Revisiting Pose Sensitivity in Splat-based Computed Tomography under Sparse-view Reconstruction

Supplementary Material

1. Computation Time

Table 1 compares the computation times of the baseline method, NeAT [3], Thies et al. [6], and ours. Our approach outperforms state-of-the-art joint calibration and reconstruction methods. The baseline exhibits either lower or higher runtimes depending on the scene. Even though our method includes an additional calibration step, the computation time does not notably increase, since our method does not include the volumetric TV regularization that requires costs, and the calibration stage significantly reduces reconsruction errors, which saves computational costs.

Table 1. Computation time comparison in minutes.

	Computation time (min)			
Scene	Baseline [8]	NeAT [3]	Thies et al. [6]	Ours
Chest	34.59	53.71	31.25	**25.44**
Foot	20.54	49.17	31.77	**19.62**
Head	26.03	51.43	31.01	**22.10**
Jaw	**14.85**	50.82	31.00	17.99
Pancreas	24.64	50.70	31.00	**22.17**
Beetle	**9.27**	36.41	31.00	13.58
Bonsai	20.64	45.15	30.76	**19.17**
Broccoli	29.21	50.16	31.25	**23.18**
Kingsnake	**13.55**	43.66	31.78	16.14
Pepper	38.50	50.97	31.00	**29.10**
Backpack	**13.70**	47.86	31.16	16.08
Engine	40.22	50.37	31.50	**29.10**
Mount	37.88	51.11	31.08	**28.69**
Present	**12.85**	50.29	30.99	15.80
Teapot	**11.41**	43.45	30.76	15.23
Mean	23.19	48.35	31.15	**20.89**

2. Effect of View Count on FDK

We model cone-beam CT reconstruction via the FDK algorithm [2] as

$$\mathbf{x} = \mathcal{R}_{\mathrm{FDK}}^{-1}\left(\{\mathbf{y}_i\}_{i=1}^{N}\right), \tag{1}$$

where $\mathbf{y}_i \in \mathbb{R}^M$ denotes the projection data for view i, $\mathbf{x} \in \mathbb{R}^V$ is the reconstructed volume, and N is the number of projection views. The projection image $\mathbf{y}_i$ is given by

$$\mathbf{y}_i = P(\boldsymbol{\theta}_i, \mathbf{t}_i)\,\mathbf{f} + \boldsymbol{\epsilon}_i, \tag{2}$$

where $P(\boldsymbol{\theta}_i, \mathbf{t}_i)$ is the forward projection operator parameterized by the rotation $\boldsymbol{\theta}_i$ and translation $\mathbf{t}_i$, $\mathbf{f}$ is the ground-truth object, and $\boldsymbol{\epsilon}_i$ is a measurement noise. We now introduce geometry calibration errors:

$$\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \delta\boldsymbol{\theta}_i, \quad \hat{\mathbf{t}}_i = \mathbf{t}_i + \delta\mathbf{t}_i, \tag{3}$$

where $\delta\boldsymbol{\theta}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\theta)$ and $\delta\mathbf{t}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_t)$ are zero-mean Gaussian perturbations, independent across views. The reconstruction with perturbed geometry is then

$$\hat{\mathbf{x}} = \mathcal{R}_{\mathrm{FDK}}^{-1}\left(\{\mathbf{y}_i, \hat{\boldsymbol{\theta}}_i, \hat{\mathbf{t}}_i\}_{i=1}^{N}\right). \tag{4}$$

For sufficiently small perturbations, the reconstruction error can be linearized by a first-order Taylor expansion:

$$\hat{\mathbf{x}} - \mathbf{x} \;\approx\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{J}_i\,\delta\mathbf{p}_i, \tag{5}$$

where $\delta\mathbf{p}_i = \begin{bmatrix}\delta\boldsymbol{\theta}_i^T & \delta\mathbf{t}_i^T\end{bmatrix}^T$ is the pose error vector for a view i, and $\mathbf{J}_i$ is the Jacobian of the reconstruction operator with respect to the geometry parameters at the view i. Based on the independence of the Gaussians, we suppose $\mathbb{E}[\delta\mathbf{p}_i] = 0$ and $\mathrm{Cov}(\delta\mathbf{p}_i) = \boldsymbol{\Sigma}_p$, then the mean reconstruction error is zero:

$$\mathbb{E}[\hat{\mathbf{x}} - \mathbf{x}] \approx 0. \tag{6}$$

The covariance of the reconstruction error is

$$\mathrm{Cov}(\hat{\mathbf{x}} - \mathbf{x}) \approx \frac{1}{N^2}\sum_{i=1}^{N} \mathbf{J}_i\,\boldsymbol{\Sigma}_p\,\mathbf{J}_i^T. \tag{7}$$

If we further assume $\mathbf{J}_i \approx \mathbf{J}$ for all i, this simplifies to

$$\mathrm{Cov}(\hat{\mathbf{x}} - \mathbf{x}) \approx \frac{1}{N}\,\mathbf{J}\,\boldsymbol{\Sigma}_p\,\mathbf{J}^T. \tag{8}$$

The mean squared reconstruction error (MSE) is proportional to the trace of this covariance:

$$\mathrm{MSE} \propto \frac{1}{N}, \tag{9}$$

and therefore the root mean squared error (RMSE) scales as

$$\mathrm{RMSE} \propto \frac{1}{\sqrt{N}}. \tag{10}$$

Thus, under the assumption of independent, zero-mean Gaussian geometry perturbations, the RMSE of reconstruction decreases proportionally to $1/\sqrt{N}$ as the number of views increases.

3. Validation of Synthetic Perturbation

In the dataset generation process, we model 3D rotation matrix $\mathbf{R}$ by sampling a Lie algebra $\phi = \begin{bmatrix}\phi_1 & \phi_2 & \phi_3\end{bmatrix}^T \in \mathfrak{so}(3)$ from a normal distribution in the tangent space at $\mathbf{I}_{3\times3}$, given by:

$$\mathbf{R} = \exp\left(\phi^\wedge\right) \in \mathrm{SO}(3), \tag{11}$$

where $\phi \in \mathfrak{so}(3)$ is a Lie algebra component, exp denotes the exponential map from the Lie algebra $\mathfrak{so}(3)$ to the Lie group $\mathrm{SO}(3)$ and the hat operator $\wedge$ maps a 3D vector to a skew-symmetric matrix [5]. Then, we generate $\mathbf{R}$ by sampling ϕ following the zero-mean Gaussian distribution below:

$$\phi \sim \mathcal{N}(0, \sigma_{\text{rot}}^2 \mathbf{I}_{3\times3}). \tag{12}$$

To compute the expected value of the rotation matrix, we leverage the Taylor series expansion of $\exp\left(\phi^\wedge\right)$:

$$\mathbb{E}\left[\mathbf{R}\right] = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\phi^\wedge\right)^n. \tag{13}$$

For small σ_{rot}, we can approximate $\mathbf{R} \approx \mathbf{I}_{3\times3} + \phi^\wedge + \frac{1}{2}\left(\phi^\wedge\right)^2$. Since

$$\left(\phi^\wedge\right)^2 = \begin{bmatrix} -\phi_2^2 - \phi_3^2 & \phi_1\phi_2 & \phi_1\phi_3 \\ \phi_1\phi_2 & -\phi_1^2 - \phi_3^2 & \phi_2\phi_3 \\ \phi_1\phi_3 & \phi_2\phi_3 & -\phi_1^2 - \phi_2^2 \end{bmatrix}, \tag{14}$$

$\mathbb{E}\left[\phi_i^2\right] = \sigma_{\text{rot}}^2$ for $i \in \{1,2,3\}$ and $\mathbb{E}\left[\phi_i\phi_j\right] = 0$ for $i, j \in \{1,2,3\}$ and $i \neq j$, we obtain the expectation of second order term as:

$$\mathbb{E}\left[\phi^\wedge\right] = \mathbf{0}, \tag{15}$$

$$\mathbb{E}\left[\left(\phi^\wedge\right)^2\right] = -2\sigma_{\text{rot}}^2 \mathbf{I}_{3\times3}, \tag{16}$$

$$\mathbb{E}\left[\mathbf{R}\right] \approx (1 - \sigma_{\text{rot}}^2)\mathbf{I}_{3\times3}. \tag{17}$$

Therefore, the expectation of $\mathbf{R}$ is approximately the identity matrix with small σ_{rot}. In conclusion, we can generate *unbiased* random translation and rotation perturbation via sampling from a normal distribution in the Euclidean and tangent space, respectively.

4. Comparison Details

In our evaluation experiments, we directly compare against NeAT [3] and Thies et al. [6], which are state-of-the-art joint calibration and reconstruction methods.

NeAT NeAT [3] introduces a hierarchical neural rendering pipeline that supports both CBCT reconstruction and system calibration. To remain consistent with their original training configurations, we transform the projection images into intensities using the Beer–Lambert law, with the maximum source intensity set to an unsigned 16-bit integer value (65535). Notably, we experimentally confirmed that applying min–max normalization in log-scale, as done in the original implementation, degrades the output quality. Therefore, we omit this step by disabling it in the code. Using these intensity images, we follow the training schedule of 40 epochs as specified in the original paper, while other configurations—such as the octree update strategy, learning rates, and regularization coefficients—are directly adopted from the released implementation.

Thies et al. Thies et al. [6] jointly reconstruct and calibrate the CBCT using the gradient descent based on the filtered backprojection (FBP). In the paper, a score network is used to serve as a loss function to measure the quality of reconstruction. However, in their implementation, the score network is not given and the CBCT code is partially opened; only the differentiable FBP code is publicly available. We instead define the loss function that computes the L2-loss between the reconstructed volume and the reference volume created using the FBP and the dense (721) projections without geometric errors for fair comparison. Additionally, we implement the codes to import perspective projection matrices for each view, and the Ram-Lak filter to pre-filter the projections for the FBP. We set the number of iterations to 100 and the step-size to 5 that are empirically chosen that show the best performance.

With these adjustments, we were able to successfully run both baseline methods and conduct a direct comparison against our approach.

5. Impact of Pose Calibration

To disentangle the effect of pose calibration from that of the underlying reconstruction model, we conduct experiments evaluating NeAT [3], Thies et al. [6], and our method under pose-perturbed inputs, both with and without pose calibration enabled. As reported in Table 2, incorporating pose calibration consistently improves reconstruction quality across all evaluated methods. Notably, our method with pose calibration yields substantially larger gains, indicating that the performance improvement arises not only from the splat-based representation, but also from the effectiveness of the proposed pose calibration framework.

Table 2. Pose calibration effect (PSNR)

	NeAT		Thies et al.		Ours	
Scene	Cal. Off	Cal. On	Cal. Off	Cal. On	Cal. Off	Cal. On
Beetle	31.64	32.99	28.12	29.03	33.15	40.48
Broccoli	15.94	16.58	16.69	19.80	22.21	30.20
Engine	16.81	17.69	16.66	19.39	24.69	31.60
Pancreas	24.31	25.48	19.90	22.09	26.51	30.04

6. Comparisons with CT Novel View Synthesis Methods

We compare with X-Gaussian [1] and X-Field [7] by adapting them to the CT reconstruction setting. We provide 75 input projections and synthesize 721 novel views, followed by ASD-POCS [4] reconstruction as in the original protocols. Quantitative results are reported in Table 3.

Table 3. Additional comparison results against CT NVS methods.

	X-Gaussian			X-Field			Ours		
Scene	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
Beetle	18.88	0.673	0.1065	30.30	0.923	0.0286	**40.48**	**0.990**	**0.0095**
Broccoli	14.96	0.502	0.1787	18.99	0.686	0.1123	**30.20**	**0.931**	**0.0309**
Engine	15.36	0.418	0.1706	18.66	0.585	0.1167	**31.60**	**0.874**	**0.0263**
Pancreas	13.03	0.268	0.2230	23.06	0.620	0.0703	**30.04**	**0.890**	**0.0315**

7. Reproducibility of Pose-induced Artifacts

To validate that the observed needle-like artifacts originate from pose inaccuracies, we perform a controlled reproduction experiment by estimating pose errors from real sparse-view data and applying them to generate simulated projections from the ground-truth volume, which are then reconstructed using the same baseline R^2-Gaussian [8] pipeline (Figure 1). As shown in Figure 1(b), the simulated reconstruction reproduces artifact patterns consistent with the real result (Figure 1(a)), establishing a direct causal link between pose errors and needle-like artifacts.

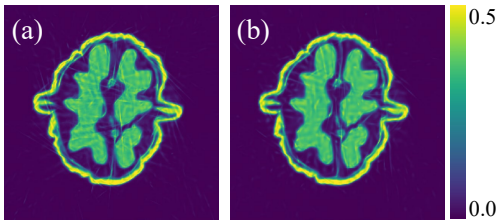

Figure 1. Reproducibility of a real artifact.

8. Limitation Analysis

We conduct an additional experiment under extremely sparse-view conditions by applying total variation (TV) regularization. As shown in Figure 2, the needle-like artifacts are effectively suppressed; however, the volumetric textures become noticeably over-smoothed. This result illustrates a clear trade-off between artifact suppression and structural detail preservation, which is particularly critical in applications such as medical imaging. We emphasize that this experiment is not intended to advocate the use of TV regularization in all cases, but rather to highlight the limitation of purely image-space regularization under severe view sparsity. Geometry-aware regularization remains an important direction for future work.

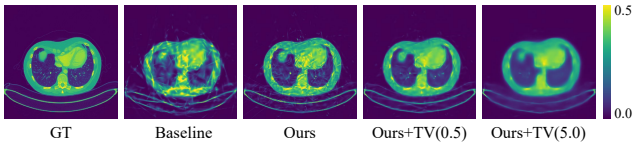

Figure 2. TV term impact under extreme sparse-view conditions.

9. Additional Results on Synthetic Dataset

We put additional figures for qualitative evaluation of ours against the SOTA joint reconstruction and calibration methods (NeAT [3], Thies et al. [6]) as Figures 3, 4, and 5. Ours reconstructs detailed edge and homogeneous region better than others although the geometry misalignment is involved in input projection images.

References

[1] Yuanhao Cai, Yixun Liang, Jiahao Wang, Angtian Wang, Yulun Zhang, Xiaokang Yang, Zongwei Zhou, and Alan Yuille. Radiative gaussian splatting for efficient x-ray novel view synthesis. In *ECCV*, 2024. 3

[2] Lee A Feldkamp, Lloyd C Davis, and James W Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A*, 1(6):612–619, 1984. 1

[3] Darius Rückert, Yuanhao Wang, Rui Li, Ramzi Idoughi, and Wolfgang Heidrich. Neat: Neural adaptive tomography. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 1, 2, 3, 4, 5, 6

[4] Emil Y Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 53 (17):4777, 2008. 3

[5] Joan Sola, Jeremie Deray, and Dinesh Atchuthan. A micro lie theory for state estimation in robotics. *arXiv preprint arXiv:1812.01537*, 2018. 2

[6] Mareike Thies, Fabian Wagner, Noah Maul, Haijun Yu, Manuela Goldmann Meier, Linda-Sophie Schneider, Mingxuan Gu, Siyuan Mei, Lukas Folle, Alexander Preuhs, et al. A gradient-based approach to fast and accurate head motion compensation in cone-beam ct. *IEEE Transactions on Medical Imaging*, 2024. 1, 2, 3, 4, 5, 6

[7] Feiran Wang, Jiachen Tao, Junyi Wu, Haoxuan Wang, Bin Duan, Kai Wang, Zongxin Yang, and Yan Yan. X-field: A physically informed representation for 3d x-ray reconstruction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3

[8] Ruyi Zha, Tao Jun Lin, Yuanhao Cai, Jiwen Cao, Yanhao Zhang, and Hongdong Li. R^2-gaussian: Rectifying radiative gaussian splatting for tomographic reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 3

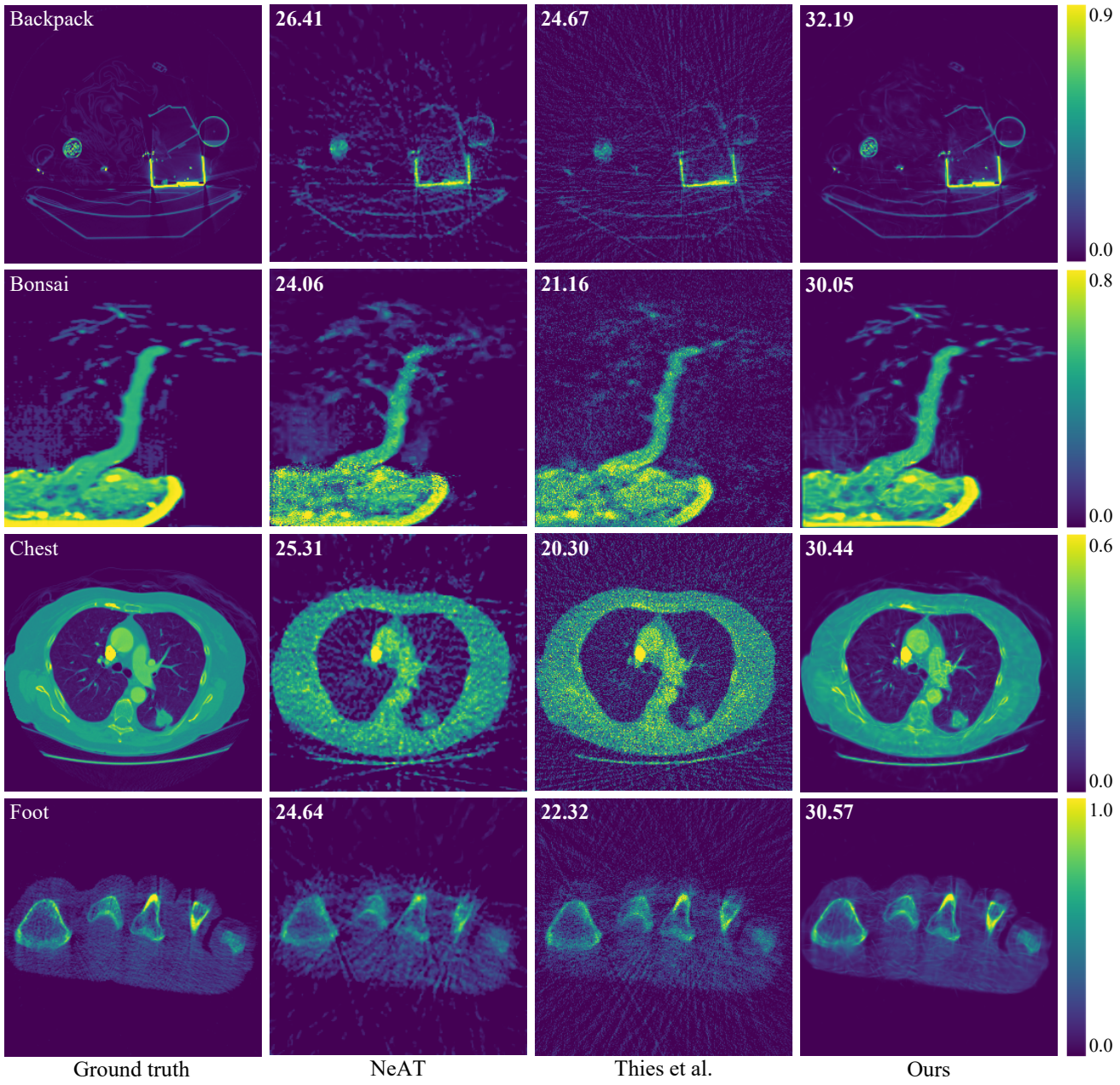

Figure 3. Comparison of reconstructed volume slices on the synthetic dataset with the pose perturbation. We compare our method against joint reconstruction and calibration approaches, including NeAT [3], and Thies et al. [6] across multiple objects. The numbers in the upper-left corners of each image indicate the PSNR values (dB) relative to the respective ground truth volume. Our method achieves higher PSNR values and better visual quality.

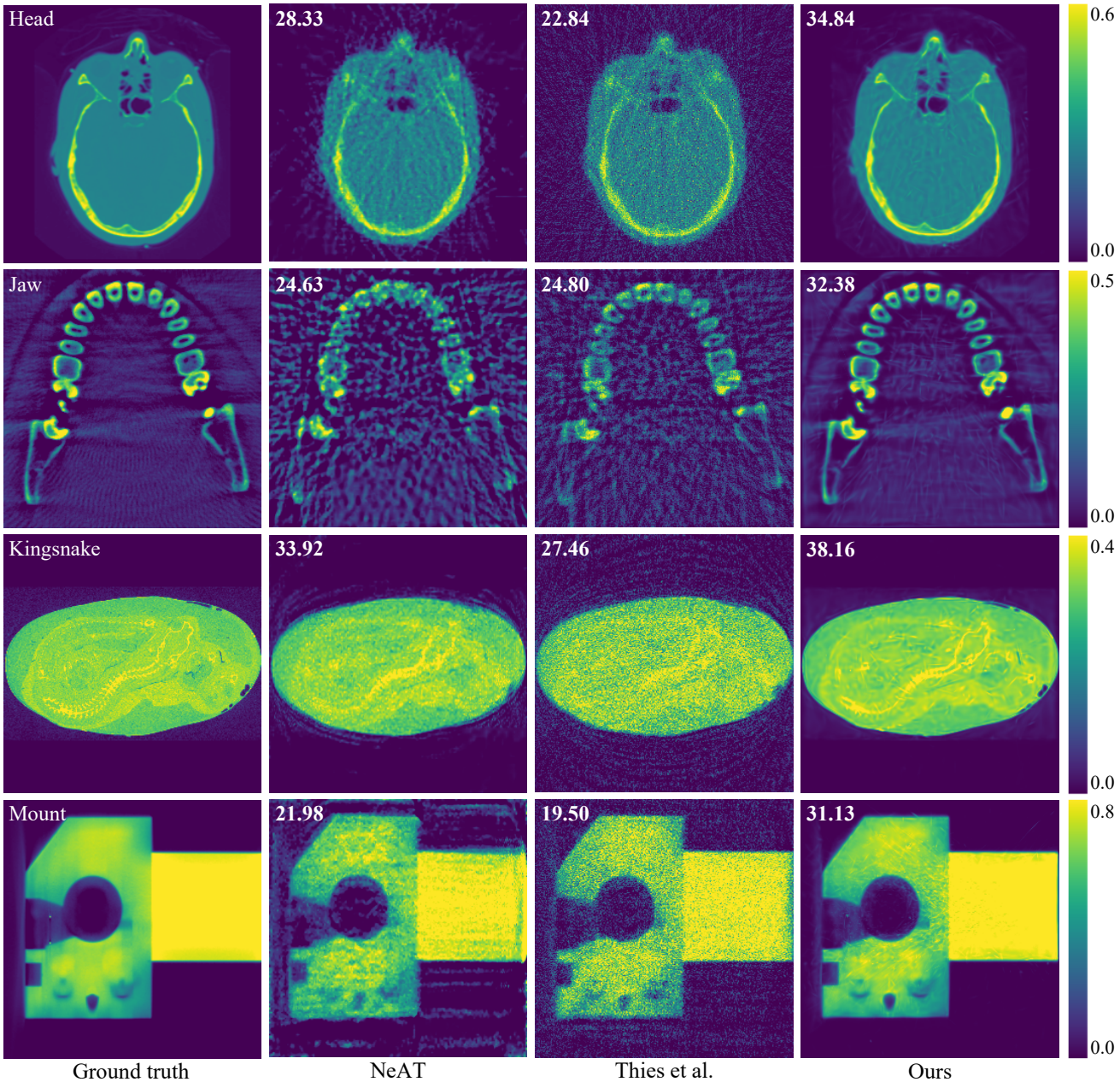

Figure 4. Comparison of reconstructed volume slices on the synthetic dataset with the pose perturbation. We compare our method against joint reconstruction and calibration approaches, including NeAT [3], and Thies et al. [6] across multiple objects. The numbers in the upper-left corners of each image indicate the PSNR values (dB) relative to the respective ground truth volume. Our method achieves higher PSNR values and better visual quality.

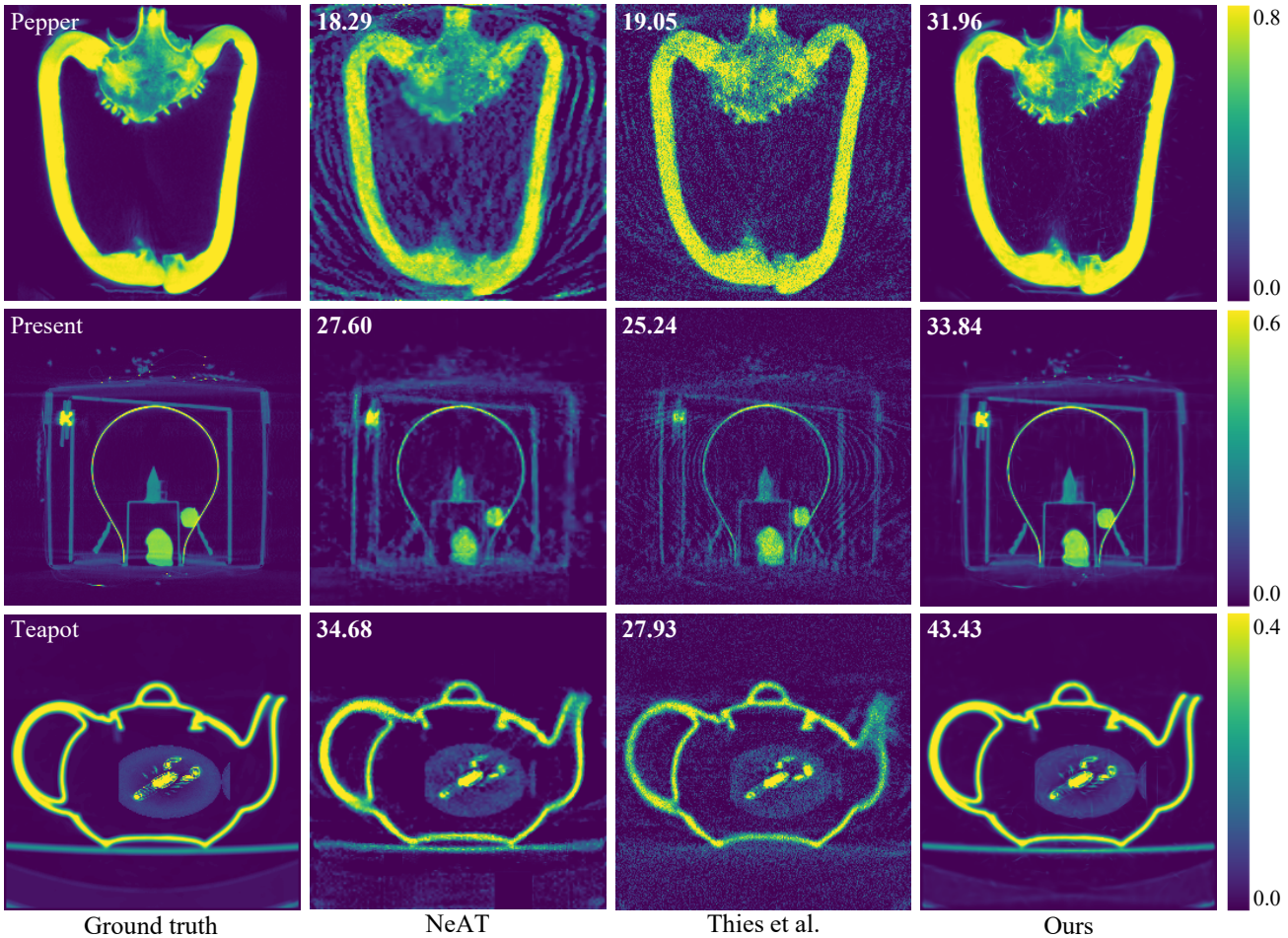

Figure 5. Comparison of reconstructed volume slices on the synthetic dataset with the pose perturbation. We compare our method against joint reconstruction and calibration approaches, including NeAT [3], and Thies et al. [6] across multiple objects. The numbers in the upper-left corners of each image indicate the PSNR values (dB) relative to the respective ground truth volume. Our method achieves higher PSNR values and better visual quality.