications.

B. Evaluation on Outdoor Datasets

To evaluate the performance of our approach on other challenging, but completely different environments, we also run experiments on the real-world validation sequence of the 2018 3DRMS Challenge [29] as well as on one sequence extracted from the Aerial Depth Dataset [1]. Here, given a set of RGB images with a known camera pose, the goal is to label the provided 3D point cloud of the scene.

Results of these experiments are summarized in Table III and shown in Figures 4 and 1. Again, the proposed Diffuser exhibits significant improvement with respect to the Bayesian fusion approach by providing a more accurate

TABLE III: Global 3D point-cloud semantic segmentation results on challenging robotic outdoor datasets.

Model + Fusion Method	**3DRMS** [29] **3D mIoU**	**Aerial** [1] **3D mIoU**
Cityscapes (PSPNet) + Bayesian [21], [22]	–	30.1
Cityscapes (PSPNet) + Diffuser (Ours)	–	**34.3**
MSeg (HRNet-W48) + Bayesian [21], [22]	88.4	59.6
MSeg (HRNet-W48) + Diffuser (Ours)	**94.5**	**67.8**

3D segmentation, independently of the 2D segmentation architecture being used. It is worth noting that state-of-the-art end-to-end frameworks for 3D segmentation might be hard to apply in such scenarios, where the lack of enough annotated 3D data becomes a limiting factor. By coupling a multi-domain 2D semantic segmentation network with our multi-view fusion scheme, we provide a flexible solution that is able to achieve reliable performance in various types of environments and eliminates the need for retraining.

V. Conclusion

In this paper, we presented Diffuser, a novel and efficient approach to semantic 3D segmentation from multiple views. Given a set of images processed by any 2D semantic segmentation framework, the multi-view fusion problem is solved via a graph-based label diffusion scheme that exploits geometric context in order to propagate predicted class labels from the 2D image space to the 3D scene map. The proposed method is not restricted to any particular semantic classes or types of environments and does not require 3D training data. Therefore, it can benefit from multi-domain 2D segmentation networks trained on massive datasets for wide applicability.

Evaluations on challenging benchmark datasets demonstrate that the proposed approach is effective both locally and globally, achieving superior accuracy compared to widely used probabilistic fusion strategies. In addition, the incremental nature of the algorithm and its computational efficiency make it especially suitable for online applications running on computationally restricted platforms.

REFERENCES

[1] L. Teixeira, M. R. Oswald, M. Pollefeys, and M. Chli, "Aerial single-view depth completion with image-guided uncertainty estimation," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, 2020.

[2] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, "MSeg: A composite dataset for multi-domain semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[3] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M. H. Yang, and J. Kautz, "SPLATNet: Sparse lattice networks for point cloud processing," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] B. Graham, M. Engelcke, and L. Van Der Maaten, "3D semantic segmentation with submanifold sparse convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal ConvNets: Minkowski convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[7] Q. H. Pham, B. S. Hua, D. T. Nguyen, and S. K. Yeung, "Real-time progressive 3D semantic segmentation for indoor scenes," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[8] C. Zhang, Z. Liu, G. Liu, and D. Huang, "Large-scale 3D semantic mapping using monocular vision," in *IEEE International Conference on Image, Vision and Computing (ICIVC)*, 2019.

[9] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[11] Q. H. Pham, T. Nguyen, B.-S. Hua, G. Roig, and S.-K. Yeung, "JSIS3D: Joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] L. Zhao and W. Tao, "JSNet: Joint instance and semantic segmentation of 3D point clouds," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[14] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Z. Hu, M. Zhen, X. Bai, H. Fu, and C.-l. Tai, "JSENet: Joint semantic segmentation and edge detection network for 3D point clouds," in *European Conference on Computer Vision (ECCV)*, 2020.

[16] A. Hermans, G. Floros, and B. Leibe, "Dense 3D semantic mapping of indoor scenes from RGB-D images," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[17] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, and P. H. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[18] A. Dai and M. Nießner, "3DMV: Joint 3D-multi-view prediction for 3D semantic scene segmentation," in *European Conference on Computer Vision (ECCV)*, 2018.

[19] M. Jaritz, J. Gu, and H. Su, "Multi-view PointNet for 3D scene understanding," in *IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019.

[20] F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, and M. Felsberg, "Deep projective 3D semantic segmentation," in *International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2017.

[21] A. Boulch, J. Guerry, B. Le Saux, and N. Audebert, "SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks," *Computers & Graphics*, vol. 71, 2018.

[22] J. Guerry, A. Boulch, B. Le Saux, J. Moras, A. Plyer, and D. Filliat, "SnapNet-R: Consistent 3D multi-view semantic labeling for robotics," in *IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017.

[23] A. Kundu, X. Yin, A. Fathi, D. Ross, B. Brewington, T. Funkhouser, and C. Pantofaru, "Virtual multi-view fusion for 3D semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2020.

[24] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena, "Semantic labeling of 3D point clouds for indoor scenes," in *Conference on Neural Information Processing Systems (NIPS)*, 2011.

[25] Y. Wang, R. Ji, and S.-F. Chang, "Label propagation from ImageNet to 3D point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[26] B. H. Wang, W. L. Chao, Y. Wang, B. Hariharan, K. Q. Weinberger, and M. Campbell, "LDLS: 3-D object segmentation through label diffusion from 2-D images," *IEEE Robotics and Automation Letters (RA-L)*, vol. 4, no. 3, 2019.

[27] X. Zhu, "Semi-supervised learning with graphs," Ph.D. dissertation, Carnegie Mellon University, 2005.

[28] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] R. Tylecek, T. Sattler, H. A. Le, T. Brox, M. Pollefeys, R. B. Fisher, and T. Gevers, "The Second Workshop on 3D Reconstruction Meets Semantics: Challenge results discussion," in *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[30] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv:1904.04514*, 2019.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[33] L. Ma, J. Stuckler, C. Kerl, and D. Cremers, "Multi-view deep learning for consistent semantic mapping with rgb-d cameras," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.