Semantic Segmentation using Implicit Representation

Mohammad Rashed Karim ElGhandour

Abstract

This paper proposes a method for improving the performance of semantic segmentation models using implicit representation called Plenoxels. The method involves using Plenoxels as an auxiliary input to enhance the performance of the models by combining the implicit representation with RGB information. Additionally, this work proposes an automated pipeline for novel synthetic views generation for tail class sampling using Plenoxels radiance field. The results showed that the combination of rendered images and spherical harmonics effectively captures the information present in camera images and incorporating projected spherical harmonics can potentially enhance the performance of semantic segmentation models over using color-only rendered images. However, the results showed no improvement when generating novel views for training, due to the used segmentation model's small size and the drop in render quality when deviating from the original camera trajectory.

1. Introduction

Semantic segmentation is a vital part of environment perception and scene understanding [13, 19, 20]. Not only can it aid in producing more realistic and seamless Augmented Reality (AR) applications, but it also has several potential applications in the area of Autonomous Vehicles. Training semantic segmentation usually faces class-imbalance issues, which is apparent even in large-scale datasets such as ScanNet [4]. Even though generating synthetic views can enhance the performance of semantic segmentation models [12], this approach requires sensor data. Recent advancements in 3D scene reconstruction facilitated the process of synthesizing novel views without the need of depth information [15,24]. The scenes are stored using an implicit representation that is not only more storage-efficient than point clouds [9], but could be used to provide additional valuable information for training as well. Therefore, in this work we propose utilizing the implicit Plenoxels representation [9,24] to generate class-balanced semantic views for semantic segmentation training, in addition to combining implicit representation with RGB information. This approach helps us achieve several goals:

1. **Novel angles:** Performing 3D augmentation showcases the object in additional angles, providing a more robust object representation for training.
2. **Scale variance:** By adding some depth randomness when generating novel views, the object is captured at a variety of sizes while maintaining its integrity.

2. Related Work

A recent direction in 3D reconstruction is the use of implicit surface representations such as Signed Distance Fields [10], Occupancy Grids [6], local light field [14] and Radiance Fields [2, 15]. However, these methods have not been researched for 2D semantic segmentation.

2.1. Radiance Fields

Radiance Fields are implicit 3D representations that model geometry and appearance, enabling view-dependant effects and allows for reconstruction simply using multi-view images unlike methods that require different sensors [8, 17, 21]. Neural Radiance Field (NeRF) [15] uses a differentiable volume rendering formula to train a deep neural network to model the radiance (brightness and colour) as a function of 3D position and 2D viewing direction. Variants of Neural Radiance Fields have been proposed [1,2,5,7] that can generate high-fidelity renders, however, they all suffer from the same problems which is taking several days to optimize a single scene. Additionally, implicit features encoded in the weights are scene-specific and cannot be transferred between scenes.

Yu et al. [24] proposed Plenoxels which represent the scene as a sparse 3D grid with Spherical Harmonic Coefficients and are optimized using differentiable rendering loss without neural networks offering fast optimization. PeRFception [9] used Plenoxels to generate a large-scale implicit representation dataset based on ScanNet [4].

2.2. 2D Segmentation

Due to its importance for environment perception, semantic segmentation was thoroughly investigated by researchers on both 2D [18, 22] and 3D [3, 16] levels. Kundu

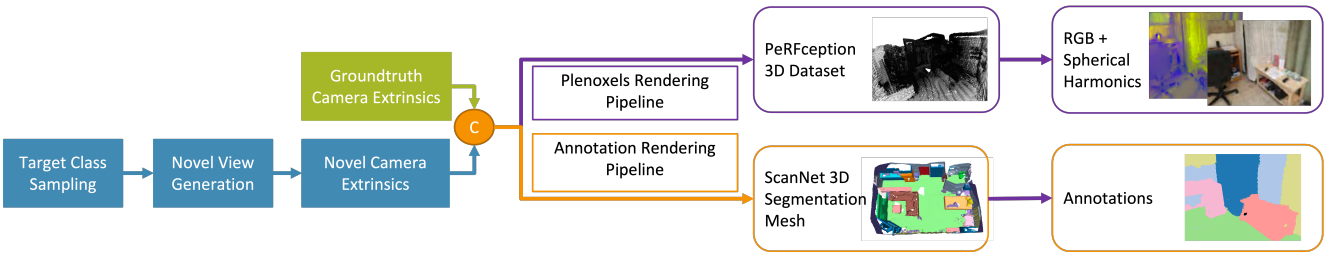


Figure 1. Pipeline overview. Target Class Sampling outputs camera poses with direct line of sight to the target class. This is then used to generate novel camera poses. The original camera poses, and the novel camera poses are input into two rendering pipelines: the Plenoxels renderer generating color and spherical harmonics images and the Annotation renderer which generates the ground truth annotations.

et al. [12] showed that synthesizing novel views can enhance the performance on both 2D and 3D semantic segmentation tasks. However, the approach did not use radiance fields-based rendering, which not only limited both the rendering quality and performance gain but it requires sensor data as well. In [23], the authors suggested using the density fields for training a semantic segmentation network. This approach, however, does not utilize the color information of the scene.

3. Method

Our approach leverages implicit representations to take advantage of spherical harmonics for view dependant colors as well as generate novel views to improve class imbalance. An overview of our entire pipeline is shown in Figure 1.

3.1. Rendering Color and Spherical Harmonics Using Plenoxels

We adopt the Plenoxels representation similar to [9, 24]. Plenoxels is a sparse voxel grid representation in which each occupied voxel store opacity (σ) and spherical harmonics coefficients ($\mathcal{S}$). In order to obtain a continuous plenoptic function, trilinear interpolation is used.

Following [15, 24], the volumetric rendering model approximates the colors by integrating over samples taken along the ray at fixed intervals.

$$\hat{C}(r) = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))p_i$$
$$\text{where} \quad T_i = exp(-\sum_{j=1}^{i-1} \sigma_j\delta_j) \tag{1}$$

T_i is defined as the accumulated transmittance along the ray r to the i^{th} sample in comparison with preceding samples. $(1 - exp(-\sigma_i\delta_i))$ calculates how much light is contributed by sample i and δ_i computes the next sample. Finally, $p_i \in \{c_i, \mathcal{S}_i\}$ denotes the value to be rendered into image plane with distance, which can be either color (c_i) or spherical harmonics ($\mathcal{S}$). When rendering the RGB image c_i is computed as the weighted sum of the spherical harmonics coefficients, where the weights are optimizable/trainable parameters. On the other hand when rendering spherical harmonics we directly render the 27 dimensional vector $\mathcal{S}_i$.

3.2. Rendering Ground truth Semantic Segmentation

A custom Segmentation Shader was implemented which returns the base label and raw color value irrespective of shadows. This is done by retrieving faces corresponding to each pixel, selecting the corresponding label value through a majority vote, and finally obtaining the label of that face. This Segmentation Shader is then fed into the rendering pipeline.

3.3. Generating Novel Views

Implicit representations offer the advantage of being able to generate views of an object from novel angles. This advantage can be used to tackle frequent dataset issues such as unbalanced classes.

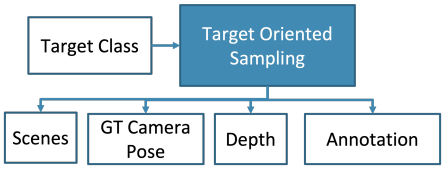


Figure 2. Target Sampling

3.3.1 View Selection

The procedure of view selection relied on the camera pointing toward the target class. This could be achieved by selecting the poses and scenes in which the target class covered 50% or more of the center of the image, as shown in figure 2. Additionally, our custom segmentation shader was extended to export the depth values. Having the depth, target class, and camera extrinsics, it is then possible to create a virtual dome around the object such that the camera is always pointing toward the target pixel as shown in Figure

3. Furthermore, a random shift is added to both the novel camera's height and the distance to the target object.

The 3D view selection provides augmentation methods that simple 2D augmentation is not capable of. The output of our view selection functionality is then filtered further.

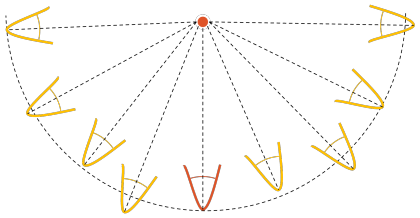


Figure 3. Novel Dome Generation

3.3.2 Filtering Novel Views

In order to avoid redundant views; as the case of the virtual camera being located behind a wall. The pipeline is designed to ensure that the target class covers at least 20% of the generated segmentation image. This ensures that the additional novel views provide useful input to the dataset rather than uselessly exploding its size. The pipeline works by feeding the output camera extrinsic matrices into the segmentation pipeline, where the renders are generated and filtered if necessary.

4. Experiments

The experiments were conducted using an NVIDIA GeForce RTX 3080. As for the semantic segmentation model, ENet [18] was implemented due to its time and computationally-efficient design. The proposed approach, however, is not restricted to a specific model and could be applied to any semantic segmentation model.

The results were obtained on the ScanNet dataset and its plenoptic representation PeRFception-ScanNet. The datasets comprise over 1500 indoor scenes across 20 classes, providing diverse and extensive data for the experiments. The scenes exhibit variations in illumination and contrast, making them suitable for the investigation. The datasets are introduced in [4] and [9].

4.1. Quantitative Results

Table 1 displays the results obtained when training ENet using different types of inputs. Since the performance drop when switching from real images to synthetic ones was minimal (only 1.2%) this indicates that the rendered images are of high quality and they preserve the semantic content of the original images. Another take from the results is the fact that we were able to gain back the performance drop by using spherical harmonics as an additional input modality. It could be claimed that the combination of rendered images and spherical harmonics are able to represent the information of the original camera images.

Our approach of novel view synthesis acts as a set of strong augmentations on the original color images in order to enhance the performance. However, as Table 1 displays, the performance on the $mIoU$ dropped after using the novel views. Since this was accompanied by the fact that training loss when using novel views was higher than without using it, we suspect that this occurred due to the limited size of the segmentation model, which has only 0.37 million parameters. Even though smaller models are faster to train, the small model size limits its learning capacity and hinders it from benefiting from strong augmentations. Additionally, the rendered image quality could decrease when generating views that are far from the original trajectory and artifacts could arise in some parts of the image. Therefore, developing a metric that asses the quality of the images could help in masking artifacts and enhance the training process.

Image Type	Novel Views	Input	$mIoU_{\%}$
Camera	×	RGB	41.2
Rendered	×	RGB	40.0
Rendered	×	RGB + SH	41.0
Rendered	✓	RGB	39.5
Rendered	✓	RGB + SH	39.0

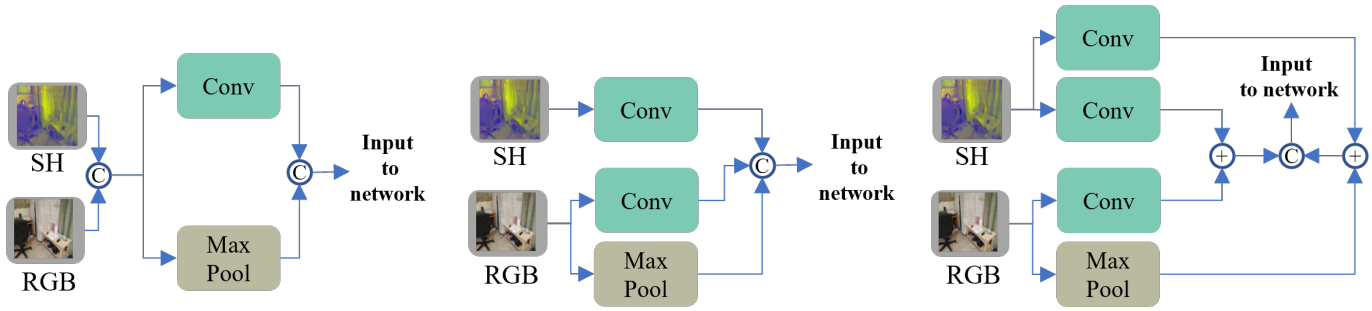
Table 1. Overview of the results comparing using original camera images against different cases of rendered data. SH: Spherical Harmonics

4.2. Visual Results

Figure 5 showcases three examples from the rendered images. The first image was generated using a pose from the original camera path, the second image was generated with a pose that's close to the trajectory, and the last one was generated with a pose that's farther away from the original trajectory. It's evident that the closer the pose is to the original trajectory, the higher the render quality is. This highlights the need for a metric to quantify the quality of the rendered images. We additionally provide a sample of the semantic segmentation visual results in Figure 6.

4.3. Ablation: Impact of Spherical Harmonics and RGB fusion approach

We investigate the effect of various data fusion methods on the performance of the model. The first method involves simple concatenation of spherical harmonics with RGB input. The second approach applies a separate convolution layer to the spherical harmonics and RGB before concatenation. The final approach involves applying two convolutional layers to the spherical harmonics and then summing one feature with the convolved RGB features and the other

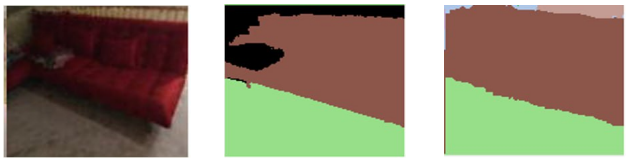


(a) Direct concatenation approach (b) Convolve and concatenate approach (c) Convolve and sum approach

Figure 4. Different approaches for merging spherical harmonics (SH) with color (RGB). C: short for concatenate, +: summation

(a) Original trajectory (b) Closer (c) Farther

Figure 5. The novel synthesized views are of higher quality when they are sampled closer to the original camera trajectory



(a) Color (b) Ground Truth (c) Prediction

Figure 6. Sample of the visual results of our model, which uses both RGB and spherical harmonics input

with downsampled RGB values. The three approaches are illustrated in Figure 4. Table 2 showcases the performance gain when using the third approach (transformer and sum) in comparison with the first two approaches.

Approach	$mIoU_{\%}$
Direct Concatenation	39.6
Convolve and Concatenate	40.6
Convolve and Sum	41.0

Table 2. Ablation of the impact of spherical harmonics and RGB fusion approach

4.4. Ablation: Improvement over the Base Model

We compare our results after hyperparameters optimization with a base implementation of ENET-ScanNet [11], which this work was built upon. Table 3 demonstrates that our implementation performs better even when using fewer modalities.

Approach	$mIoU_{\%}$
RGB [11, 18]	20.1
RGB + Depth [11, 18]	37.1
RGB (ours)	41.2
Rendered RGB (ours)	40.0

Table 3. Ablation on the improvement over the base model implementation

5. Conclusion

The feasibility of using implicit representation for improving the performance of 2D semantic segmentation models was investigated in two directions. The first involved utilizing the implicit representation, specifically Plenoxels, as an auxiliary input to improve the performance of semantic segmentation models. The second direction aimed to leverage the implicit representation to generate novel views that expand the training data. The findings indicated that incorporating projected spherical harmonics can potentially enhance the performance of semantic segmentation models. Furthermore, the combination of rendered images and spherical harmonics effectively captures the information present in camera images. Finally, the results showed no improvement when generating novel views for training. This can be because the model is too small to benefit from strong augmentations, or it might be because deviations from the original trajectory result in lower-quality rendered images. Future work should thus concentrate on training models with more capacity than ENet and creating a metric that measures the render quality in various areas of the images.

References

[1] Di Chang, Aljaž Božič, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Süsstrunk, and Matthias Nießner. Rc-mvsnet: Unsupervised multi-view stereo with neural rendering. *arXiv preprint arXiv:2203.03949*, 2022. 1

[2] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021. 1

[3] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[4] 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, 3

[5] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields, 2021. 1

[6] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020. 1

[7] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021. 1

[8] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 1

[9] Yoonwoo Jeong, Seungjoo Shin, Junha Lee, Chris Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Perfception: Perception using radiance fields. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1, 2, 3

[10] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020. 1

[11] Murthy Krishna. Enet-scannet. https://github.com/krrish94/ENet-ScanNet, 2019. 4

[12] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3d semantic segmentation, 2020. 1, 2

[13] Felix Järemo Lawin, Martin Danelljan, Patrik Tosteberg, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Deep projective 3d semantic segmentation. In *International Conference on Computer Analysis of Images and Patterns*, pages 95–107. Springer, 2017. 1

[14] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 1

[15] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[16] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3D: Out-of-Context Data Augmentation for 3D Scenes. In *International Conference on 3D Vision (3DV)*, 2021. 1

[17] Diana Pagliari, Fabio Menna, R Roncella, Fabio Remondino, and Livio Pinto. Kinect fusion improvement using depth camera calibration. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 45, 2014. 1

[18] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation, 2016. 1, 3, 4

[19] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1

[20] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 1

[21] Julián Tachella, Yoann Altmann, Nicolas Mellado, Aongus McCarthy, Rachael Tobin, Gerald S Buller, Jean-Yves Tourneret, and Stephen McLaughlin. Real-time 3d reconstruction from single-photon lidar data using plug-and-play point cloud denoisers. *Nature communications*, 10(1):1–6, 2019. 1

[22] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, jul 2019. 1

[23] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi S. M. Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes, 2021. 2

[24] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5491–5500, 2022. 1, 2