IMPLICIT NEURAL REPRESENTATIONS (INRS) FOR AUDIO DENOISING

Aniruddh Bansal, Purva Chiniya, Sukriti Paul
Project and Course Advisor: Prof. Ming Lin

ABSTRACT

We investigate the use of Implicit Neural Representations (INR) for denoising audio signals corrupted by Gaussian noise. By leveraging frameworks such as HyperSound and SIREN, we demonstrate the capacity of INR to encode and reconstruct clean audio signals. Our findings show that HyperSound effectively reduces both high-frequency and low-frequency components in noise, significantly improving speech intelligibility (STOI), background noise quality (CBAK), and perceptual evaluation of speech quality (PESQ). In our experiments, HyperSound not only removes high-frequency noise but also mitigates low-frequency noise components. While SIREN exhibits some denoising ability, it is less consistent and effective, particularly in terms of signal-to-noise ratio (SSNR). Overall, the research highlights the potential of INR for robust audio denoising, with HyperSound showing superior performance in maintaining audio clarity and quality. [1]

1 INTRODUCTION

Implicit Neural Representations (INRs) are coordinate-based models of multimedia signals that use neural networks to represent signals. INRs allow the signal to be resampled at a arbitrary frequency while keeping memory requirements constant, irrespective of the spatial resolution. INRs have shown significant potential in applications like super-resolution, compression, and 3D rendering. However, their application in the audio domain is restricted to compression and representation. In this work, we explore the paradigm of using INRs not just for signal reconstruction, but also for auxiliary tasks like denoising.

We explore two broad techniques for audio-INRs. A direct signal overfitting approach where each signal is modelled using a small network Sitzmann et al. (2020) and a hypernetwork based method Szatkowski et al. (2023) which relies on directly predicting network parameters for a given audio.

2 RELATED WORKS

2.1 INRS

Neural Networks as implicit functions for signal parameterization find early inspiration in learning resolution-independent 3D shape templates Genova et al. (2019), utilizing the implicit network bias for template learning as network models weights. Sitzmann et al. (2020) utilized periodic activation functions to improve learning for finer-level signal representation, including images, video, 3D shapes, audio, wavefields. The implicit learning inspired use of INRs for reconstructing denoised images Kim et al. (2022a), leveraging their inductive ability for zero-shot image reconstruction from background noise induced image sample.

2.2 INRS FOR AUDIO

Implicit Neural Representation hold potential to represent audio-data as shown in Sitzmann et al. (2020). Zuiderveld et al. (2021) utilize INRs conditioned as backbones for audio generation leverag-

[1] project code: `https://github.com/sukritipaul5/AudioDenoise-INR`

ing resolution-independent learning requirements. Meta-learning for generalizing INRs Szatkowski et al. (2023) improves audio representation for accurate speech synthesis. Learning compressed audio representation using lightweight INRs Lanzendörfer & Wattenhofer (2023), also solves the problem of added network noise by applying siamese based network to obtain twin model outputs and signal subtraction for noise removal. INRs for audio super-resolution Kim et al. (2022b) promise scale-invariant generation hence enabling lightweight INRs to locally parameterize audio chunks over time along with auto-regressive latent representation.

Our key contribution involves the extension of INRs to Audio Denoising. We pioneer the application of INRs for denoising audio signals, demonstrating their capability to effectively reconstruct clean audio from noisy inputs using frameworks such as HyperSound and SIREN. Through our experiments, we show that HyperSound effectively reduces both high-frequency and low-frequency noise components, thereby facilitating robust audio denoising.

3 Methodology

In this section, we detail the experiments conducted to evaluate the performance of INRs for audio denoising. We specifically focus on two approaches: the HyperSound approach Szatkowski et al. (2023) and the SIREN Sitzmann et al. (2020) approach.

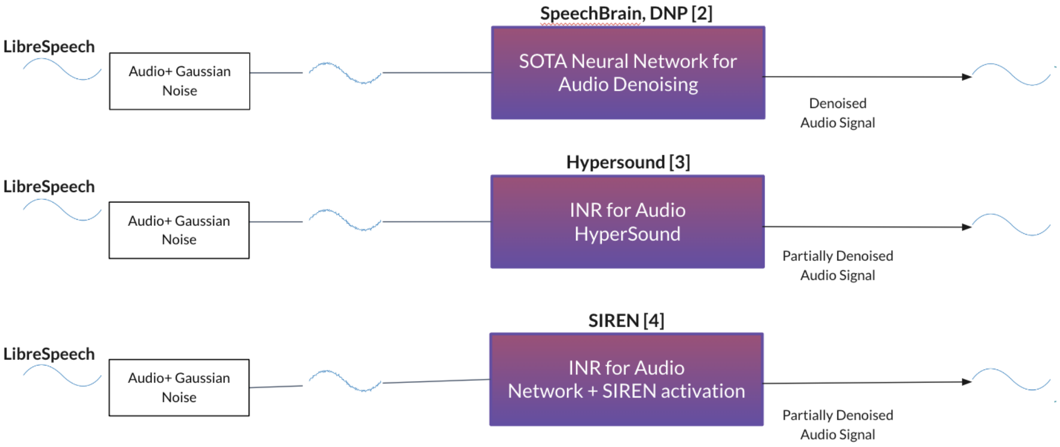

Figure 1: Audio Denoising workflows on our baselines (SpeechBrain, DNP), HyperSound, and SIREN

3.1 SIREN approach

In Sitzmann et al. (2020), the authors consider the problem of training a small neural network to represent the underlying signal as a function of its “implicit” coordinates. Basically, the network learns a functional mapping between the coordinate space and the signal space.

$$f_\theta : X \to Y$$

where f is a neural network with sinusoidal activation that maps coordinate space X to signal values Y. Sinusoidal or periodic activation is essential to construct a continuous functional representation.

3.2 HyperSound approach

The core idea of HyperSound is to use a hypernetwork to produce weights for a smaller target network that serves as the INR for the given audio signal. The hypernetwork, which is trained on a diverse set of noisy audio samples, can generalize to produce effective denoising INRs for new, unseen audio signals. Training the HyperSound model involves optimizing the hypernetwork to minimize the reconstruction error between the denoised output and the ground truth clean audio.

The loss function combines time-domain and frequency-domain components to ensure perceptual quality:

$$L(x, \hat{x}) = \lambda_{\text{SL1}} \cdot L_{\text{SL1}}(x, \hat{x}) + \lambda_{\text{STFT}} \cdot L_{\text{STFT}}(x, \hat{x})$$

Smooth L1 Loss (L_{SL1}): A robust loss function that penalizes large errors less than the standard L1 loss, helping to stabilize training.

STFT Loss (L_{STFT}): Multi-resolution Short-Time Fourier Transform loss that ensures the reconstructed audio retains its perceptual quality across different frequencies. This loss is particularly effective at reducing high-frequency noise, which is critical for maintaining the clarity and quality of the denoised audio.

4 Experiments

4.1 baselines

The SpeechBrain toolkit Ravanelli et al. (2021) is an open-source conversational AI toolkit by HuggingFace that includes state-of-the-art models for both speech and text processing. We use the SepFormer model Subakan et al. (2021), which employs a transformer-based approach to capture both short-term and long-term dependencies. This model demonstrates robust performance, even when the encoded representation is downsampled by a factor of 8.

The Denoising Neural Process (DNP) model Michelashvili & Wolf (2019) offers a completely unsupervised approach to audio denoising. It mirrors the training paradigm of INRs by training a neural network to fit the signal implicitly, thereby eliminating the need for clean audio data. The DNP model accumulates a fitting score for each time-frequency bin and applies time-frequency domain filtering based on these scores.

4.2 Observations

We test the reconstruction quality of our model on the LibriSpeech Panayotov et al. (2015) audio samples. We have used 40 randomly selected audio samples for over 60 INR experiments.

4.2.1 MelSpec Analysis

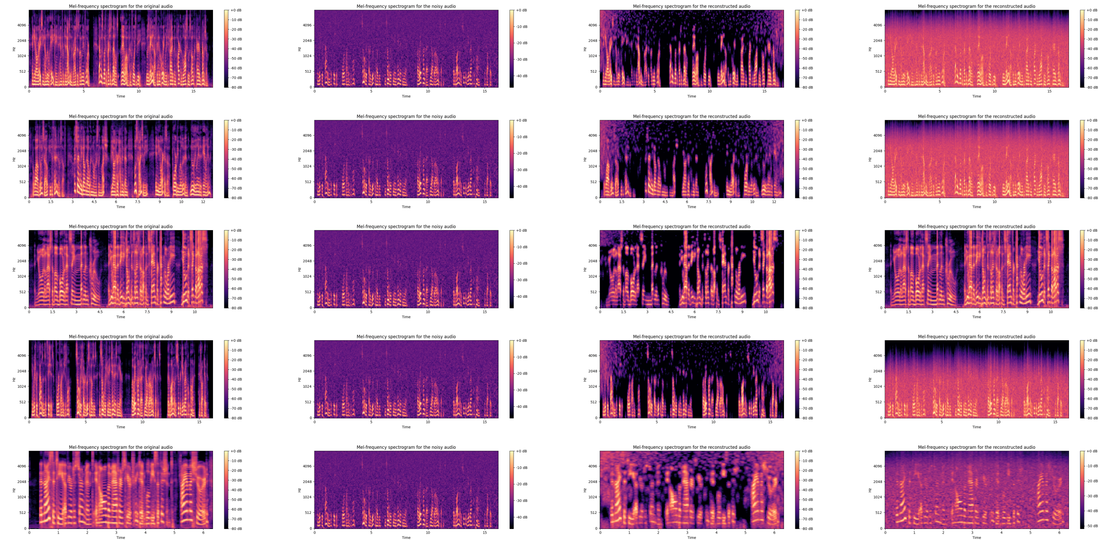

Figure 2: Columns 1 and 2 represent MelSpec for the clean and noisy audio samples, while columns 3 and 4 display MelSpec for the HyperSound and SIREN reconstructions, respectively.

The MelSpec analysis in Figure 2 reveals that the INR frameworks are able to effectively reduce the Gaussian noise while preserving the original audio characteristics. In the original MelSpec, energy bands are concentrated in specific frequency ranges corresponding to the harmonic and formant structures of the audio signal. The noisy spectrogram shows a more dispersed energy distribution due to the added Gaussian noise, resulting in blurred harmonic structures and less distinct temporal patterns. The reconstructed MelSpec for the reconstructed audio via HyperSound closely resembles the original, with energy bands more focused and noise significantly reduced. This indicates that the reconstructed audio has clearer temporal patterns and distinct harmonic structures. The energy bands in the SIREN MelSpec are concentrated more clearly than in the noisy spectrogram, but they appear slightly less focused compared to the HyperSound spectrogram. This indicates that while SIREN reduces noise, it may not preserve the energy bands as effectively as HyperSound. The spectrograms corroborate the performance metrics, confirming SIREN and HyperSound as enablers of denoising and the superiority of HyperSound in audio denoising.

4.2.2 Amplitude vs. Index Graph

The amplitude vs. index graph for the SIREN and HyperSound in Figure 3 indicate that both frameworks reduce Gaussian noise in the signal. HyperSound shows a smoother reconstructed speech pattern with fewer residual noise components, indicating better noise reduction and signal preservation capabilities. SIREN, while effective, may have minor residual noise, resulting in less stable amplitude distribution compared to HyperSound and more signal spikes.

4.2.3 Quantitative Metrics

	SSNR(↑)	**STOI(↑)**	**COVL(↑)**	**CBAK(↑)**	**PESQ(↑)**	**CSIG(↑)**
Noisy Audio	-3.820	0.813	1.000	1.560	1.024	1.000
SpeechBrain	**3.500**	**0.914**	**1.777**	1.619	**1.359**	**2.282**
DNP	-2.518	0.757	1.000	**1.703**	1.034	1.000
SIREN	-3.910	0.803	1.000	1.539	1.035	**1.095**
HyperSound	**0.195**	**0.828**	1.000	**1.647**	**1.081**	1.000

Table 1: Performance Comparison of Audio Denoising Methods Across Various Metrics. The first row references values for Noisy Audio vs Clean Audio followed by baselines (SpeechBrain, DNP.) Among INRs (SIREN, Hypersound): Hypersound performs better for all but COVL and CSIG.

We observe distinct differences in the performance metrics of SIREN and HyperSound. HyperSound demonstrates superior performance in intelligibility (STOI), background noise reduction (CBAK), and perceptual quality (PESQ), with average scores of 0.8279, 1.6465, and 1.0813, respectively. These metrics indicate that HyperSound more effectively reduces background noise and enhances the overall perceptual quality of the audio. Conversely, while SIREN shows a marginally higher signal distortion quality (CSIG) score of 1.095, it has a lower SSNR of -3.910, indicating poorer performance in signal-to-noise ratio compared to HyperSound's score of 0.1948. This suggests that HyperSound not only preserves the integrity of the original signal better but also provides a more balanced and effective denoising performance overall, making it the more robust choice for audio denoising tasks. SIREN, albeit not as good as Hypersound, does show some denoising ability, evidenced by improvements in STOI, CBAK, and PESQ scores.

Overall, HyperSound appears to provide a more precise reconstruction of the original audio signal.

5 Conclusion

Our experiments demonstrate that INRs can effectively be used for audio denoising. Specifically, the HyperSound approach, which utilizes hypernetworks to dynamically generate INRs, showed significant improvements in key audio quality metrics such as SSNR, STOI, and PESQ. The HyperSound model was able to effectively reduce noise while preserving the clarity and intelligibility of the audio signals.

The SIREN approach, which employs sinusoidal activation functions, also demonstrated improvements in denoising performance. However, it introduced some high-frequency noise artifacts that

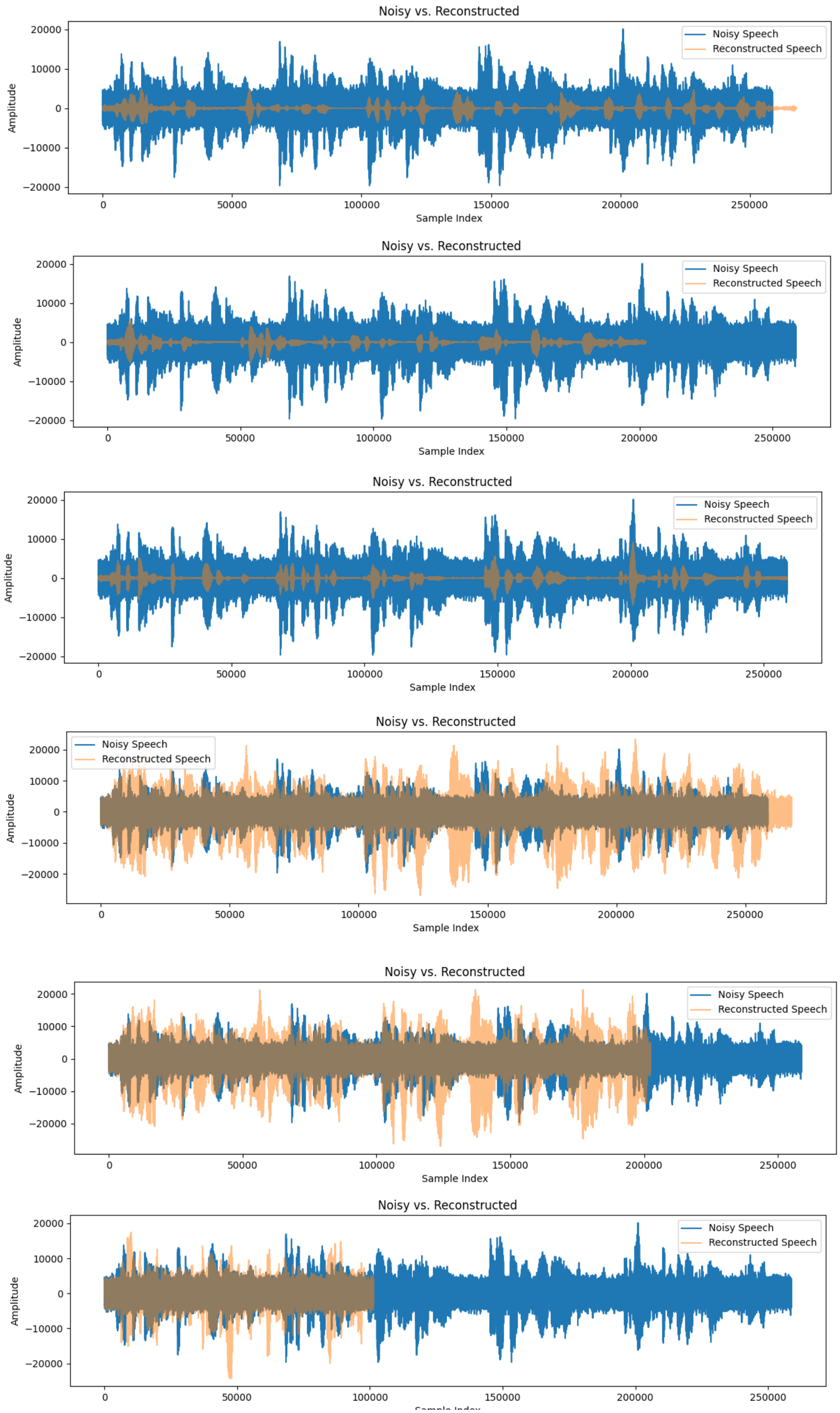

Figure 3: Amplitude vs. Index graphs for various audio samples denoised by HyperSound (rows 1-3) and SIREN models (rows 4-6.)

require further optimization. Overall, the use of STFT Loss in both approaches was crucial for reducing high-frequency noise and ensuring the perceptual quality of the reconstructed audio.

These results validate the potential of INRs for audio denoising and highlight the effectiveness of hypernetwork-based approaches in handling high-dimensional and high-variance audio data.

6 Future Work

Our future work will focus on optimizing hyperparameters and network architectures to further enhance denoising performance and reduce high-frequency noise artifacts. Additionally, we plan to test the models on diverse noise types and larger, more varied datasets to evaluate their generalizability across different conditions. Exploring the application of INRs in other audio processing tasks, such as dereverberation and audio source separation, and integrating these methods with existing neural network frameworks will be a priority. Furthermore, optimizing the models for real-time audio processing will enable live denoising applications, broadening the scope and impact of INRs in advanced audio technologies.

7 Audio Denoising for XR: Relation to the Course

In Extended Reality (XR), audio denoising is essential for enhancing user immersion and experience. Denoising removes unwanted noise from audio signals, ensuring that sounds within virtual and augmented environments remain clear and realistic. This clarity is crucial for maintaining the integrity of spatial audio, which relies on precise sound cues to create a sense of direction and space. Additionally, denoising improves voice communication in social and collaborative XR applications, making conversations more intelligible and natural. Overall, audio denoising significantly contributes to a more immersive and comfortable XR experience, allowing users to engage more fully with their virtual surroundings.

We find multiple applications where INRs for audio denoising is useful:

1. **Enhanced Audio Quality:** Improved denoising techniques, such as those demonstrated by HyperSound, can significantly enhance audio clarity and quality in XR/AR/VR environments, leading to more immersive and realistic experiences.
2. **Improved Speech Intelligibility:** By effectively reducing noise and improving metrics such as STOI, CBAK, and PESQ, these INR-based methods can enhance the intelligibility of speech in virtual environments, which is crucial for effective communication and interaction.
3. **Real-time Noise Reduction:** The robust performance of HyperSound in mitigating both high-frequency and low-frequency noise components can be utilized for real-time noise reduction in AR/VR headsets, improving user experience in noisy environments.
4. **Application in Virtual Meetings:** Enhanced audio denoising can benefit virtual meetings and conferences held in XR/AR environments, ensuring clear and uninterrupted audio, which is vital for professional and educational applications.
5. **Support for Accessibility:** Improved audio denoising techniques can support accessibility features in XR/AR/VR, such as providing clearer audio for users with hearing impairments, thereby making these technologies more inclusive.

References

Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7154–7164, 2019.

Chaewon Kim, Jaeho Lee, and Jinwoo Shin. Zero-shot blind image denoising via implicit neural representations. *arXiv preprint arXiv:2204.02405*, 2022a.

Jaechang Kim, Yunjoo Lee, Seunghoon Hong, and Jungseul Ok. Learning continuous representation of audio for arbitrary scale super resolution. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3703–3707. IEEE, 2022b.

Luca A Lanzendörfer and Roger Wattenhofer. Siamese siren: Audio compression with implicit neural representations. *arXiv preprint arXiv:2306.12957*, 2023.

Michael Michelashvili and Lior Wolf. Speech denoising by accumulating per-frequency modeling fluctuations. *arXiv preprint arXiv:1904.07612*, 2019.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25. IEEE, 2021.

Filip Szatkowski, Karol J Piczak, Przemysław Spurek, Jacek Tabor, and Tomasz Trzciński. Hypernetworks build implicit neural representations of sounds. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 661–676. Springer, 2023.

Jan Zuiderveld, Marco Federici, and Erik J Bekkers. Towards lightweight controllable audio synthesis with conditional implicit neural representations. *arXiv preprint arXiv:2111.08462*, 2021.