Neural Interferometry: Image Reconstruction from Astronomical Interferometers using Transformer-Conditioned Neural Fields

Benjamin Wu,[1,2*] **Chao Liu,**[2] **Benjamin Eckart,**[2] **Jan Kautz**[2]

[1] National Astronomical Observatory of Japan
[2] NVIDIA
benwu.astro@gmail.com, chaoliu@nvidia.com, beckart@nvidia.com, jkautz@nvidia.com

Abstract

Astronomical interferometry enables a collection of telescopes to achieve angular resolutions comparable to that of a single, much larger telescope. This is achieved by combining simultaneous observations from pairs of telescopes such that the signal is mathematically equivalent to sampling the Fourier domain of the object. However, reconstructing images from such sparse sampling is a challenging and ill-posed problem, with current methods requiring precise tuning of parameters and manual, iterative cleaning by experts. We present a novel deep learning approach in which the representation in the Fourier domain of an astronomical source is learned implicitly using a neural field representation. Data-driven priors can be added through a transformer encoder. Results on synthetically observed galaxies show that transformer-conditioned neural fields can successfully reconstruct astronomical observations even when the number of visibilities is very sparse.

Introduction

Improvements in astronomical imaging have continuously transformed humanity's understanding of physics and the universe. Currently, the highest angular resolutions are achieved using the technique of astronomical interferometry, which combines measurements from multiple telescopes to approximate that of a single telescope with a much larger aperture. Astronomical observations using radio interferometry have generated numerous scientific discoveries, including the first resolved images of protoplanetary disks [2], circumplanetary disks [4], and the event horizon of a supermassive black hole [14].

A single telescope's angular resolution θ relates to its diameter as $\theta \propto \frac{\lambda}{D}$, where λ is the observed wavelength and D is the diameter of the aperture. Interferometers, on the other hand, employ many telescopes at many different locations, and do not image an object directly. Instead, each pair of telescopes in the array measures a point in the spatial frequency domain of the celestial object. The effective angular resolution of a pair of telescopes is $\theta \propto \frac{\lambda}{B}$, where B is the projected pairwise distance orthogonal to the line of sight to the source. Given this pairwise distance relationship to effective angular resolution, B can be maximized on Earth by utilizing a global network of telescopes, a class of interferometry referred to as very-long-baseline interferometry (VLBI). One such example is the Event Horizon Telescope (EHT), which observed the M87 supermassive black hole with 25 μarcsec resolution at 1.3 mm wavelength achieved through VLBI baselines up to 10,700 km [14].

However, as longer baselines are added to an interferometric array, the relative sparsity of the spatial frequency domain samples also increases. Recovering the original celestial image based only on these sparse measurements is a highly ill-posed problem which depends strongly on appropriate priors to constrain an infinite solution space of valid images. This is the primary challenge of interferometric image reconstruction.

Astronomical interferometry falls into a broader class of inverse problems, which generally involve the reconstruction of an unknown signal based on limited observations from a typically non-invertible forward process. Many computational imaging tasks, such as deblurring, deconvolution, and inpainting, also fit under this framework and have achieved success through neural networks.

The work presented in this paper incorporates recent advancements in the computer vision and deep learning communities into a novel method for interferometric image reconstruction. Our main contributions are outlined as follows:

- We introduce a "neural interferometry" approach based on coordinate-based neural fields to perform sparse-to-dense inpainting in the spectral domain
- We show that learned priors can be incorporated through conditioning via a transformer encoder in a data-driven fashion

We demonstrate the effectiveness of our approach on a large dataset of astronomical images. We hope this work further inspires additional cross-pollination between the astrophysics and computer vision communities.

Interferometry Background

A cornerstone of astronomical interferometry is the Van Cittert-Zernike theorem, which states that the Fourier transform of the spatial intensity distribution $I(l, m)$ of a distant, incoherent source is equal to a measurable complex value

*The author contributed this work while at NVIDIA.

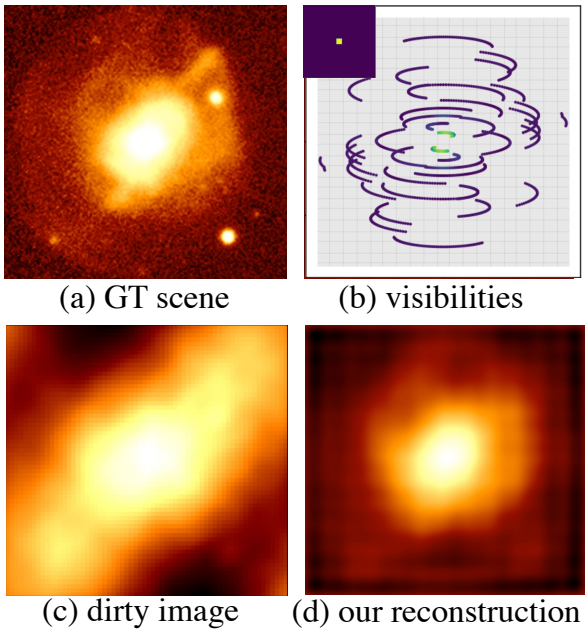


Figure 1: We reconstruct the image using the sparse measurements in the spectral domain. (a)The ground truth scene being imaged using Very-Long-Baseline Interferometry (VLBI). (b) The measured sparse visibility. Due to small field-of-view spanned by the imaged scene, the spectral measurements span a limited central region of the full spectrum, as shown in the inset. (c) The image reconstructed directly using the sparse visibilities in (b). (d) The image reconstructed using our proposed method.

called the "visibility", $V(u,v)$. Thus, it is possible to convert to and from the image domain (l,m) and the Fourier domain (u,v) via the following transform pairs,

$$V(u,v) = \int_l \int_m e^{-2\pi i(ul+vm)} I(l,m) dl dm \quad (1)$$

$$I(l,m) = \int_u \int_v e^{2\pi i(ul+vm)} V(u,v) du dv. \quad (2)$$

The telescope baselines lie in (u,v) space, which is dimensionless (normalized by observed wavelength), and each baseline corresponds to a unique coordinate which samples one value of the complex visibility.

A sharp image $I(l,m)$ can be recovered with high fidelity only if the full spectral domain has been densely sampled. In reality, however, dense sampling becomes exceedingly difficult when imaging celestial objects of smaller angular size at higher angular resolution.

Figure 1 illustrates the challenge in image reconstruction based on the sparse samplings in the Fourier domain. Even with large baselines, the angular resolution is still rather limited due to the vast distance between the imaged scene and the Earth. As a result, the spectral measurements only span a limited range of the full spectrum, as shown in the inset in Figure 1 (b). In addition, the samples are sparse in the spectral domain due to limited number of viable observatories built on the Earth. The small sampling range and sparse sampling density make it difficult to faithfully recover even the low frequency component of the scene, as shown in Figure 1 (c). In this paper, we propose a learning-based method to recover the image from sparse spectral measurements within a limited range.

Related Work

Interferometric Image Reconstruction

Two primary classes of imaging algorithms are used in the field of interferometry: the standard CLEAN approach (e.g., [16, 9]) and regularized maximum likelihood (RML) approaches (e.g., [25, 41]).

CLEAN The traditional approach used throughout radio interferometry is the CLEAN algorithm. CLEAN is an inverse-modeling approach for performing a deconvolution to recover the original image, i.e., $\mathcal{F}^{-1}[V(u,v)W(u,v)] = \mathcal{F}^{-1}[V(u,v)] * \mathcal{F}^{-1}[W(u,v)] = I(l,m) * \mathcal{F}^{-1}[W(u,v)]$. The particular configuration of the telescope array $W(u,v)$ forms a corresponding point-spread function in the image plane, $\mathcal{F}^{-1}[W(u,v)]$ or "dirty beam". The image constructed directly from the complex visibilities ("dirty image") is equivalent to the dirty beam convolved with the original celestial image $I(l,m)$. Deconvolution is approximated by iteratively subtracting the peak emission (convolved with the dirty beam) from the dirty image while building up a model of the "clean" emission.

CLEAN approximates the astronomical image using a collection of point sources, so this method is limited in its ability to reconstruct objects with extended emission. Additionally, self-calibration is generally required in cases where time-varying noise affects the visibility phase and amplitude.

RML RML is a forward-modeling approach that looks for an image that is consistent with the complex visibilities while optimizing for other properties such as smoothness or sparsity. RML methods can perform image reconstruction using either the complex visibilities or closure quantities that are more robust to atmospheric noise at the expense of fewer independent input values. Many RML methods have been developed for the EHT, through approaches such as sparse modeling [17, 1], Bayesian patch priors [5], and maximum entropy methods [7].

Neural Fields

Neural fields, also sometimes referred to as coordinate-based neural representations, have recently gained popularity in the computer vision and computer graphics communities with the advent of Neural Radiance Fields (NeRF) [24], DeepSDFs [27], and Occupancy Networks (OccNet) [23]. Neural field techniques approximate some (typically low-dimensional) function by learning a set of network weights that successfully reproduce a given set of input/output mappings. For the case of DeepSDFs and OccNets, 3D geometry is implicitly learned through level set supervision. NeRF and its follow-up works accumulate occlusion densities along rays inside of a volumetric rendering paradigm.

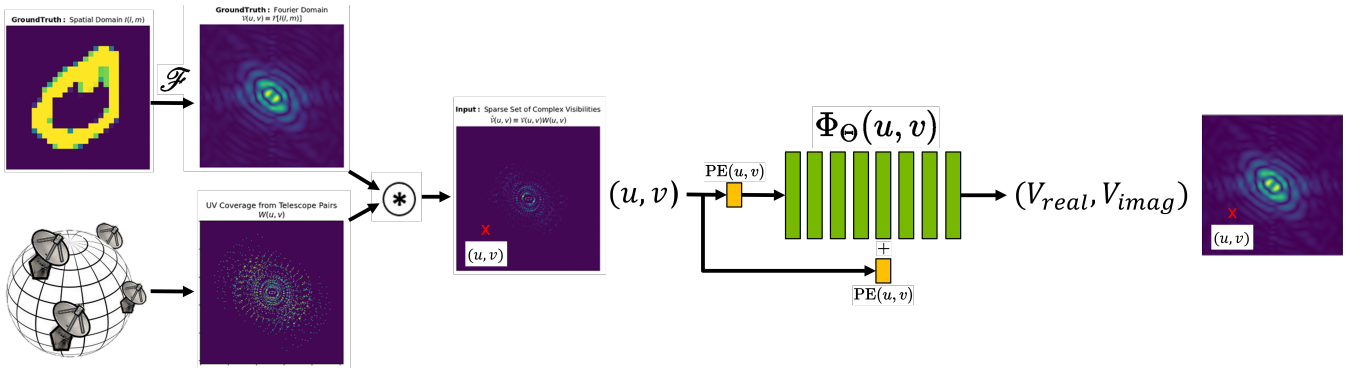


Figure 2: Training a base model to fit the neural field. Given the ground truth image and frequency coverage, we simulate the reference corresponding visibility using Eq. 1 based on the Van Cittert-Zernike theorem. The neural field $\Phi(u, v; \Theta)$ utilizes an architecture similar to that proposed by DeepSDF [27] and NeRF [24], shown here as an 8 layer MLP with a single skip connection and input positional encoding (PE). The MLP weights Θ are learned by minimizing the reconstruction loss within the UV coverage in the frequency space. Note that, like NeRF, an unconditional model as shown here must be re-trained from scratch every time the measured scene changes. Thus, we investigate additional conditioning mechanisms to allow learned data priors and enable efficient single-shot feed-forward inference.

Though conceptually quite simple, this general technique has emerged as a leading performer for tasks like novel view synthesis [3, 40], superresolution [8], and 3D reconstruction [26, 32].

One main benefit of using a neural field representation as opposed to an explicit, discrete representation (*e.g.* pixel-based or voxel-based) is that the neural field can have continuously varying input coordinates. Handling continuous coordinates in a pixel or voxel approach typically requires some sort of interpolation scheme, with accuracy limited by the density of the pixels/voxels. Coordinate-based neural representations, on the other hand, are more naturally suited to handle applications needing continuous coordinates, with effective resolution only limited by the capacity of the network [10]. Given that interferometric measurements produce complex visibilities that do not fall neatly onto integer coordinates, we found it natural to model the (u, v) spectrum with a neural field.

Transformers

Starting with the seminal paper, *Attention is All You Need* [38], transformers and attention mechanisms have become ubiquitous in the field of deep learning. More recently, transformer-based approaches have been successfully applied to computer vision, where convolutional network architectures previously tended to dominate [15]. The effective application of transformers in computer vision domains such as classification (ViT [12]), detection (DETR [6]), dense prediction (DPT [30]), iterative perception (Perceiver [19, 18]), and autoregressive generation (PixelTransformer [36], DALL-E [29]) show that transformer architectures can be extremely general.

In particular for neural interferometry, we are attracted to the transformer architecture due to its natural *permutation invariance*. That is, the visibility patterns that result from the interferometric measurements have no ordering. As such, the transformer is well-suited to ingest this kind of data as a means of conditioning our neural field representation.

Method

Instead of modifying the image domain directly, we present a deep learning approach designed to learn the neural field within the Fourier domain based only on the sparsely sampled visibilities $\hat{V}(u, v)$.

Base Model

Given the sparsely sampled visibilities $\hat{V}(u, v)$, rather than directly estimate the complex visibilities $V(u, v)$, we aim to find a neural field $\Phi(u, v)$ such that it satisfies the constraint defined by a functional F that relates Φ with the sparse measurements:

$$F(\Phi(u, v), \hat{V}(u, v)) = 0 \tag{3}$$

One straightforward selection for functional F is the norm-2 function which encourages the reconstruction of the measurements. We propose approximating the implicit function $\Phi(u, v)$ with a Multi-Layer Perceptron (MLP) parameterized by its weights Θ. The intuition of this approximation is that the continuous coordinate input of the MLP makes it possible to learn the latent manifold from data more efficiently than traditional models [10]. Based on this property, there have been several recent successful applications of MLP-based models for learning image synthesis and 3D representations [33, 24].

In our case, we aim to learn the representation in the frequency domain given the sparse visibility measurements. More specifically, during training we optimize the MLP weights Θ to minimize the reconstruction loss in the frequency domain:

$$\min_{\Theta} \sum_{u,v \in \Omega_M} |\Phi(u, v; \Theta) - \hat{V}(u, v)|^2 \tag{4}$$

where Ω_M is the subset of the frequency domain where the frequency coverage function $W(u, v) = 1$. In implementa-

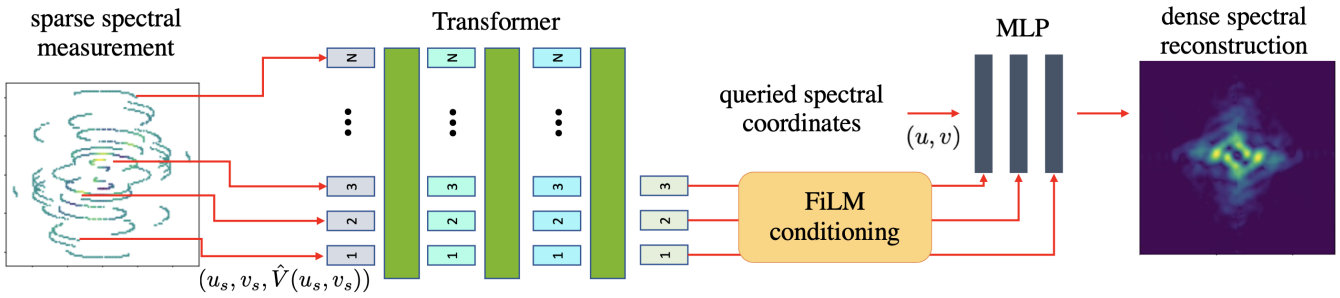


Figure 3: Proposed model. The Transformer maps the sparse spectral measurements to output tokens that condition the layers of the MLP via FiLM [28, 13]. During training, the weights of the Transformer and MLP are jointly learned. During testing, given the queried spectral coordinates (u_q, v_q) on a dense grid, the dense spectral reconstruction can be estimated with a single forward pass. Note that we use 8 output tokens to condition the 8 MLP layers. We include only three output tokens here for simplicity of illustration.

tion, we use a NeRF-style [24] positional encoding (axis-aligned, powers-of-two frequency sinusoidal encodings) to make it easier for the MLP to learn high frequency information [34].

Figure 2 illustrates the process of training a neural field using the sparse measurements of a single synthetic scene. Like the NeRF-style methods, this requires retraining from scratch every time the measured scene changes. However, the interferometric situation is even more challenging: the sparse visibility patterns common to many VLBI configurations represent an extreme case of spectral inpainting. Unlike the strong "deep image prior" inductive biases shown to be present in convolutional networks for spatial inpainting [37], the natural inductive biases of MLP-based neural fields for spectral inpainting were found to be unsuitable under such sparse guidance. Thus, we aim to extend the learned neural field by adding a learning-based prior.

Learning Priors from Data

With the proposed setup above, the weights of this network will be overfit to a single set of observations and cannot generalize. Further, when the sampling pattern becomes very sparse, such as in the case of VLBI, the reconstruction capability becomes poor. Due to these challenges, we propose a Transformer-based encoder that can learn appropriate data priors, given many observations. An added benefit of this approach is that, opposed to a computationally expensive NeRF-style optimizer, our proposed data-conditional construction needs only a single forward pass at inference time.

Transformer Encoded MLP Our chosen approach for learning prior knowledge from data is through a Transformer encoder. Compared with the CNN-based architecture, the Transformer architecture does not require the the input data to be defined on a regular grid such as discrete 2D image or 3D voxels. Instead, with continuous positional encoding such as a sinusoidal function, we can map the inputs defined on a continuous domain to the output tokens with a Transformer. As a result, a Transformer encoder is more suitable to encode the sparse visibilities measured on the continuous spectral (u, v) coordinates.

Our Transformer architecture is based on similar encoding architectures from [39, 11]. We feed in individual tuples of spectral (u, v) coordinates and complex visibilities as the inputs. The spectral coordinates are positionally encoded before being concatenated with the complex visibilities to form the input tokens. Then, the input tokens are mapped to the latent tokens via multi-headed self-attention layers.

Conditioning the Neural Field Each output token modulates a corresponding layer of the MLP through a data-dependent scale and bias to the ith layer's activation $\mathbf{x}_i$. This type of conditioning is known as FiLM [28, 13] and can be defined as a modulation of $\mathbf{x}_i$ in terms of its corresponding token $\mathbf{t}_i$,

$$\mathbf{FiLM}(\mathbf{x}_i) = \gamma(\mathbf{t}_i) \odot \mathbf{x}_i + \beta(\mathbf{t}_i) \tag{5}$$

In our experiments, we use an 8-layer MLP and 8 output tokens for the Transformer. Both $\gamma(\cdot)$ and $\beta(\cdot)$ are implemented as simple affine layers with non-linearities.

The architecture of the Transformer Encoder that encodes the visibilities into FiLM conditioning variables is shown in Figure 4. The Transformer takes as input the continuous spectral coordinates (u, v) and the complex measurements as tuples. The spectral coordinates are mapped into positional encoding (PEs) while the complex measurements are treated as 2D input and linearly embedded. The dimensions of the PE and linear embedding are both 512, with PE being the Random Fourier Embedding [35]. The embedded measurements and the PEs are concatenated to form the input tokens to the Transformer layers.

The multi-headed self-attention layers all have five heads. The two-layer MLPs between two neighboring self-attention layers share weights. The 1024-dimension output tokens are used as the conditioning variables in the following FiLM layers to condition the MLP layers. We use the same token reduction methods as in [12] - (1) either the output tokens from the final attention layer are linearly weighted or (2) only the first 8 tokens are used for FiLM conditioning. We found these two reduction methods perform equally well.

The Transformer-based architecture is suitable for encoding the spectral visibility measurements since visibilities are

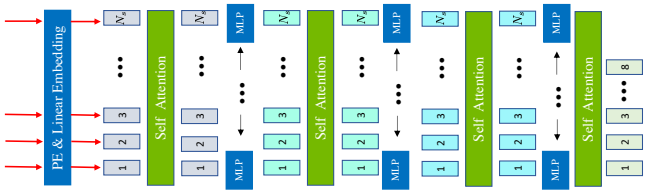


Figure 4: Architecture of the Transformer Encoder that encodes the visibilities into FiLM conditioning variables. The Transformer takes as input the continuous spectral coordinates (u, v) and the complex measurements as tuples. The spectral coordinates are mapped into positional encodings (PEs) while the complex measurements are treated as 2D input and linearly embedded. The embedded measurements and the PEs are concatenated to form the input tokens to the transformer layers. The MLPs between two neighboring Multiheaded Self-Attention layers share the same weights. The output tokens are used as the conditioning variables in the following FiLM layers to condition the MLP layers.

over the continuous, rather than grid-coordinate, spectral domain. The sparsity of the measurements makes it computational feasible to use each spectral sampling point as one token, given that the complexity of the transformer is quadratic in the number of tokens. In addition, the dense connections in the layers of the transformer (*i.e.* the self-attention layers) make it possible to learn the long range correlation among points that are far away in the spectral domain. This is a vital property to enable better image reconstruction since each sample visibility in the spectral domain will influence the entire image reconstruction, and thus the operations in the spectral domain have non-local effects in the image domain.

During training, we jointly optimize the weights of the MLP Θ_m and the Transformer Θ_t to minimize the reconstruction loss in the frequency domain over random u, v samples from a continuous bounded domain Ω, with bounds determined by the maximum baseline of the telescope array:

$$\begin{aligned} \min_{\Theta_m, \Theta_t} \quad & \sum_{u,v \in \Omega} |\Phi(u, v; \{\mathbf{t}_i\}; \Theta_m) - V_{\text{gt}}(u, v)|^2 \\ \text{with} \quad & \{\mathbf{t}_i\} = \Psi\left(\{u_s, v_s, \hat{V}(u_s, v_s)\}; \Theta_t\right), \end{aligned} \tag{6}$$

where $\{u_s, v_s, \hat{V}(u_s, v_s)\}$ is the set of sparse spectral samplings, and $\{\mathbf{t}_i\}$ is the set of output tokens from the Transformer.

Training Data

Datasets

To learn priors from a large amount of data, we synthesized interferometric observations of the Galaxy10 (SDSS) and Galaxy10 (DECals) datasets[20]. The two Galaxy10 datasets each contain approximately 20,000 colored galaxy images from the Sloan Digital Sky Survey and DESI Legacy Imaging Survey, respectively [22, 21]. Galaxy10 (SDSS) images are 69x69 while Galaxy10 (DECals) images are 256x256.

Each image was converted to grayscale, scaled to 200x200 via cubic interpolation, and then synthetically observed using the *eht-imaging toolkit* [7]. The observing parameters were set to match an 8-telescope EHT configuration for observations of M87* from 2017 and can be seen in Figure 5. Each synthetic observation returns a set of sparse continuous visibilities $\{u_s, v_s, \hat{V}(u_s, v_s)\}$, which acts as input to our Transformer model. We assume that the visibilities are well-calibrated (e.g., no atmospheric phase errors) in order to focus on the core reconstruction problem.

For ground truth, we sample the Fourier domain of our dataset images in a 256x256 grid pattern set within the maximum baseline of the telescope array. These dense visibilities are used for supervision during training. We note that the image used for ground truth is one reconstructed from the densely sampled grid of complex visibilities, and is thus lower resolution than the latent Galaxy10 sample due to an upper limit on the range of frequencies the array can access.

Results and Discussion

During testing, given the learned weights Θ, we can estimate the visibility $\hat{V}(u_q, v_q)$ for any queried frequency (u_q, v_q) even if (u_q, v_q) is not in the domain Ω_M where the measurements are available. This process is analogous to the extrapolation or inpainting operation, where the dense continuous data is recovered from discrete and possibly sparse measurements with analytical basis. However, we have replaced this with the neural field in our case.

Comparison of Methods

We compare the images reconstructed by our transformer-conditioned neural field model against standard baseline methods. Figure 6 illustrates how each method performed in reconstructing images from the Galaxy10 (DECals) synthetic observations. The challenge of reconstruction is highlighted by the dirty images, with the sparse visibilities producing strong side-lobes where real emission is not present.

Interpolation Baseline We show results from a naive approach of inpainting the real and imaginary spectral domains through cubic interpolation. The images reconstructed from these interpolated visibilities exhibit strong patterns of non-physical artefacts.

CLEAN The CLEAN algorithm is able to reduce the amount of spurious emission. However, the method's assumption of point-like sources is evident, as extended emission is not smoothly reconstructed.

Deep Learning Baseline In order to provide a deep learning baseline, we also use a U-Net approach [31] in which the inputs are the sparse visibilities discretized to a regular 256x256 domain and the supervised outputs are the ground truth reconstruction. We trained the U-Net on Galaxy10 (SDSS) and tested on Galaxy10 (DECals). The U-Net struggles to generalize to the slight variations between the two datasets, reconstructing galaxies with jagged features.

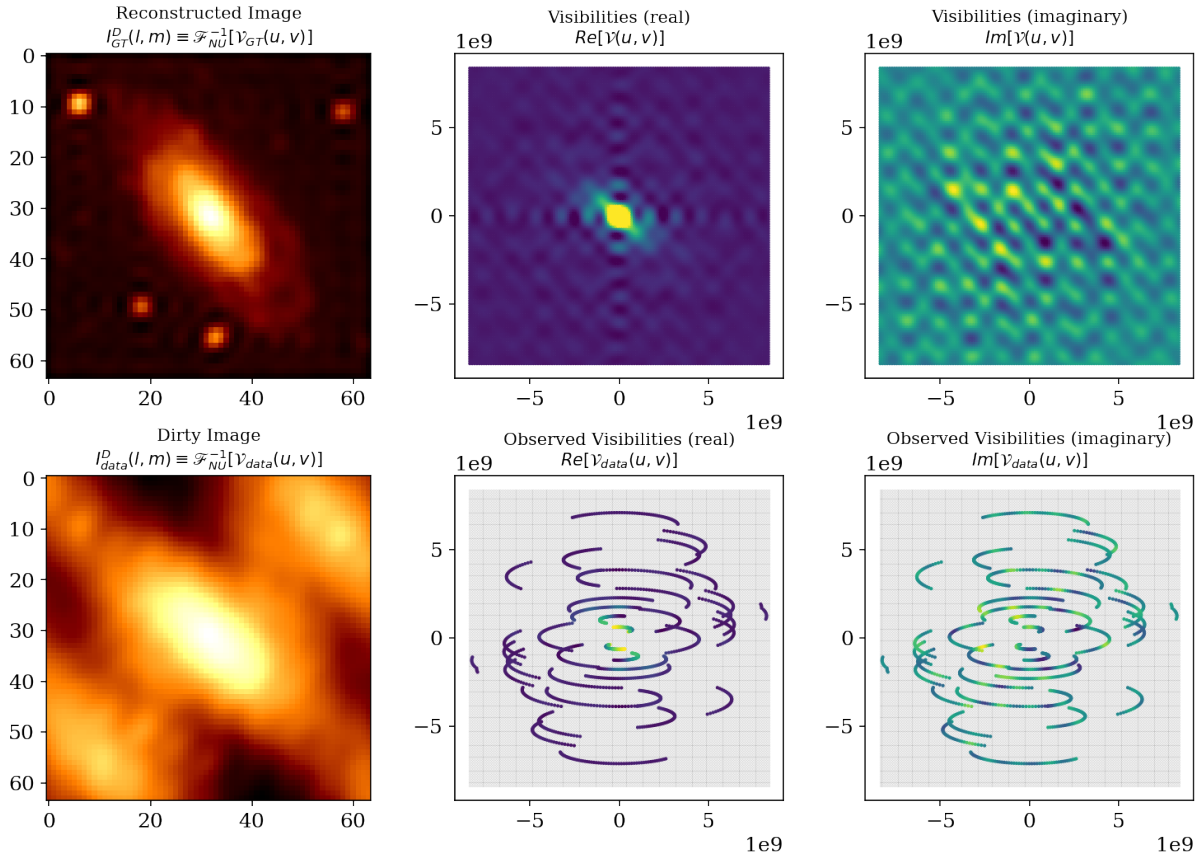


Figure 5: Dataset sample. The ground truth data are shown in the top row while the sparsely sampled data are shown on the bottom. (Left) Reconstructed image based the complex visibilities; (Middle) the corresponding real visibility component; (Right) the corresponding imaginary visibility component.

Results Our transformer model was similarly trained on Galaxy10 (SDSS) and tested on Galaxy10 (DECals). As shown in the inset of Figure 1 (b), the visible measurements only span a limited portion of the spectral domain in the center, which only includes the low-frequency information in the image. This makes the recovery of the celestial objects of interest in the image difficult, given their small sizes in FOV (e.g., galaxies that only span a small region in the image). In spite of the challenge, our method faithfully recovers the main objects in the image, as shown in Figure 6 (c). The resulting image reconstruction matches well with the ground truth reconstruction, reproducing the shape, orientation, and even multiplicity to a high degree. The quantitative results are shown in Table 1. Our method performs better than all the compared methods in terms of both MSE and SSIM score. Our method is not only more accurate but also much faster than the standard traditional method (CLEAN), given that only one single forward pass of the network is needed to get the queried spectral measurement values.

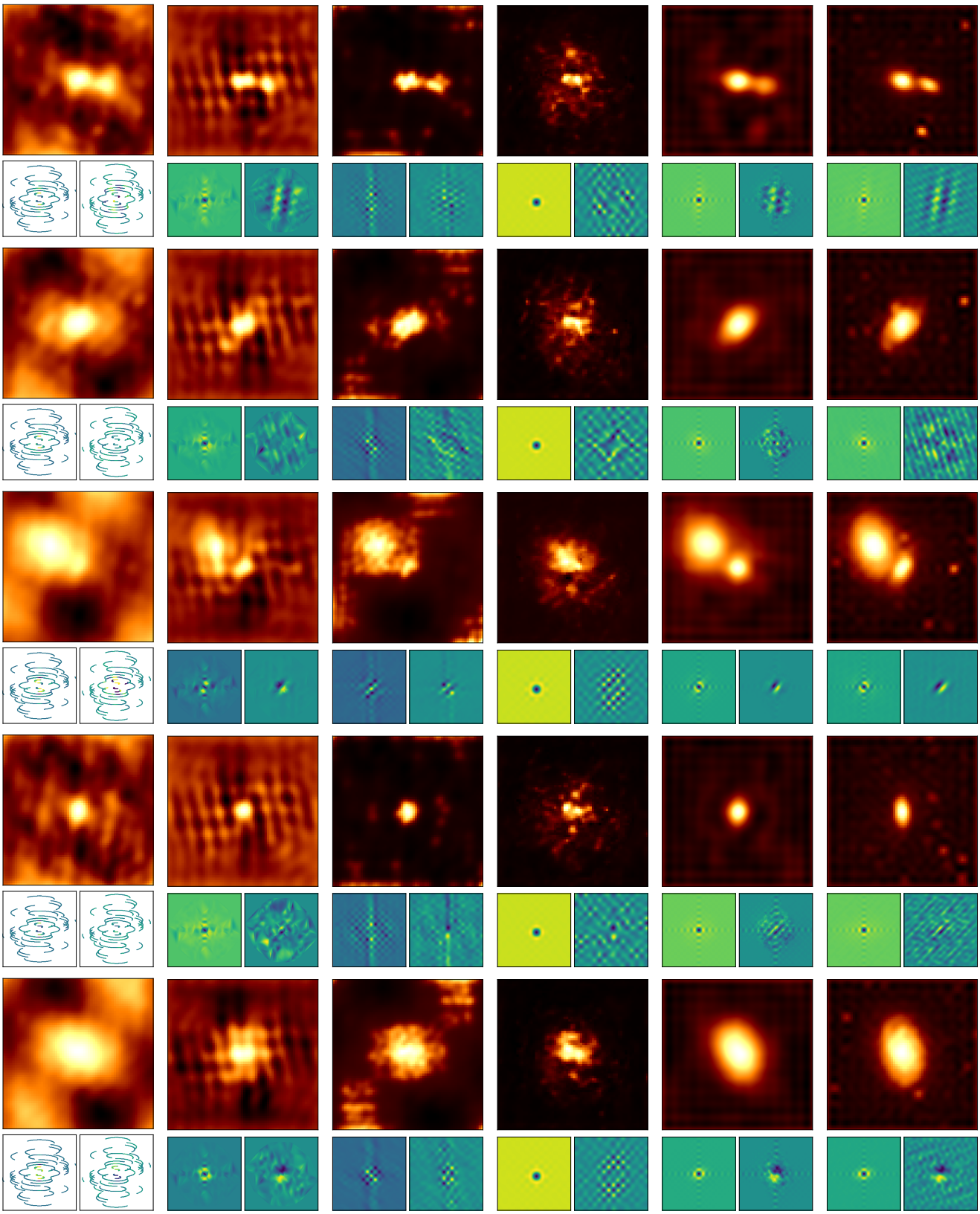
Table 1: Quantitative Metrics on Test Set Observations. Our proposed transformer outperforms both traditional methods (CLEAN) and also a strong deep learning baseline (U-Net).

Model	MSE ↓	SSIM ↑
Dirty	$2.66 \times 10^{-3} \pm 1.37 \times 10^{-3}$	0.358 ± 0.077
Cubic Interp.	$9.85 \times 10^{-4} \pm 1.18 \times 10^{-4}$	0.717 ± 0.043
CLEAN	$2.40 \times 10^{-3} \pm 1.39 \times 10^{-4}$	0.174 ± 0.007
U-Net	$1.46 \times 10^{-3} \pm 6.64 \times 10^{-4}$	0.786 ± 0.096
Transformer	$\mathbf{4.46 \times 10^{-7} \pm 2.55 \times 10^{-7}}$	$\mathbf{0.968 \pm 0.018}$

Conclusion and Future Work

Very long baseline interferometry (VLBI) presents an extremely challenging and ill-posed inverse problem in the spectral domain. We demonstrate how a neural field approach combined with a transformer-based encoding and conditioning mechanism can outperform established and widely used interferometry techniques like CLEAN.

Innovations in the deep learning and computer vision community are likely to bear fruit in related scientific fields, like astronomy. In this work, we show one such example by exploring a cross-pollination of techniques from deep learning with astronomical imaging. We hope this work can potentially serve as an inspiration for further research into the application of neural networks to astronomical imaging.



(a) Dirty Image (b) Cubic Interp. (c) CLEAN (d) U-Net (e) Ours (f) Ground Truth

Figure 6: Visual comparison of different methods. Results from left to right: (a) Dirty image, (b) Cubic interpolation, (c) CLEAN, (d) U-Net, (e) Ours, (f) Ground truth reconstruction. Different validation samples from the Galaxy10 (DECals) dataset are shown in each row. The smaller panels below each image show the real (left) and imaginary (right) components of the visibilities corresponding to each image.

References

[1] Akiyama, K.; Kuramochi, K.; Ikeda, S.; Fish, V. L.; Tazaki, F.; Honma, M.; Doeleman, S. S.; Broderick, A. E.; Dexter, J.; Mościbrodzka, M.; Bouman, K. L.; Chael, A. A.; and Zaizen, M. 2017. Imaging the Schwarzschild-radius-scale Structure of M87 with the Event Horizon Telescope Using Sparse Modeling. , 838(1): 1.

[2] ALMA Partnership; Brogan, C. L.; Pérez, L. M.; Hunter, T. R.; Dent, W. R. F.; Hales, A. S.; Hills, R. E.; Corder, S.; Fomalont, E. B.; Vlahakis, C.; Asaki, Y.; Barkats, D.; Hirota, A.; Hodge, J. A.; Impellizzeri, C. M. V.; Kneissl, R.; Liuzzo, E.; Lucas, R.; Marcelino, N.; Matsushita, S.; Nakanishi, K.; Phillips, N.; Richards, A. M. S.; Toledo, I.; Aladro, R.; Broguiere, D.; Cortes, J. R.; Cortes, P. C.; Espada, D.; Galarza, F.; Garcia-Appadoo, D.; Guzman-Ramirez, L.; Humphreys, E. M.; Jung, T.; Kameno, S.; Laing, R. A.; Leon, S.; Marconi, G.; Mignano, A.; Nikolic, B.; Nyman, L. A.; Radiszcz, M.; Remijan, A.; Rodón, J. A.; Sawada, T.; Takahashi, S.; Tilanus, R. P. J.; Vila Vilaro, B.; Watson, L. C.; Wiklind, T.; Akiyama, E.; Chapillon, E.; de Gregorio-Monsalvo, I.; Di Francesco, J.; Gueth, F.; Kawamura, A.; Lee, C. F.; Nguyen Luong, Q.; Mangum, J.; Pietu, V.; Sanhueza, P.; Saigo, K.; Takakuwa, S.; Ubach, C.; van Kempen, T.; Wootten, A.; Castro-Carrizo, A.; Francke, H.; Gallardo, J.; Garcia, J.; Gonzalez, S.; Hill, T.; Kaminski, T.; Kurono, Y.; Liu, H. Y.; Lopez, C.; Morales, F.; Plarre, K.; Schieven, G.; Testi, L.; Videla, L.; Villard, E.; Andreani, P.; Hibbard, J. E.; and Tatematsu, K. 2015. The 2014 ALMA Long Baseline Campaign: First Results from High Angular Resolution Observations toward the HL Tau Region. , 808(1): L3.

[3] Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. arXiv:2103.13415.

[4] Benisty, M.; Bae, J.; Facchini, S.; Keppler, M.; Teague, R.; Isella, A.; Kurtovic, N. T.; Pérez, L. M.; Sierra, A.; Andrews, S. M.; Carpenter, J.; Czekala, I.; Dominik, C.; Henning, T.; Menard, F.; Pinilla, P.; and Zurlo, A. 2021. A Circumplanetary Disk around PDS70c. , 916(1): L2.

[5] Bouman, K. L.; Johnson, M. D.; Zoran, D.; Fish, V. L.; Doeleman, S. S.; and Freeman, W. T. 2016. Computational Imaging for VLBI Image Reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, 913.

[6] Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229. Springer.

[7] Chael, A. A.; Johnson, M. D.; Bouman, K. L.; Blackburn, L. L.; Akiyama, K.; and Narayan, R. 2018. Interferometric Imaging Directly with Closure Phases and Closure Amplitudes. , 857(1): 23.

[8] Chen, Y.; Liu, S.; and Wang, X. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8628–8638.

[9] Clark, B. G. 1980. An efficient implementation of the algorithm 'CLEAN'. , 89(3): 377.

[10] Davies, T.; Nowrouzezahrai, D.; and Jacobson, A. 2020. On the effectiveness of weight-encoded neural implicit 3D shapes. *arXiv preprint arXiv:2009.09808*.

[11] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

[12] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[13] Dumoulin, V.; Perez, E.; Schucher, N.; Strub, F.; Vries, H. d.; Courville, A.; and Bengio, Y. 2018. Feature-wise transformations. *Distill*. Https://distill.pub/2018/feature-wise-transformations.

[14] Event Horizon Telescope Collaboration. 2019. First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole. *The Astrophysical Journal, Letters*, 875(1): L1.

[15] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Högbom, J. A. 1974. Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines. *Astronomy & Astrophysics, Supplement*, 15: 417.

[17] Honma, M.; Akiyama, K.; Uemura, M.; and Ikeda, S. 2014. Super-resolution imaging with radio interferometry using sparse modeling. *Publications of the Astronomical Society of Japan*, 66(5). 95.

[18] Jaegle, A.; Borgeaud, S.; Alayrac, J.-B.; Doersch, C.; Ionescu, C.; Ding, D.; Koppula, S.; Zoran, D.; Brock, A.; Shelhamer, E.; et al. 2021. Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv preprint arXiv:2107.14795*.

[19] Jaegle, A.; Gimeno, F.; Brock, A.; Zisserman, A.; Vinyals, O.; and Carreira, J. 2021. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*.

[20] Leung, H. 2021. Galaxy10 Dataset. Accessed: 2021-09-09.

[21] Lintott, C.; Schawinski, K.; Bamford, S.; Slosar, A.; Land, K.; Thomas, D.; Edmondson, E.; Masters, K.; Nichol, R. C.; Raddick, M. J.; Szalay, A.; Andreescu, D.; Murray, P.; and Vandenberg, J. 2011. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. , 410(1): 166–178.

[22] Lintott, C. J.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D.; Raddick, M. J.; Nichol, R. C.; Szalay, A.; Andreescu, D.; Murray, P.; and Vandenberg, J. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. , 389(3): 1179–1189.

[23] Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4460–4470.

[24] Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

[25] Narayan, R.; and Nityananda, R. 1986. Maximum entropy image restoration in astronomy. , 24: 127–170.

[26] Niemeyer, M.; Mescheder, L.; Oechsle, M.; and Geiger, A. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3504–3515.

[27] Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 165–174.

[28] Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; and Courville, A. C. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI*.

[29] Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

[30] Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. *arXiv preprint arXiv:2103.13413*.

[31] Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

[32] Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2304–2314.

[33] Sitzmann, V.; Martel, J. N.; Bergman, A. W.; Lindell, D. B.; and Wetzstein, G. 2020. Implicit Neural Representations with Periodic Activation Functions. In *Proc. NeurIPS*.

[34] Tancik, M.; Srinivasan, P. P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J. T.; and Ng, R. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*.

[35] Tancik, M.; Srinivasan, P. P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J. T.; and Ng, R. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *NeurIPS*.

[36] Tulsiani, S.; and Gupta, A. 2021. PixelTransformer: Sample Conditioned Signal Generation. *arXiv preprint arXiv:2103.15813*.

[37] Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454.

[38] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

[39] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

[40] Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021. IBRNet: Learning Multi-View Image-Based Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4690–4699.

[41] Wiaux, Y.; Jacques, L.; Puy, G.; Scaife, A. M. M.; and Vandergheynst, P. 2009. Compressed sensing imaging techniques for radio interferometry. *Monthly Notices of the Royal Astronomical Society*, 395(3): 1733–1742.