3D Reconstruction Approaches: A Comprehensive Review

Abstract

Vision systems have evolved into the norm and one of the most useful sensory assets in commercial and everyday applications during the past decade. Numerous applications tend to use unique vision systems in home, working, industrial, and other environments since vision offers a number of essential qualities. A vision system must effectively and robustly rebuild the working space and 3D surface in order to accomplish these objectives. One of the core problems in image-based modeling and computer vision is producing a 3D model from 2D photos that is as realistic as feasible. The best option is to recreate the scene automatically with minimal to no user input. There are currently several methods for 3D reconstruction from 2D images; each algorithm has its own execution conditions, strengths, and weaknesses. This study proposes a novel comparison of 3D reconstructions obtained using Artificial Intelligence, specifically NeRF Neural Networks, and Neural 3d Scene Reconstruction with Manhattan-world assumption and Finite Aperture Stereo: 3D Reconstruction for Macro Scale Scenes. Although the approaches are traditionally used with different objectives and in different contexts, they can also be used for similar purposes, such as 3D reconstruction.

Keywords

3D reconstruction, view synthesis, comparative analysis, photogrammetry, volume rendering, manhattan-world

1. Introduction

Three-dimensional reconstruction from visual data is a long-standing computer vision problem with real-world applications in fields such as robotic surgery, autonomous vehicles, and virtual and augmented reality. Three-dimensional object and surface reconstruction from images is also an important topic in various application areas, such as quality inspection, clinical photography, robotics, agriculture, and archaeology. In the domain of quality inspection, a large number of inspection tasks depend on three-dimensional reconstruction techniques, such as the surface measurement applied to high-precision engineered products such as aircraft wings. Tasks of this kind usually require the accurate measurement of depth on small surfaces. Other tasks depend on the precise measurement of a sparse set of well defined points, for example to determine if an assembly process has been completed with the required accuracy, or measurement of the relative movement between important parts during a crash test. Three-dimensional clinical photographs have the potential to provide quantitative measurements that reduce subjectivity in assessing the surface anatomy of the subject before and after a surgical intervention by providing numeric scores for the shape, symmetry and longitudinal change of anatomic structures. Furthermore, the vast majority of nowadays mobile robots are equipped with one, two or more cameras in order to provide visual feedback in applications like maze exploration, map navigation and obstacle avoidance. Besides, in the field of agriculture, new applications emerged recently, such as a mobile robotic system that uses cameras to reconstruct the surface of the plants to find parasites and report them. One more interesting application area includes the archaeological excavations and historic objects, where three-dimensional surface reconstruction is applied to many archaeological sites in order to preserve crucial details of the site and use them afterwards for 3D presentation and tourist attraction.

With encouraging outcomes, deep learning has recently supplanted the conventional computer vision algorithms for the challenge of 3D reconstruction. It has been demonstrated that deep learning networks are more resistant to noise and fluctuations in input data than conventional techniques. They also have the capacity to autonomously learn complex features from data. They are therefore ideally suited for

applications like this when the input photos vary greatly. Recently, a number of algorithms for 3D reconstruction of common imagery have been presented using learning-based techniques.

The process of producing three-dimensional (3D) representations of an object from a single or multiple images of the object is known as three-dimensional (3D) reconstruction of 2D images. It is the task of inferring a scene's geometric structure from a set of two-dimensional images. We want to know the 3D shape and position in space of all the objects in a scene given one or more 2D views of the scene. For each 2D point (x, y) in the image, we wish to recover the corresponding 3D point (x, y, z) in world coordinates, where z is the point's distance from the camera. This may be thought of as mapping the 2D points in the image space to the 3D points in real-world space. There are various versions of the 3D reconstruction problem in terms of input type and quantity of photos, such as 3D reconstruction from a single image (shape-from-X), multiple images (stereo), or 3D reconstruction from RGB-D data. Only the second and third types of problems namely, 3D reconstruction from numerous images are discussed in this study.

2. Related Work

MVS: Multi-View Stereo (MVS) is a well-known 3D reconstruction method that relies on feature congruence between a collection of images which are projections of the same 3D object from different angles as the primary indicator of 3D structure. For multi-view 3D reconstruction, several techniques use a two-stage pipeline that involves first estimating the depth map for each picture based on MVS and then performing depth fusion to get the final reconstruction results. Traditional MVS techniques, such as novel view synthesis, have been used because they can rebuild 3D shapes quite accurately.

Structure from Motion(SFM), a photogrammetric technique, is increasingly popular for producing high resolution mapping outputs (such as point clouds and orthoimages) from pictures captured with minimal resources, consumer-grade cameras with adequate end lap and sidelap. The automatic extraction of important features from the obtained images is the first step in processing. The multidimensional maximum likelihood of descriptors and the outlier rejection criteria are used to match the retrieved features in multidimensional descriptors like SIFT. In order to produce a sparse point cloud, the technique is followed by bundle adjustment, which simultaneously solves for the intrinsic and extrinsic orientation parameters of the camera. The intrinsic orientation (IO) parameters list the camera's primary point, focal length, and skew coefficient along with other optical details. and radial and tangential lens distortion coefficients. The extrinsic orientation (EO) parameters are the 3D position and orientation of the camera when the images were acquired.

COLMAP is widely regarded as one of the most effective conventional MVS methods[5]. It uses pinhole images as input and combines a structure from motion (SFM) calibration pipeline with a view dependent reconstruction pipeline to generate high-quality 3D models. 1) Feature detection and extraction, 2) Feature matching and geometric verification, and 3) Structure and motion reconstruction comprise the SFM sequential processing pipeline. Following that, the MVS implementation in COLMAP is used to generate the mesh. Multi-View Stereo (MVS) in COLMAP computes depth and/or normality information for each pixel in an image using the output of SFM. After combining the depth and normal maps from multiple 3D images, a dense point cloud of the scene is generated. Using the depth and normal information from the merged point cloud, algorithms such as Poisson surface reconstruction can recover the 3D surface geometry of the scene.

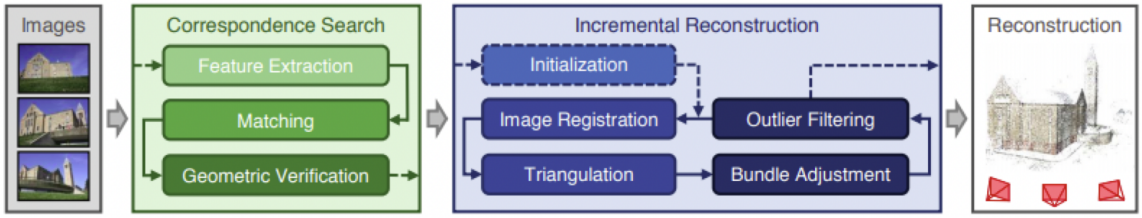

Figure 1: COLMAP an incremental SFM technique for efficient reconstruction[5]

Semantic segmentation: Semantic segmentation has made great progress recently because of learning-based techniques. To obtain pixel-level image semantic segmentation findings, FCN applies fully convolutional to the entire image[8]. In an effort to preserve the rich spatial information in the deep layers, recent approaches aim to aggregate high-resolution feature maps using a learnable decoder[9]. For large receptive fields, dilated convolutions are used in another line of research[12]. Many studies try to obtain semantic segmentation from 3D space in addition to 2D segmentation approaches[13]. They create networks to process various 3D data representations, such as point clouds and voxels.

Neural Implicit Representation: The representation of geometric information in neural networks in a parametrized manner is commonly referred to as neural implicit representation. Early attempts to create implicit representations in the field of 3D shape representation used deep networks that map coordinates to a signed distance function. It describes geometrical shapes, calculates the distance of point X from a surface's boundary, and determines whether a point is inside or outside the boundary. Early models, such as DeepSDF, were constrained by the need for ground truth 3D geometry, limiting these functions to only approximating synthetic scenes[7]. Working with real-world scenes presented a significant challenge in capturing "ground truth" geometry. Later work, such as NeRF, was able to operate by differentiable rendering functions that can efficiently employ 2D images to improve their representation because this prerequisite had been fixed. Niemeyer et al[10] 's depiction of surfaces as 3D occupancy fields and Sitzman et al[11] .'s use of a scene representation network are two specific examples of prior work (SRN). Scene representations in SRNs take the form of continuous functions that translate global coordinates into feature representations of regional scene characteristics. At each continuous 3D coordinate, this approach outputs an RGB color and a feature vector. This method of employing SRNs has the significant benefit that it can be trained from beginning to end using only 2D photos and associated camera postures, without access to depth information (or shape). However, these earlier methods had difficulty dealing with complex geometry and view-dependent lighting effects.

3. Overview

3.1 Nerf Theory:

Recent deep learning-based methodologies have significantly improved the development of unique view synthesis that is photorealistic. One of the most cutting-edge learning techniques now being used is called neural radiance fields (or NeRFs), which synthesizes new views of complicated scenes by optimizing an underlying continuous volumetric scene function using a limited number of input views. NeRF, short for non-convolutional fully connected deep network, is an implicit MLP-based model that maps 5D vectors—3D coordinates and 2D viewing directions—to output volume density and view dependent emitted radiance at that spatial location. It is computed by fitting the model to a set of training views. After that, the resulting 5D function can be utilized to create new perspectives using standard volume rendering techniques. The 5D scene representation is approximated with a deep fully- connected neural network, also known as multilayer perceptron (MLP): $F\Theta : (x, d) \rightarrow (c, \sigma)$, where Θ are the weights to be optimized.

The first component of the network predicts the volume density as a function of simply the position x in order to produce a multiview consistent representation. This portion of the network outputs a 256-dimensional feature vector in addition to having 8 fully connected layers, ReLU activations, and 256 channels per layer. The previously determined feature vector and the camera looking direction are concatenated in the second section of the network, where they are then transferred to a further fully-connected layer, where the view-dependent RGB color is the final output.

Broadly speaking, novel view synthesis using a trained NeRF model is as follows:

- For each pixel in the image being synthesized, march camera rays through the scene and generate a set of sampling points (see (a) in Fig. 1).
- For each sampling point, use the viewing direction and sampling location to extract local color and density, as computed by NeRF MLP(s) (see (b) in Fig. 1).
- Use volume rendering to produce the image from these colors and densities (see (c) in Fig. 1).

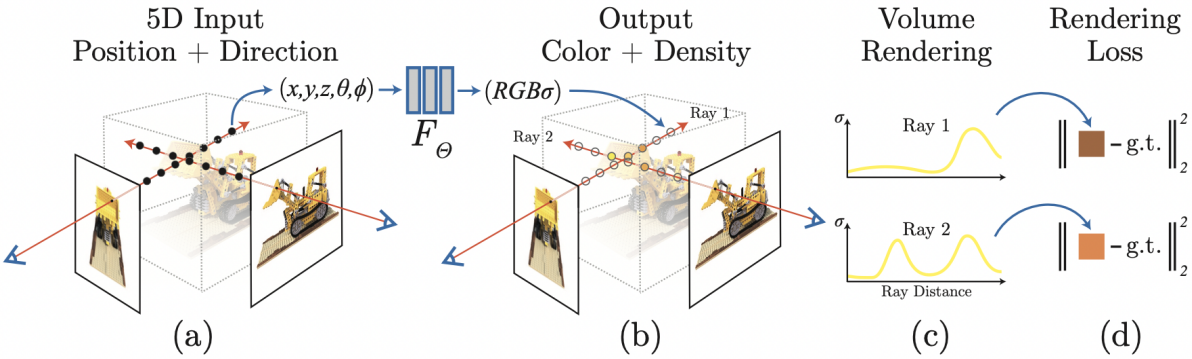

Figure 2: NeRF volume rendering and training process. Image sourced from [1]. (a) illustrates the selection of sampling points for individual pixels in a to-be-synthesized image. (b) illustrates the generation of densities and colors at the sampling points using NeRF MLP(s). (c) and (d) illustrate the generation of individual pixel color(s) using colors and densities along the associated camera ray(s) via volume rendering, and the comparison to ground truth pixel color(s), respectively.[1]

In more detail, given volume density and color functions, volume rendering is used to obtain the color C(r) of any camera ray r(t) = o + td, with camera position o and viewing direction d using where where T(t) is the accumulated transmittance, representing the probability that the ray travels from t_l to t without being intercepted. Also note that the model follows a pinhole camera assumption.

$$C(\mathbf{r}) = \int_{t_1}^{t_2} T(t) \cdot \sigma(\mathbf{r}(t)) \cdot \mathbf{c}(\mathbf{r}(t), \mathbf{d}) \cdot dt, \qquad (2)$$

For each pixel, a square error photometric loss is used to optimize the MLP parameters. NeRF models also employ positional encoding, to greatly improve fine detail reconstruction in the rendered views.

3.2 Finite Aperture Stereo:

Multi-view stereo is still a popular option when reconstructing 3D geometry despite performance varying greatly depending on the scene material. Additionally, pinhole cameras can only be used for larger scenes with diffuse reflectance due to the shallow depth of field that macro scale pictures inherently have. On the other hand, defocus blur may be considered a useful reconstruction signal in and of itself, especially when view-dependent components are present. In light of this, this study explored the relationship between stereo and defocus cues in the context of multi-view 3D reconstruction[2]. Additionally, a thorough pipeline that comprises steps for image creation, camera calibration, and reconstruction was proposed for scene modeling from a finite aperture camera. Studies have demonstrated how each cue helps to get improved performance observed over a range of complex materials and geometries. It is basically a hybrid approach to combine stereo and defocus in an MRF (Markov Random Field) framework.

MVS recovers 3D structure by identifying corresponding features from images of the scene taken at different viewpoints. Using geometric constraints arising from the pinhole camera model, 3D points can be triangulated from two or more of these features according to the pose of each view. Broadly speaking, the quality of reconstruction largely depends on three factors which include Scene Representation, Feature Matching and Regularisation. Depth from Defocus directly recovers the spatial scene structure using a monocular camera. The depth of the tracked feature points is calculated by measuring the amount of defocus, expressed e.g. by the standard deviation σ of the Gaussian-shaped point spread function (PSF) that blurs the image. By modeling the point spread function (PSF) of the camera, depth information of the scene can be leveraged from the formation of defocus on the image plane, depth information of the scene can be leveraged from the formation of defocus on the image plane.

3.3 Neural 3D Scene Reconstruction with Manhattan-World Assumption:

The objective of the paper is to solve for three-dimensional reconstruction of low texture, specular and reflective regions[3]. The key idea is to apply the Manhattan-World assumption which states that floors and walls of indoor scenes generally align with three dominant directions. Regularization is applied on implicit neural representations . The method consists of the following steps:

- An off-the-shelf two-dimensional semantic segmentation network is used to segment floors and walls of indoor scenes (DeepLabV3+)

- Geometric constraints are applied to these regions. Specifically, normal directions of floor regions are enforced to be straight up; $L_f(r) = |1 - n(x_r) \cdot n_f|$ A learnable normal is introduced and normal directions of wall regions. They are enforced such that they are either parallel or orthogonal with the learnable normal. $L_w(r) = \min |i - n(x_r) \cdot n_w|$

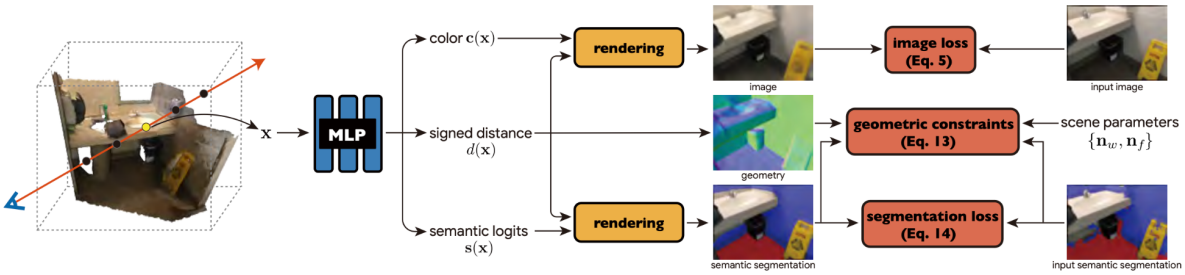

Figure 3. Overview of method: Using implicit brain representations, we learn the geometry, appearance, and semantics of 3D scenes. The pixel color and semantic probabilities of an image pixel are rendered using differentiable volume rendering and are supervised with input photos and semantic labels in 2D. We introduce geometric constraints in planar regions based on the Manhattan-world assumption, which enhances both the reconstruction and segmentation accuracy, in order to jointly maximize the geometry and semantics.[3]

Applying geometric constraints can significantly improve the reconstruction quality in most cases. However, two-dimensional segmentation results predicted by the network could be wrong in some regions. To solve this issue, the neural scene representations are augmented by additionally predicting semantic logic for each point in three-dimensional space. This is used to render two-dimensional segmentation under a particular camera view. As next steps in the pipeline, the geometric constraints are improved to joint optimization loss to optimize semantics together with geometry.

The joint optimization strategy can correct some pixels which are misclassified to planar regions by reducing the corresponding probabilities. A possible trivial solution is that both the probabilities of floor and wall vanish. In order to avoid this, the semantics are supervised using the cross entropy loss.

4. Comparisons

4.1 NerF and MVS based methods:

Compared with the state-of-the-art methods using the MVS and photogrammetric approach; the Finite Aperture stereo method clearly supersedes in multiple scenarios. The resulting geometry is much more stable and the output reconstructions are complete and consistent. The test cases in the study highlight that, in spite of limited quality of input values, NeRF networks return more views, while the MVS based methods return better mesh resolution[4]. On the contrary, it also appeared in studies that NeRF can be computationally expensive as opposed to MVS based methods.

4.2 NerF and Neural 3d reconstruction MW assumption:

The performance of both algorithms can vary given different contexts. This analysis has been observed from the lens of indoor scene reconstruction. The dataset used is ScanNet. The performance of NeRF is poor in comparison to [3] since the volume density representation has not sufficient constraint on geometry. However, other volume rendering techniques like VolSDF perform better in this scenario[14]. This is demonstrated in the following image.

Figure 7. Novel view synthesis results. We select novel views relatively far from training views for the qualitative comparison. Our method produces better rendering results compared to NeRF and VolSDF. Due to the lack of ground truth images in novel views, we render GT mesh in these views for reference.[3]

4.3 MVS and 3d reconstruction based methods:

As it eliminates reconstructed points that are inconsistent between different perspectives during the fusion step, COLMAP can reach extraordinarily high precision. However, recall is sacrificed in this method. The consistency of calculated depth maps, however, cannot be guaranteed even after optimization, leading to noisy reconstructions. The performance of MVS-based approaches is still subpar in this case. It should be highlighted that only the Manhattan-world assumption is taken into account in this paper. While most artificial scenes conform to this presumption, some situations call for a broader presumption, such as the Atlanta-world assumption. By altering the way geometric restrictions are expressed in the loss function, the proposed framework could be expanded to incorporate other suppositions.

5. Conclusion

We live in a 3D environment even if an image is a 2D array. Despite the fact that 3D reconstruction has been studied for almost 50 years, current advancements in the subject have only focused on very low-level reconstructions, including obtaining an accurate depth map, creating a 3D point cloud, or displaying scenes. We have shown that reconstruction is a task that includes all low-level, mid-level, and high-level representation, much like recognition.

In this paper, we provided an exploratory investigation of comparison of three quite distinct methodological strategies for producing 3D reconstructions from a small number of photos. We can infer that there isn't a single solution that works for all reconstruction issues. All of our methods rely on image datasets as inputs, whereas photogrammetry and multi-view stereo are focused on 3D reconstruction, NeRF networks are made to produce volumetric representations of the scene and produce novel, high-resolution photorealistic views of actual objects and scenes. However, using traditional computer vision methods, 3D reconstructions can also be obtained starting from the output of properly trained NeRF networks. In our analysis of the comparison's results, we've shown how NeRF networks can create a 3D reconstruction, although one of poor quality, even when photogrammetry techniques are unable to do so. Overall, the investigations have provided intriguing new information about our particular area of interest—3D reconstruction. In reality, it is well recognized that NeRF networks can be an alternate and complementary tool for the most challenging scenarios in situations where the quantity and quality of available pictures can be very restricted.

The third paper in our comparative analysis can be considered as a special case reconstruction problem. It aimed to solve the issue of poor performance in reconstruction of low-textured planar regions. This has been a recurring issue with the state-of -the-art approaches. As next steps, the constraints framework discussed in the approach can be integrated in reconstruction in problems which include indoor imagery.

6. Future Scope

The speed, quality, and training view requirements of NeRF models have significantly improved since the initial study was released, addressing all of the original model's flaws. NeRF models have been used for 3D reconstruction and view synthesis of human avatars and urban landscapes, as well as for urban mapping, modeling, and photogrammetry. Due to the active problem-solving and study of every component and issue, it is advised to pursue insightful follow-up work in the domain as a potential future strategy.

MVSNeRF is an intriguing area to investigate since our study clearly addressed the benefits and drawbacks of MVS and NeRF techniques[6]. They also used a pretrained CNN to extract 2D picture characteristics. Then, using plane sweeping and a variance-based cost, these 2D features were translated to a 3D voxelized cost volume. A 3D neural encoding volume was extracted using a pre-trained 3D CNN, and utilizing interpolation, per-point latent codes were produced. The NeRF MLP then created point density and color using these latent characteristics as input, point coordinate, and viewing direction, when conducting point sampling for volume rendering. The NeRF MLP and the 3D feature volume are jointly optimized during the training process. During analysis of the DTU dataset, within 15 minutes of training, MVSNeRF could achieve similar results to hours of baseline NeRF training.

References:

[1] 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 *European conference on computer vision*. Springer, 2020, pp. 405–421.

[2] M. Bailey, A. Hilton and J. -Y. Guillemaut, "Finite Aperture Stereo: 3D Reconstruction of Macro-Scale Scenes," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 2474-2484, doi: 10.1109/ICCVW54120.2021.00280.

[3] Guo, H., Peng, S., Lin, H., Wang, Q., Zhang, G., Bao, H., & Zhou, X. (2022). Neural 3D Scene Reconstruction with the Manhattan-world Assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5511-5520).

[4] Condorelli, F., Rinaudo, F., Salvadore, F., and Tagliaventi, S., “A Comparison Between 3d Reconstruction Using Nerf Neural Networks and Mvs Algorithms on Cultural Heritage Images”, ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 43B2, pp. 565–570, 2021. doi:10.5194/isprs-archives-XLIII-B2-2021-565-2021.

[5] Johannes L Schonberger and JanMichael Frahm. “Structure From Motion revisited”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4104–4113.

[6] Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., & Su, H. (2021, August 24). *MVSNERF: Fast generalizable radiance field reconstruction from multi-view stereo*. arXiv.org. Retrieved October 21, 2022, from https://arxiv.org/abs/2103.15595

[7] Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019, January 16). *DEEPSDF: Learning continuous signed distance functions for shape representation*.

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015

[9] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *T-PAMI*, 2017.

[10] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., & Geiger, A. (2019, April 30). *Occupancy networks: Learning 3D reconstruction in Function Space*.

[11] Sitzmann, V., Zollhöfer, M., & Wetzstein, G. (2020, January 28). *Scene representation networks: Continuous 3D-structure-aware neural scene representations*.

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous con- volution, and fully connected crfs. *T-PAMI*, 2017

[13]Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[14] Yariv, L., Gu, J., Kasten, Y., & Lipman, Y. (2021, December 1). *Volume rendering of neural implicit surfaces*. arXiv.org. Retrieved October 21, 2022, from https://arxiv.org/abs/2106.12052