MORE: Multi-Organ Tomographic REconstruction Dataset *Appendix*

Shaokai Wu[1], **Yapan Guo**[2*], **Yanbiao Ji**[1], **Jing Tong**[1], **Yue Ding**[1]
Yuxiang Lu[1], **Mei Li**[1], **Suizhi Huang**[1], **Hongtao Lu**[1*]
[1]Shanghai Jiao Tong University [2]Suzhou Xiangcheng People's Hospital
https://more-med.github.io/

A Project Page

We provide a webpage for our **MORE** dataset.

B Code for Reproducibility

We provide the codes for all the benchmarks in our benchmark. The code is available in the attachment, including all compared methods, namely RED-CNN [1], AUTOMAP [9], MCG [3], DiffusionMBIR [2], SWORD [8], Score-MRI [4], NeRP [7], and the proposed GIFT. Now these codes are under the process of organizing and will be publicly available on GitHub soon.

C Dataset Details

Scan Parameters The CT scans were acquired using a Siemens SOMATOM Definition AS+ scanner, and the MRI scans were acquired using a Siemens MAGNETOM Avanto 1.5T scanner. The detailed scan parameters are included in the attachment 'CT_Parameters.csv' and 'MRI_Parameters.csv'.

Storage We have uploaded the dataset to Huggingface. The dataset is organized into two parts: the CT dataset and the MRI dataset. Both CT and MRI parts have been split by body parts for storage, which follows Figure 1 in the main paper.

Metadata We include the metadata of the dataset in the attachment 'Metadata.csv'. The metadata includes the patient ID, body part, lesion, and the number of scans for each instance.

D Experiment Details

Data Preprocessing We preprocess the raw DICOM files to 2D image slices following the standard practice in SimpleITK. These image slices are used as the ground truth, i.e., the target volume for the reconstruction task. These image slices are then projected to the measurement domain to simulate the acquisition process. This simulation process is implemented by the Radon transform in the CT case and the Fourier transform in the MRI case. Note that this is the common practice in medical image reconstruction research [8, 5, 3, 2, 7].

Implementation Details RED-CNN [1] is a denoising method that uses a residual encoder-decoder network to remove noise from the input image. Therefore, the input of RED-CNN is the noisy image reconstructed by FBP (for CT) or IFFT (for MRI), and then outputs a denoised image. On the other hand, the training process of AUTOMAP involves implementing a deep neural network with a feed-forward architecture composed of fully connected layers followed by sparse convolutional autoencoders. The network is trained to map sparse measurements to high-quality images.

The diffusion-based methods MCG [3], DiffusionMBIR [2], and SWORD [8] learn the diffusion process, and the reconstruction iteratively updates the image by the diffusion process. For the

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

Algorithm 1 Adaptive Densification

Require: $\mu \in \mathbb{R}^{n \times d}$: mean of Gaussians ($d = 3$ for 3D space)
Require: $\sigma \in \mathbb{R}^{n \times d}$: standard deviation of Gaussians
Require: $I \in \mathbb{R}^{n \times 1}$: intensity of Gaussians
Require: n_{max}: Maximum allowed quantity of Gaussians (500K by default)
Require: τ: Minimum gradient value (2e-4 by default)
Require: θ: Threshold determining Gaussian classification as small or large (0.005 of body diagonal length by default)
Require: $size$: Box size of each Gaussian (17 by default)
for $iteration \leftarrow 100$ **to** max_iter **step** 100 **do**
$avg\_grad \leftarrow \mu.grad/iteration$
$mask_{clone} \leftarrow (avg\_grad \geq \tau) \ \wedge (||\sigma||_2 \leq \theta)$
$available\_gaussians \leftarrow n_{max} - n$ ▷ Ensure the total number of Gaussians does not exceed the limit
$n_{clone} \leftarrow \min(available\_gaussians, \sum_{mask_{clone}})$
$sorted\_indices \leftarrow \texttt{argsort}(avg\_grad[mask_{clone}])$
$mask_{clone} \leftarrow mask_{clone} \ \wedge \ \texttt{top\_k}(sorted\_indices, n_{clone})$ ▷ Select top n_{clone} Gaussians to clone
$I[mask_{clone}] \leftarrow I[mask_{clone}]/2$ ▷ Halve the intensity of the cloned Gaussians to maintain the total intensity
$\mu_{clone} \leftarrow zero\_grad(\mu[mask_{clone}]), \ \sigma_{clone} \leftarrow zero\_grad(\sigma[mask_{clone}]), \ I_{clone} \leftarrow zero\_grad(I[mask_{clone}])$
$\mu \leftarrow \mu \cup \mu_{clone}, \ \sigma \leftarrow \sigma \cup \sigma_{clone}, \ I \leftarrow I \cup I_{clone}, \ n \leftarrow n + n_{clone}$
$mask_{split} \leftarrow (avg\_grad \geq \tau) \ \wedge (||\sigma||_2 > \theta)$
$available\_gaussians \leftarrow n_{max} - n$
$n_{split} \leftarrow \min(available\_gaussians, \sum_{mask_{split}})$
$sorted\_indices \leftarrow \texttt{argsort}(avg\_grad[mask_{split}])$
$mask_{split} \leftarrow mask_{split} \ \wedge \ \texttt{top\_k}(sorted\_indices, n_{split})$ ▷ Select top n_{split} Gaussians to split
$\mu_{new} \leftarrow PDF(\mu[mask_{split}], \sigma[mask_{split}], 2)$ ▷ Initialize new Gaussians by PDF sampling
$\sigma_{new} \leftarrow repeat(\sigma[mask_{split}], 2)/\sqrt[3]{2}, \ I_{new} \leftarrow repeat(I[mask_{split}], 2)$ ▷ Divide σ by $\sqrt[3]{2}$ to maintain the total volume
$\mu \leftarrow \mu \cup \mu_{new} - \mu_{split}, \ \sigma \leftarrow \sigma \cup \sigma_{new} - \sigma_{split}, \ I \leftarrow I \cup I_{new} - I_{split}, \ n \leftarrow n + n_{split}$
$mask_{prune} \leftarrow (avg\_grad) \leq \tau \ \vee (||\sigma||_2 > 3 \times size)$
$\mu \leftarrow \mu[\neg mask_{prune}], \ \sigma \leftarrow \sigma[\neg mask_{prune}], \ I \leftarrow I[\neg mask_{prune}], \ n \leftarrow n - \sum_{mask_{prune}}$
end for
return μ, σ, I

NeRF-based method NeRP [7], which does not need training data, directly uses a neural network to implicitly model the 3D volume and update the parameters under the supervision of the sparse measurements.

Our GIFT also does not need training data. Compared to the implicit modeling of NeRF, GIFT explicitly models the 3D volume as a set of Gaussians and updates the parameters of the Gaussians under the supervision of the sparse measurements. The Gaussians are Initialized with random means and standard deviations, and the intensity is initialized with the average intensity of the sparse measurements. Every 100 iterations, we adaptively densify the Gaussians to ensure the reconstruction quality inspired by 3DGS [6]. The detailed algorithm is shown in Algorithm 1.

More details about the implementation of the benchmarks are provided in the code repository.

E Datasheet

Motivation

Q1. For what purpose was the dataset created?

Answer: Mainly for two purposes: (1) to provide a comprehensive dataset of multiple body parts and lesions for medical image reconstruction research instead of focusing on a single organ or disease in the existing datasets; (2) to evaluate the performance of different reconstruction methods on a diverse dataset to ensure the robustness of the methods.

Q2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Answer: The dataset was created by the BCMI Lab at Shanghai Jiao Tong University and the Radiology Department at Suzhou Xiangcheng People's Hospital.

Q3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Answer: The dataset was created without any specific funding or associated grants.

Q4. Any other comments?

Answer: Examined by three experienced radiologists during the data collection process, we ensure the distribution of the dataset is consistent with the real-world distribution of medical images and thus suitable for evaluating the performance of medical image reconstruction methods in clinical practice.

Composition

Q5. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Answer: Each patient can be viewed as an instance in the dataset, and each instance contains a 3D volume of CT or MRI scans. Each scan is a DICOM file that represents a 2D image slice of the 3D volume.

Q6. How many instances are there in total (of each type, if appropriate)?

Answer: There are 135 CT instances and 54 MRI instances in total, with 65,575 CT scans and 7,498 MRI scans.

Q7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

Answer: This dataset contains all possible instances of body parts and lesions that are commonly seen in clinical practice and have been carefully selected by three experienced radiologists to ensure the diversity and representativeness of the dataset.

Q8. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?

Answer: Yes, we provide both the unprocessed DICOM files and the reconstructed 3D volumes for each instance. However, the private information of patients in the raw data has been removed to protect privacy.

Q9. Is there a label or target associated with each instance?

Answer: Yes, each instance has been labeled with the specific body part and lesion it represents.

Q10. Is any information missing from individual instances?

Answer: No.

Q11. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Answer: Yes, each scan is associated with a specific patient and the corresponding body part and lesion.

Q12. Are there recommended data splits (e.g., training, development/validation, testing)?

Answer: Yes, we provide a training set and a testing set for each type of scan. The training set is used to train the reconstruction models (for those training-based methods), and the testing set is used to evaluate the performance of the models.

Q13. Are there any errors, sources of noise, or redundancies in the dataset?

Answer: No. We aimed to provide a high-quality dataset for medical image reconstruction research, and the dataset has been carefully examined to ensure the quality and reliability of the data.

Q14. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

Answer: This dataset is self-contained since all the data was collected from Suzhou Xiangcheng People's Hospital.

Q15. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?

Answer: No. All the confidential information of patients has been removed to protect privacy, i.e., one cannot identify the patients from the dataset.

Q16. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

Answer: No.

Q17 Does the dataset identify any subpopulations (e.g., by age, gender)?

Answer: Yes, the dataset contains scans from patients and thus the subpopulations can be recognized. For example, the pelvic structures of males and females are different. However, these data are essential for medical image reconstruction research and do not contain any private information that can be used to identify individuals.

Q18 Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

Answer: No. The private information of patients has been removed to protect privacy, and the dataset cannot be used to identify the patients.

Q19 Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

Answer: No, and the dataset has been approved by the ethics committee of Suzhou Xiangcheng People's Hospital.

Q20 Any other comments?

Answer: This dataset is intended for research purposes only. All DICOM data has been anonymized to protect patient privacy and comply with the Helsinki Declaration.

Collection Process

Q21. How was the data associated with each instance acquired?

Answer: First, the patients underwent CT or MRI scans at Suzhou Xiangcheng People's Hospital. Then, the radiologists examined the scans and selected the representative scans for each body part and lesion. Finally, the selected scans were collected and stored in the dataset.

Q22. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

Answer: The data was collected using CT and MRI scanners at Suzhou Xiangcheng People's Hospital. The DICOM files were then extracted from the scanners and stored in the dataset.

Q23. If the dataset is a sample from a larger set, what was the sampling strategy?

Answer: This dataset is newly collected and does not sample from a larger set. However, during the data collection process, the radiologists selected the representative scans for each body part and lesion.

Q24. Who was involved in data collection process (e.g., students, crowd-workers, contractors) and how were they compensated (e.g., how much were crowd-workers paid)?

Answer: The collection process of the raw data was conducted by the radiologists at Suzhou Xiangcheng People's Hospital, and then the data was processed and stored by the BCMI Lab at Shanghai Jiao Tong University. No compensation was involved in the data collection process.

Q25. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?

Answer: The data was collected over a period of 2 months, from April 2024 to June 2024. The creation timeframe of the data associated with the instances matches the data collection timeframe.

Q26. Were any ethical review processes conducted (e.g., by an institutional review board)?

Answer: Yes, this dataset has been reviewed and approved by the ethics committee of Suzhou Xiangcheng People's Hospital. The approval number is 2024-KY-03.

Q27. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Answer: No. All data was collected directly from the patients who underwent CT or MRI scans at Suzhou Xiangcheng People's Hospital.

Q28 Were the individuals in question notified about the data collection?

Answer: This research is a retrospective study, for which a Waiver of Informed Consent Application has been signed and approved by the Ethics Committee.

Q29 Did the individuals in question consent to the collection and use of their data?

Answer: The same as the answer to the previous question.

Q30 If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

Answer: N/A; This dataset is intended for research purposes only, and should not be used for any other purposes.

Q31 Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

Answer: We have anonymized the data to protect patient privacy and comply with the Helsinki Declaration.

Q32 Any other comments?

Answer: This dataset has been collected in compliance with the Helsinki Declaration and the regulations of the ethics committee of Suzhou Xiangcheng People's Hospital.

Preprocessing, Cleaning, and/or Labeling

Q33. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Answer: Yes, and we provide both the raw DICOM files and the processed image files in the dataset. In the preprocessing step, the raw DICOM files were converted to 2D image slices following the standard practice in SimpleITK.

Q34. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Answer: Yes, both are provided to facilitate future research and unanticipated uses.

Q35. Is the software used to preprocess/clean/label the instances available?

Answer: Yes, all the programs are open-source and free to use.

Q36. Any other comments?

Answer: The raw data is stored in the DICOM format, and the processed image files are stored in the PNG format. Notably, the raw data is unique but there can be multiple ways to preprocess the data. For reproducibility and future research, we provide the processed image files in the dataset.

Uses

Q37. Has the dataset been used for any tasks already?

Answer: No, this dataset is novel and has not been used for any tasks yet.

Q38. Is there a repository that links to any or all papers or systems that use the dataset?

Answer: No, this dataset is novel and has not been used in any papers or systems yet.

Q39. What (other) tasks could the dataset be used for?

Answer: Besides medical image reconstruction, this dataset can also be used for the classification task of different lesions because each instance is labeled with the specific body part and lesion it represents.

Q40. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Answer: No, since both the raw data and the processed image files are provided in the dataset, the dataset will be easy to use for future research.

Q41. Are there any tasks for which the dataset should not be used?

Answer: No. However, this dataset is intended for research purposes only.

Q42. Any other comments?

Answer: The license of the dataset is CC BY-NC 4.0, which means the dataset can be used for non-commercial purposes with proper attribution.

Distribution

Q43. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Answer: Yes, our dataset will be made publicly available for research purposes.

Q44. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)

Answer: The dataset is hosted at Huggingface, and may also be uploaded to other platforms later. The latest news will be updated on our `https://more-med.github.io/`webpage.

Q45. When will the dataset be distributed?

Answer: It will be distributed after the publication of the paper.

Q46. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

Answer: Yes, the dataset will be distributed under the CC BY-NC 4.0 license.

Q47 Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

Answer: Since this dataset is intended for research purposes only, there are no restrictions on the data associated with the instances.

Q48. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

Answer: No.

Q49. Any other comments?

Answer: The dataset is intended for research purposes only and should not be used for any other purposes.

Maintenance

Q50. Who will be supporting/hosting/maintaining the dataset?

Answer: Currently, this dataset is hosted on the Hugging Face Datasets platform. To avoid data loss, we will also upload the dataset to other platforms such as Google Drive. The latest news will be updated in our webpage.

Q51. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Answer: Please contact us via email or the issue page on Hugging Face or GitHub.

Q52. Is there an erratum?

Answer: The erratum will be maintained on the webpage of this dataset.

Q53. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Answer: Yes, we will update the dataset if there are any labeling errors or new instances to be added.

Q54. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

Answer: The dataset is anonymized and does not contain any private information of patients. As long as the dataset is used for research purposes only, there are no limits on the retention of the data.

Q55. Will older versions of the dataset continue to be supported/hosted/maintained?

Answer: Yes, we will maintain the older versions of the dataset to ensure reproducibility and traceability of the research results.

Q56. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Answer: If so, they should also follow the CC BY-NC 4.0 license and provide proper attribution.

Q57. Any other comments?

Answer: We will keep maintaining the dataset and provide the latest information on the webpage.

References

[1] Hu Chen, Yi Zhang, Mannudeep K Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36(12):2524–2535, 2017.

[2] Hyungjin Chung, Dohoon Ryu, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Solving 3d inverse problems using pre-trained 2d diffusion models. ieee. In *CVF Conference on Computer Vision and Pattern Recognition*, volume 1, page 6, 2023.

[3] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

[4] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2022.

[5] Bing Guan, Cailian Yang, Liu Zhang, Shanzhou Niu, Minghui Zhang, Yuhao Wang, Weiwen Wu, and Qiegen Liu. Generative modeling in sinogram domain for sparse-view ct reconstruction. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2023.

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[7] Liyue Shen, John Pauly, and Lei Xing. Nerp: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[8] Kai Xu, Shiyu Lu, Bin Huang, Weiwen Wu, and Qiegen Liu. Stage-by-stage wavelet optimization refinement diffusion model for sparse-view ct reconstruction. *IEEE Transactions on Medical Imaging*, 2024.

[9] 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):487–492, 2018.