Grid, Set, or Graph? Representations for Learned VLBI Reconstruction Under Distribution Shift

Dmitrii S. Zagorulia
Lebedev Physical Institute of the Russian Academy of Sciences
Leninsky prospekt 53, Moscow, 119991, Russia

Abstract

Radio interferometers measure complex visibilities at irregularly sampled spatial frequencies, with sampling patterns that vary between observations. We investigate how the representation of these measurements affects learned image reconstruction by comparing three approaches: a regular grid processed by a convolutional neural network (CNN), an unordered set processed by DeepSets, and a k-nearest-neighbour graph processed by a graph neural network (GNN). The models are trained to reproduce CLEAN images from Astrogeo X-band observations and evaluated without adaptation on the MOJAVE 15 GHz survey. The GNN achieves the lowest median normalized mean squared error (NMSE) on both held-out X-band observations (0.202, compared with 0.228 for DeepSets and 0.376 for the CNN) and MOJAVE (0.386, compared with 0.448 and 0.707). However, all three models exhibit similar relative increases in NMSE under the survey shift (88–96%), providing no clear evidence of differences in relative robustness. An oracle-scaled mean-image baseline outperforms all three models, while diagnostics on selected observations suggest that amplitude underprediction contributes substantially to the degradation on MOJAVE. Because the survey shift simultaneously changes observing frequency, array configuration, and source characteristics, these results do not isolate the effects of sampling geometry.

1 Introduction

Very-long-baseline interferometry (VLBI) measures complex visibilities at irregular positions in the spatial-frequency (uv) plane rather than directly observing the sky brightness distribution. Recovering an image from these incomplete measurements is an ill-posed inverse problem, traditionally addressed using CLEAN [11] or regularized maximum-likelihood (RML) methods [3, 1, 5]. Unlike many conventional image reconstruction problems, the sampling pattern varies between observations due to differences in array configuration and observing coverage. Consequently, learned reconstruction methods must generalize not only across source morphologies but also across measurement operators.

A central question is how to represent irregularly sampled visibilities for neural networks. Gridding the measurements enables the use of conventional convolutional neural networks (CNNs), but introduces discretization and aggregation. Set-based architectures such as DeepSets [20] process individual measurements without imposing an ordering, while graph neural networks (GNNs) additionally model relationships between nearby samples in the uv plane. We compare these three representations using models of comparable size and a shared decoder, trained on Astrogeo X-band observations [16] and evaluated without adaptation on the MOJAVE 15 GHz survey [14].

Our study focuses on the effect of input representation on learned reconstruction rather than proposing an alternative to established imaging algorithms. Since the models are trained using CLEAN images as targets, we evaluate their ability to reproduce these reconstructions rather than recover the underlying sky brightness distribution. We find that the graph-based model achieves the lowest reconstruction

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Representations for the Physical Sciences Workshop 2026.

error on both surveys, followed by DeepSets and the grid-based CNN. However, all three models exhibit similar relative degradation under the survey shift, with additional diagnostics suggesting that amplitude estimation plays an important role in their generalization performance.

2 Related work

CLEAN [11] and regularized maximum-likelihood (RML) methods [3, 1] are widely used for very-long-baseline interferometric imaging, including observations with the Event Horizon Telescope [5]. Recent work has also explored learned regularization for interferometric reconstruction [18]. In contrast, we consider direct visibility-to-image prediction and focus on the representation of irregularly sampled measurements. DeepSets [20] and PointNet [17] use shared feature extractors and permutation-invariant pooling to process unordered sets, while graph neural networks [13, 10, 9] additionally model relationships between individual measurements. Following the relational perspective of Battaglia et al. [2], we compare grids, sets, and graphs as alternative representations with different assumptions about spatial organization and local interactions.

3 Data and task

Each observation consists of a UVFITS visibility file paired with a CLEAN image.

Visibilities. We retain measurements with positive weights w_i and estimate their uncertainties as $\sigma_i = w_i^{-1/2}$. Each visibility is augmented with its Hermitian conjugate. The median number of original visibilities per observation is 8.7×10^3 for X-band and 3.0×10^4 for MOJAVE. We use the first polarization product in each file and normalize the uv coordinates by a fixed extent of 2.6×10^8 λ. Visibility amplitudes and uncertainties are normalized by the maximum visibility amplitude within each observation.

Targets. CLEAN images are resampled to 256×256 pixels using bilinear interpolation and normalized by the 99th percentile of their absolute pixel values. The original angular field of view is preserved, so the angular pixel scale varies between observations. The median field of view is 76.8 mas for X-band and 51.2 mas for MOJAVE. All models predict images in this normalized pixel space, in which reconstruction metrics are also computed.

Datasets. We use Astrogeo X-band observations [16] for training and in-distribution (ID) evaluation, with a 90/10 split by source to prevent repeated observations of the same source from appearing in both subsets. The resulting datasets contain 50,811 training and 6,025 validation observations. For out-of-distribution (OOD) evaluation, we use 10,986 observations of 620 sources from the MOJAVE 15 GHz survey [14], without adaptation or model selection on MOJAVE data. The surveys differ in observing frequency, array configuration, and source sample. In particular, the median maximum uv radius increases from 2.3×10^8 λ at X-band to 4.4×10^8 λ for MOJAVE, placing 4.5% of MOJAVE visibilities outside the fixed gridding extent, compared with 0.3% at X-band. All 620 MOJAVE sources also occur in the X-band dataset, including 554 in the training split. Thus, the OOD evaluation measures generalization across surveys rather than to previously unseen sources.

Metrics. We evaluate reconstruction quality using normalized mean squared error, $\mathrm{NMSE} = \|\hat{x} - x\|_2^2 / \|x\|_2^2$, where $\hat{x}$ is the predicted image and x is the CLEAN target. We also report structural similarity (SSIM) [19], computed with a 7×7 window and a data range determined by each target. Because reconstruction errors exhibit heavy-tailed distributions, we report median metrics and paired per-observation comparisons.

4 Representations and models

We compare three visibility representations using models with a shared transposed-convolution decoder that maps a $16 \times 16 \times 128$ latent representation to a 256×256 image. The models differ in their input representations and encoders.

Table 1: Median reconstruction quality on X-band validation (ID) and MOJAVE (OOD). Δ is the absolute and "Rel." the relative increase in median NMSE from ID to OOD, with 95% intervals from a bootstrap over sources. The last two rows are input-independent baselines: the pixel-wise mean of the training targets, and the same image multiplied by the scalar that best fits each target (an oracle, since it uses the target).

	NMSE ↓				SSIM ↑		
Model	ID	OOD	Δ	Rel. [95% interval]	ID	OOD	Parameters
Grid CNN	0.376	0.707	+0.331	+88% [76, 101]	0.9804	0.9956	413,233
DeepSets	0.228	0.448	+0.220	+96% [82, 112]	0.9822	0.9971	527,153
GNN	0.202	0.386	+0.185	+92% [77, 108]	0.9823	0.9973	515,461
Mean image	0.606	0.707	+0.101	+17%	0.9806	0.9965	–
with oracle scale	0.196	0.159	−0.037	−19%	0.9815	0.9960	–

Grid CNN. Visibilities are binned onto a 64×64 grid spanning $|u|, |v| \leq 2.6 \times 10^8\,\lambda$ using inverse-variance-weighted averaging. The resulting three-channel input (real part, imaginary part, and sampling mask) is processed by a convolutional encoder. This representation aggregates measurements within each grid cell, discarding information about their individual coordinates and values. The model has 413,233 trainable parameters.

DeepSets. Each normalized visibility tuple is independently encoded by a shared multilayer perceptron with layer normalization. Mean pooling produces a permutation-invariant, 128-dimensional representation, which is projected and upsampled to the decoder input dimensions. The model has 527,153 trainable parameters.

GNN. Visibilities form the nodes of a k-nearest-neighbor graph ($k = 10$) constructed in the uv plane. Nodes use the same features as DeepSets, while edges encode relative coordinates and distances $(\Delta u_{ij}, \Delta v_{ij}, \|\Delta \mathbf{u}_{ij}\|)$. Three `GINEConv` layers [12] with mean aggregation incorporate edge features into message passing, each followed by `GraphNorm` [4] and ReLU activation. Node embeddings are mean-pooled and decoded as in DeepSets. The model has 515,461 trainable parameters. To limit memory usage, observations exceeding 30,000 nodes, including Hermitian conjugates, are randomly subsampled for the GNN only.

Although the models have comparable parameter counts, their architectures and preprocessing differ. In particular, gridding discards individual measurements, while the GNN applies additional subsampling to large observations. The experiment therefore compares complete reconstruction pipelines rather than isolating their inductive biases.

5 Experimental setup

All models are trained to minimize pixel-wise mean squared error (MSE) between predicted and target CLEAN images using the Adam optimizer with an initial learning rate of 10^{-3}. The learning rate is reduced on a validation-loss plateau (factor 0.3, patience 3), and early stopping is applied with patience 5. We select the checkpoint with the lowest validation loss and use an effective batch size of 16 for all models. Each model is trained once with a fixed random seed (42). The grid CNN and DeepSets are trained on two GPUs, while the GNN is trained on one. Training hyperparameters are summarized in Table 3 (Appendix A).

6 Results

Reconstruction quality. The GNN achieves the lowest median NMSE on both X-band and MOJAVE, followed by DeepSets and the grid CNN (Table 1; Figure 1, Appendix B). Paired comparisons show the same ordering (Table 2; Figure 3, Appendix B): on MOJAVE, the GNN outperforms DeepSets on 71.1% of observations (median paired NMSE difference of -0.051) and the grid CNN on 90.0%. The GNN also outperforms DeepSets despite having slightly fewer parameters, suggesting that model size alone does not explain their performance difference. However, differences in

Table 2: Paired win rates on NMSE: the fraction of observations on which the first model has lower NMSE than the second, with 95% source-level bootstrap intervals.

Comparison	X-band (ID), $n = 6{,}025$	MOJAVE (OOD), $n = 10{,}986$
GNN vs. DeepSets	59.4% [58.0, 60.8]	71.1% [68.4, 73.6]
GNN vs. Grid CNN	76.1% [74.5, 77.8]	90.0% [88.2, 91.7]
DeepSets vs. Grid CNN	71.3% [69.5, 73.2]	85.6% [83.8, 87.5]

architecture and preprocessing prevent us from attributing the advantage specifically to graph-based interactions.

Reference baselines. Predicting the mean training image for every observation gives a median NMSE of 0.606 on X-band and 0.707 on MOJAVE (Table 1). All three models outperform this baseline on X-band, whereas only DeepSets and the GNN retain an advantage on MOJAVE. The grid CNN achieves the same median NMSE as the constant baseline (0.707), indicating no improvement in this metric under the survey shift.

Amplitude estimation. Allowing the mean training image to be rescaled by an oracle-derived scalar for each target reduces median NMSE to 0.196 on X-band and 0.159 on MOJAVE, outperforming all three learned models. Although this baseline is not a usable predictor, it suggests that much of the variation in the normalized targets can be captured by scaling a common spatial template. The median fitted scale differs substantially between surveys (0.64 for X-band and 3.75 for MOJAVE). For each model and split, we analyse 33 saved observations with NMSE closest to the corresponding full-split median. On these selected MOJAVE observations, the models tend to underestimate peak brightness. Oracle rescaling of the selected GNN and DeepSets predictions reduces NMSE from 0.386 to 0.133 and from 0.448 to 0.116, respectively. These findings suggest that amplitude underprediction contributes substantially to the observed degradation. A possible explanation is the mismatch between visibility normalization by maximum amplitude and CLEAN target normalization by the 99th percentile of absolute pixel intensities.

Cross-survey generalization. Median NMSE increases by 88% for the grid CNN, 96% for DeepSets, and 92% for the GNN when evaluated on MOJAVE. The source-level bootstrap intervals for these relative increases overlap substantially (Table 1), providing no clear evidence that one model is more robust in relative terms. The GNN has the smallest absolute increase in NMSE (+0.185), although this partly reflects its lower initial error. The survey shift simultaneously changes observing frequency, array configuration, and source characteristics, preventing us from isolating the effects of sampling geometry. Moreover, prior exposure to the same sources at X-band does not correspond to lower MOJAVE reconstruction error: for the GNN, median NMSE is 0.349 on the 66 sources absent from training and 0.390 on the 554 previously observed sources, with overlapping intervals.

Structural similarity and failure cases. Despite the increase in NMSE, median SSIM exceeds 0.98 for every model and is higher on MOJAVE than on X-band (Figure 2, Appendix B). The input-independent mean-image baseline achieves SSIM values of 0.9806 on X-band and 0.9965 on MOJAVE, exceeding the grid CNN on both surveys and approaching the GNN. This suggests that global SSIM is not sufficiently sensitive to differences in reconstruction quality in these images, potentially because of the large fraction of low-intensity background pixels. All three models exhibit occasional large reconstruction errors ($\mathrm{NMSE} > 5$ for 0.7–1.2% of MOJAVE observations). The GNN's advantage is therefore primarily observed across the bulk of the error distribution rather than in the extreme tail. Additional reconstruction diagnostics, including the frequency of high-error predictions and inactive embedding dimensions, are reported in Table 4 (Appendix B). Selected qualitative reconstructions illustrating these failure modes are shown in Figure 4 (Appendix C).

7 Limitations

Confounded distribution shift. The transition from X-band to MOJAVE changes observing frequency, array configuration, calibration, and source sample simultaneously. Beyond differences in uv coverage, the normalized visibility uncertainty σ has a median of 0.167 on X-band and 0.027 on

MOJAVE, while the restoring beam is narrower in the resized MOJAVE images. These differences may contribute to the observed degradation independently of the input representation. Controlled perturbations of the sampling pattern, such as visibility dropout or reduced observing coverage, are needed to isolate the effects of sampling geometry.

Experimental design. Each model was trained with a single random seed, and the evaluated pipelines differ in parameter count, encoder architecture, preprocessing, and computational constraints. In particular, the grid CNN aggregates visibilities, whereas the GNN randomly subsamples large observations. We have not performed ablations of graph connectivity, edge features, or neighbourhood size, so the GNN's advantage cannot yet be attributed specifically to local uv-space relationships. Furthermore, relative degradation depends on in-distribution performance and should not be interpreted as a stand-alone measure of robustness.

Targets and evaluation. The models are trained to reproduce CLEAN images rather than recover the underlying sky brightness distribution, and we do not compare against other established reconstruction methods. The normalization of CLEAN targets may contribute to the observed amplitude shift between surveys, but this hypothesis remains untested. Moreover, the amplitude diagnostics are based on 33 selected observations per model and split rather than the full evaluation datasets. The use of pixel-wise MSE for training and checkpoint selection may also favour high-intensity targets, motivating further evaluation with alternative normalization and loss functions.

8 Conclusion

We compared grid-, set-, and graph-based representations for learned reconstruction from very-long-baseline interferometric visibilities. The GNN achieved the lowest reconstruction error on both held-out X-band observations and the MOJAVE 15 GHz survey, followed by DeepSets and the grid CNN. However, all three models exhibited similar relative degradation under the survey shift, with no clear evidence of differences in relative robustness. Additional diagnostics suggest that amplitude underprediction contributes substantially to this degradation, potentially due to differences in input and target normalization. While these results demonstrate consistent performance differences between the evaluated pipelines, further experiments with controlled sampling perturbations, graph ablations, and alternative normalization strategies are needed to identify their underlying causes.

Acknowledgments and Disclosure of Funding

Computational resources were provided by Kaggle's free GPU environment and the Lebedev Physical Institute of the Russian Academy of Sciences. This work used data from the Astrogeo VLBI FITS image database [16] and the MOJAVE program [14]. This research received no external funding. The author declares no competing interests.

References

[1] K. Akiyama, K. Kuramochi, S. Ikeda, et al. Imaging the Schwarzschild-radius-scale Structure of M87 with the Event Horizon Telescope Using Sparse Modeling. *The Astrophysical Journal*, 838(1):1, 2017.

[2] P. Battaglia, J. B. C. Hamrick, V. Bapst, et al. Relational Inductive Biases, Deep Learning, and Graph Networks. *arXiv*, 2018. doi: 10.48550/arXiv.1806.01261.

[3] K. L. Bouman, M. D. Johnson, D. Zoran, et al. Computational Imaging for VLBI Image Reconstruction. In *Computer Vision and Pattern Recognition*, pages 913–922, 2016.

[4] T. Cai, S. Luo, K. Xu, et al. GraphNorm: A Principled Approach to Accelerating Graph Neural Network Training. In *International Conference on Machine Learning*, volume 139, pages 1204–1215, 2021.

[5] The Event Horizon Telescope Collaboration. First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole. *The Astrophysical Journal Letters*, 875(1):L4, 2019.

[6] W. Falcon and The PyTorch Lightning team. PyTorch Lightning, 2019. URL `https://github.com/Lightning-AI/lightning`.

[7] M. Fey and J. E. Lenssen. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[8] M. Fey, J. Sunil, A. Nitta, et al. PyG 2.0: Scalable Learning on Real World Graphs. In *KDD Workshop on Temporal Graph Learning*, 2025.

[9] J. Gilmer, S. S. Schoenholz, P. F. Riley, et al. Neural Message Passing for Quantum Chemistry. In *International Conference on Machine Learning*, volume 70, page 1263–1272, 2017.

[10] W. Hamilton, Z. Ying, and J. Leskovec. Inductive Representation Learning on Large Graphs. In *Neural Information Processing Systems*, pages 1024–1034, 2017.

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

[12] W. Hu, B. Liu, J. Gomes, et al. Strategies for Pre-Training Graph Neural Networks. In *International Conference on Learning Representations*, 2020.

[13] T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*, 2017.

[14] M. L. Lister, D. C. Homan, K. I. Kellermann, et al. Monitoring Of Jets in Active Galactic Nuclei with VLBA Experiments. XVIII. Kinematics and Inner Jet Evolution of Bright Radio-loud Active Galaxies. *The Astrophysical Journal*, 923(1):30, 2021.

[15] A. Paszke, S. Gross, F. Massa, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv*, 2019. doi: 10.48550/arXiv.1912.01703.

[16] L. Petrov. The astrogeo VLBI FITS image database. `https://astrogeo.org/vlbi_images/`, 2021. doi: 10.25966/kyy8-yp57.

[17] C. R. Qi, H. Su, K. Mo, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Computer Vision and Pattern Recognition*, pages 77–85, 2017.

[18] M. Terris, A. Dabbech, C. Tang, et al. Image reconstruction algorithms in radio interferometry: From handcrafted to learned regularization denoisers. *Monthly Notices of the Royal Astronomical Society*, 518 (1):604–622, 2023.

[19] Z. Wang, A.C. Bovik, H.R. Sheikh, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[20] M. Zaheer, S. Kottur, S. Ravanbakhsh, et al. Deep Sets. In *Neural Information Processing Systems*, volume 30, 2017.

A Data preprocessing and experimental details

Visibility preprocessing. Only the first polarization product in each UVFITS file is used. This corresponds to RR in 99% of X-band files and 90% of MOJAVE files, with LL used in the remaining MOJAVE files. Although 88% of MOJAVE files contain four polarization products, the other products are not included in our experiments. The number of original visibilities per observation ranges from 4 to 6.7×10^5.

Image properties. Most native CLEAN images have dimensions of 512×512 pixels (70% for X-band and 60% for MOJAVE). The remaining MOJAVE images are predominantly 1024×1024 (33%) or 2048×2048 (7%) pixels. The median native pixel scale is 0.2 mas for X-band and 0.1 mas for MOJAVE. The median angular field of view is 76.8 mas (5th–95th percentile: 51–102 mas) for X-band and 51.2 mas (51–205 mas) for MOJAVE. After resizing, the restoring beam spans a median of 5.8 pixels for X-band and 3.7 pixels for MOJAVE, measured using the geometric mean of its major and minor full widths at half maximum.

Decoder design. In preliminary experiments, broadcasting the pooled DeepSets or GNN representation across all spatial positions produced nearly constant reconstructions, with $\mathrm{NMSE} \approx 1$ for DeepSets. We therefore introduced a learned projection and upsampling stage to generate a spatially varying decoder input.

The training configurations for all three models are summarized in Table 3.

Bootstrap confidence intervals are estimated using 2,000 source-level resamples with replacement, retaining all observations associated with each sampled source. Models are implemented using PyTorch [15], PyTorch Geometric [7, 8], and PyTorch Lightning [6].

Table 3: Training hyperparameters, common to all three models unless otherwise specified.

Loss	Pixel-wise MSE, 256×256 targets
Optimizer	Adam, learning rate 10^{-3}
LR scheduler	ReduceLROnPlateau, factor 0.3, patience 3
Early stopping	Patience 5, validation loss
Checkpoint	Lowest validation loss
Maximum epochs	100
Best epoch	CNN: 6; DeepSets: 36; GNN: 31
Random seed	42
Grid CNN batch	8/GPU $\times$ 2 GPUs = 16
DeepSets batch	2/GPU $\times$ 4 steps $\times$ 2 GPUs = 16
GNN batch	2 $\times$ 8 steps $\times$ 1 GPU = 16
GNN node cap	30,000 (random subsampling)

B Additional quantitative results

Figure 1 summarizes the median reconstruction errors across the two surveys, while Figure 2 shows the distributions of NMSE and SSIM on MOJAVE. The paired comparisons in Figure 3 further illustrate the differences in reconstruction performance between models.

Table 4 provides additional information on reconstruction failures and embedding variability. The grid CNN produces substantially more near-trivial predictions and has more inactive embedding dimensions than either DeepSets or the GNN.

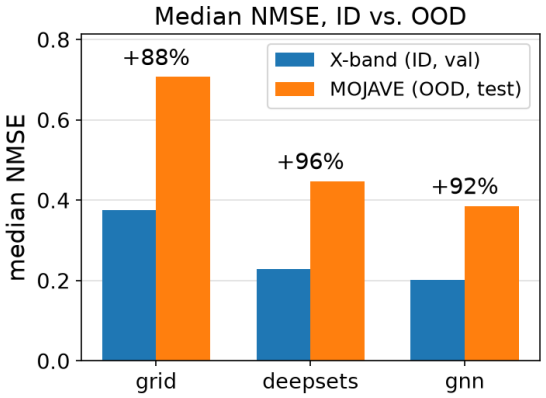

Figure 1: Median NMSE on held-out Astrogeo X-band observations (ID) and MOJAVE (OOD) for the three models.

Table 4: Reconstruction diagnostics on MOJAVE. Near-trivial predictions satisfy $|\mathrm{NMSE}-1| < 0.05$; very poor predictions have $\mathrm{NMSE} > 5$. Inactive dimensions have standard deviation below 10^{-6} across all 17,011 evaluated observations.

	Grid CNN	DeepSets	GNN
Near-trivial predictions	19.0%	0.4%	0.3%
Very poor ($\mathrm{NMSE} > 5$)	0.7%	1.2%	1.0%
$\mathrm{NMSE} \geq 1$	17.4%	4.1%	4.4%
Inactive dimensions	43	3	0

A preliminary GNN variant using `SAGEConv` [10], without explicit edge features, achieved median NMSE values of 0.261 on X-band and 0.436 on MOJAVE. Although its relative degradation was smaller (67%), its reconstruction error was higher on both datasets and training stopped substantially earlier. This illustrates why relative degradation alone is insufficient to characterize generalization performance.

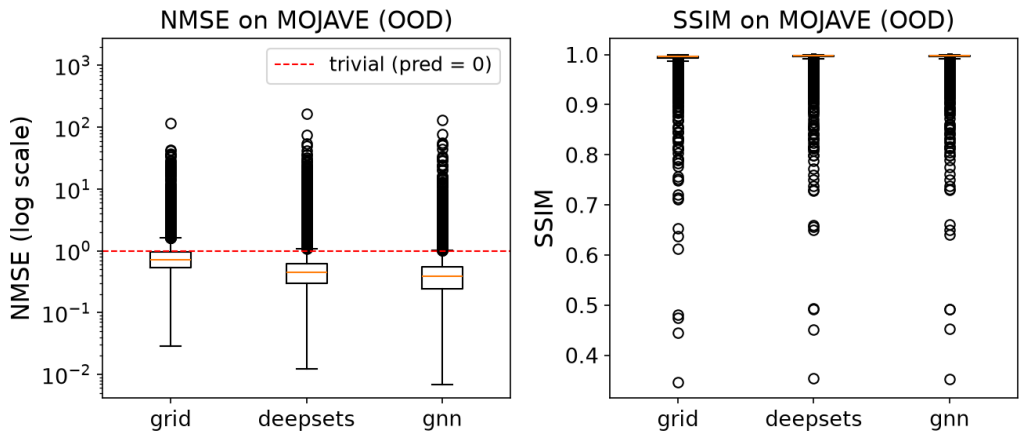

Figure 2: Distribution of per-observation NMSE and SSIM on MOJAVE. Mean NMSE is 0.851 for the grid CNN, 0.653 for DeepSets, and 0.571 for the GNN, exceeding the corresponding median values.

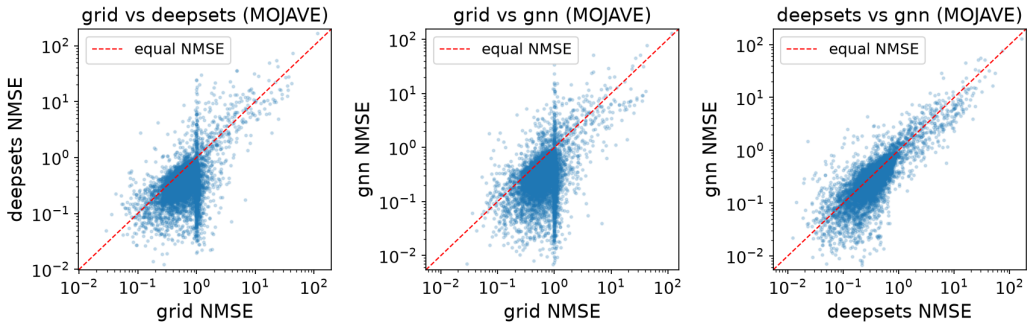

Figure 3: Paired per-observation NMSE on MOJAVE. Points below the diagonal indicate observations for which the model on the vertical axis achieves lower reconstruction error.

C Qualitative reconstruction examples

Figure 4 presents selected best, median, and worst reconstruction examples from MOJAVE for each model. The examples illustrate both the reconstruction of compact sources and characteristic failure cases. In particular, the GNN fails to recover a resolved double source in one high-error example, instead producing a single compact component.

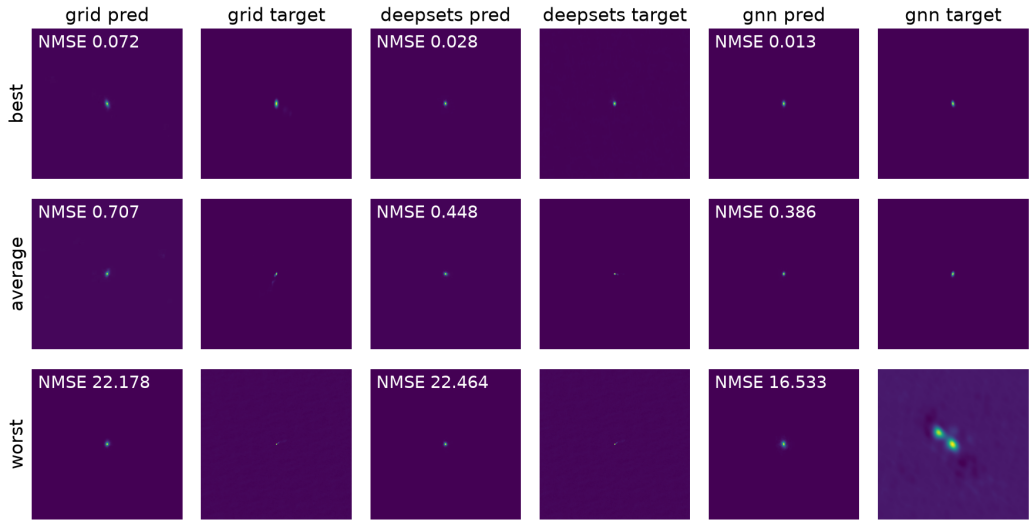

Figure 4: Selected best, median, and worst reconstructions on MOJAVE for each model, using representative examples from the corresponding error groups. The targets differ between models. The grid CNN and DeepSets overpredict brightness for a faint worst-case target, while the GNN reconstructs a resolved double source as a single unresolved component.