Faithful, Generalizable, and Dynamic Neural Fields

Aditya Chetan

Neural Field (NF) [13, 15, 19] research has garnered increasing attention, leading to remarkable progress in 3D Shape Modeling. Neural Fields serve as invaluable tools for succinctly representing complex scenes [25]. Consider, for instance, a scenario where a user wishes to capture and distribute a reconstructed version of a 3D environment, intended for utilization by others in virtual reality headsets. In this context, sharing a compact NF like NeRF [7, 13] is much more memory efficient compared to bulky explicit representations like meshes and point clouds. This remarkable compressive capability has ignited a fascination with NFs for 3D shape modeling. However, it is worth noting that some important challenges still exist. While NFs represent the input signal (like a 3D shape) quite well, as noted by recent works [11, 12, 26], they are often not ***faithful***, failing to represent the underlying geometric properties like derivatives of the signal accurately. This limits their usage in downstream applications like physical simulation and rendering. Furthermore, obtaining useful high-fidelity reconstructions from sparse observations of the scene is still an elusive goal. To produce such reconstructions, we would need ***generalizable*** priors that require careful design. Lastly, modeling ***dynamic*** 3D scenes, especially from monocular videos is still largely unsolved, with past works making strong assumptions like human-centric scenes. Thus, my research endeavors center around the pursuit of ***faithful*** and ***generalizable*** representation techniques for ***dynamic scenes*** using Neural Fields.

1. My Prior Work

Faithfulness. Recently, Hybrid Neural Fields (HNFs) [6, 14, 16, 21] like Instant NGP have gained focus as an efficient technique for signal representation (like learning SDFs). These models typically consist of a small neural network along with a data structure that stores local features about the shape. The representation power of the neural network is distributed to the features in the data structure, allowing a smaller network with increased performance and similar quality of results. In order for HNFs to qualify as first-class shape representations, we should be able to utilize them directly in applications like rendering and physical simulation. These applications require accurate spatial differential operators (like gradient and laplacian) of the HNF which we found to be rife with high-frequency noise, as also highlighted recently by Neuralangelo [11] and illustrated in Figure 1. We reason that this is because HNFs contain small magnitudes of high-frequency noise, which gets accentuated when we compute spatial derivatives of the HNF using automatic differentiation (Autodiff).

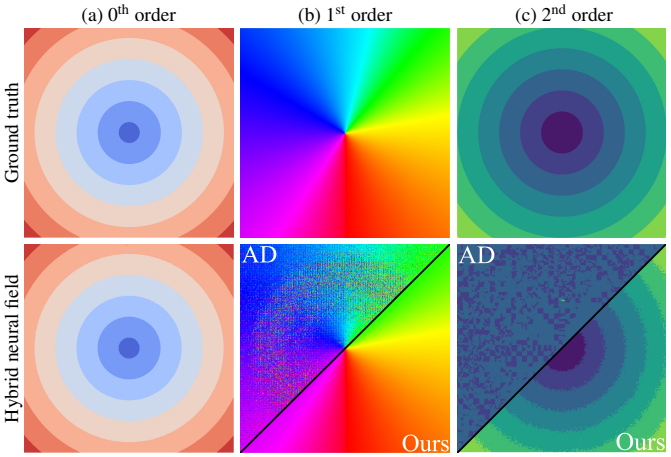

Figure 1: **Differential operators of HNFs.** Even accurate HNFs may provide inaccurate differential operators, as illustrated through the SDF of a circle in 2D. Note the noisy Autodiff operators and our more accurate approach.

To mitigate this, we propose to apply Autodiff to a *local low-order polynomial approximation* of the signal. Approximating the signal locally by a low-order polynomial gets rid of high-frequency noise while preserving fine-grained details. Our operator can be applied post hoc to any given pre-trained HNF. Apart from this post hoc operator, we also propose a *fine-tuning approach* that aligns the Autodiff gradients of the network with our post hoc operator. We showed how our approaches result in more accurate gradients and higher-order differential operators for any given pre-trained neural field compared to vanilla Autodiff or even finite-difference-based methods over a variety of shapes and provide benefits in applications like rendering, collision simulation, and solving PDEs using HNFs.[1]

2. Future Plans and Research Directions

Faithful NFs enable us to focus on more specific challenges of representing dynamic scenes and generalizability, which I plan to tackle in my ongoing and future projects. While scene representation using NFs has made a lot of progress in the past few years, it still struggles with generalization and modeling dynamic scenes with deformable objects. For instance, say we are recording a monocular video of a busy crosswalk in an urban setting. Such scenes would contain multiple moving objects, novel types of objects as well as unconstrained and complex movements, as shown in Figure 2. Existing methods struggle to model such dynamic scenes without strong assumptions typically on object categories that may not generalize in the wild. In my research, I aim to address these challenges with general priors that are capable of handling such scenes with general priors in a category-agnostic fashion.

[1] This work is currently under submission at ICLR 2024.

Figure 2: **Research Thrusts.** *Dynamic Scenes* (Left): Reconstructing *monocular dynamic* scenes with *multiple* moving objects (shaded in different colors), with *unconstrained* motions (arrows denote motions), and *generalizing* to novel object types (unseen trash can shape, circled) is still largely unsolved. Frames from video of a busy crosswalk. *Generalization* (Right): While past works can generalize well enough for static scenes to get a coarse reconstruction, they miss fine-grained details and glue distinct objects together (cars glued to the road). Figure from NKSR [10].

Dynamic Scenes. Reconstruction of dynamic 3D scenes from monocular videos is an extremely challenging task due to the under-constrained nature of the problem and occlusions in the scenes. Existing methods ranging from classical Non-Rigid Surface-from-Motion [4, 9, 22] to more recent NF-based methods [5, 18, 20, 28], rely on accurate tracking of 3D points in 2D, to learn a 3D deformation field between frames. However, in order to compute deformation fields, they either make strong assumptions about the deformable objects in the scene, such as assuming humanoid or quadruped structures [20, 28], or assume more supervision, like accurate depth maps [5]. This limits their usage on in-the-wild videos.

In my ongoing work, I am developing methods that assume general priors that are agnostic to object categories and still capable of providing 3D reconstructions. Recently, Omnimotion [23] has addressed this issue in part, using general priors with monocular videos to learn a global quasi-3D NF representation of the scene that provides occlusion-aware tracks. However, it does not assume access to camera poses, making geometry extraction difficult. Recent work [29] has shown that *cues from the stationary background* and an *initial monocular depth* can be used to compute a *consistent depth field* for dynamic scenes up to a global scaling factor. But [29] relies on foreground masks that can be inaccurate in case of occlusions. Currently, I am working on integrating occlusion-aware background masks from Omnimotion's tracks with monocular depth priors to learn a temporally consistent depth field. The resulting representation will give us a globally consistent depth field and geometry in visible regions, enabling applications like video editing and generalizing novel view synthesis to a wider range of views.

Generalization. Completion of non-visible regions or reconstruction from sparse observations using NFs requires some prior over 3D shapes that generative models can provide. However, 3D generative models, suffer from poor generalization capabilities, stemming from data scarcity. While self-supervised [8] and unsupervised [27] methods exist, they make strong assumptions about input data used for training such as points being sampled uniformly from the shapes' surface, availability of full-view point clouds, and so on. However, these assumptions are too strong to hold in practice leading to poor generalization results for unseen sampling densities and categories. While recent works like NKF [24] and NKSR [10] attempt to solve this problem, they are focused solely on reconstruction at the local scale (on the level of patches) [10] or global scale (on the level of the object or scene) [24]. Local-scale reconstruction is sensitive to incomplete sampling (common for example, in LiDAR scans of scenes with occlusions), whereas a completely global approach is sensitive to noisy sampling. Furthermore, current scene reconstruction approaches often group semantically distinct objects together as one surface. This can be a challenge for downstream applications like robotic manipulation where clean geometry is necessary for reasoning. In future work, I aim to explore *multi-scale approaches* for generalizable reconstruction that *incorporate priors from past videos* [17] of object interactions to reason about scene compositionality and surface separation.

3. Conclusion

With improvements in the quality of NFs as a scene representation, we are seeing a shift in how visual data is represented, with NFs increasingly being used as first-class shape representations [1, 2, 3]. My goal is to explore the properties of NFs to ensure that we can fully leverage them for visual data representation and understanding and address their limitations. More broadly, I believe that there is a wealth of general priors that can be utilized to develop a more holistic understanding of visual data. While my focus is primarily on 3D data representation using Neural Fields, I hope to produce studies that offer insights that are generally applicable for visual understanding and can be utilized for developing any system that seeks to utilize general priors in a self-supervised and label-efficient fashion.

References

[1] Instant nerf artists. https://www.nvidia.com/en-us/research/ai-art-gallery/instant-nerf/. Accessed: 2023-09-06.

[2] New ways maps is getting more immersive and sustainable. https://blog.google/products/maps/sustainable-immersive-maps-announcements/. Accessed: 2023-09-06.

[3] Reconstructing indoor spaces with nerf. https://blog.research.google/2023/06/reconstructing-indoor-spaces-with-nerf.html. Accessed: 2023-09-06.

[4] Akhter, I., Sheikh, Y., Khan, S., and Kanade, T. Trajectory space: A dual representation for nonrigid structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 7 (2011), 1442–1456.

[5] Cai, H., Feng, W., Feng, X., Wang, Y., and Zhang, J. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)* (2022).

[6] Chen, A., Xu, Z., Geiger, A., Yu, J., and Su, H. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)* (2022).

[7] Chen, Z., Funkhouser, T., Hedman, P., and Tagliasacchi, A. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *CVPR* (2023).

[8] Chou, G., Chugunov, I., and Heide, F. Gensdf: Two-stage learning of generalizable signed distance functions. In *Proc. of Neural Information Processing Systems (NeurIPS)* (2022).

[9] Fragkiadaki, K., Salas, M., Arbeláez, P., and Malik, J. Grouping-based low-rank trajectory completion and 3d reconstruction. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1* (Cambridge, MA, USA, 2014), NIPS'14, MIT Press, p. 55–63.

[10] Huang, J., Gojcic, Z., Atzmon, M., Litany, O., Fidler, S., and Williams, F. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023).

[11] Li, Z., Müller, T., Evans, A., Taylor, R. H., Unberath, M., Liu, M.-Y., and Lin, C.-H. Neuralangelo: High-fidelity neural surface reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).

[12] Mehta, I., Chandraker, M., and Ramamoorthi, R. A level set theory for neural implicit evolution under explicit flows. In *European Conference on Computer Vision* (2022), Springer, pp. 711–729.

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

[14] Müller, T., Evans, A., Schied, C., and Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph. 41*, 4 (July 2022), 102:1–102:15.

[15] Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[16] Sara Fridovich-Keil and Alex Yu, Tancik, M., Chen, Q., Recht, B., and Kanazawa, A. Plenoxels: Radiance fields without neural networks. In *CVPR* (2022).

[17] Sharma, P., Tewari, A. K., Du, Y., Zakharov, S., Ambrus, R., Gaidon, A., Freeman, W. T., Durand, F., Tenenbaum, J. B., and Sitzmann, V. Seeing 3d objects in a single image via self-supervised static-dynamic disentanglement. *ArXiv abs/2207.11232* (2022).

[18] Sidhu, V., Tretschk, E., Golyanik, V., Agudo, A., and Theobalt, C. Neural dense non-rigid structure from motion with latent space constraints. In *European Conference on Computer Vision (ECCV)* (2020).

[19] Sitzmann, V., Martel, J. N., Bergman, A. W., Lindell, D. B., and Wetzstein, G. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS* (2020).

[20] Song, C., Yang, G., Deng, K., Zhu, J.-Y., and Ramanan, D. Total-recon: Deformable scene reconstruction for embodied view synthesis. *arXiv* (2023).

[21] Takikawa, T., Litalien, J., Yin, K., Kreis, K., Loop, C., Nowrouzezahrai, D., Jacobson, A., McGuire, M., and Fidler, S. Neural geometric level of detail: Real-time rendering with implicit 3D shapes.

[22] Torresani, L., Hertzmann, A., and Bregler, C. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 5 (2008), 878–892.

[23] Wang, Q., Chang, Y.-Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., and Snavely, N. Tracking everything everywhere all at once. *arXiv:2306.05422* (2023).

[24] Williams, F., Gojcic, Z., Khamis, S., Zorin, D., Bruna, J., Fidler, S., and Litany, O. Neural fields as learnable kernels for 3d reconstruction, 2021.

[25] Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., and Sridhar, S. Neural fields in visual computing and beyond. *Computer Graphics Forum* (2022).

[26] Yang, G., Belongie, S., Hariharan, B., and Koltun, V. Geometry processing with neural fields. In *Thirty-Fifth Conference on Neural Information Processing Systems* (2021).

[27] Yang, G., Huang, X., Hao, Z., Liu, M.-Y., Belongie, S., and Hariharan, B. Pointflow: 3d point cloud generation with continuous normalizing flows. *arXiv* (2019).

[28] Yang, G., Vo, M., Neverova, N., Ramanan, D., Vedaldi, A., and Joo, H. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR* (2022).

[29] Yoon, J. S., Kim, K., Gallo, O., Park, H. S., and Kautz, J. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).