Upgraded Semantic NeRF: 3D Consistent Data Annotation on Sparse Labels

Shen Zhehao
ShanghaiTech University
2021533110
shenzhh@shanghaitech.edu.cn

Hong Yu
ShanghaiTech University
2021533148
hongyu@shanghaitech.edu.cn

Abstract

This paper proposes a method designed to mitigate the burden of manual annotation. We introduce a mechanism that integrates sparse labels to facilitate the segmentation of novel views, effectively obviating the need for comprehensive manual annotation. We present an enhanced version of the Semantic NeRF algorithm that guarantees three-dimensional consistency and concurrently expedites the computation process beyond the capabilities of the basic version.

1 Introduction

In computer vision and computer graphics research, there exists an extensive demand for large, annotated image datasets. This necessity consequently accentuates the criticality of developing techniques for annotating 3D scenes under sparse labels. Numerous methodologies have been probed to tackle this issue, ranging from co-training [1], multi-view semi-supervised learning to video-based segmentation approaches such as MiVOS [2] and XMem [3]. However, these prior works are frequently encumbered by a lack of frame-to-frame consistency, often leading to divergences in object attribute recognition across various perspectives.

This paper introduces an advanced variant of Neural Radiance Fields (NeRF)[4] - Semantic NeRF[5]. This innovation incorporates semantic data into its design, thereby enhancing the precision of 3D object segmentation. Semantic NeRF extends the foundational NeRF model by integrating semantic inputs and outputs, modifying the loss function to implement semantic constraints in addition, thereby ensuring more accurate segmentation. Our model undertakes 3D reconstruction and semantic segmentation in tandem, significantly enhancing the efficiency and efficacy of the process.

In order to augment training speeds and mitigate computational expenses, we have integrated the Instant-NGP [6] into our Semantic NeRF model. By incorporating hash encoding, the model accelerates the process of feature vector retrieval during training through trilinear interpolation, thereby enhancing the overall efficiency of the model.

To summarize, our main contributions include:

- Applying Semantic NeRF to solve 3D consistent data annotation on sparse labels.
- Applying hash encoding method to accelerate the training, which greatly saves the time and cost of training models and generating data annotations.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Related Works

2.1 Segmantic Methods

Recent developments have witnessed the introduction of a variety of segmentation algorithms, including video-based segmentation methodologies such as MiVOS and XMem. These techniques leverage temporal information to enhance segmentation precision within video sequences. However, when these methods are implemented in practical applications, they often exhibit a lack of consistency within 3D scenes. This realization incited us to consider NeRF. By capitalizing on NeRF's attributes, we incorporated an additional semantic information dimension to the neural radiance fields, thereby facilitating semantic segmentation while preserving 3D geometric consistency.

2.2 3D Scene Reconstruction

3D object reconstruction, a longstanding field within computer graphics, has seen significant advancements through recent developments in neural networks, enabling innovative methods for dense 3D scene representation. NeRF, a method utilizing implicit neural representations, accomplishes dense scene reconstruction from sparse viewpoints, a feat previously unheard of. Moreover, the incorporation of neural network representations considerably minimizes storage requirements for scene representation.

Notable contributions have been made in the realm of neural radiance fields [7][8][9]. Among them, Instant-NGP, which integrates the original 8-layer MLP into a hash encoding and 2-layer MLP network, dramatically reduces training time. This architecture is more than 100 times faster than vanilla NeRF. Hash encoding serves as a cornerstone within this architecture, fostering a more streamlined and expedited computation of inputs. Broadly speaking, the Instant-NGP methodology embodies a remarkable advancement in the domains of 3D scene comprehension and computer graphics.

2.3 Semantic NeRF

Semantic NeRF augments the base NeRF approach by infusing an additional semantic layer, which enhances the precision of the reconstructed 3D environment. This is due to our belief that objects with similar shapes and colors tend to have the same or similar semantic information. This enhancement is actualized by the introduction of a semantic plane to the process, with a semantic segmentation network generating semantic data for every 3D point.

The integration of semantic data and improvement of reconstruction accuracy is accomplished through the advent of semantic consistency loss, an alteration to the original loss function in Semantic NeRF. This loss assesses the uniformity between forecasted and actual segmentation outcomes, subsequently refining the precision of the reconstructed scene.

Nonetheless, as Semantic NeRF is anchored in the foundational NeRF structure, its training velocity is relatively sluggish, prompting a need for structural improvements to expedite the process.

3 Methodology

We drew inspiration from Instant-NGP and utilize hash encoding to replace the vanilla NeRF's encoding approach. In our architecture, 3D position (x, y, z) and viewing direction (θ, ϕ) are fed into the network after hash encoding. Volume density σ and semantic logits $\mathbf{s}$ are functions of 3D position while colours $\mathbf{c}$ additionally depend on viewing direction.

3.1 Loss Combining Photometric and Semantic

We train the whole network from scratch under photometric loss L_p and semantic loss L_s. For the color image synthesis, we adopt the mean square error (MSE):

$$L_p = \sum_{\mathbf{r}\in\mathcal{R}} \left[\left\| \hat{\mathbf{C}}_c(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|_2^2 + \left\| \hat{\mathbf{C}}_f(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|_2^2 \right]$$

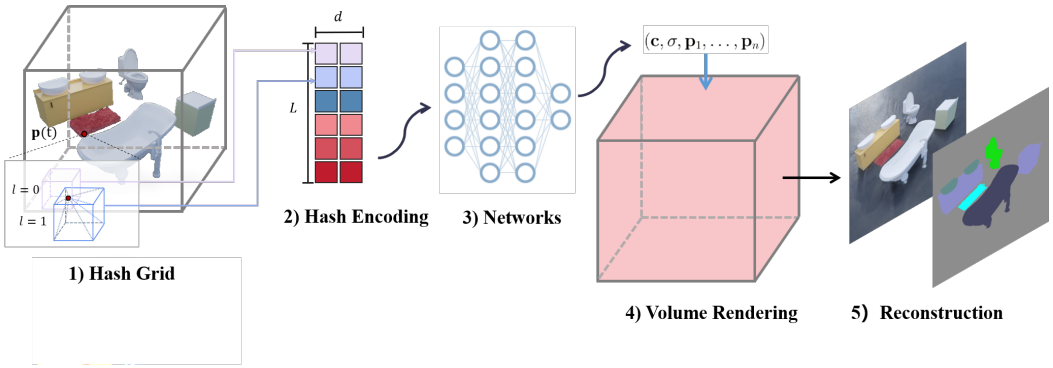

Figure 1: Pipeline: (1) Hash Grid: for a 3D input position $\mathbf{x}$, we first find L surrounding voxels in a 3D multiresolution hash grid. (2) Hash Encoding: compute d dimension feature with trilinear interpolation of eight corners for each voxel and query the feature vector from the hash table.(3) Network: the network takes the generated feature and view direction as inputs, and outputs color $\mathbf{c}$, density σ and semantic classes $\mathbf{p}_1, \ldots, \mathbf{p}_n$ (4) Volume Rendering. (5) Reconstruction: we can achieve reconstruction and semantic segmentation results. [10].

Where R are the sampled rays within a training batch, and $C(r)$, $\hat{C}_c(r)$ and $\hat{C}_f(r)$ are the ground truth, coarse volume predicted and fine volume predicted RGB colours for ray r, respectively. For semantic label prediction, we use the cross entropy loss function:

$$L_s = -\sum_{\mathbf{r}\in\mathcal{R}} \left[\sum_{l=1}^{L} p^l(\mathbf{r}) \log \hat{p}_c^l(\mathbf{r}) + \sum_{l=1}^{L} p^l(\mathbf{r}) \log \hat{p}_f^l(\mathbf{r}) \right]$$

$$L = L_p + \lambda L_s$$

are the multi-class semantic probability at class l of the ground truth map, coarse volume and fine volume predictions for ray r, respectively.

Hence the total training loss L is:

$$L = L_p + \lambda L_s$$

3.2 Acceleration Based on Instant-NGP

Given a fully connected neural network $\mathrm{m}(\mathbf{y}; \phi)$, we want to find an encoding of its inputs $\mathbf{y} = enc(\mathbf{x}; \theta)$. We apply hash encoding: for a point in space, we leverage the feature vectors of its surrounding eight points for trilinear interpolation, thereby obtaining its feature vector. The feature vectors of these eight points are retrieved from the hash table based on the point's index. The inclusion of additional trainable variables significantly enhances the speed of feature querying and training, thus effectively accelerating the process.

4 Experiments

In this section, we evaluate our approach in a variety of challenging scenarios. Besides, various rendering results of Upgraded Semantic NeRF are shown in Figure 2.

4.1 Comparisons

In this section, we present a comparative analysis among various methodologies. From a visual perspective, our outcomes evidently demonstrate superior preservation of scene structure, as compared to other evaluated techniques.

We did the training with 10 frames that have been annotated. The training time for each method is 15 minutes, 8 hours, 21 minutes and 22 minutes respectively. We can easily find out that our method (Upgraded Semantic NeRF) has both the best training time and the best performance.

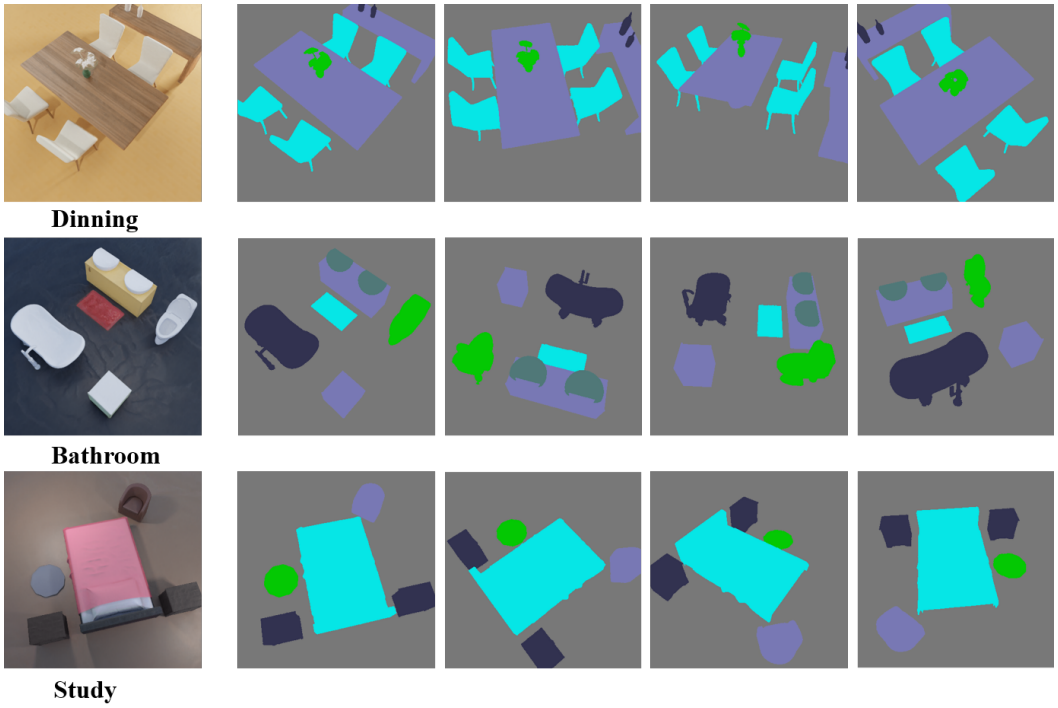

Figure 2: Our render semantic segmentation results in three multi-object scenes, including "Dinning","Bathroom" and "Study".

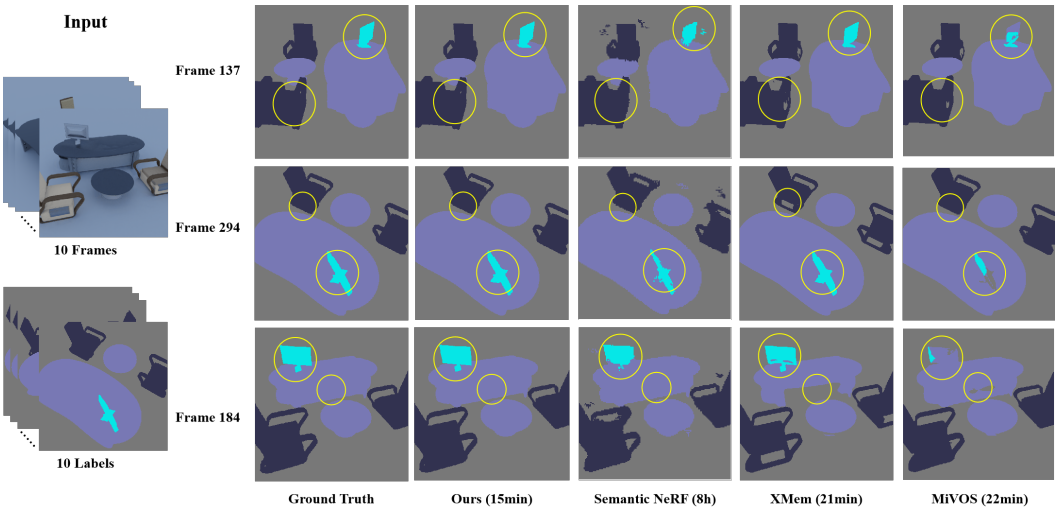

Figure 3: A comparison between GT, our results, Semantic NeRF, XMem and MiVOS under different training time.

5 Conclusion

In this paper, we used Semantic NeRF to annotate data with only sparse labeling. Compared to current methods, our approach utilizes the 3D consistency of NeRF and significantly improves the accuracy of data annotation and segmentation. Also, we modified the original framework of Semantic NeRF by utilizing the hash encoding method in Instant-ngp, which greatly accelerates training time and improves training efficiency. Additionally, our approach has a wide range of applications, including artificial intelligence, robotics [11], and other related fields.

References

[1] Yi Zhang Shuwei Qian Chongjun Wang Mingcai Chen, Yuntao Du. Semi-supervised learning with multi-head co-training. *arXiv preprint arXiv:2107.04795*, 2021.

[2] Chi-Keung Tang Ho Kei Cheng, Yu-Wing Tai. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. *arXiv preprint arXiv:2103.07941*, 2021.

[3] Alexander G. Schwing Ho Kei Cheng. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. *arXiv preprint arXiv:2207.07115*, 2022.

[4] Matthew Tancik Jonathan T.Barron Ravi Ramamoorthi Ren Ng Ben Mildenhall, Pratul P.Srinivasan. Nerf: Representing scenes as neural radiance fields for view synthesis. *arXiv preprint arXiv:2003.08934*, 2003.

[5] Stefan Leutenegger Andrew J. Davison Shuaifeng Zhi, Tristan Laidlow. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2103.15875*, 2021.

[6] Christoph Schied Alexander Keller Thomas Muller, Alex Evans. Instant neural graphics primitives with a multiresolution hash encodin. *arXiv preprint arXiv:2201.05989*, 2022.

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022.

[8] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, June 2022.

[9] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.

[10] Bo Yang Bing Wang, Lu Chen. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *arXiv preprint arXiv:2208.07227*, 2022.

[11] Zhipeng Bao Martial Hebert Yu-Xiong Wang Mingtong Zhang, Shuhong Zheng. Beyond rgb: Scene-property synthesis with neural radiance fields. *arXiv preprint arXiv:2206.04669*, 2022.