Optics EXPRESS

Space-time reconstruction for lensless imaging using implicit neural representations

TIFFANY CHIEN,[1,*] RUIMING CAO,[2] FANGLIN LINDA LIU,[1] LEYLA A. KABULI,[1] AND LAURA WALLER[1]

[1] *Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA*
[2] *Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA*
[*] *tiffany_chien@berkeley.edu*

Abstract: Many computational imaging inverse problems are challenged by noise, model mismatch, and other imperfections that decrease reconstruction quality. For data taken sequentially in time, instead of reconstructing each frame independently, space-time algorithms simultaneously reconstruct multiple frames, thereby taking advantage of temporal redundancy through space-time priors. This helps with denoising and provides improved reconstruction quality, but often requires significant computational and memory resources. Designing effective but flexible temporal priors is also challenging. Here, we propose using an implicit neural representation to model dynamics and act as a computationally tractable and flexible space-time prior. We demonstrate this approach on video captured with a lensless imager, DiffuserCam, and show improved reconstruction results and robustness to noise compared to frame-by-frame methods.

1. Introduction

Computational imaging systems often encounter bottlenecks due to challenging inverse problems that are ill-posed, underdetermined, or difficult to model accurately. Regularization can mitigate these challenges and constrain the solution space by biasing reconstructions towards certain types of objects, and most regularizers enforce spatial assumptions like sparsity or smoothness. However, relying heavily on these spatial priors can be tenuous, as they may not be consistent with the true structure of objects. The time domain offers an additional perspective to incorporate prior information and constrain solutions based on *temporal* dynamics. A vast range of real-world samples are dynamic in nature, and there is often prior knowledge available about their temporal characteristics. For example, a moving worm undergoes smooth deformable motion, a heart beats mostly periodically, and neurons fire transient spikes [1].

While algorithms that treat each time point independently neglect this valuable knowledge, space-time algorithms can take advantage of it to improve reconstruction quality. The main challenge for space-time algorithms is computational tractability, as solving for many time points simultaneously is inherently more challenging than solving for each one independently. Thus, a successful space-time algorithm must parametrically model the dynamics in a way that creates a tractable optimization landscape, while still maintaining flexibility to accommodate different kinds of motion. Existing approaches include solving for smooth deformation matrices between time points [2] and restricting the overall space-time matrix to be low-rank [3].

In this work, we replace matrix-based methods with an implicit neural representation (INR) to parameterize sample dynamics. INRs were first popularized as a representation of static scenes: instead of representing an image or 3D volume as a discrete grid of pixels or voxels, an INR is a neural network trained to take in (x, y) or (x, y, z) coordinates and output the value of a particular scene at that coordinate. The INR thus acts as an *implicit* and continuous representation of that particular scene. Because information from all parts of the scene are inherently mixed together, it also acts as an implicit spatial regularizer. It has also been proposed that INRs induce a more

#530480 https://doi.org/10.1364/OE.530480
Journal © 2024 Received 24 May 2024; revised 27 Aug 2024; accepted 28 Aug 2024; published 19 Sep 2024

convex loss landscape than explicit pixel grids which create many local minima [4]. These regularization and optimization properties allowed these INRs to achieve excellent reconstruction results on both computer vision problems and other computational imaging systems [5–10]. Like conventional reconstruction methods, an INR reconstructs the scene directly from the captured measurements without requiring a large training dataset.

In this work, we use INRs to represent dynamic scenes by taking in both space and time coordinates (x, y, t) as inputs, allowing for flexible implicit regularization in both space and time. Compared to static scenes, the memory compression benefits of INRs become crucial for optimizing large videos. Recent work has shown impressive results on a variety of dynamic inverse problems, including novel view synthesis [11,12], tomography [13], super-resolution microscopy [14,15], imaging through scattering [4], and MRI [16]. Existing work focuses on multi-shot imaging systems that take multiple partial measurements sequentially in time. This lowers their temporal resolution, but an INR that takes advantage of temporal redundancy can help recover that temporal information. In this work, we demonstrate that even for videos taken from single-shot imaging systems that capture a full measurement at every time point, jointly reconstructing space and time is still highly advantageous for denoising and improving reconstruction quality beyond methods that reconstruct one frame at a time.

We apply space-time INRs to DiffuserCam, a lensless imaging system consisting of only a thin diffuser placed close to a sensor (Fig. 1), which allows for highly compact photography

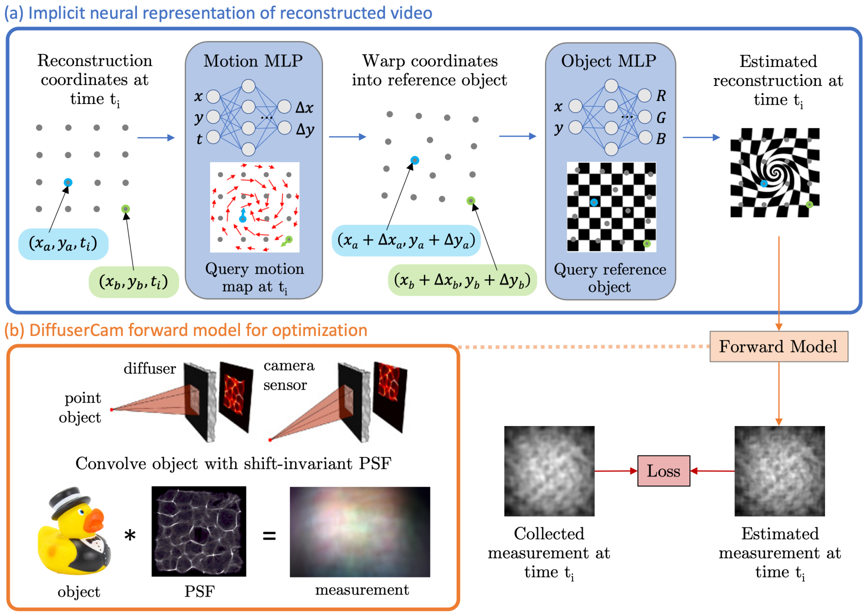
(a) Implicit neural representation of reconstructed video
Reconstruction coordinates at time t_i
Motion MLP
x
y
t
Δx
Δy
Warp coordinates into reference object
Object MLP
x
y
R
G
B
Estimated reconstruction at time t_i
(x_a, y_a, t_i)
(x_b, y_b, t_i)
Query motion map at t_i
$(x_a + \Delta x_a, y_a + \Delta y_a)$
$(x_b + \Delta x_b, y_b + \Delta y_b)$
Query reference object
(b) DiffuserCam forward model for optimization
Forward Model
diffuser
camera sensor
point object
Convolve object with shift-invariant PSF
*
=
object
PSF
measurement
Loss
Collected measurement at time t_i
Estimated measurement at time t_i

Fig. 1. Implicit neural representation (INR) for DiffuserCam lensless imaging video reconstruction. (a) The INR takes in coordinates (x, y, t) and outputs the value of the video at that coordinate. It is composed of two multi-layer perceptrons (MLPs): the Motion MLP represents the motion mapping of each video frame to a single reference object, which is modeled by the Object MLP. The networks are optimized (without any training data) by passing their estimated reconstruction through a physical forward model and comparing estimated measurements with collected measurements. Here we illustrate querying for the coordinates of the reconstruction at a single time t_i, and this optimization process is repeated for all time points. (b) DiffuserCam is a lensless imaging system that consists of a thin diffuser placed close to a camera sensor. We model DiffuserCam with a shift-invariant point spread function (PSF), so the captured measurement is modeled as a convolution of the object with a pre-calibrated PSF.

[17] and microscopy systems [18]. Unlike a lensed imaging system, DiffuserCam's multiplexed point spread function (PSF) allows higher-dimensional information to be extracted from 2D measurements through compressed sensing, such as 3D [17], high-speed video [19], and hyperspectral information [20]. In this work, we focus on the non-compressive case of reconstructing one 2D scene per 2D measurement. Although this situation is better conditioned than the 3D case, even 2D lensless imagers are plagued by reconstruction artifacts that limit their uptake in photography and scientific applications.

2. Methods

2.1. INR architecture and optimization

Inspired by related work on dynamic inverse problems, our INR consists of a series of two multi-layer perceptrons (MLPs) [11,12,14,15]. As shown in Fig. 1, the first network ("Motion MLP") takes in (x, y, t) and outputs a spatial displacement $(\Delta x, \Delta y)$ that represents how the location (x, y) moved at time t relative to the reference object. The second MLP ("Object MLP") represents the reference object by taking in $(x + \Delta x, y + \Delta y)$ and outputting the RGB values of that point in the reference object. This two-network decomposition constrains the temporal dynamics to nonrigid deformation, which does not apply to all scenes (see Fig. S2) but helps the optimization converge.

The optimization process for our INR is identical to a conventional gradient-based reconstruction algorithm, only with the estimated reconstruction parameterized by neural networks instead of an explicit pixel grid. As shown in Fig. 1, during each batch of optimization, all the (x, y) coordinates at a certain time t_i are passed into the INR, which outputs its estimated reconstructed frame at t_i. The estimated reconstruction is passed through the system's physical forward model (described below for DiffuserCam) to get an estimated measurement, and the mean squared error with the true measurement is computed. The gradient of this loss is then used to update the estimated reconstruction by updating the weights of the networks. This process is repeated for all time points. Just like conventional methods, there is no training data besides the measurements, and a new INR is optimized for each new reconstruction.

2.2. Forward model

As shown in Fig. 1(b), we model DiffuserCam with a shift-invariant point spread function (PSF), meaning that a point source at different locations in the scene will cast the same PSF, just shifted. This assumption allows the image y formed from a scene x to be modeled as a convolution with the PSF h: $y = x * h$. This model is computationally efficient and simple to calibrate by measuring a single PSF before imaging other scenes [17]. Replacing this forward model in the optimization pipeline allows our space-time INR to be easily adapted to other computational imaging systems (see Supplement 1 for a discussion of memory management for different forward models).

2.3. Implementation details

There are a number of implementation and hyperparameter choices to make when using INRs that allow us to adapt the spatial and temporal biases of the representation to different kinds of scenes. Because the effects of these choices (and their interactions with each other) can be difficult to predict or tune rigorously, here we will present the ones that we found to have the most significant impact on our results, along with our practical intuitive understanding of how they affect optimization. Other work applying space-time INRs to different systems also conduct similar empirical studies [14].

First, past work has shown that passing (x, y, t) coordinates directly as the inputs to the INR prevents it from learning to represent high-frequency content. One way to fix this is to pass the coordinates through some input encoding before being passed into the INR; we use the fixed sinusoidal input encoding from [21] that maps the coordinates to samples of sinusoids of different frequencies, and encode the time coordinate t before the Motion MLP and the spatial coordinates x and y before the Object MLP. One important hyperparameter is the maximum frequency of the sampled sinusoids in the encoding: higher values bias the INR to represent more higher frequencies, but also give it the degrees-of-freedom to overfit to noise. We find that this hyperparameter sometimes needs to be tuned depending on the complexity of the data. Figure 4 sweeps a range of maximum spatial frequencies, showing that lower values impose more dramatic smoothing, while higher values can become noisy. For our data, we did not find that changing the maximum temporal frequency had much of an effect, implying that the motion is quite smooth and relatively easy to solve.

Another important choice to make is how many frames of video to reconstruct at once with the INR. Having access to many frames theoretically gives the model more information to achieve strong denoising, but practically, it may struggle with solving for the dynamics in a longer video. The optimal point in this tradeoff will depend strongly on the level of noise (or other imperfections) in the data and the complexity of the dynamics to reconstruct. Figure 4 shows reconstructions of between 5 and 100 frames of video, showing that too few frames are still challenged by noise, while too many frames eventually exceed the model's ability to solve the dynamics.

3. Results

Figure 2 shows reconstruction results on DiffuserCam data simulated at different noise levels for a microscopic 25-frame video of a hydra waving its tentacles. The simulated measurements were

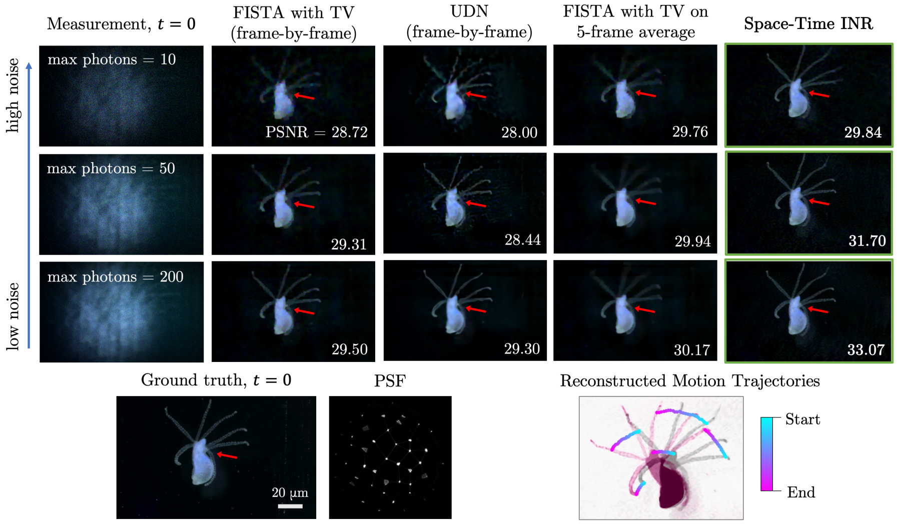

Fig. 2. Simulation results comparing reconstruction quality at different noise levels, showing one frame of a reconstructed video. Both frame-by-frame methods we compare to, FISTA and UDN, show perceptual artifacts at high noise levels, such as the loss or distortion of the tentacle shown by the red arrows. Averaging measurement frames before applying FISTA gives some robustness to noise, but incurs motion blur. Our space-time INR reconstructs all frames of the video at once, resulting in strong robustness to noise without sacrificing temporal resolution. See Visualization 1 for full video comparisons. Another advantage of our space-time INR is the ability to easily retrieve the full motion trajectory of any point directly from the Motion MLP, as shown in the bottom right (colored points represent different time points, plotted on the overlaid first (gray) and last (magenta) frames).

generated by convolving with an experimentally captured PSF from a random microlens phase mask [17,22]. Poisson noise was added with varying maximum photon counts that correspond to realistic experimental conditions in fluorescence microscopy. Figure 3 shows experimental results from a photography-scale DiffuserCam setup, with a toy duck moving closer to the camera over the course of 10 frames and a hand waving rigidly over 20 frames.

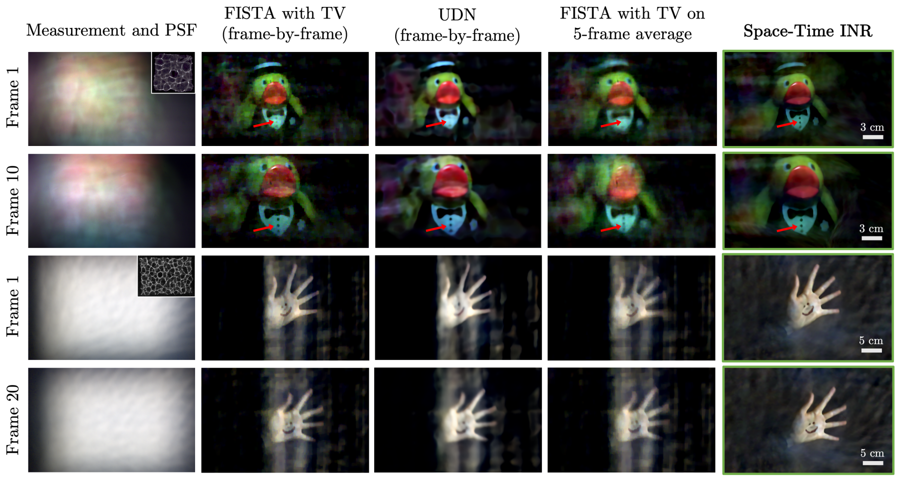

Fig. 3. Experimental results on a photography-scale DiffuserCam setup with a rubber duck moving towards the camera over time and a waving hand. The first and last frames of each reconstructed video are shown. The perceptual quality of our space-time INR is significantly better than both frame-by-frame methods, with both smooth blocks of color and well-resolved fine features such as the buttons on the duck's shirt (shown by red arrows). Our method also removes some but not all background artifacts, such as the vertical streaks in the hand data. Like in Fig. 2, averaging can help frame-by-frame methods at the cost of motion blur. In the duck video, the blur is particularly pronounced in the later frames, and in the hand video, the fingers are slightly blurred throughout; our method does not sacrifice temporal resolution to combine information from multiple frames. See Visualization 2 for a full video reconstruction comparison.

We compare with two frame-by-frame algorithms that utilize different spatial regularizers. First, we apply the fast iterative shrinkage-thresholding (FISTA) algorithm with 2D total variation (TV) regularization, a conventional pixel grid-based algorithm [23]. We also compare to an untrained deep network (UDN), which represents the reconstruction with a convolutional neural network [24]. More comparisons are shown in Fig. S1.

Figure 2 shows in simulation that at high noise levels, both frame-by-frame methods suffer from major artifacts, such as loss or distortion of the hydra's tentacles. In experiment (Fig. 3), the two frame-by-frame methods also struggle, with FISTA looking more grainy and UDN tending to blur out fine features. On the other hand, our space-time INR reconstructs both smooth blocks of color and well-resolved fine features under challenging simulated and experimental conditions.

We also compare our method to simply averaging measurement frames to suppress noise before applying frame-by-frame methods. As expected, this reduces noise artifacts, but comes at the cost of motion blur for dynamic objects: in Fig. 2, for example, the hydra's static body looks better with averaging, but the moving tentacles are blurred, while in Fig. 3, because the duck makes a large movement in the later frames, its face gets blurred out. Our space-time method achieves the benefits of averaging without trading off temporal resolution.

Our method achieves these improvements in reconstruction quality while remaining computationally tractable in part because of the compressive properties of INRs, which allow us to use relatively small networks: our Motion MLP has two 64-unit layers, and our Object MLP has eight 256-unit layers, giving a total of about 48k learnable parameters. Compared to the number of explicit RGB values reconstructed, this gives a compression ratio of about 10-20x for the data shown in Fig. 2 and 3. The results shown here took 1-3 hours to optimize on a NVIDIA GeForce RTX 3090 GPU. Recent work on view synthesis has found that using a learned input encoding instead of the fixed sinusoidal encoding we use here can substantially speed up INR optimization [25,26].

4. Discussion and conclusion

In this work, we have demonstrated a computationally tractable approach to space-time reconstruction using an implicit neural representation, successfully taking advantage of information from all frames of a captured video to improve reconstruction quality. Our method relies on the space-time prior of an implicit neural representation to model dynamics and regularize the optimization landscape. While our method can be adapted to many inverse problems, we test it on 2D DiffuserCam, and in future work we hope to extend it to the compressive higher-dimensional applications where lensless imaging shows unique advantages. In such cases, a strong prior is even more important for high reconstruction quality.

Our method uses no explicit regularizers, relying exclusively on the implicit priors induced by the structure and optimization of the neural networks. This raises the open question of what those priors are exactly—what types of scenes and motion is the INR biased towards representing? Based on our experience, this prior broadly encourages smoothness and continuity in space and time. One avenue to understand its biases further is by looking at failure cases (see Supplement 1 for a few examples). From that experience, we would describe the INR's aesthetic preferences to be: in space, it seems to prefer smooth distortions over graininess or blockiness (which are common with traditional regularizers, e.g., TV), and in time it seems to prefer highly deformable motion over rigid body motion. Figure 4 shows empirically that adjusting the hyperparameters of the model gives us some control over its regularization choices. Existing theoretical analysis of

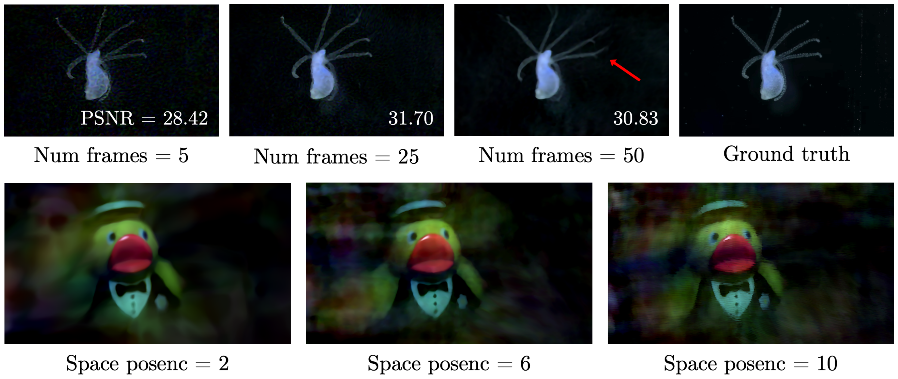

Fig. 4. Tuning the hyperparameters of the INR can adapt its spatial and temporal behavior to different data. Results shown on simulated data from Fig. 2 at noise level = 50 max photons. First row: one way to adapt to different time dynamics: the number of video frames chosen for the INR to jointly reconstruct must trade off having more redundant information for denoising versus having too much motion for the INR to solve successfully (red arrow shows failed motion registration). Second row: one way to adapt to different spatial characteristics: the maximum frequency of the sinusoidal encoding on the spatial coordinates affects the INR's representation of spatial frequencies, with lower values smoothing out details and higher values overfitting to noise.

spatial INRs shows bias towards lower spatial frequencies [21], and extending this analysis to the temporal domain could help reveal innate biases of space-time INRs towards certain dynamics.

Although our method relies on neural networks, it does not learn from a training dataset like supervised machine learning, and is conceptually more closely related to conventional gradient-based reconstruction algorithms that optimize directly from the captured measurements. While past work shows that supervised learning can achieve high-quality and faster results at test-time than methods like ours that learn from scratch for every new reconstruction [27], these approaches rely on large datasets with ground truth, which are hard to gather for computational imaging problems, and furthermore, the question of generalization to novel data remains a major open problem in machine learning. Ultimately, the tradeoffs between these methods will likely depend on the problem at hand.

In conclusion, space-time methods are a powerful avenue for improving reconstruction quality when we approach the limits of spatial information, and we have found implicit neural representations to be a flexible, practical, and promising approach.

Funding. National Science Foundation (DGE 2146752, DMR 1548924).

Acknowledgments. The authors would like to thank Kristina Monakhova for providing experimental data and helpful discussions, Clara Hung for helping to build the experimental setup, and Henry Pinkard for support writing the manuscript. This work was supported by STROBE: A National Science Foundation Science & Technology Center under Grant No. DMR 1548924. The work of Leyla A. Kabuli was supported by the National Science Foundation's Graduate Research Fellowship under Grant DGE 2146752. Laura Waller is a Chan-Zuckerberg Biohub SF Investigator.

Disclosures. The authors declare no conflicts of interest.

Data availability. Experimental data is available from [24].

Supplemental document. See Supplement 1 for supporting content.

References

1. E. A. Pnevmatikakis, D. Soudry, Y. Gao, *et al.*, "Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data," Neuron **89**(2), 285–299 (2016).
2. G. Zang, R. Idoughi, R. Tao, *et al.*, "Space-time tomography for continuously deforming objects," ACM Trans. Graph. **37**(4), 1–14 (2018).
3. F. Ong, X. Zhu, J. Y. Cheng, *et al.*, "Extreme MRI: Large-scale volumetric dynamic imaging from continuous non-gated acquisitions," Magn. Reson. Med. **84**(4), 1763–1780 (2020).
4. B. Y. Feng, H. Guo, M. Xie, *et al.*, "NeuWS: Neural wavefront shaping for guidestar-free imaging through static and dynamic scattering media," Sci. Adv. **9**(26), eadg4671 (2023).
5. V. Sitzmann, J. N. Martel, A. W. Bergman, *et al.*, "Implicit Neural Representations with Periodic Activation Functions," in *Proc. NeurIPS* (2020).
6. B. Mildenhall, P. P. Srinivasan, M. Tancik, *et al.*, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *ECCV* (2020).
7. H. Zhou, B. Y. Feng, H. Guo, *et al.*, "Fourier ptychographic microscopy image stack reconstruction using implicit neural representations," Optica **10**(12), 1679–1687 (2023).
8. Y. Sun, J. Liu, M. Xie, *et al.*, "CoIL: Coordinate-Based Internal Learning for Tomographic Imaging," IEEE Trans. Comput. Imaging **7**, 1400–1412 (2021). Conference Name: IEEE Transactions on Computational Imaging.
9. G. Zang, R. Idoughi, R. Li, *et al.*, "IntraTomo: Self-supervised Learning-based Tomography via Sinogram Synthesis and Prediction," in *IEEE/CVF International Conference on Computer Vision* (2021), pp. 1940–1950.
10. R. Liu, Y. Sun, J. Zhu, *et al.*, "Recovery of continuous 3D refractive index maps from discrete intensity-only measurements using neural fields," Nat. Mach. Intell. **4**(9), 781–791 (2022). Publisher: Nature Publishing Group.
11. A. Pumarola, E. Corona, G. Pons-Moll, *et al.*, "D-NeRF: Neural Radiance Fields for Dynamic Scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 10313–10322.
12. K. Park, U. Sinha, J. T. Barron, *et al.*, "Nerfies: Deformable Neural Radiance Fields," in *IEEE/CVF International Conference on Computer Vision* (2021), pp. 5845–5854.
13. T. Chien, C. Ophus, and L. Waller, "Space-Time Implicit Neural Representations for Atomic Electron Tomography on Dynamic Samples," in *NeurIPS Workshop on Deep Learning and Inverse Problems* (2023).
14. R. Cao, F. L. Liu, L.-H. Yeh, *et al.*, "Dynamic Structured Illumination Microscopy with a Neural Space-time Model," in *International Conference on Computational Photography* (IEEE, 2022), pp. 1–12.
15. R. Cao, N. Divekar, J. Nuñez, *et al.*, "Neural space-time model for dynamic scene recovery in multi-shot computational imaging systems" (2024), pp. 2024.01.16.575950, Section: New Results.
16. J. F. Kunz, S. Ruschke, and R. Heckel, "Implicit Neural Networks with Fourier-Feature Inputs for Free-breathing Cardiac MRI Reconstruction," ArXiv (2024).

Optics EXPRESS

17. N. Antipa, G. Kuo, R. Heckel, *et al.*, "DiffuserCam: lensless single-exposure 3D imaging," Optica **5**(1), 1–9 (2018).
18. G. Kuo, F. L. Liu, I. Grossrubatscher, *et al.*, "On-chip fluorescence microscopy with a random microlens diffuser," Opt. Express **28**(6), 8384–8399 (2020).
19. N. Antipa, P. Oare, E. Bostan, *et al.*, "Video from Stills: Lensless Imaging with Rolling Shutter," in *International Conference on Computational Photography* (IEEE, 2019), pp. 1–8.
20. K. Monakhova, K. Yanny, N. Aggarwal, *et al.*, "Spectral DiffuserCam: lensless snapshot hyperspectral imaging with a spectral filter array," Optica **7**(10), 1298–1307 (2020).
21. M. Tancik, P. Srinivasan, B. Mildenhall, *et al.*, "Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains," in *Advances in Neural Information Processing Systems* (2020), pp. 7537–7547.
22. F. Linda Liu, G. Kuo, N. Antipa, *et al.*, "Fourier DiffuserScope: single-shot 3D Fourier light field microscopy with a diffuser," Opt. Express **28**(20), 28969 (2020).
23. A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," SIAM J. Imaging Sci. **2**(1), 183–202 (2009).
24. K. Monakhova, V. Tran, G. Kuo, *et al.*, "Untrained networks for compressive lensless photography," Opt. Express **29**(13), 20913 (2021).
25. S. Fridovich-Keil, G. Meanti, F. R. Warburg, *et al.*, "K-Planes: Explicit Radiance Fields in Space, Time, and Appearance," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), pp. 12479–12488.
26. T. Müller, A. Evans, C. Schied, *et al.*, "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding," ACM Trans. Graph. **41**(4), 1–15 (2022). Place: New York, NY, USA Publisher: ACM.
27. K. Monakhova, J. Yurtsever, G. Kuo, *et al.*, "Learned reconstructions for practical mask-based lensless imaging," Opt. Express **27**(20), 28075–28090 (2019).