A Survey of NeRF for Sparse Views

Yuxuan Kuang
School of EECS, Peking University
kuangyuxuan@stu.pku.edu.cn

1. Introduction

Neural radiance fields (NeRF) [8] encode a scene into a neural representation that enables photo-realistic rendering of novel views. However, a successful reconstruction from RGB images requires a large number of input views taken under static conditions — typically up to a few hundred images for room-size scenes. Moreover, the speed of NeRF is disastrously slow because of the hundreds of views it needs to process. As a result, how to use NeRF for 3D reconstruction leveraging very few images is quite important. In this survey, I will introduce the variants of NeRF which are able to leverage sparse views — even single view to reconstruct 3D scenes.

2. Related Works

To tackle this tricky problem, many methods have been proposed. In Figure 1, selected methods are grouped into different perspectives that these methods are based on. For each perspective, a few works will be mentioned and their contributions will be explained.

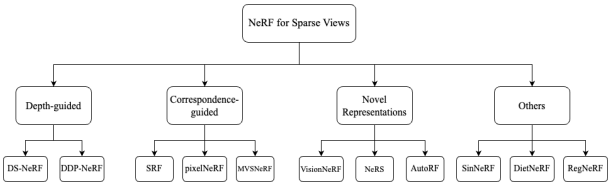

Figure 1. Taxonomy of NeRF for Sparse Views

2.1. Using Depth to Guide Reconstruction

It is very common to use depth information to guide 3D reconstruction, since in multi-view stereo, coarse 3D information can be easily obtained by Structure from Motion (SfM) [11].

In DS-NeRF [3], researchers use SfM to produce sparse 3D points that can be used as "free" depth supervision during training and add a loss to encourage the distribution of a ray's terminating depth matches a given 3D keypoint, incorporating depth uncertainty.

While in DDP-NeRF [13], researchers first utilize the sparse point cloud reconstructions from SfM preprocessing, which to feed into a depth completion network. They then impose these depth estimates as constraints to the NeRF optimization according to the estimated uncertainty.

Both of the works facilitate novel view synthesis results at significantly higher image quality and lower depth error compared to NeRF, and more importantly, they both use depth information to help constraint NeRF and achieve sparse-view reconstruction.

2.2. Leveraging Correspondences between 2D and 3D

Since we know NeRF is the "neural" bridge between 2D and 3D, then it is natural to use the correspondences between 2D and 3D to help the reconstruction.

In SRF [2], researchers predict color and density for each 3D point given an encoding of its stereo correspondence in the input images. The encoding is implicitly learned by an ensemble of pair-wise similarities — emulating classical stereo. Experiments show that SRF learns structure instead of over-fitting on a scene.

In [15], researchers propose pixelNeRF, a learning framework that enables predicting NeRFs from one or several images in a feed-forward manner. Unlike the original NeRF network, which does not make use of any image features, pixelNeRF takes spatial image features aligned to each pixel as an input. This image conditioning allows the framework to be trained on a set of multi-view images, where it can learn scene priors to perform view synthesis from one or few input views.

And in MVSNeRF [1], researchers leverage plane-swept cost volumes (widely used in multi-view stereo) for geometry-aware scene reasoning, and combines this with physically based volume rendering for neural radiance field reconstruction. They first construct a cost volume by warping 2D image features onto a plane sweep, then apply 3D CNN to reconstruct a neural encoding volume with per-voxel neural features, which they use to interpolate features to regress volume density and RGB radiance by an MLP.

2.3. Applying Novel Representations

NeRF [8] proposed a neural representation of 3D scenes, but it also has a few drawbacks, such as unable to capture the global structure of the scene, etc. A few works addressed these drawbacks by applying novel representations which are inspired by the original architecture of NeRF.

In VisionNeRF [7], researchers propose to leverage both the global and local features to form an expressive 3D representation. The global features are learned from a Vision Transformer [4], while the local features are extracted from a 2D convolutional network. Therefore, this novel 3D representation allows the network to reconstruct unseen regions without enforcing constraints like symmetry or canonical coordinate systems.

While NeRS [16] learns a neural shape representation of a closed surface that is diffeomorphic to a sphere, guaranteeing water-tight reconstructions. Moreover, surface parameterizations allow NeRS to learn (neural) bidirectional surface reflectance functions (BRDFs) that factorize view-dependent appearance into environmental illumination, diffuse color (albedo), and specular "shininess".

And in [9], AutoRF is proposed to learn a normalized, object-centric representation whose embedding describes and disentangles shape, appearance, and pose. Each encoding provides well-generalizable, compact information about the object of interest, which is decoded in a single-shot into a new target view, thus enabling novel view synthesis.

2.4. Others

Of course, apart from the above perspectives, there are methods that make use of other elaborate ideas to achieve sparse-view reconstruction.

The core idea of learning scene representation from sparse input is the idea of semi-supervised learning, where psuedo labels play an important role. In [14], researchers present a Single View NeRF (SinNeRF) framework consisting of thoughtfully designed semantic and geometry regularizations. Specifically, SinNeRF constructs a semi-supervised learning process, where geometry pseudo labels and semantic pseudo labels are introduced to guide the progressive training process.

Using semantic consistency is also a novel perspective to look at. In DietNeRF [5], an auxiliary semantic consistency loss is introduced to encourage realistic renderings at novel poses and allow it to get supervision from arbitrary poses. They leverage CLIP [12] vision encoder to get semantic similarities and use the consistency for supervision, thus improving the perceptual quality of few-shot view synthesis when learned from scratch.

Finally, in RegNeRF [10], researchers observe that the majority of artifacts in sparse input scenarios are caused by errors in the estimated scene geometry, and by divergent behavior at the start of training. Therefore they address this by regularizing the geometry and appearance of patches rendered from unobserved viewpoints, and annealing the ray sampling space during training. They additionally use a normalizing flow model to regularize the color of unobserved viewpoints.

Method	Taxonomy	Single View	PSNR↑
NeRF [8]	/	×	8 (Dense)
DS-NeRF [3]	DG	×	12.17
DDP-NeRF [13]	DG	×	/
SRF [2]	CG	×	15.68
pixelNeRF [15]	CG	✓	18.95
MVSNeRF [1]	CG	×	18.54
VisionNeRF [7]	NR	✓	/
NeRS [16]	NR	×	/
AutoRF [9]	NR	✓	/
SinNeRF [14]	O	✓	16.52
DietNeRF [5]	O	×	11.85
RegNeRF [10]	O	×	18.89

Table 1. Comparison of NeRF Variants. DG, CG, NR, O refer to depth-guided, correspondence-guided, novel representation, others respectively. PSNR is tested on DTU dataset [6] in 3 views.

3. Relationship between NeRF Variants

Although these variants are separated into different classes, they also have inner relations. Most of them are designed to reconstruct 3D scenes, whereas [2] and [9] are specifically designed to reconstruct 3D instances.

Also, some of them ([15], [7], [9], [14]) have the ability to reconstruct 3D scene only from one single view, making it extremely useful in real life and in computational photography in real production.

From above, we can see that the tricky issue mainly lies in how to effectively supervise the reconstruction since views are sparse. In the works above, depth, correspondence, psuedo labels are used, and novel representations are proposed to avoid the sparse supervision. It shows that to successfully train a NeRF, either we can add supervision to the system or we can use a more efficient way to model the problem, just like NeRF to light field.

In Table 1, the comparison between NeRF variants are summarized. Notice that although all the works claim that they achieve the SotA in this field, they still slightly differ in performance, which is evaluated by PSNR.

4. Future Development

As we can see from the above, NeRF variants are still in the early stage of development. There are still many problems to be solved, such as how to append more annotations to them, how to make the training process as fast as possible, etc. Maybe in the future, we can see more and more NeRF variants that can be used in real life.

References

[1] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.

[2] J. Chibane, A. Bansal, V. Lazova, and G. Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021.

[3] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[5] A. Jain, M. Tancik, and P. Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, October 2021.

[6] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014.

[7] K.-E. Lin, L. Yen-Chen, W.-S. Lai, T.-Y. Lin, Y.-C. Shih, and R. Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[9] N. Müller, A. Simonelli, L. Porzi, S. R. Bulò, M. Nießner, and P. Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[10] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. M. Sajjadi, A. Geiger, and N. Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[11] O. Ozyesil, V. Voroninski, R. Basri, and A. Singer. A survey of structure from motion, 2017.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[13] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, and M. 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)*, pages 12892–12901, June 2022.

[14] D. Xu, Y. Jiang, P. Wang, Z. Fan, H. Shi, and Z. Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. 2022.

[15] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021.

[16] J. Y. Zhang, G. Yang, S. Tulsiani, and D. Ramanan. NeRS: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *Conference on Neural Information Processing Systems*, 2021.