Risk Quantification in Deep MRI Reconstruction

Vineet Edupuganti[1], **Morteza Mardani**[1],
Shreyas Vasanawala[2], **John Pauly**[1],
Department of Electrical Engineering[1] and Radiology[2], Stanford University
{ve5,morteza,vasanawala,pauly}@stanford.edu

Abstract

Reliable medical image recovery is crucial for accurate patient diagnoses, but little prior work has centered on quantifying uncertainty when using non-transparent deep learning approaches to reconstruct high-quality images from limited measured data. In this study, we develop methods to address these concerns, utilizing a VAE as a probabilistic recovery algorithm for pediatric knee MR imaging. Through our use of SURE, which examines the end-to-end network Jacobian, we demonstrate a new and rigorous metric for assessing risk in medical image recovery that applies universally across model architectures.

1 Introduction

Artificial intelligence has introduced a paradigm shift in medical image reconstruction in the last few years, offering significant improvements in speed and image quality [1, 2, 3, 4, 5, 6]. Yet despite the importance of assessing risk in medical image reconstruction, little work has explored the robustness of deep learning (DL) architectures in inverse problems, and there is a lack of established methods for quantifying model uncertainty [7]. Given the pernicious effects of inaccurate reconstruction, these reliable uncertainty quantification methods could have utility both as an evaluation metric and as a way of gaining interpretability regarding risk factors for a given model and dataset [8].

As such, this work introduces procedures that can provide insights into the model robustness of DL MR reconstruction schemes. We develop a variational autoencoder (VAE) model for MR image recovery, which is notable for its low error and probabilistic nature. We then introduce Stein's Unbiased Risk Estimator (SURE) as a means of assessing uncertainty of the DL model, which we find effectively approximates MSE and serves as a valuable tool for assessing risk when the ground truth reconstruction (i.e. fully sampled image) is unavailable.

2 Preliminaries and Problem Statement

In MR imaging, the goal is to recover the true image $x_0 \in \mathbb{C}^n$ from undersampled k-space measurements $y \in \mathbb{C}^m$ with $m \leq n$ that admit

$$y = \Phi x_0 + v. \tag{1}$$

Here, Φ in general includes the acquisition model with the sampling mask Ω, the Fourier operator F, as well as coil sensitivities. The noise term v also accounts for measurement noise and unmodeled dynamics.

Given the ill-posed nature of this problem, it is necessary to incorporate prior information to obtain high-quality reconstructions. This prior spans across a manifold of realistic images ($\mathcal{S} \subset \mathbb{C}^n$). However, since not all points on this manifold are consistent with the measurements, we must consider the intersection of the prior manifold $\mathcal{S}$ with a data consistent subspace $\mathcal{C}_y := \{x \in \mathbb{C}^n : y = \Phi x + v\}$. Note that there might be multiple admissible solutions $x_0, x_1, \ldots, x_n$ at the intersection with different likelihoods.

While DL models can be effectively used for learning the projection onto the intersection $\mathcal{S} \cap \mathcal{C}_y$, performance can be limited on data unlike the training examples. In particular, one risk is the

NeurIPS 2020 Workshop on Deep Learning and Inverse Problems, virtual.

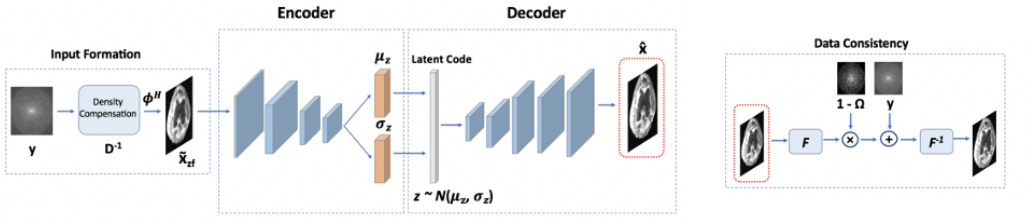

Figure 1: The model architecture, with aliased inputs feeding into the VAE encoder, the latent code feeding into the VAE decoder, and data consistency applied to obtain the output reconstruction.

introduction of realistic artifacts, or so-termed "hallucinations," which can prove costly in a domain as sensitive as medical imaging by misleading radiologists and resulting in incorrect diagnoses [9, 10]. Hence, analyzing the uncertainty and robustness of DL techniques in MR imaging is essential.

3 VAEs for Medical Image Recovery

For image recovery, we consider a probabilistic VAE architecture shown in Fig. 1. At test time, latent code vectors are sampled from a normal distribution $z \sim \mathcal{N}(\mu_z, \sigma_z)$ to generate new reconstructions. To ensure reconstructions did not deviate from physical measurement, an affine projection based on the undersampling mask (i.e. data consistency) was applied, which we found essential for high SNR [7].

The loss function in training was based on the mixture of pixel-wise ℓ_2 with a KL-divergence term (weighted by constant η) that constrains the latent code. As η increased, the integrity of the latent code was preserved at the expense of reconstruction quality. The training cost is formed as:

$$\min \mathbb{E}_{x,y}\left[\|\hat{x} - x_0\|_2^2\right] + \eta KL\big(\mathcal{N}(\mu_z, \sigma_z)\|\mathcal{N}(0,1)\big). \tag{2}$$

4 SURE for Uncertainty Analysis

4.1 Denoising SURE

In a clinical setting, it is not possible to know the fully-sampled ground truth corresponding to a given reconstruction, which motivates the use of SURE [11, 12]. Despite being well-established, SURE has not been widely used for uncertainty analysis in imaging or DL problems. With zero-filled (aliased) input x_{zf} and reconstruction $\hat{x}$ with dimension n, SURE can be expressed as follows (where $r \sim \mathcal{N}(0, \sigma^2 I)$ is the noise process that describes the difference between the input and the ground truth):

$$SURE = -n\sigma^2 + \underbrace{\|\hat{x} - x_{zf}\|^2}_{\text{RSS}} + \sigma^2 \underbrace{\operatorname{tr}\left(\frac{\partial \hat{x}}{\partial x_{zf}}\right)}_{\text{DOF}}. \tag{3}$$

This form importantly does not depend on ground truth image x_0, and approximates the DOF with the trace of the end-to-end network Jacobian $J = \partial\hat{x}/\partial x_{zf}$. This direct dependence is of note, since the Jacobian represents the network sensitivity to small input perturbations and has been previously utilized to analyze robustness in computer vision tasks [13]. Additionally, this formulation is agnostic to model architecture and can be applied to both deterministic and probabilistic networks.

By making the experimentally-validated assumption that error in the output reconstruction is not large, σ^2 can be estimated by setting the sum of the first two terms in the SURE expression to zero (i.e. $\sigma^2 = \|\hat{x} - x_{zf}\|^2/n$), yielding the following expression

$$SURE = \sigma^2 \operatorname{tr}(J). \tag{4}$$

4.2 Gaussian residuals with density compensation

The key assumption behind SURE is that the residual model is Gaussian with zero mean. However, it is not safe to assume that the undersampling noise in MRI reconstruction inherits this property. For

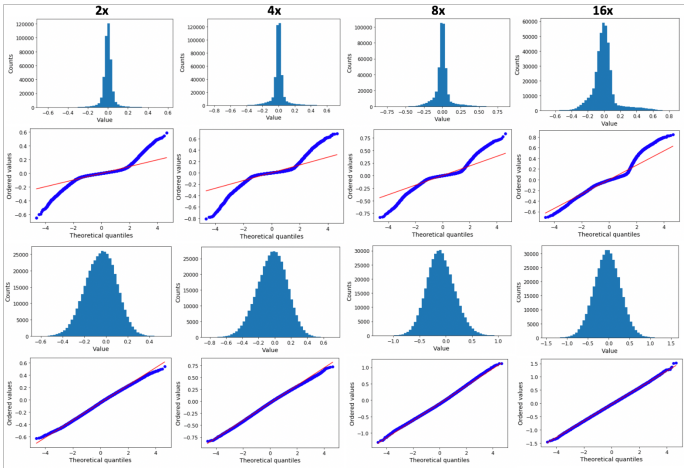

Figure 2: Histograms and quantile-quantile plots of the residuals between zero-filled and ground truth images without (top two rows) and with density compensation with 2, 4, 8, and 16 fold undersampling, respectively.

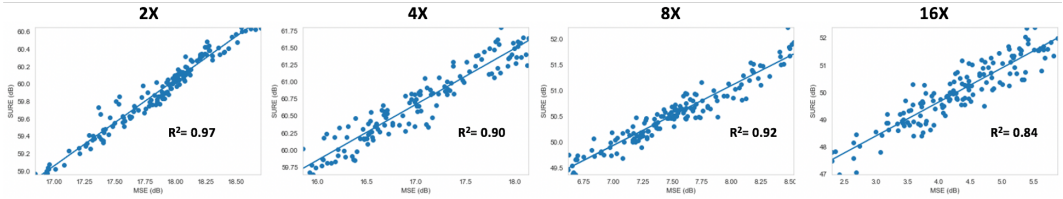

Figure 3: SURE vs. MSE at various undersampling rates.

this reason, we introduce density compensation on the input image as a way of enforcing zero-mean residuals. This approach has the added benefit of making artifacts independent of the underlying image and we find that it significantly increases residual normality (see Fig. 2).

More specifically, given a 2D sampling mask Ω, we can treat each element of the mask as a Bernouli random variable with a certain probability $D_{i,j}$, where $\mathbb{E}[\Omega] = D$ (this is dependent on the sampling approach). With this formulation, we can define a density compensated zero-filled image as follows: $\tilde{x}_{zf} = F^{-1}D^{-1}\Omega F x_0$. We can rewrite this expression using x_0 as

$$\tilde{x}_{zf} = x_0 + \underbrace{(F^{-1}D^{-1}\Omega F - I)x_0}_{:=r}. \tag{5}$$

First, we observe that r has zero mean since $\mathbb{E}[D^{-1}\Omega] = I$. In addition, the residual variance obeys

$$\sigma^2 = \mathrm{tr}\big(x_0^{\mathsf{H}}(F^{-1}D^{-1}F - I)x_0\big). \tag{6}$$

In practice we do not have access to the ground truth image x_0, and instead rely on the approximation for σ^2 for the residual variance. Given these main properties, the density compensation method that this work introduces represents an important step that can allow denoising SURE to be used effectively in medical imaging and other inverse problems.

5 Empirical Evaluations

We assess our model and methods on a dataset of Knee MR images (reconstructions and Monte Carlo uncertainty results are shown in Appendix). The TensorFlow source code for implementation is publicly available via Github[1].

[1] `https://github.com/MortezaMardani/GAN-Hallucination/tree/VAE-GAN`

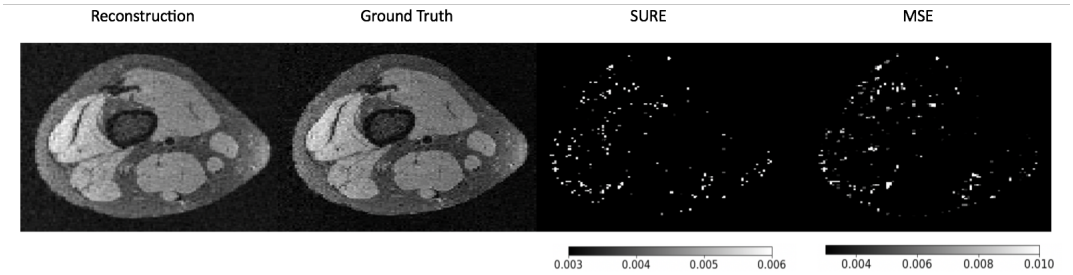

Figure 4: Pixel-wise SURE and MSE images for a given reference slice with 2-fold undersampling. Note that MSE depends on the ground truth while SURE does not.

Dataset. The Knee dataset used was obtained from 19 patients with a 3T GE MR750 scanner [14]. Each volume consisted of 320 2D slices of dimension 320×256 that were divided into training, validation, and test examples, and a 5-fold variable density undersampling mask with radial view ordering (designed to preserve low-frequency structural elements) was used to produce aliased input images x_{zf} for the model to reconstruct [15].

Network architecture. The VAE encoder was composed of 4 layers formed through a sequence of strided convolution operations followed by ReLU activations and batch normalization [16]. Latent space mean, μ, and standard deviation, σ, were represented by fully connected layers. The VAE decoder also had 4 layers and utilized transpose convolution operations for upsampling [17]. Skip connections were utilized to improve gradient flow through the network [18].

5.1 Residual distribution with density compensation

As described earlier, denoising SURE builds on the Gaussian assumption for the residuals $r = x_{zf} - x_0$. To validate this assumption, we produce histograms and Q-Q plots of the residuals at various undersampling rates, by considering the differences for individual pixels across test images. From the top two rows of Fig. 2 one can observe the residual distribution is not perfectly normal in any of the cases, which can limit the effectiveness and accuracy of SURE.

To overcome the lack of normality in the residuals, we apply our density compensation method. As Fig. 2 (bottom two rows) shows, the distribution of residuals (for various undersampling rates) better matches the normal distribution, and the mean of the distribution lies very close to zero.

5.2 SURE results

To evaluate the effectiveness of the density-compensated SURE approach, we produce correlations of SURE versus MSE (which depends on the ground truth and is a standard metric for assessing model error) using the results from our test images. Fig. 3 shows the strong linearity of the correlations under all conditions. The linear relationship is strongest for higher undersampling rates ($R^2 = 0.97$ for 2-fold while $R^2 = 0.84$ for 16-fold). Nonetheless, the results show that even with relatively high undersampling, SURE can be used to effectively estimate risk in medical image reconstruction.

Fig. 4 shows pixel-wise SURE and MSE maps for a given reference slice. The SURE approximation is reasonably effective, with substantial overlap in the areas with highest reconstruction error.

6 Conclusions

This work introduces methods to analyze uncertainty in deep-learning based medical image recovery. The strong correlations between SURE and MSE at both the global and pixel level indicate that, with our density compensation modification, this new approach for uncertainty quantification has great potential as a general-purpose risk evaluation tool in imaging problems (and is by nature effective across arbitrary deep learning architectures).

Future work will address data uncertainty related to the acquisition model and training set size. Additionally, it would be useful to analyze uncertainty for abnormal and pathological cases specifically.

References

[1] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[2] Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. Compressed sensing mri reconstruction using a generative adversarial network with a cyclic loss. *IEEE Transactions on Medical Imaging*, 37(6):1488–1497, 2018.

[3] Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487, 2018.

[4] Dongwook Lee, Jaejun Yoo, and Jong Chul Ye. Deep residual learning for compressed sensing MRI. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 15–18. IEEE, 2017.

[5] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. *arXiv preprint arXiv:1703.03208*, 2017.

[6] Ali Mousavi, Ankit B Patel, and Richard G Baraniuk. A deep learning approach to structured signal recovery. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 1336–1343. IEEE, 2015.

[7] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. Deep generative adversarial neural networks for compressive sensing (GANCS) MRI. *IEEE Transactions on Medical Imaging*, 2018.

[8] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[9] Li Xu, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural networks for image deconvolution. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2014.

[10] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, et al. DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1310–1321, 2018.

[11] Christopher A Metzler, Ali Mousavi, Reinhard Heckel, and Richard G Baraniuk. Unsupervised learning with Stein's unbiased risk estimator. *arXiv preprint arXiv:1805.10531*, 2018.

[12] Ryan J Tibshirani and Saharon Rosset. Excess optimism: How biased is the apparent error of an estimator tuned by SURE? *Journal of the American Statistical Association*, pages 1–16, 2018.

[13] Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 514–529, 2018.

[14] `http://mridata.org/`.

[15] Joseph Y Cheng, Tao Zhang, Marcus T Alley, Michael Lustig, Shreyas S Vasanawala, and John M Pauly. Variable-density radial view-ordering and sampling for time-optimized 3D cartesian imaging. In *Proceedings of the ISMRM Workshop on Data Sampling and Image Reconstruction*, 2013.

[16] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.

[17] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1175–1183. IEEE, 2017.

[18] Adji B Dieng, Yoon Kim, Alexander M Rush, and David M Blei. Avoiding latent variable collapse with generative skip models. *arXiv preprint arXiv:1807.04863*, 2018.

[19] Morteza Mardani, Qingyun Sun, Shreyas Vasanawala, Vardan Papyan, Hatef Monajemi, John Pauly, and David Donoho. Neural proximal gradient descent for compressive imaging. In *Advances in Neural Information Processing Systems*, pages 9573–9583, 2018.

[20] Robert Tibshirani. *Bias, variance and prediction error for classification rules*. Citeseer, 1996.

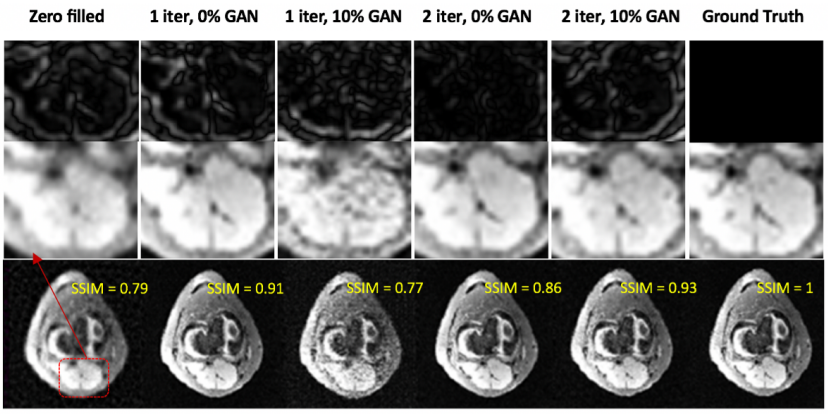

Figure 5: The aliased input, reconstructions with one recurrent block (RB) and pure MSE loss, one RB and 10 percent GAN loss, two RBs and pure MSE loss, and two RBs and 10 percent GAN loss, and the ground truth for a representative slice (5 fold undersampling with radial view ordering). SSIM between the given image and the ground truth is shown for all cases. The top row shows absolute error for the highlighted ROI.

7 Appendix

7.1 Reconstructions with VAE model

In assessing the VAE model, we considered two modifications. First, we incorporated an adversarial component, by using an 8-layer CNN as the discriminator to provide feedback to the VAE generator (this GAN setup has been shown to better model high frequencies important in MR imaging) [7]. The influence of the discriminator was weighted by constant λ. Secondly, we explored a recurrent architecture, in which multiple recurrent blocks (RBs) are used (i.e. the model generator and data consistency layers repeat) [19]. Figure 5 shows model reconstructions and errors under combinations of these hyperparameters. GAN loss results in sharper images, though error is introduced since MSE has less weight in the loss function. Multiple RBs seem to improve image quality and can serve as a simple and useful tool for promoting robustness in reconstruction.

7.2 Monte Carlo Uncertainty Results

As another technique for assessing uncertainty, we used a Monte Carlo approach to statistically analyze outputs corresponding to a given input under various hyperparameter settings. Utilizing the probabilistic nature of the VAE model, for a given zero-filled image $x_{zf} = \Phi^{\mathsf{H}} y$ (i.e. the aliased input to the model), we can use our encoder to find the mean μ and variance σ^2 of the latent code z, which we use to draw samples. We can then use our decoder to produce reconstructions of latent code samples z_i and then aggregate the results over k samples to produce pixel-wise mean and variance maps for the reconstructions.

This Monte Carlo sampling approach allows one to evaluate variance as well as higher order statistics, which can be very useful in understanding the extent and impact of model uncertainty. However, despite the information the Monte Carlo approach can provide, important statistics such as bias are dependent on knowledge of the ground truth, which is where SURE proves to be a superior method.

Figures 6 and 7 show Monte Carlo results derived from analyzing 1K outputs for a given reference test slice (obtained by latent code sampling). Mean reconstruction, pixel-wise variance, squared bias (difference between mean reconstruction and ground truth), and squared error are shown, utilizing the common relation $error^2 = bias^2 + variance$ [20].

The results indicate that variance, bias, and error increase as the GAN loss weight λ increases and as the number of RBs decreases. Furthermore in all cases with GAN loss, the variance extends to structural components of the image, which poses the most danger in terms of diagnosis. Nevertheless, with a reasonably conservative choice of GAN weight, the risk is substantially lower. More RBs can lower variance as well as error, and can be a useful tweak to improve robustness.

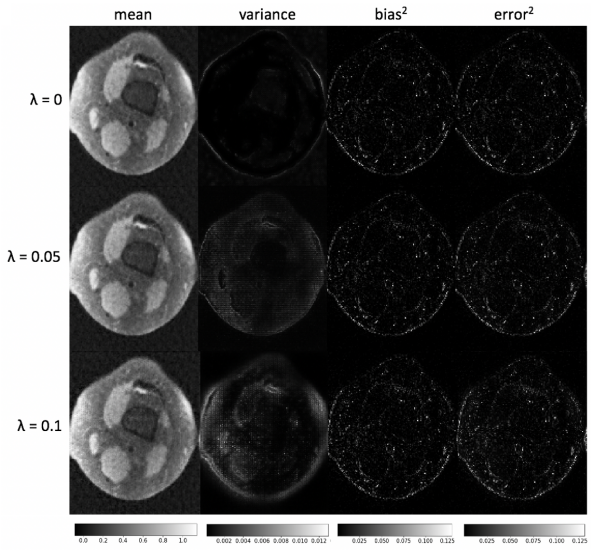

Figure 6: Mean reconstruction, pixel-wise variance, bias^2, and error^2 for a given reference slice across all realizations (5 fold undersampling and one recurrent block). Row 1: 0% GAN loss ($\lambda = 0$) . Row 2: 5% GAN loss ($\lambda = 0.05$). Row 3: 10% GAN loss ($\lambda = 0.10$).

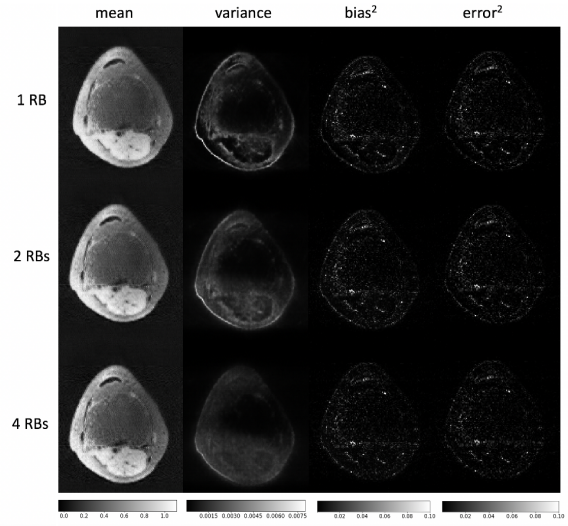

Figure 7: Mean reconstruction, pixel-wise variance, bias^2, and error^2 for a given reference slice across all realizations (5 fold undersampling and no adversarial loss). Row 1: one RB. Row 2: two RBs. Row 3: four RBs.