Depth-Supervised Dyanamic NeRF

Chang Luo
chang.luo@tum.de

Abstract

The reconstruction of non-rigid scenes from sparse image data is an under-constrained problem. Our goal is to leverage depth information of RGB-D images to supervise the ill-constrained NeRF optimization process. The depth guidance is realized by expanding dynamic NeRF with the depth loss function and depth-guided points sampling strategy. Our method outputs a more reliable depth prediction and slightly improved color image prediction.

1. Introduction

Synthesizing novel views from dynamic scenes at arbitrary world positions is very crucial for VR and AR applications. NeRF [7] encodes static scenes with a multi-layer preceptron (MLP) and synthesizes images by taking advantage of Neural Rendering technology. However, the problem becomes more challenging for non-rigid namely dynamic scenes since the model should not only learn the spatial information but also the temporal relationship across all dynamic frames. Significant improvements were achieved by decoupling the geometry and time of the scene with the use of two separate MLPs [12, 18] for disentangling the problem. However, the problem is still under-constrained which results in poor image synthesis performance as well as ambiguous depth prediction. Our method aims to leverage depth information of RGB-D images to guide the NeRF optimization process in order to improve model for recovering the color and geometry information of non-rigid scenes. In summary, we propose a new method with including of depth supervision in NeRF training process to make the problem better constrained, enabled by the following contributions:

- Introducing MSE and GNLL as Depth loss function for Depth-Supervised Training
- Implementing Depth-Guided Sampling to guide NeRF to get more accurate depth prediction
- Comparing the metrics of the method both on Synthesis Dataset and Real-World Dataset

2. Related Work

Classical approaches. The task of Synthesizing images from new view angles is a well-studied problem and can be tackled with classical scene reconstruction approaches like Structure-from-Motion (SfM) [16] or SLAM [3], more recently, DynamicFusion [5] comes up with a method which is adapted for non-rigid scenes. Once the 3D world of the scene is reconstructed, it is easy to adjust the virtual camera pose to generate viewing shots from arbitrary viewing angles.

Implicit scene representation. The coordinate-based methods, also known as neural implicit representation [2, 8, 9, 14, 15] establish a new aspect by using the deep learning based method for representing a scene. These methods aim to train a Multiple Layer Perceptron (MLP) to regress an implicit scene representation. While DeepSDF [10] addresses the MLP to learn a signed distance function, Neural Radiance Field (NeRF) [7] learns an implicit function for representing the whole scene. The coordinate input and viewing angle are assumed to be mapped into the color and density value.

Dynamic NeRF and Depth. More recent works [11, 12, 18] decompose the dynamic scenes into a canonical model and a deformation model, the other network follows the pattern of NeRF to regress an implicit representation function for outputting density and color of the query coordinate. Several recent works have also explored ways to incorporate depth observations for the reconstruction of static scenes. NerfingMVS [20] trains a monocular depth network for inducing depth priors to guide the NeRF sampling. Roessle et al. [13] also proposed a method that involves dense depth completion and exploits an additional depth loss for the NeRF geometry.

3. Methods

Based on D-NeRF [12] implementation, our method attempts to improve Non-Rigid NeRF methods by additionally supervising a depth term, with a dense input of depth frame, which could make the depth supervision more accurate and make the problem well-constrained.

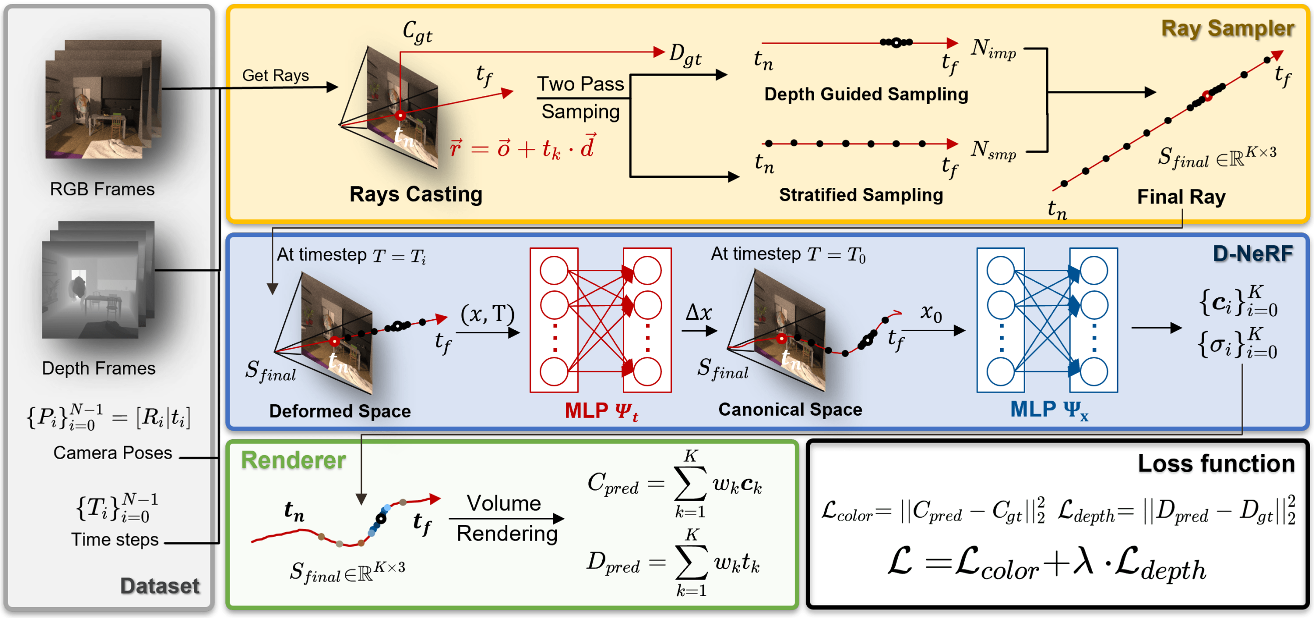

Figure 1. Overview of our optimization pipeline. Given an RGBD-Dataset and aligned camera poses and time steps, Ray Sampler randomly generates some rays cast back into the scene, and two path sampling is applied for sample query points along the ray. D-NeRF [12] backbone will then transfer the points back to canonical space and estimate its' color and density. The final output will be further integrated by Renderer as the prediction color and depth, which can be supervised with the groundtruth of the observation.

Starting with an RGBD-Dataset, which is composed of RGB images $C_i{}_{i=0}^{N-1}$, $C_i \in [0,1]^{H\times W\times 3}$ and Depth images $D_i{}_{i=0}^{N-1}$, $D_i \in [t_n, t_f]^{H\times W\times 1}$, where t_n and t_f describes the near and far plane distance of the whole scene. Since the scene is the non-static scene, the time steps $T_i{}_{i=0}^{N-1}$, $T_i \in [0,1]$ and the camera pose $P_i \in SE(3)$ at the respective time step. Fig 1 presents the overview of the whole approach.

As a part of our pipeline, the dataset will be fed into Ray Sampler for retrieving multiple rays back into the three-dimensional scene. With the two-path sampling strategy enabled, the ray will be not only sampled uniformly among the whole ray but also more densely in a range around the depth which can be controlled by standard deviation. Subsequently, the sampled points will be passed into the D-NeRF [12] backbone to recover the color and density of query points at the time step t. Finally, the points with predicted color and density can be integrated according to the Volumetric Rendering, the output will be the final predicted color and depth or the pixel corresponding to the casting ray, which can be guided by our loss function to ensure that the network can be optimized.

3.1. Ray Sampler

For each iteration, the Ray Sampler will pick an arbitrary image pair C_i, D_i and its corresponding camera pose P_i as well as the time step T_i from the RGBD-Dataset, by further randomly choosing N_{rand} indices of the image pixels, which we will cast the ray back. In more detail, the rays $\vec{r} = \vec{o} + t_k \cdot \vec{d}$ can be retrieved, where $t_k \in [t_n, t_f]$, $\vec{o}$ refers to the camera origin coordinate and $\vec{d}$ represents the ray casting direction. By controlling the factor k, we can easily distribute samples to make them spread out along the whole ray range with $k \in [t_n, t_f]$. For sample them uniformly, the equidistant sampling distance can be calculated by $(t_f - t_n)/N_{smp}$. As the result of the Stratified Sampling, we will get a collection of points with N_{smp} samples.

D-NeRF [12] ray sampler implements a built-in hierarchical sampling strategy, which by analyzing the final prediction output of stratified to come out a rough density distribution along the ray, which is called *corase* pass. By exploiting it, it is able to sample more densely around the position that has a high density, namely where the ray is very possible to terminate. The above processing step is named *fine* pass and the number of fine samples is defined by hyperparameter N_{imp}.

In our case, since the depth of the pixel where the ray is cast into is already given, we can directly draw samples from a Gaussian distribution $\mathcal{N}(D_{gt}, D_{std})$, where D_{gt} is the groundtruth depth and S is a hyperparameter refers to the standard deviation to restrict sampling range around the D_{gt}. As the final stage of two-pass sampling, we prepare a final point sets $S_{final} \in \mathbb{R}^{K\times 3}$ with $K = N_{smp} + N_{imp}$, which is combined with the sample points from two passes.

3.2. D-NeRF backbone

Following by the idea of D-NeRF [12], we also have two network separated, the deformation MLP Ψ_t and the canonical MLP Ψ_x. The former one will read a pair of (x_T, T), and try to learn an offset for such a point to transfer it back into the position in canonical space ($T = 0$), denoted as,

$$\Psi_t(x_T, T) \rightarrow \Delta x \tag{1}$$

where $T \in [0, 1]$ satisfies $x_T + \Delta x = x_0$. The canonical MLP will take the transferred point and predict its color and density in time step $T = 0$, denoted as,

$$\Psi_x(x) \rightarrow (c_i, \sigma_i) \tag{2}$$

where $i \in [0, S_{final}]$. We pass all samples which are drawn from the Ray Sampler into the D-NeRF [12] backbone to estimate the RGB color and density for each query sample.

3.3. Volumetric Rendering

Once the color and density of the samples are predicted by D-NeRF [12] network, the volumetric integration can be employed to calculate the final output color or depth of the ray, namely where the camera can see from this pixel. The final integrated color and depth can be rendered out by the following formulas.

$$C_{pred} = \sum_{k=1}^{K} w_k \mathbf{c}_k, \tag{3}$$

$$D_{pred} = \sum_{k=1}^{K} w_k t_k, \tag{4}$$

$$\text{where} \quad w_k = I_k(1 - \exp(-\sigma_k \delta_k)), \tag{5}$$

$$I_k = \exp(-\sum_{k'=1}^{k} \sigma_{k'} \delta_{k'}), \tag{6}$$

$$\delta_k = t_{k+1} - t_k. \tag{7}$$

The term I_k describes the possibility of the ray is still alive at sample k, which decreases in a very fast manner when the ray reaches the region that has high density. Therefore, the weights w_k will be low either in the region where density is low or in the region the ray is very likely to have terminated.

3.4. Loss Function

The objectives to be optimized are the two MLPs Ψ_t and Ψ_x. Since the prediction of color and depth can be calculated by taking advantage of Volumetric Rendering. A simple MSE loss function can be applied to both color loss and depth loss with

$$\mathcal{L}_{color} = ||C_{pred} - C_{gt}||_2^2, \tag{8}$$

$$\mathcal{L}_{depth} = ||D_{pred} - D_{gt}||_2^2. \tag{9}$$

As an alternative of the depth loss function, we also implement Gaussian negative log likelihood (GNLL), which will be activated if conditions Q and G are fulfilled, where $Q = |D_{pred} - D_{gt}| > S_{pred}$ constrains the difference between groundtruth and predicted depth should be larger than it's standard deviation and $G = S_{pred} > S$ restricts that the predicted standard deviation should be larger than the given standard deviation from hyperparameter.

$$\mathcal{L}_{depth} = \begin{cases} \log(S_{pred}^2) + \frac{(D_{pred} - D_{gt})^2}{S_{pred}^2} & \text{if } Q \text{ or } G, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

We can get the standard deviation of predicted depth by using the following formula

$$S_{pred}^2 = \sum_{k=1}^{K} w_k (t_k - D_{pred})^2. \tag{11}$$

In this way, the NeRF is inclined to have the ray termination within the depth range, which may results in a more accurate depth prediction.

4. Results

We evaluate our methods and compare it D-NeRF [12] baseline to check the qualitative outcome and quantitative metrics. Besides, we also conduct an ablation study to check how the performance will be influenced by disabling Depth-Guided Sampling (DGS) or switching the loss function to GNLL loss. Considering testing the stability and generality of the approach, we test our methods on two datasets, one is a synthetic dataset from TöRF paper [1], and the other one is a real-world dataset captured from Kinect camera [6] with some invalid depth existing.

4.1. Training Hyperparamters

We basically follow the setup of the D-NeRF paper to initialize our D-NeRF backbone [12], namely the deformation MLP Ψ_t and canonical MLP Ψ_x. For each randomly picked image, we cast $N_{rand} = 512$ rays back into the scene, and for each ray, we take $N_{smp} = 64$ for *coarse* pass and $N_{imp} = 128$ for *fine* pass. The weighting factor λ of depth loss function is kept as 0.001 for experiments with depth-supervision enabled. Especially for DGS, the standard deviation S is set to 0.01, which means N_{imp} samples will be sampled from range $D_{gt} \pm 0.01$.

4.2. Qualitative Result

Fig. 2 represents the qualitative outcome of our approach performed on both datasets we mentioned before. We can observe that our method can estimate a more meaningful depth compared to D-NeRF [12] baseline, which only predicts a very mean depth. Besides, the improvement of the color image is also perceivable. According to the Toss

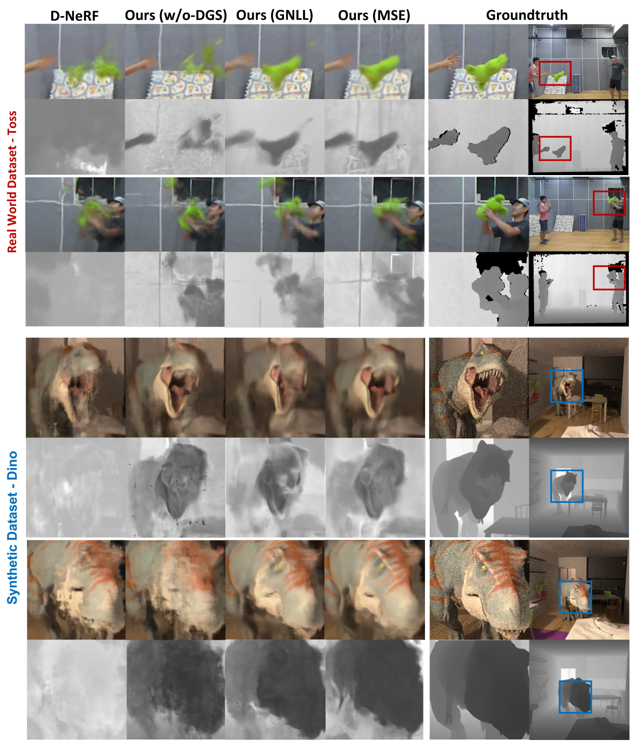

Figure 2. Qualitative result of the predicted test images both on Toss Dataset and Dino Dataset.

dataset, the baseline was unable to recover the rough shape of the green doll, while our MSE-trained methods rendered a very similar result with respect to the groundtruth. Furthermore, experiments performed on the Dino dataset have shown performance improvements, compared to the baseline, there are fewer artifacts can be observed in the output of our MSE-trained network, and the local details can be better identified.

4.3. Quantitative Result

We also measured our method with several metrics referring to the measurement of Depth Priors NeRF [13], which includes four metrics, namely Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [19], Learned Perceptual Image Patch Similarity (LPIPS) [21] and Depth RMSE, for final evaluation. Tab. 1 shows the metrics comparison between several experiment setups on two datasets. The quantitative result also shows some extent of consistency with the qualitative result, the MSE-trained one presents the best PSNR metric among all methods. Moreover, we can also observe that the method works more stable within the Dino dataset since all metrics indicate that the MSE-trained one has the best performance. The most conspicuous contrast is shown in the RMSE item, i.e., the prediction of depth. Therefore, we are able to assert that the introduction of depth supervision can significantly improve the reliability and meaningfulness of the depth information inferred by NeRF.

	Method	PSNR↑	SSIM↑	LPIPS↓	RMSE↓
Toss	D-NeRF [12]	20.31	0.703	0.315	2.348
	Ours (w/o DGS)	20.45	**0.715**	0.318	0.501
	Ours (w/ GNLL)	20.54	0.710	**0.313**	**0.415**
	Ours (w/ MSE)	**20.87**	0.703	0.335	0.468
Dino	D-NeRF [12]	22.13	0.330	0.692	1.298
	Ours (w/o DGS)	22.27	0.328	0.703	0.088
	Ours (w/ GNLL)	22.26	0.336	0.688	**0.022**
	Ours (w/ MSE)	**22.73**	**0.339**	**0.685**	0.094

Table 1. Quantitative result of the predicted test images both on Toss Dataset and Dino Dataset.

4.4. Ablation Study

We also performed two ablation studies to check out the influence of Depth-Guided Sampling and GNLL loss. Removing DGS and training the network with normal MSE will lead to a generally worse outcome, sometimes it even downgrades the output (see Fig. 2 Row 1 and Row 4). We also tried to replace MSE with GNLL depth loss, which outputs more flat depth prediction with less noise hence result in lower RMSE. However, the color output is poor when compared with the MSE-trained one.

4.5. Limitations and future work

The main limitation can be summarized as the degradation of the quality of the output image. Although we observe that the dynamic part of the scene can be better recovered compared to the D-NeRF [12] baseline. However, for some static parts of images, the method is inclined to be unstable and will render out some noise or unmeaningful local deformations hence hurting the global performance. Another limitation is inherited from NeRF, the learned MLP will be overfitted and scene-specific without generality and takes an age to obtain a usable model. Considering future work, we may replace MLP with Voxel Grid from paper DVGO [17] and TiNeuVox [4] aiming to reach a faster overfitting speed.

5. Conclusion

We presented an approach for taking advantage of depth information for better constraining the optimization with improved quality of both color and depth observed. In summary, this approach outperforms the D-NeRF [12] baseline, with the introduction of depth supervision, the decrease in depth error is significant, which helps one to obtain a more meaningful geometric structure in space.

References

[1] Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew O'Toole. Törf: Time-of-flight radiance fields for dynamic scene view synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[2] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1

[3] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006. 1

[4] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 4

[5] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6639–6648, 2019. 1

[6] Haram Kim, Pyojin Kim, and H Jin Kim. Moving object detection for visual odometry in a dynamic environment based on occlusion accumulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8658–8664. IEEE, 2020. 3

[7] 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. In *ECCV*, 2020. 1

[8] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[9] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 1

[10] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[11] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[12] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 4

[13] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 1, 4

[14] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 1

[15] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 1

[16] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[17] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 4

[18] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 1

[19] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[20] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, 2021. 1

[21] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4