Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A foundation model for three-dimensional anatomical reconstruction from sparse X-ray projections

Yiqun Lin
University College London
Jiayang Xu
The Hong Kong University of Science and Technology
Lie Ju
University College London
Jiarong Guo
The Hong Kong University of Science and Technology
Weiru Wang
University College London
Hualiang Wang
Zhejiang University
Huifeng Yao
Allvista Technology (Hong Kong) Limited
Haoran Sun
Koln 3D Technology (Medical) Limited
Bin Pu
Hunan University
Kenli Li
Hunan University
Daniel Alexander
University College London
Hejie Cui
Stanford University
Yukun Zhou
yukun.zhou.19@ucl.ac.uk

University College London

Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11354431/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A foundation model for three-dimensional anatomical reconstruction from sparse X-ray projections

Yiqun Lin[1,2,3,4,5], Jiayang Xu[6], Lie Ju[1], Jiarong Guo[5,6], Weiru Wang[1,2,3,4], Hualiang Wang[7], Huifeng Yao[5], Haoran Sun[8], Bin Pu[9], Kenli Li[9], Daniel C. Alexander[3,4], Hejie Cui[10], and Yukun Zhou[1,2,3,4,*]

[1]Institute of Ophthalmology, University College London, London, UK
[2]NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, UK
[3]UCL Hawkes Institute, University College London, London, UK
[4]Department of Computer Science, University College London, London, UK
[5]Allvista Technology (Hong Kong) Limited, Hong Kong SAR, China
[6]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[7]Zhejiang University, Hangzhou, China
[8]Koln 3D Technology (Medical) Limited, Hong Kong SAR, China
[9]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[10]Stanford University, Stanford, CA, USA
*Corresponding author: yukun.zhou.19@ucl.ac.uk

Abstract

Computed tomography (CT) provides three-dimensional anatomical information but requires costly scanners and exposes patients to higher radiation doses than conventional X-ray imaging. Reconstructing three-dimensional anatomy from a few X-ray projections could extend widely available X-ray systems beyond two-dimensional imaging. We present SPARC, a foundation model pretrained once on 47,149 chest scans across diverse acquisition settings, enabling a single model to be adapted to different X-ray imaging configurations and anatomical regions. Across eight datasets covering five body regions, seven mainly outside the chest, SPARC reconstructed volumes with greater fidelity than five existing methods, exceeding the strongest in peak signal-to-noise ratio by 0.8–2.6 dB with eight projections and 1.4–2.7 dB with four. CT-based tools for abnormality classification and organ segmentation retained most of their performance when applied directly to SPARC's reconstructions from eight projections, fewer than 3% of the hundreds recorded by a CT scanner. They reached 98% of their area under the receiver operating characteristic curve on CT for 18 chest abnormalities, and 86% and 93% of their Dice score on CT for abdominal organs and lung lobes, with Dice scores 8.8% and 4.4% higher, in relative terms, than with the best competing method. Pretraining considerably increased the adaptation data efficiency. With ten patients, SPARC reconstructed chest CT more accurately than the same architecture trained from scratch on 100 patients. With the models and code released, SPARC could support the clinical translation of sparse-view reconstruction and expand access to three-dimensional imaging where CT scanners are unavailable.

Introduction

Computed tomography (CT) is the reference standard for imaging anatomy in three dimensions. A CT scanner rotates an X-ray source and a detector around the patient, recording hundreds of X-ray projections. Each projection measures how strongly the X-rays are attenuated along their paths through the body. From these projections, the scanner reconstructs a CT volume, a three-dimensional map of attenuation in Hounsfield units (HU). CT scanners are costly and fixed in place; their availability differs widely between and within health systems. CT also accounts for a large share of the radiation dose from medical imaging. X-ray systems that record projections are much more widely installed: radiography units, mobile and bedside units, interventional C-arms, and dental or orthopaedic cone-beam systems. They measure the same X-ray attenuation as CT, but record one or a few projections instead of hundreds. Reconstructing three-dimensional anatomy from a few X-ray projections would allow these systems to provide volumetric images, which they were not designed to produce. With far fewer projections, the radiation dose would also be lower than that of a CT scan at the same exposure per projection.

Reconstructing a three-dimensional CT volume from a few X-ray projections is a severely ill-posed inverse problem: many different volumes would produce the same few X-ray projections, so prior knowledge of anatomy is needed to identify the most plausible one. Per-scan optimisation methods represent each scan with an implicit neural representation[1–6] or with three-dimensional Gaussians[7–9] and optimise it separately for every scan. Because they use no prior knowledge from other patients, they require many projections unless an earlier scan of the same patient is available[2], and their reconstructions degrade severely with only a few projections. Learned methods instead learn this prior knowledge from paired X-ray projections and CT volumes. Feed-forward networks[10–15] map the projections of a patient to a three-dimensional volume in a single pass, but each is usually trained for one body region and one acquisition setting. Generative methods use diffusion models, either as a learned prior of CT images combined with a projection-consistency step[16–19] or conditioned directly on the X-ray projections[20–22], and they can produce anatomical structures that are not present in the patient. Post-processing methods[23–25] remove streak artefacts from a conventional reconstruction but were designed for many more projections than sparse acquisitions provide.

Large reconstruction models train one network on many CT volumes. DeepSparse was pretrained on 8,407 volumes from the head to the knee and is finetuned for each new dataset and number of projections[26]. X-LRM was developed on about 16,600 torso volumes and X-GRM on about 15,000 volumes of the torso and teeth; both were trained with 6, 8 or 10 projections[27,28]. In all three, the projections covered a full or half circle and each trained model assumed one acquisition geometry, whereas installed X-ray systems differ in angular range and geometry. Their evaluation also centred on image similarity to the reference CT, measured by the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM). These pixel-wise metrics can miss incomplete reconstruction of small or thin anatomical structures[29], and a reconstruction with a high PSNR can still shift the Hounsfield units on which CT-based analysis tools rely. Some of these studies also evaluated whether CT-based segmentation tools still segment the lungs or bones on the reconstructions[26,27], how robust reconstruction is to errors in the assumed geometry[27] and how much data are needed to adapt to a new dataset[26]. However, these evaluations were limited in depth and covered a narrow range of tasks and conditions. More comprehensive evaluation is needed before clinical use, to establish the reliability of CT-based analysis tools on reconstructed volumes, to assess robustness to the photon noise and calibration errors of installed X-ray systems, and to determine the requirements for projections, paired training data and computation.

Here we present SPARC, a foundation model that reconstructs three-dimensional anatomy from a few X-ray projections (Fig. 1a), together with the evaluation and analyses needed for its potential clinical use. SPARC was pretrained once, in about 1,700 GPU-hours, on 47,149 chest scans with varied

projection patterns (e.g., the number, angles and geometry of the projections; Fig. 1b), to learn anatomical knowledge that is not tied to one acquisition setting. We evaluated SPARC on eight datasets spanning five body regions (Fig. 1c) and compared it with five existing reconstruction methods. Beyond image fidelity, we evaluated whether CT-based analysis tools for abnormality classification and organ segmentation remain effective when applied directly to the reconstructed volumes, which indirectly measures how well clinically relevant abnormalities and anatomical structures are preserved. We also evaluated factors relevant to deployment, including computational cost, robustness to acquisition errors, the number of input projections and the amount of paired data required for adaptation. SPARC achieved the highest reconstruction fidelity among the evaluated methods on every dataset, and the CT-based analysis tools retained most of their performance on the reconstructed volumes. These findings suggest that SPARC could enable existing X-ray systems to recover clinically relevant three-dimensional information. We release the code together with the pretrained and finetuned models, so that others need not repeat the pretraining and can adapt SPARC to other X-ray systems with limited paired data and computing resources, a step towards extending three-dimensional imaging to settings where access to CT is limited.

Results

SPARC and study design

This study assessed how much of the anatomical and diagnostic information of CT can be recovered from a few X-ray projections (Fig. 1a). SPARC was pretrained once on 47,149 chest CT volumes from the CT-RATE training split[30]. Pretraining took 1,703 GPU-hours on NVIDIA GH200 GPUs. X-ray projections were simulated from each volume, as is common in learned sparse-view reconstruction[10,26,27], with a number of projections, an angular pattern and an acquisition geometry drawn at random for each sample (Fig. 1b). Besides a reconstruction loss, pretraining used CT-feature supervision, in which SPARC learns to predict from the projections the features that a frozen CT encoder extracts from the corresponding CT (Methods; Supplementary Fig. 1a,e). SPARC was evaluated on eight datasets from five body regions (Tab. 1; Fig. 1c), seven of which mainly cover body regions other than the chest used for pretraining. No test volume duplicates a training volume or, where patients are identified, comes from a training patient (Methods). SPARC was compared with five sparse-view methods, DIF-Net[10], DIF-Gaussian[12], C^2RV[11], SVCT[31] and DeepSparse[26], all trained on the same data under one training budget and stopping rule. The pretrained SPARC model was adapted to each of the eight datasets using that dataset's training set, with a separate model for each number of projections (Methods). To assess clinical utility, we applied two CT-based analysis tools trained on full CT, without modification, to the volumes reconstructed by each method: a classifier for 18 abnormalities, trained on CT-RATE, and a segmentation model for ten structures, trained on TotalSegmentator[32] (Supplementary Fig. 2a,b). We compared each tool's accuracy on the reconstructions with that on the full CT of the same patients (Fig. 1a).

Reconstruction fidelity was measured against the full CT by the PSNR and SSIM, and Hounsfield-unit accuracy by the mean absolute error within intensity bands. Classification was evaluated by the macro-averaged area under the receiver operating characteristic curve (macro-AUROC) and under the precision–recall curve (macro-AUPRC) over the 18 abnormalities, and by the sensitivity and specificity at operating points set on full CT. Segmentation was evaluated by the Dice coefficient, the normalised surface Dice and the 95th-percentile Hausdorff distance. All metrics were computed on the full test sets, with bootstrap confidence intervals and Holm-adjusted P values (Methods).

Tab. 1 | Datasets. Each evaluation dataset is split into training, validation and test sets; the validation set is used only to select checkpoints, to stop training early and to set the classifier's thresholds. Counts are CT volumes. MSD, Medical Segmentation Decathlon. Classification covers 18 abnormalities, and segmentation covers five abdominal organs and five lung lobes. Pretraining used chest CT only; the other seven datasets mainly cover body regions other than the chest.

Dataset	Body region	Use	Train / val. / test
CT-RATE[30], training split	Chest	Pretraining	47,149 / – / –
CT-RATE[30], validation cohort	Chest	Reconstruction, classification	2,134 / 303 / 602
TotalSegmentator[32]	Multiple regions	Reconstruction, segmentation	1,006 / 52 / 79
MSD[33]	Abdomen, chest	Reconstruction	986 / 212 / 209
AbdomenCT-1K[34]	Abdomen	Reconstruction	744 / 159 / 158
AMOS[35]	Abdomen	Reconstruction	350 / 75 / 75
CQ500[36]	Head	Reconstruction	330 / 71 / 71
ToothFairy3[37–39]	Maxillofacial	Reconstruction	372 / 80 / 77
VerSe[40]	Spine	Reconstruction	262 / 56 / 50

Reconstruction fidelity across eight datasets

Fig. 2 compares the reconstructions of SPARC and the five competing methods with eight projections. SPARC had the highest PSNR and SSIM on every dataset, including the seven datasets that mainly cover body regions other than the chest (Fig. 2a,b). SPARC's PSNR ranged from 29.48 dB (95% confidence interval (CI): [28.73, 30.25]) on VerSe to 33.30 dB (95% CI: [32.94, 33.66]) on ToothFairy3, and its SSIM from 86.23% (95% CI: [84.48, 87.90]) on VerSe to 96.89% (95% CI: [96.14, 97.37]) on CQ500 (Supplementary Tab. 2). The strongest competing method was DeepSparse on every dataset and metric except SSIM on CT-RATE, where it was C^2RV. SPARC's paired advantage over the strongest competing method in PSNR ranged from +0.83 dB (95% CI: [0.72, 0.94]) on AMOS to +2.62 dB (95% CI: [2.50, 2.73]) on TotalSegmentator. In SSIM, the advantage ranged from +0.82% (95% CI: [0.63, 1.08]) on ToothFairy3 to +7.13% (95% CI: [6.88, 7.40]) on CT-RATE. All 16 paired differences were significant (all Holm-adjusted $P < 0.001$; Supplementary Tab. 3). On every dataset, SPARC's PSNR exceeded that of the strongest competing method in at least 98.7% of test volumes. Reconstructions by the competing methods showed streak artefacts, blocking or over-smoothing, most clearly in the head and abdomen (Fig. 2c). Examples from the other three datasets are shown in Supplementary Fig. 3. With four projections, SPARC again achieved the highest fidelity on every dataset (Supplementary Fig. 4), with every paired difference from the strongest competing method significant (all $P < 0.001$; Supplementary Tab. 3); examples from all eight datasets are shown in Supplementary Fig. 5 and 6.

Hounsfield-unit accuracy, spatial error and computational cost

Fig. 3a,b compares the Hounsfield-unit errors of the reconstructions on the CT-RATE test set, with eight projections. SPARC's mean absolute error was 8.6 HU (95% CI: [8.3, 8.9]) in air, 65.0 HU (95% CI: [63.1, 67.1]) in lung, 41.3 HU (95% CI: [40.8, 41.8]) in soft tissue and 142.9 HU (95% CI: [140.4, 145.5]) in bone (Fig. 3a). The strongest competing method in each band reached 12.8, 80.4, 51.6 and 196.6 HU, so SPARC had the lowest error in every band. Over the whole volume, SPARC's mean absolute error was 25.2 HU against 34.1 HU for DeepSparse, a paired difference of -8.9 HU (95% CI: [-9.2, -8.6]; $P < 0.001$; Supplementary Tab. 4). In the signed error (Fig. 3b), all methods underestimated the attenuation of soft tissue and bone. SPARC's signed error was -7.9 HU (95% CI: [-8.4, -7.4]) in soft tissue and -100.6 HU (95% CI: [-103.7, -97.6]) in bone, against -17.2 to -31.9 HU and -169.9 to -218.3 HU for the competing methods. SPARC therefore had the smallest underestimation of all methods. With four projections, SPARC again had the lowest mean absolute error in every band (Supplementary Tab. 4). On TotalSegmentator, SPARC had the lowest error in every intensity band and in all ten segmented structures (Supplementary Tab. 4 and 5).

Fig. 3c shows SPARC's errors in three zones around the central axis of each volume, defined by the body radius, the distance from the axis within which almost all of the body lies (Methods). In each of the eight datasets, the mean absolute error was largest in the central zone, smaller in the peripheral zone and smallest outside the body radius. On VerSe, for example, the error was 66 HU in the central zone, 39 HU in the peripheral zone and 11 HU outside (Supplementary Tab. 6). Supplementary Fig. 7a–d shows the error as a continuous function of the distance from the axis and by intensity band. Supplementary Fig. 7e,f shows the PSNR of each volume with eight projections against body width and against the PSNR of the same volume with four projections.

Fig. 3d–f compares the computational cost. With eight projections and a 256^3 output, SPARC reconstructed a volume in 3.89 s on one NVIDIA GH200 GPU, against 0.03 to 9.52 s for the competing methods (Fig. 3d). Because its point decoder predicts the attenuation at any query point, the same model can produce a volume at any output resolution: with eight projections, a 128^3 volume took 0.60 s and a 512^3 volume 29.7 s (Supplementary Fig. 8a; Supplementary Tab. 8). Adapting SPARC to a dataset

required 3 to 64 GPU-hours on one GPU, depending on the size of the dataset, compared with 1,703 GPU-hours for pretraining (Supplementary Tab. 10). With eight projections and a 256^3 output, the backbone took 123 ms and the point decoder 3.97 s, so most of the reconstruction time was spent in the point decoder (Fig. 3e,f). The point decoder's time grew almost linearly with the number of output voxels, whereas the backbone's time did not change. For the models trained on AMOS, with a 256^3 output, the point decoder's time also grew with the number of projections, from 2.18 s with two projections to 7.04 s with sixteen projections, whereas the backbone's time increased only from 79 ms to 179 ms (Supplementary Fig. 8d; Supplementary Tab. 9).

Robustness to acquisition errors

Fig. 4 compares the methods under photon noise and under errors in the assumed source-to-detector distance, with every eight-projection model re-evaluated on perturbed projections without retraining. At 10^4 incident photons per detector pixel, the highest noise level in this comparison, SPARC's PSNR decreased by at most 3.4 dB (Fig. 4a–d). SPARC's PSNR remained significantly higher than that of the strongest competing method on every dataset. On TotalSegmentator, CQ500 and VerSe, SPARC's PSNR advantage over the strongest competing method was larger at the lowest number of photons: 2.6, 1.1 and 2.2 dB at 10^6 photons and 3.3, 2.6 and 3.2 dB at 10^4 photons, respectively. On CT-RATE, the advantage remained at about 2 dB. A 5% error in the assumed source-to-detector distance (Fig. 4i–l) or an angular range limited to 90° (Supplementary Fig. 9q–t) reduced SPARC's PSNR by up to 13.3 dB. The reduction was larger than under any noise level and larger than for the competing methods. Supplementary Fig. 9e–h,m–p also shows the results under angular jitter and detector offsets. Across all 64 combinations of the 16 perturbation settings and four datasets, SPARC's PSNR was significantly higher than that of the strongest competing method in 54 combinations (Supplementary Tab. 11). A competing method had a significantly higher PSNR in five combinations, by at most 0.23 dB. All five combinations involved a source-to-detector distance error or a 90° angular range. In the remaining five combinations, the difference was not significant. SPARC's SSIM was significantly higher than that of the strongest competing method in all 64 combinations (Fig. 4e–h,m–p; Supplementary Tab. 13). Supplementary Fig. 10 shows the same analysis with four projections (Supplementary Tab. 12 and 14). Supplementary Fig. 11a–h shows the change in PSNR of SPARC alone at more levels of noise and geometric error. Supplementary Fig. 11i also shows the PSNR of SPARC on AMOS when a fixed photon budget was divided among 2, 4, 8 or 16 projections. Overall, SPARC was robust to photon noise but required accurate geometry.

Abnormality classification from sparse projections

Fig. 5 shows the classifier trained on full CT applied, without modification, to each method's test reconstructions. With eight projections, SPARC's reconstructions reached a macro-AUROC of 0.756 (95% CI: [0.730, 0.779]), against 0.772 (95% CI: [0.746, 0.795]) on full CT, a paired difference of -0.016 (95% CI: [-0.027, -0.005]; $P < 0.05$). The competing reconstructions reached 0.717 to 0.745 (Supplementary Tab. 18). SPARC outperformed DIF-Net, DIF-Gaussian, C^{2}RV and DeepSparse (paired differences +0.015 to +0.039, all $P < 0.05$), whereas its paired difference from SVCT, +0.011 (95% CI: [0.000, 0.023]; $P > 0.05$), was not significant (Supplementary Tab. 19). SPARC's reconstructions also had the highest macro-AUPRC among the reconstructions (0.403, against 0.430 on the CT), although the paired difference was significant only against DIF-Gaussian and DeepSparse (both $P < 0.01$). The classifier's thresholds were fixed for each finding on the full-CT validation set. At these thresholds, the sensitivity on SPARC's reconstructions was maintained (0.678, compared with 0.672 on the CT) and the specificity decreased by 0.056. The reconstructions by DIF-Net, DIF-Gaussian and C^{2}RV decreased the specificity by 0.111 to 0.131, and those by SVCT and DeepSparse decreased the sensitivity by 0.031

to 0.046 (Supplementary Tab. 20). With four projections, SPARC again had the highest macro-AUROC (0.741; Supplementary Fig. 12), although its differences from SVCT and DeepSparse were not significant (both $P > 0.05$; Supplementary Tab. 19). The largest per-finding differences between SPARC's reconstructions and full CT were for lung opacity and consolidation (Fig. 5b). Fig. 5c shows the classifier's Grad-CAM[41] maps for four findings, on full CT and on each reconstruction. We analysed all 1,311 positive volume–finding pairs of ten findings with an anatomical location. The classifier's Grad-CAM map on SPARC's reconstruction correlated with its map on the CT at 0.669 on average, against 0.507 to 0.585 for the competing reconstructions. Compared with the map on each competing reconstruction, the map on SPARC's reconstruction had a higher correlation with the map on the CT in 94 to 99% of the volume–finding pairs (Supplementary Tab. 21). Overall, the classifier trained on full CT retained most of its accuracy on SPARC's reconstructions, and its Grad-CAM maps were consistent with those on CT.

Multi-organ segmentation from sparse projections

Fig. 6 shows the segmentation model trained on full CT applied, without modification, to each method's reconstructions. With eight projections, SPARC's reconstructions reached a mean Dice of 0.740 (95% CI: [0.705, 0.772]) on the five abdominal organs and 0.819 (95% CI: [0.788, 0.846]) on the five lung lobes. On full CT, the mean Dice was 0.860 (95% CI: [0.831, 0.884]) and 0.881 (95% CI: [0.853, 0.906]). The competing reconstructions reached 0.389 to 0.680 on the abdominal organs and 0.607 to 0.784 on the lung lobes. SPARC's reconstructions outperformed those of every competing method, by +0.060 to +0.351 on the abdominal organs and +0.035 to +0.211 on the lung lobes (all $P < 0.001$). With SPARC's reconstructions, the segmentation model reached the highest normalised surface Dice at 2 mm, 0.471 on the abdominal organs and 0.682 on the lung lobes (full CT: 0.776 and 0.828). The 95th-percentile Hausdorff distance with SPARC's reconstructions was also the smallest, 22.4 mm and 23.5 mm (full CT: 16.8 mm and 14.4 mm), as was the average symmetric surface distance. The paired differences from the competing reconstructions in these three measures were significant (all $P < 0.05$), except those from DeepSparse in average symmetric surface distance on the abdominal organs and lung lobes (both $P > 0.05$). Structure volumes measured on SPARC's reconstructions had the smallest absolute error among the reconstructions (Supplementary Tab. 24). With four projections (Supplementary Fig. 14), SPARC's reconstructions reached a mean Dice of 0.685 on the abdominal organs and 0.775 on the lung lobes. SPARC's reconstructions again outperformed those of every competing method in both groups (Supplementary Tab. 23 and 25). The largest differences from full CT were for the left and right kidneys (Fig. 6b and Supplementary Tab. 22). Fig. 6c shows four test volumes segmented on the CT and on each reconstruction, next to the reference labels.

Effect of the number of projections

Fig. 7 shows how reconstruction and the clinical tasks change with the number of projections. Increasing the number of projections from four to eight improved SPARC's PSNR in every dataset, by 1.87 to 2.48 dB (Fig. 7a). Lung vessels and bone edges also became sharper (Fig. 7f). With eight instead of four projections, the mean Dice of the segmentation model on SPARC's reconstructions rose from 0.685 to 0.740 on the abdominal organs (Fig. 7e) and from 0.775 to 0.819 on the lung lobes. The macro-AUROC of the classifier on SPARC's reconstructions rose from 0.741 to 0.756 (Fig. 7c). SPARC and DeepSparse, the competing method with the highest reconstruction fidelity, were also trained on CT-RATE and TotalSegmentator with two or sixteen projections, with a separate model for each number of projections. On TotalSegmentator, SPARC's PSNR rose from 27.66 dB with two projections to 32.83 dB with sixteen projections, and the abdominal Dice of the segmentation model on its reconstructions from 0.641 to 0.780. SPARC outperformed DeepSparse with every number of projections, by 0.072 to 0.192 in

Dice (all $P < 0.001$), with the largest difference at two projections. On CT-RATE, SPARC's PSNR rose from 27.77 dB with two projections to 32.86 dB with sixteen projections, and the macro-AUROC of the classifier on its reconstructions from 0.734 to 0.767. The macro-AUROC on SPARC's reconstructions was higher than on DeepSparse's reconstructions with every number of projections, by 0.007 to 0.019, with the largest difference at two projections. However, the difference was significant only with eight and sixteen projections (both $P < 0.05$; Fig. 7b–e). Overall, reconstruction and both clinical tasks improved from two to sixteen projections.

Contribution of pretraining

Fig. 8a,b compares SPARC with the same architecture trained from random initialisation as the number of CT-RATE patients used for adaptation was varied from 10 patients (28 volumes) to all 913 patients (2,134 volumes). The adaptation protocol and the test set were the same for every number of patients. SPARC's PSNR was higher than that of the architecture trained from scratch by +4.21 dB (95% CI: [4.15, 4.28]; $P < 0.001$) with 10 patients and by +0.84 dB (95% CI: [0.81, 0.87]; $P < 0.001$) with all 913 patients, so the benefit of pretraining was largest with the fewest patients. With 10 patients, SPARC's PSNR and SSIM (28.17 dB and 84.59) already exceeded those of the architecture trained from scratch on 100 patients (27.68 dB and 80.66; Supplementary Tab. 27). SPARC's reconstructions from 10 patients already showed the lung vessels that the architecture trained from scratch blurred (Fig. 8d). DeepSparse and SVCT were also trained on the same patient subsets (Supplementary Tab. 27). With every number of patients, SPARC's PSNR and SSIM were significantly higher than those of both methods (all $P < 0.001$).

Fig. 8c compares three starting points for adapting the same architecture: no pretraining (random initialisation), pretraining with the reconstruction loss only, and SPARC's pretraining with both losses. The adaptation protocol was the same for all three starting points. Without pretraining, the reconstruction PSNR was 30.56 dB on CT-RATE and 30.22 dB on TotalSegmentator. Pretraining with the reconstruction loss added 0.30 dB (95% CI: [0.26, 0.33]) on CT-RATE and 1.09 dB (95% CI: [0.99, 1.19]) on TotalSegmentator. CT-feature supervision added a further 0.54 dB (95% CI: [0.52, 0.56]) on CT-RATE and 0.07 dB (95% CI: [0.05, 0.09]) on TotalSegmentator (all $P < 0.001$; examples in Fig. 8e). The classifier trained on full CT reached a macro-AUROC of 0.753 on the CT-RATE reconstructions without pretraining. Neither the reconstruction loss (+0.001) nor CT-feature supervision (a further +0.002) changed the macro-AUROC significantly (both $P > 0.05$). On TotalSegmentator, the segmentation model trained on full CT reached a mean Dice of 0.682 on the abdominal organs and 0.802 on the lung lobes without pretraining (Supplementary Tab. 28). The reconstruction loss raised the Dice by 0.051 (95% CI: [0.045, 0.058]; $P < 0.001$) on the abdominal organs and 0.018 (95% CI: [0.014, 0.022]; $P < 0.001$) on the lung lobes. CT-feature supervision changed the Dice by a further +0.007 (95% CI: [0.005, 0.011]; $P < 0.001$) on the abdominal organs and -0.002 (95% CI: $[-0.005, 0.001]$; $P > 0.05$) on the lung lobes. Overall, both pretraining losses improved reconstruction. The better reconstructions raised the Dice of the segmentation model but did not significantly change the macro-AUROC of the classifier.

Discussion

In this study, we developed SPARC, a foundation model pretrained once on chest CT to reconstruct three-dimensional anatomy from a few X-ray projections, and evaluated both the fidelity of its reconstructions in five body regions and their use with CT-based analysis tools. Adapted to each dataset, the pretrained model reconstructed anatomy with greater fidelity than competing methods in all five regions, and pretraining reduced the paired data (projections with the corresponding CT) that adaptation required. These results suggest that sparse-view reconstruction need not be developed separately for each body region and acquisition setting. Classification and segmentation models trained on CT also retained most of their accuracy when applied unchanged to SPARC's reconstructions, which complements image similarity by showing the clinical value of the reconstructions more directly. For device repurposing, these findings suggest that an installed X-ray system could gain three-dimensional imaging by adapting our shared model and calibrating the system's geometry, without new reconstruction models or analysis tools.

SPARC's strong performance beyond the chest may reflect its pretraining across diverse acquisition geometries rather than a single fixed geometry. Previous sparse-view methods are trained or finetuned for one dataset and number of projections[10–12,26,31], and recent large reconstruction models assume one acquisition geometry with projections over a full or half circle[27,28]. Such models may relate each detector pixel to a fixed position in the body, a relation that no longer holds when the geometry or the body region changes. SPARC instead receives the geometry of every ray as input and was pretrained with varied numbers of projections, angular patterns and source-to-detector distances. The model therefore has to learn how a volume is projected onto the detector, a relation that is the same for every body region. During adaptation, the model then only has to learn the anatomy of each body region, which may explain why SPARC, pretrained on chest CT only, also reconstructed the head, the spine and the abdomen with the greatest fidelity (Fig. 2). On the abdominal datasets, SPARC also outperformed DeepSparse, although DeepSparse had been pretrained on CT that included the abdomen[26]. The architecture also contributed: without pretraining, it already outperformed every competing method on CT-RATE and TotalSegmentator when adapted with the full training set (Fig. 2a and 8c). SPARC's design is therefore a better reconstruction model in itself, independent of pretraining.

The errors of SPARC were largest near the centre of the body, in patients with a wider body cross-section and in dense bone (Fig. 3a–c; Supplementary Fig. 7). Rays through the centre of a wide body pass through the most tissue. In each projection, many structures in this region overlap, and a few projections cannot separate them, so the model relies more on the anatomical knowledge learned in training. Patients with a wider body cross-section may therefore require more projections or a higher exposure. All methods also underestimated the attenuation of dense bone. A possible reason is that a model trained with a squared-error loss predicts an average of the volumes that fit the projections. When the exact position of a thin, dense structure is uncertain, this average blurs the structure and lowers its peak value. Bone measurements on the reconstructions should therefore be interpreted with caution. Accurate Hounsfield units also matter for reading the images. Clinicians view CT with fixed display windows and distinguish tissues such as fat, fluid, blood and calcification by their attenuation. Regional errors in Hounsfield units could therefore alter both the appearance and the interpretation of a structure.

CT-based analysis tools retained most of their accuracy when applied directly to SPARC's reconstructions (Fig. 5 and 6). Existing tools developed for CT could therefore be used on the reconstructions without new labelled data. The tools performed worse on reconstructions with more streaks, blurring and intensity bias, which differ more from CT (Fig. 2c and 3a,b; Supplementary Tab. 19 and 23). When labelled data are available, SPARC can also be used in two other ways. First, to further improve tasks that need fine detail, such as segmentation, a task network can be trained on SPARC's reconstructions instead of on CT (Supplementary Fig. 13a). Second, the tasks can be performed by heads on SPARC's encoder,

which skip reconstruction. Classification then becomes much faster with similar accuracy, although the abdominal organs are segmented less accurately (Supplementary Fig. 13b). For some findings, such as cardiomegaly, pleural effusion and consolidation, the classification head also attributed its predictions more to the involved organ than a classifier trained on the reconstructions (Supplementary Fig. 13c), possibly because the encoder was pretrained to predict the CT features at each location.

Compared with organ segmentation, abnormality classification was less affected by the loss of detail. Across reconstruction methods, numbers of projections and pretraining settings, the accuracy of the classifier varied less than that of the segmentation model (Fig. 5–8). A finding labelled for the whole volume, such as a pleural effusion or an enlarged heart, can be detected from coarse changes in attenuation. In contrast, segmenting a small, low-contrast organ adjacent to other organs requires fine detail, which a few projections determine poorly. Improvements in reconstruction detail therefore benefit segmentation more than classification. Consistent with this explanation, the kidneys, which are small, low-contrast and surrounded by fat and bowel, showed the largest performance degradation in segmentation relative to full CT (Fig. 6b).

Pretraining was most beneficial when limited paired data were available for adaptation (Fig. 8a,b). A few projections are consistent with many volumes, and a reconstruction model must learn from training data which of them are anatomically plausible. Without pretraining, this anatomical knowledge and how the projections relate to the volume must both be learned from the adaptation data alone; pretraining on chest CT from other CT-RATE patients probably provided much of both. Pretraining had a similar effect for DeepSparse, which, when finetuned on 20% of a chest CT training set, reconstructed about as accurately as when trained from scratch on all of it[26]. The two pretraining losses indicate which part of this knowledge carries over to other body regions (Fig. 8c). Besides chest anatomy, the reconstruction loss teaches the model how the projections relate to the volume, a relation that does not depend on the body region; it accounted for nearly all of the improvement from pretraining on the multi-region TotalSegmentator dataset. CT-feature supervision instead teaches the model to predict the features of a CT encoder trained only on chest CT, which may explain why it improved reconstruction much more on CT-RATE than on TotalSegmentator. A CT encoder pretrained on other body regions may extend this benefit to those regions. Because real projections are rarely acquired together with CT of the same patient, needing fewer paired examples may be particularly useful when SPARC is adapted to real projections.

Deploying SPARC on existing X-ray systems mainly requires accurate geometric calibration. The model was very sensitive to errors in the assumed geometry: under every tested geometric error, it lost more accuracy than any competing method, although it usually remained the most accurate (Fig. 4i–p; Supplementary Fig. 9e–p). For each point of the volume, SPARC uses the given geometry to find the matching position in every projection and combines the features found there. With a wrong geometry, these positions show different anatomy in different projections, so features of different structures are mixed and fine detail is probably blurred. All methods lost accuracy under such errors, but the more accurate a method was without errors, the more it generally lost, probably because finer detail depends more on precise alignment. This reliance on the given geometry is also what allows one pretrained model to be adapted to many acquisition settings; training with perturbed geometry may reduce the resulting sensitivity. By contrast, SPARC was relatively robust to photon noise: its accuracy decreased only gradually, and it remained the most accurate method at every noise level tested (Fig. 4a–h). Noise, unlike a geometric error, varies randomly between pixels and does not move structures, so features are still combined at the correct positions. Accurate geometry therefore matters more for SPARC than low noise, and each X-ray system should undergo a one-time geometric calibration before clinical use, as is routine for C-arm and cone-beam CT systems.

This study has several limitations. First, all projections were simulated from CT. This limitation is common to learned sparse-view reconstruction rather than specific to SPARC: the compared methods and

most previous studies were also trained and evaluated on simulated projections[10,11,26,27], because large datasets of real X-ray projections paired with CT are not available. Real acquisitions also contain scatter, beam hardening, detector blur and patient motion, which were absent during training and may reduce the reported accuracy; SPARC therefore requires adaptation and evaluation on real projections before clinical use. Second, as with any learned reconstruction method, and particularly with generative methods[20–22], SPARC may produce structures that are not present in the patient or miss structures that are present. High average fidelity does not exclude such errors in individual patients, where they could change the diagnosis, so clinical use would require a more fine-grained examination pipeline to identify unreliable regions. Third, clinical utility was assessed using research analysis tools and report-derived labels rather than direct evaluation by radiologists; whether radiologists or approved clinical software would reach the same conclusions from the reconstructions remains to be established. Finally, pretraining used chest CT only, and the clinical tasks covered only the chest and abdomen; the diagnostic value of reconstructions in other body regions, such as for head injuries or spinal fractures, remains to be studied.

In conclusion, by recovering much of the information that CT provides from a few simulated X-ray projections, SPARC offers a potential route to obtaining three-dimensional anatomical information from existing X-ray infrastructure. By sharing the models, code and data splits, we aim to lower the barriers to adapting SPARC to other imaging systems and evaluating it on real projections, supporting further development towards three-dimensional imaging in settings without access to CT.

Methods

Model description

SPARC reconstructs a three-dimensional attenuation volume from a few X-ray projections with known acquisition geometry (Supplementary Fig. 1). Each projection I_v, $v = 1, \ldots, V$, records the line integrals of the linear attenuation coefficient along the rays of a cone-beam geometry π_v, which is given by the poses of the source and detector together with the detector's intrinsic parameters. Training has three stages. A CT encoder is pretrained and frozen to supply the target for CT-feature supervision (Stage 1); the reconstruction backbone and point decoder are pretrained once on simulated projections (Stage 2); both are then adapted to each dataset and number of projections (Stage 3).

The CT encoder of Stage 1 is the ViT-L/16 encoder[42] of a three-dimensional masked autoencoder[43] (Supplementary Fig. 1g). The autoencoder splits a 256^3 CT volume into $16^3 = 4{,}096$ patches of 16^3 voxels, masks 75% of them and reconstructs the masked patches from the rest with a normalised-pixel loss. After pretraining, the decoder is discarded and the encoder is frozen; it maps a CT volume X to 4,096 tokens of dimension 1,024, $g_\phi(X)$. A frozen CT encoder, rather than an exponential moving average of the network being trained, is used because the target has to describe the CT from the start of Stage 2, when the backbone is still untrained.

The backbone of Stage 2 applies the same two-dimensional encoder to each projection (Supplementary Fig. 1c): a five-scale ConvNeXt[44] pyramid, from 512×512 to 32×32 pixels, with a six-layer transformer at the coarsest scale. The acquisition geometry is given to the two-dimensional network as Plücker ray embeddings. For each pixel, the direction $\mathbf{d}$ and moment $\mathbf{m}$ of its ray form a six-dimensional vector. The moment is divided by the source-to-axis distance so that both parts have similar magnitude. Each of the 32×32 positions of the coarsest scale forms one transformer token, which corresponds to a patch of 16×16 pixels in the projection. The six-dimensional vectors are averaged within each patch, embedded with a small multilayer perceptron and added to the tokens before the transformer. The output of the transformer is added back to the coarsest scale of the pyramid. A grid of 16^3 points, aligned with the token grid of the CT encoder, is then projected onto each X-ray projection (Supplementary Fig. 1d). The multi-scale features at the projected positions are sampled bilinearly, as in pixel-aligned neural rendering[45], and combined by taking their maximum across projections, as in DIF-Gaussian and C^{2}RV[11,12]; this maximum does not depend on the order or number of projections. An eight-layer three-dimensional transformer refines the grid into the feature volume $\mathbf{F}$, with 768 channels per grid point.

Two heads take $\mathbf{F}$ as input during pretraining (Supplementary Fig. 1e,f). The point decoder (Supplementary Fig. 1f) predicts the attenuation $\mu(\mathbf{q})$ at any query point $\mathbf{q}$. Its input is $\mathbf{F}$, sampled trilinearly at $\mathbf{q}$, concatenated with the multi-scale two-dimensional features at the projected positions of $\mathbf{q}$. These features are sampled bilinearly and combined by taking their maximum across projections, in the same way as for the grid points. A residual multilayer perceptron with three hidden layers of 512, 512 and 256 units maps this input to $\mu(\mathbf{q})$. Each hidden layer includes batch normalisation. Every layer after the first receives the output of the previous layer concatenated with this input. The point decoder is trained with the reconstruction loss $\mathcal{L}_{\text{rec}}$, the mean squared error against the CT. Because it predicts one point at a time, the same decoder produces any output grid. The CT-feature head (Supplementary Fig. 1e), which consists of four transformer blocks and a linear layer, maps $\mathbf{F}$ to the token space of the CT encoder. The head is trained to predict $g_\phi(X)$ with a smooth-L_1 loss, $\mathcal{L}_{\text{feat}}$, and the gradient is stopped at the target. We refer to training with $\mathcal{L}_{\text{feat}}$ as CT-feature supervision. CT-feature supervision follows joint-embedding predictive architectures, which predict the features of a target encoder rather than the input pixels[46–48]. Self-supervised encoders trained on CT[49,50], on radiographs[51] or on CT sinograms[52] alone do not learn this relation between X-ray projections and three-dimensional anatomy. X-WIN transfers CT knowledge to an encoder for single chest radiographs by predicting CT projections[53], but it does not reconstruct a

volume from projections with known geometry. The pretraining loss is $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{feat}}$ with $\lambda = 1$. The CT-feature head is discarded after pretraining; the backbone and the point decoder are then adapted to each dataset and number of projections.

Model training

Both pretraining stages used the CT-RATE training split (47,149 volumes) with AdamW[54] and a warm-up followed by cosine decay of the learning rate, on 16 NVIDIA GH200 GPUs in four nodes. The learning rate was not scaled with the number of GPUs, which we found necessary to prevent the collapse of the latent representation, a common problem in three-dimensional CT pretraining. Both stages resampled each training volume from the CT to 256^3 voxels at an isotropic spacing drawn at random between 1.0 and 2.0 mm, so that the CT encoder was trained on volumes similar to those it received in Stage 2. Parts of a training volume that extended beyond the scan were filled with air. Stage 2 centred each training volume on the bounding box of the body (voxels above -500 HU). Stage 1 centred each training volume on a random point inside this bounding box. Stage 2 rendered the projections during training at 512×512 detector pixels from CT volumes of 256^3 voxels. Because the number of projections changed between samples, each GPU held one volume, for a global batch of 16 volumes. The point decoder was trained on 16,384 query points per step, sampled uniformly at random within the CT volume (Supplementary Fig. 1f). To train one backbone for many acquisition settings, the geometry was sampled at random during training: the number of projections (2 to 16), their angular pattern and the distances of the source and detector from the axis, with the detector sized to the patient (Supplementary Fig. 1b). The five angular patterns were uniform (projections evenly spaced), random (projections randomly spaced), limited arc, bi-planar and clustered. All pretraining settings are listed in Supplementary Tab. 32 and 33.

Adaptation trained the backbone and the point decoder on one dataset and one number of projections, with AdamW, five warm-up epochs, a cosine decay of the learning rate from 10^{-4} to 10^{-6}, module-wise learning-rate decay of 0.75 and gradient clipping at 1.0, on one GPU. All reconstruction models in this study, for SPARC and the five competing methods on every dataset and number of projections, were trained under the same budget and stopping rule. Training ran for at most 400 epochs, with validation every 20 epochs and early stopping after at least 120 epochs once five consecutive validations had shown no gain of 0.02 dB in PSNR. The checkpoint with the best validation PSNR was evaluated once on the test set. Each method used its own optimiser settings (SPARC: three volumes per batch); only the budget and the stopping rule were shared. SPARC's sixteen-projection models did not fit a batch of three on one GPU; they were trained on three GPUs with one volume each, and synchronised batch normalisation made this equivalent to a batch of three.

Datasets

CT-RATE[30] contains chest CT volumes with radiology reports, from which 18 abnormalities are labelled per volume. Its training split (47,149 volumes) was used only for pretraining. Its separate validation cohort (3,039 volumes from 1,304 patients), which shares no patient with the training split, was used for reconstruction and abnormality classification. We split this cohort by patient into training (913 patients, 2,134 volumes), validation (130 patients, 303 volumes) and test sets (261 patients, 602 volumes).

TotalSegmentator[32] (version 2.0.1) contains CT volumes of many body regions with reference masks of anatomical structures. Of its 1,228 volumes, we used the 1,137 that could be read and contained at least one of the ten structures segmented in this study, for reconstruction and segmentation. These structures are the five lung lobes and five abdominal organs (the liver, spleen, left kidney, right kidney and aorta).

Six further datasets were used for reconstruction only. From the Medical Segmentation Decathlon (MSD)[33], we combined the 1,407 CT volumes of its six CT tasks (liver, lung, pancreas, hepatic vessel,

spleen and colon) into one dataset. AbdomenCT-1K[34] provided 1,061 abdominal CT volumes. AMOS[35] provided 500 abdominal CT volumes; its MRI volumes were not used. CQ500[36] provided 472 head CT studies that could be read; from each study, we used the series with the most slices. ToothFairy3[37–39] provided 529 maxillofacial cone-beam CT volumes, and VerSe[40] provided 368 spine CT volumes. Tab. 1 gives the number of volumes in each split after the removal of duplicated test volumes described below, and the Data availability section lists the sources of all datasets.

Dataset preprocessing and splits

Each CT volume was resampled to isotropic spacing matched to the size of its anatomy: 1.6 mm for the chest, abdomen, spine and multi-region scans, 1.2 mm for the head and 0.625 mm for the maxillofacial scans. It was then cropped or padded around its centre to 256^3 voxels. Every evaluation dataset was divided into training, validation and test sets at the patient level where patients are identified, and otherwise by volume. The validation set was used to select checkpoints, to stop training early and to set the classifier's thresholds, but never for reporting. We checked every test set against its training and validation sets for repeated scans. Scans were compared by the SHA-1 hash of each voxel array, by the correlation of downsampled 32^3 volumes and, where the dataset identifies patients, by patient identifier. A pair was counted as the same scan when the correlation exceeded 0.95; different CT-RATE patients reached at most 0.88. Thirteen test volumes duplicated a training volume or came from a patient in the training set: three in MSD, whose CT tasks share scans, one in AbdomenCT-1K, three in ToothFairy3 and six in VerSe, where one patient's spine can be split into several volumes. We removed them from the test sets without retraining any model, and every number in this study uses the reduced test sets. A few validation volumes repeat a training volume; this affects only checkpoint selection.

Projection simulation and acquisition geometry

Projections were rendered from the CT volumes with nanodrr, a reimplementation of the differentiable digitally reconstructed radiograph engine DiffDRR[55], which integrates the linear attenuation coefficient (water $0.0192\,\mathrm{mm}^{-1}$) along each ray. These line integrals equal the negative logarithm of the transmitted fraction of X-ray intensity and were used directly as the projections. Unless stated otherwise, projections were spaced uniformly around the patient, with a source-to-detector distance 1.5 times the source-to-axis distance and a 512×512 detector. The output was a 256^3 volume. The source-to-axis distance was matched to the anatomy: 800 mm for the torso (chest, abdomen, spine and multi-region scans), 600 mm for the head and 312 mm for the maxillofacial region. Detector pixels were 2.4 mm for the torso. For the head and the maxillofacial region, detector pixels were sized to the patient (Supplementary Tab. 33). Four and eight projections were the main settings, each trained separately. SPARC and DeepSparse were also trained with two and with sixteen projections on CT-RATE and TotalSegmentator (Supplementary Tab. 26). SPARC's two- and sixteen-projection models on AMOS were used for the analysis with a fixed total photon budget. With two projections, the projections were approximately orthogonal (bi-planar).

Competing methods

We compared SPARC with DIF-Net[10], DIF-Gaussian[12], C^{2}RV[11], SVCT[31] and DeepSparse[26], five methods that reconstruct a three-dimensional volume from sparse projections in one forward pass and generalise across patients. Each was trained with its published code on the same projections, geometry, data splits and training budget as SPARC; DeepSparse was finetuned from its published pretrained weights, and the other methods were trained from random initialisation. Methods that optimise each scan separately (Introduction), and methods restricted to two orthogonal radiographs or to two-dimensional slices,

work under different assumptions and were not compared. Two large reconstruction models for sparse projections, X-LRM[27] and X-GRM[28], were not compared because their code was not publicly available.

Evaluation of reconstruction

Reconstructions and ground truth were compared on the 256^3 grid. PSNR and SSIM were computed after clipping both to $[-1{,}024, 1{,}024]$ HU and scaling this range to $[0, 1]$. SSIM values and SSIM differences are reported as percentages. The Hounsfield-unit and spatial error analyses used reconstructions and ground truth clipped to the same range. The Hounsfield-unit errors of all methods were compared on the CT-RATE and TotalSegmentator test sets. The mean absolute and signed errors were computed over the whole volume and within intensity bands defined on the ground truth (air below -900 HU, lung -900 to -500 HU, soft tissue -100 to below 300 HU and bone at least 300 HU). For TotalSegmentator, they were also computed inside the reference masks of the ten structures. The outermost voxel layer, partly zero padding in the ground truth, was excluded. For SPARC's spatial error analysis, a two-voxel rim was excluded. For each voxel, we measured its distance r from the central longitudinal axis of the volume. For each volume, we defined the body radius R as the 99th percentile of r over the ground-truth voxels above -300 HU, that is, over soft tissue and bone. The error was averaged in three zones: a central zone ($r < R/2$), a peripheral zone ($R/2 \leq r < R$) and the zone outside the body radius ($r \geq R$), which is mostly air around the body. Because the body is not circular in cross-section, these zones are rings around the axis and do not follow the skin. For the radial profiles in Supplementary Fig. 7a, the error was averaged in 20 bins of r, each $R/10$ wide. Voxels beyond $2R$ were counted in the last bin. Within each intensity band, we also computed the normalised root mean squared error (NRMSE) and the voxel-wise Pearson correlation between the reconstruction and the ground truth (Supplementary Fig. 7c,d). The NRMSE was the root mean squared error divided by the standard deviation of the ground-truth HU values in that band. The body width of each volume was the distance between the 1st and 99th percentiles of the left-to-right positions of its ground-truth voxels above -300 HU (Supplementary Fig. 7e).

Clinical tasks

In the primary test, each method's test reconstructions were given without modification to two CT-based analysis tools, task networks trained on full CT. The classifier was a three-dimensional convolutional network that classified the 18 CT-RATE abnormalities, and the segmentation model was a three-dimensional U-Net[56] that segmented the ten TotalSegmentator structures (Supplementary Fig. 2b). Each task network was trained from random initialisation on the training set, three times with different seeds. The three resulting networks of each task were applied to every method's reconstructions and, as the reference, to the full CT volumes of the test set. The predicted probabilities (classification) or per-volume Dice (segmentation) were averaged over the three seeds. The classifier was trained for at most 60,000 iterations and validated every 2,000 iterations. Training stopped early once eight consecutive validations had shown no gain of 0.002 in macro-AUROC, but not before 20,000 iterations. The segmentation model was trained for 6,000 iterations (Supplementary Tab. 34).

With eight projections, this primary approach was compared with two other ways of using SPARC for the same tasks (Supplementary Fig. 2a and 13). In the first, SPARC*, task networks with the same architectures and settings were trained from random initialisation on SPARC's reconstructions. In the second, SPARC†, task heads were attached directly to SPARC's pretrained backbone, which then served as an encoder of the projections (Supplementary Fig. 2c,d), so that no volume was reconstructed. For classification, the mean of $\mathbf{F}$ over the 16^3 grid points was passed through a LayerNorm followed by a multilayer perceptron with one hidden layer of 512 units, which gave one score for each of the 18 findings. For segmentation, a multilayer perceptron with two hidden layers of 256 units predicted one

score for each of the ten structures and for the background at each query point. The segmentation head took the same input features as the point decoder but had no residual connections. We call the pretrained backbone, when it is used with a task head, SPARC's encoder. The encoder was either frozen or finetuned. Finetuning first trained the head with the encoder frozen and then trained the head together with the encoder. In this second phase, the encoder had a smaller learning rate than the head (Supplementary Tab. 34). Like the task networks, each head was trained three times with different seeds, and the predicted probabilities (classification) or per-volume Dice (segmentation) were averaged.

Classification was evaluated by the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC), each averaged over the 18 findings (macro-average). Sensitivity and specificity were measured at fixed operating points. For each finding, the threshold maximising Youden's index for the classifier on the full-CT validation set (303 volumes from 130 patients) was applied unchanged to every input. Segmentation was evaluated by the Dice coefficient, the normalised surface Dice at tolerances of 2 mm and 3 mm, the 95th-percentile Hausdorff distance, the average symmetric surface distance and the average surface distance from the predicted to the reference boundary. Each of these measures was computed per seed on the 1.6-mm grid before averaging over the three seeds. Structure volume was evaluated by the relative volume error, defined as the predicted minus the reference volume divided by the reference volume. Both the signed and the absolute relative volume error were reported. The volume error was computed only for structures lying wholly inside the scan: at least three voxels from the end of the scan's coverage on every axis and at least 1 mL in reference volume. For a structure cut off by the coverage, the relative volume error is not meaningful. A structure that the segmentation model did not predict at all was scored 0 in both Dice measures, assigned a volume error of -100% and excluded from the distances. Supplementary Tab. 24 and 25 give the number of such structures for each input, summed over the three seeds.

To examine the image regions on which the classifier based each prediction, we computed Grad-CAM[41] maps of the seed-0 classifier on the CT and on each reconstruction. We did so for every positive test volume of the ten findings with an anatomical location (1,311 volume–finding pairs from 203 patients). For each pair, we computed the voxel-wise Pearson correlation inside the body between the map on each reconstruction and the map on the CT. The body was defined as the CT voxels above -500 HU, with enclosed regions below this threshold, such as the lungs, filled in each axial slice. A six-voxel rim at the edge of the volume was excluded from the body. For the classifier trained on full CT (on the CT), SPARC* and SPARC† with the finetuned encoder, we measured the share of the map inside the body that lies in the involved organ (Supplementary Fig. 13c; Supplementary Tab. 21). The organs were segmented from the CT with the public TotalSegmentator model[32] and dilated by three voxels: the heart for cardiomegaly and coronary calcification, the aorta for arterial wall calcification, and the lungs for consolidation, lung opacity, atelectasis, emphysema and lung nodules. For pericardial and pleural effusion, the heart and the lungs were instead dilated by 16 mm.

Analyses of pretraining

To measure the amount of paired data required for adaptation, the training set of the CT-RATE validation cohort (913 patients) was subsampled by patient into nested subsets of 10, 25, 50, 100, 225 and 450 patients (28 to 1,079 volumes). The subsets were drawn once with a fixed seed, so that every method used the same patients at every size. Validation and test sets were unchanged, and the full set of 913 patients completes the series (Supplementary Tab. 27). SPARC, DeepSparse and SVCT were trained on each subset, as was the SPARC architecture from random initialisation (From scratch). To separate the contributions of the two pretraining losses, the same architecture was adapted to CT-RATE and TotalSegmentator from three starting points: random initialisation, Stage-2 pretraining with the reconstruction loss only, and SPARC's pretraining with both losses. The adaptation protocol was unchanged.

The resulting reconstructions were given to the CT-based analysis tools (Supplementary Tab. 28).

Robustness to acquisition errors

Every trained model was evaluated again, without retraining, on the test sets of CT-RATE, TotalSegmentator, CQ500 and VerSe with four and eight projections, under one perturbation at a time (Fig. 4; Supplementary Fig. 9 and 10). Photon noise was simulated from the line integrals p. The expected count $I_0 e^{-p}$, with I_0 set to 10^6, 10^5, 3×10^4 or 10^4 incident photons per detector pixel, received Poisson noise and Gaussian electronic noise with a standard deviation of 10 counts. The noisy count N was clipped to a minimum of one count and converted back to a line integral, $-\log(N/I_0)$. Geometric errors were simulated by rendering with a perturbed geometry while every method was given the unperturbed geometry. The perturbations were angular jitter, a source-to-detector distance error and a detector offset. Angular jitter was an angular error drawn independently for each projection (standard deviation 0.5, 1 or 2°). The source-to-detector distance error was 1, 2 or 5%, with its sign drawn for each volume. The detector offset was drawn for each volume (standard deviation 1, 2 or 4 mm in each direction). A limited angular range, with the projections spread over 180, 120 or 90°, is a change of protocol rather than an error, so the models were given the true angles. Each random draw depended only on the volume and the condition, so that all methods received the same perturbed projections and each volume received its own draw. Changes are paired differences from the unperturbed evaluation of the same model on the same volumes. For SPARC alone, the same test sets were also evaluated at eight projections under additional perturbations (Supplementary Fig. 11a–h; Supplementary Tab. 15 and 16). The additional noise levels were 3×10^5, 3×10^3 and 10^3 incident photons per detector pixel. The additional geometric errors were an angular error (standard deviation 0.25 or 3°), a source-to-detector distance error of 0.5, 3 or 10%, and a detector offset (standard deviation 8 mm). Source-to-detector distance errors with the same sign for every volume were evaluated at -5, -2, $+2$ and $+5$%; a positive error means that the true distance was longer than the distance given to the model. The additional angular ranges were 270 and 60°, again with the true angles given to the model. Combined perturbations paired 10^5 or 10^4 incident photons per detector pixel with an angular error (standard deviation 0.5, 1 or 2°), or 10^4 incident photons per detector pixel with a source-to-detector distance error of 2%. In the fixed-exposure analysis on AMOS, a total exposure of 8×10^4 or 8×10^5 incident photons per detector pixel was divided equally among 2, 4, 8 or 16 projections (Supplementary Fig. 11i; Supplementary Tab. 17). Each number of projections used its own trained model. The total exposures were simulation parameters, not measurements of patient dose.

Computational cost

Reconstruction time was measured on one NVIDIA GH200 GPU with batch size 1 on every CT-RATE test volume with eight projections, after three warm-up volumes in each timing process. The reconstruction time includes input preparation, which turns the projections and acquisition geometry into the network's inputs, and the network itself, that is, the backbone and the point decoder, with the GPU synchronised around each step. Loading the CT and rendering the simulated projections were excluded, because neither step is required when real projections are used. File input and output were also excluded. Peak allocated memory was recorded while the network ran (Supplementary Fig. 8b,c). Floating-point operations were counted with PyTorch's FlopCounterMode; operators without a registered formula, such as grid sampling, are not counted, so these values are lower bounds. Because SPARC's point decoder is resolution-free, its cost was also measured for output grids from 128^3 to 512^3 (Supplementary Fig. 8a,b). To show how the time is divided, input preparation, the backbone and the point decoder were also timed one by one on the same test volumes, for output grids from 96^3 to 512^3 (Supplementary Tab. 9). Because of the separate timing, the summed time of the three steps is 5 to 9% longer than the reconstruction time.

The time against the number of projections was measured with the models trained on AMOS with two, four, eight and sixteen projections, on its test set at a 256^3 output (Supplementary Fig. 8d). Training cost is the wall time in each run's log, summed over resumed segments, times the number of GPUs.

Confidence intervals and statistical tests

All metrics were computed once on the full test set of each dataset. The 95% confidence intervals come from a bootstrap over patients. The patients of a test set were resampled with replacement 10,000 times (random seed 0), and each drawn patient contributed all of their test volumes. The limits of the CI are the 2.5th and 97.5th percentiles of the statistic over the resamples. Resampling patients rather than volumes changes the CIs only for CT-RATE, whose 602 test volumes come from 261 patients; every other test set has one volume per patient. PSNR, SSIM and Hounsfield-unit errors were averaged over the resampled volumes. The macro-AUROC is the mean over the 18 findings of the AUROC computed on the resampled volumes, each weighted by how often its patient was drawn; a finding without both positive and negative volumes in a resample was skipped. Dice was averaged over volumes for each structure, skipping volumes in which the structure is absent, and then over the structures of each group, the five abdominal organs and the five lung lobes. Because every method was evaluated on the same test volumes, the same resampled patients were used for both members of a comparison, giving paired CIs for differences. The strongest competing method in a comparison (the comparator) is the one with the highest mean for that dataset, number of projections and metric. For Hounsfield-unit errors, where lower values are better, the comparator is the competing method with the lowest mean. The comparator is chosen separately for each intensity band in the Hounsfield-unit analysis and for each perturbation setting in the robustness analysis. The share of improved volumes is the fraction of test volumes in which SPARC's score exceeds the comparator's. For paired differences, two-sided *P* values were derived from the bootstrap CIs[57]. The standard error was taken as the width of the CI divided by 2×1.96. The *P* value was the two-sided normal tail probability of the difference divided by this standard error. Signed volume errors were tested against zero in the same way. *P* values were adjusted for multiple comparisons with Holm's method[58] within each family of comparisons, defined as the paired differences reported together in one table of the Supplementary Information; a difference reported only in the text formed a family of its own. All reported *P* values are Holm-adjusted, and a difference was considered significant at an adjusted *P* below 0.05. The text gives *P* values as levels (below 0.001, 0.01 or 0.05, or above 0.05); the tables of the Supplementary Information give their values. All statistics in the text, figures and Supplementary Information are generated by analysis scripts from the per-volume result files.

Ethics

This study used only publicly available, de-identified datasets, accessed under their respective licences and data-use agreements; no new data were collected and no participants were contacted.

Data availability

All data used in this study are publicly available; no new patient data were collected. Each dataset was obtained from the source given below under its data-use terms; where we used a mirror or a redistributed copy rather than the original release, both are given. Pretraining and the chest experiments use CT-RATE[30] (`https://huggingface.co/datasets/ibrahimhamamci/CT-RATE`): its official training split (47,149 volumes) for pretraining and its official validation cohort (3,039 volumes) for chest reconstruction and classification. For TotalSegmentator[32], we used version 2.0.1 of the dataset (1,228 CT volumes; `https://doi.org/10.5281/zenodo.10047292`). The Medical Segmentation Decathlon[33] contributes the images of its six CT tasks (liver, lung, pancreas, hepatic vessel, spleen and colon; `http://medicaldecathlon.com`). AbdomenCT-1K[34] contributes its three image parts, obtained through the download form linked at `https://github.com/JunMa11/AbdomenCT-1K`. AMOS[35] contributes the 500 CT scans of the AMOS 2022 release (`https://doi.org/10.5281/zenodo.7262581`). CQ500[36], released by Qure.ai, was obtained from its Kaggle mirror (`https://www.kaggle.com/datasets/crawford/qureai-headct`). ToothFairy3[37–39] was obtained from `https://toothfairy3.grand-challenge.org/dataset/`, which requires registration. VerSe[40], released at `https://osf.io/nqjyw/` and `https://osf.io/t98fz/`, was obtained from the copy redistributed in the CADS collection[59] (`https://huggingface.co/datasets/huggingface/CADS-dataset`, subset `0010_verse`). The files used from each source and the conversion steps are listed in `preprocessing/README.md` of the code repository. The reconstruction training, validation and test splits, with the 13 duplicated test volumes removed (Methods), and the list of the removed volumes are in the `splits` folder of the code repository.

Code availability

The code for SPARC's pretraining (Stages 1 and 2), its adaptation to each dataset (Stage 3) and the data preparation, together with the configuration files for those runs, is available at `https://github.com/HORIZONHealthcare/SPARC` under the Creative Commons Attribution-NonCommercial 4.0 International licence. The competing methods were trained from their authors' public implementations, which we do not redistribute. The pretrained Stage-1 and Stage-2 weights are available at `https://huggingface.co/lyqun/SPARC`, and the sixteen per-dataset reconstruction models behind the reported results at `https://huggingface.co/lyqun/SPARC-reconstruction`, under the same licence; access requires a short request form and is granted automatically.

Acknowledgements

The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation, and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. Y.Z. is supported by a Wellcome Trust Early-Career Award (318987/Z/24/Z) and a Moorfields Eye Charity Equipment Grant (EQR-25B-102). No other specific funding was received for this work.

Author contributions

Y.L. conceived the study, designed SPARC, implemented the pretraining, reconstruction and transfer pipelines, ran the experiments, analysed the results and wrote the manuscript. J.X. contributed to the implementation and the experimental infrastructure, trained the competing methods and carried out the robustness and efficiency evaluations. L.J., J.G., W.W., H.W. and H.Y. contributed to data curation and preprocessing, to the evaluation protocols and to the interpretation of the results. H.S., B.P. and K.L. provided guidance on the study design and the interpretation of the results, and revised the manuscript. D.C.A. provided methodological guidance and revised the manuscript. H.C. contributed to the design of the clinical evaluation and revised the manuscript. Y.Z. supervised the study, obtained the resources, guided the clinical framing and revised the manuscript. All authors approved the final manuscript.

Competing interests

Y.L., J.G. and H.Y. are part-time employees of Allvista Technology (Hong Kong) Limited, and H.S. is an employee of Koln 3D Technology (Medical) Limited; neither company provided funding, salaries or computing resources for this study. The other authors declare no competing interests.

Use of large language models

Claude (Anthropic) was used to assist with writing data analysis and figure plotting code and with editing the manuscript text. The authors drafted the manuscript, including its outline. They reviewed, ran and verified all code, checked every result against the experiment outputs, and edited all text. They take full responsibility for the content of the manuscript.

References

[1] Zha, R., Zhang, Y. & Li, H. NAF: Neural attenuation fields for sparse-view CBCT reconstruction. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 13436 of *Lecture Notes in Computer Science*, 442–452 (2022). 2209.14540.

[2] Shen, L., Pauly, J. & Xing, L. NeRP: Implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. *IEEE Transactions on Neural Networks and Learning Systems* **35**, 770–782 (2024). Early access 2022; arXiv:2108.10991, 2108.10991.

[3] Wu, Q., Feng, R., Wei, H., Yu, J. & Zhang, Y. Self-supervised coordinate projection network for sparse-view computed tomography. *IEEE Transactions on Computational Imaging* **9**, 517–529 (2023). ArXiv:2209.05483.

[4] Rückert, D., Wang, Y., Li, R., Idoughi, R. & Heidrich, W. NeAT: Neural adaptive tomography. *ACM Transactions on Graphics (SIGGRAPH)* **41** (2022). ArXiv:2202.02171.

[5] Cai, Y., Wang, J., Yuille, A., Zhou, Z. & Wang, A. Structure-aware sparse-view x-ray 3D reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). ArXiv:2311.10959.

[6] Tian, X. *et al.* Unsupervised self-prior embedding neural representation for iterative sparse-view CT reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2025). ArXiv:2502.05445.

[7] Zha, R. *et al.* R^2-Gaussian: Rectifying radiative gaussian splatting for tomographic reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)* (2024). 2405.20693.

[8] Cai, Y. *et al.* Radiative gaussian splatting for efficient x-ray novel view synthesis. In *European Conference on Computer Vision (ECCV)* (2024). ArXiv:2403.04116.

[9] Li, Y. *et al.* 3DGR-CT: Sparse-view CT reconstruction with a 3D gaussian representation. *Medical Image Analysis* **103** (2025). ArXiv:2312.15676.

[10] Lin, Y., Luo, Z., Zhao, W. & Li, X. Learning deep intensity field for extremely sparse-view CBCT reconstruction. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 14229 of *Lecture Notes in Computer Science*, 13–23 (2023). 2303.06681.

[11] Lin, Y. *et al.* C^2RV: Cross-regional and cross-view learning for sparse-view CBCT reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11205–11214 (2024). 2406.03902.

[12] Lin, Y., Wang, H., Chen, J. & Li, X. Learning 3d gaussians for extremely sparse-view cone-beam CT reconstruction. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 15007 of *Lecture Notes in Computer Science* (2024). Also known as DIF-Gaussian, 2407.01090.

[13] Kyung, D., Jo, K., Choo, J., Lee, J. & Choi, E. Perspective projection-based 3D CT reconstruction from biplanar x-rays. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023). ArXiv:2303.05297.

[14] Zhang, C. *et al.* XTransCT: Ultra-fast volumetric CT reconstruction using two orthogonal x-ray projections via a transformer network. *Physics in Medicine & Biology* (2024). ArXiv:2305.19621.

[15] Ying, X. *et al.* X2CT-GAN: Reconstructing CT from biplanar X-Rays with generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10619–10628 (2019). 1905.06902.

[16] Song, Y., Shen, L., Xing, L. & Ermon, S. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations (ICLR)* (2022). ArXiv:2111.08005.

[17] Chung, H., Ryu, D., McCann, M. T., Klasky, M. L. & Ye, J. C. Solving 3D inverse problems using pre-trained 2D diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). ArXiv:2211.10655.

[18] Xu, K., Lu, S., Huang, B., Wu, W. & Liu, Q. Stage-by-stage wavelet optimization refinement diffusion model for sparse-view CT reconstruction. *IEEE Transactions on Medical Imaging* (2024). ArXiv:2308.15942.

[19] Liu, J. *et al.* DOLCE: A model-based probabilistic diffusion framework for limited-angle CT reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 10498–10508 (2023). ArXiv:2211.12340.

[20] Jeong, Y. S., Yoo, H. B. & Chun, I. Y. DX2CT: Diffusion model for 3D CT reconstruction from bi or mono-planar 2D x-ray(s). *arXiv preprint arXiv:2409.08850* (2024).

[21] Liu, X. *et al.* DiffuX2CT: Diffusion learning to reconstruct CT images from biplanar x-rays. In *European Conference on Computer Vision (ECCV)* (2024). ArXiv:2407.13545.

[22] Sun, Y. *et al.* DIFR3CT: Latent diffusion for probabilistic 3D CT reconstruction from few planar x-rays. *arXiv preprint arXiv:2408.15118* (2024).

[23] Jin, K. H., McCann, M. T., Froustey, E. & Unser, M. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing* **26**, 4509–4522 (2017). ArXiv:1611.03679.

[24] Ma, C., Li, Z., Zhang, J., Zhang, Y. & Shan, H. FreeSeed: Frequency-band-aware and self-guided network for sparse-view CT reconstruction. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2023). ArXiv:2307.05890.

[25] Li, Z., Ma, C., Chen, J., Zhang, J. & Shan, H. Learning to distill global representation for sparse-view CT. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 21196–21207 (2023). ArXiv:2308.08463.

[26] Lin, Y. *et al.* DeepSparse: A foundation model for sparse-view CBCT reconstruction. *IEEE Transactions on Medical Imaging* (2026). IEEE Xplore document 11436116; arXiv:2505.02628, 2505.02628.

[27] Zhang, G. *et al.* X-LRM: X-ray large reconstruction model for extremely sparse-view computed tomography recovery in one second. In *International Conference on 3D Vision (3DV)* (2026). ArXiv:2503.06382, 2503.06382.

[28] Liu, Y. *et al.* X-GRM: Large gaussian reconstruction model for sparse-view x-rays to computed tomography. *arXiv preprint arXiv:2505.15235* (2025). 2505.15235.

[29] Lin, T. *et al.* Are pixel-wise metrics reliable for computerized tomography reconstruction? In *Advances in Neural Information Processing Systems (NeurIPS)* (2025). ArXiv:2506.02093.

[30] Hamamci, I. E., Er, S., Wang, C., Menze, B. *et al.* Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering* (2026). Introduces CT-CLIP and the CT-RATE dataset. Preprint arXiv:2403.17834.

[31] Liu, Z. *et al.* Geometry-aware attenuation learning for sparse-view CBCT reconstruction. *IEEE Transactions on Medical Imaging* (2024). 2303.14739.

[32] Wasserthal, J., Breit, H.-C., Meyer, M. T. *et al.* TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence* **5**, e230024 (2023).

[33] Antonelli, M., Reinke, A., Bakas, S. *et al.* The medical segmentation decathlon. *Nature Communications* **13**, 4128 (2022).

[34] Ma, J. *et al.* Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 6695–6714 (2021).

[35] Ji, Y. *et al.* AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 36722–36732 (2022).

[36] Chilamkurthy, S., Ghosh, R., Tanamala, S. *et al.* Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet* **392**, 2388–2396 (2018).

[37] Bolelli, F. *et al.* Multi-structure segmentation in CBCT volumes: the ToothFairy2 challenge. *Medical Image Analysis* **112**, 104095 (2026).

[38] Bolelli, F. *et al.* Segmenting maxillofacial structures in CBCT volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5238–5248 (2025).

[39] Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E. & Grana, C. Enhancing patch-based learning for the segmentation of the mandibular canal. *IEEE Access* **12**, 79014–79024 (2024).

[40] Sekuboyina, A., Husseini, M. E., Bayat, A. *et al.* VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Analysis* **73**, 102166 (2021). ArXiv:2001.09193.

[41] Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626 (2017).

[42] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)* (2021).

[43] He, K. *et al.* Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988 (2022). 2111.06377.

[44] Liu, Z. *et al.* A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11976–11986 (2022).

[45] Yu, A., Ye, V., Tancik, M. & Kanazawa, A. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4578–4587 (2021). 2012.02190.

[46] Assran, M. *et al.* Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15619–15629 (2023). 2301.08243.

[47] Bardes, A. *et al.* Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research (TMLR)* (2024). Introduces V-JEPA. arXiv:2404.08471, 2404.08471.

[48] Assran, M. *et al.* V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985* (2025). 2506.09985.

[49] Wu, L., Zhuang, J. & Chen, H. VoCo: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). ArXiv:2402.17300.

[50] Pai, S. *et al.* Vision foundation models for computed tomography. *arXiv preprint arXiv:2501.09001* (2025). The proposed model is named CT-FM, 2501.09001.

[51] Pérez-García, F. *et al.* Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence* **7**, 119–130 (2025). Model named RAD-DINO. Preprint arXiv:2401.10815.

[52] Gwak, S., Yang, S., Jeong, H., Park, J. & Kang, M. Efficient sparse-view medical image classification for low radiation and rapid COVID-19 diagnosis. *Biomedical Engineering Letters* **15**, 785–795 (2025). Proposes the Projection-wise Masked Autoencoder (ProMAE).

[53] Yang, Z., Wang, G., Hendler, J., Kalra, M. K. & Yan, P. X-WIN: Building chest radiograph world model via predictive sensing. *arXiv preprint arXiv:2511.14918* (2025). 2511.14918.

[54] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)* (2019).

[55] Gopalakrishnan, V. & Golland, P. Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In *Clinical Image-Based Procedures (CLIP), held in conjunction with MICCAI*, vol. 13746 of *Lecture Notes in Computer Science*, 1–11 (Springer, 2022). ArXiv:2208.12737.

[56] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9901 of *Lecture Notes in Computer Science*, 424–432 (Springer, 2016).

[57] Altman, D. G. & Bland, J. M. How to obtain the P value from a confidence interval. *BMJ* **343**, d2304 (2011).

[58] Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979).

[59] Xu, M., Amiranashvili, T., Navarro, F. *et al.* CADS: A comprehensive anatomical dataset and segmentation for whole-body anatomy in computed tomography. *arXiv preprint arXiv:2507.22953* (2025). 2507.22953.

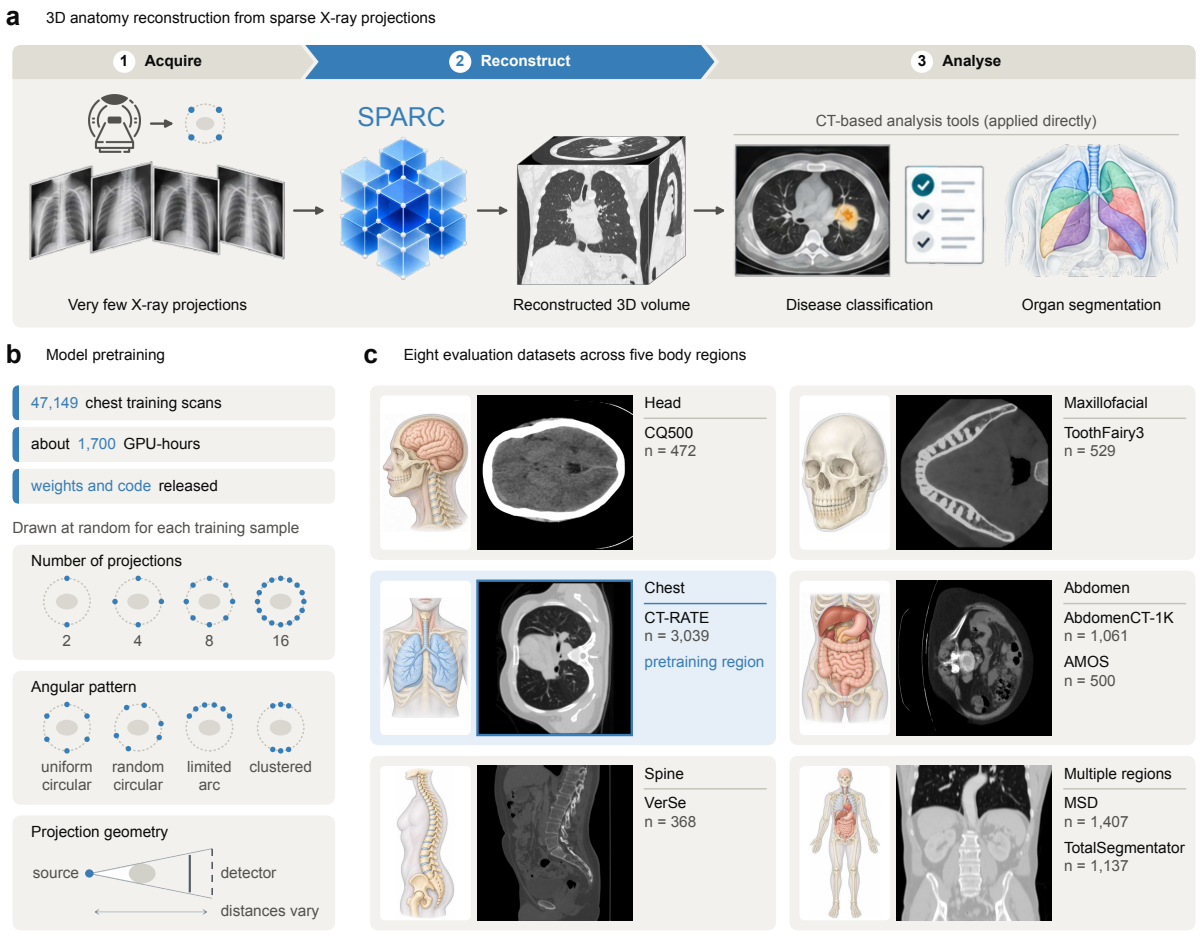


Fig. 1 | Study overview. **a.** Schematic of the approach: an existing X-ray system acquires very few X-ray projections, SPARC reconstructs a three-dimensional volume from them, and CT-based analysis tools trained on full CT classify diseases and segment organs directly on the reconstruction. **b.** Pretraining data, cost and the acquisition settings drawn at random for each training sample. **c.** Evaluation datasets by body region, each shown by a CT slice from its test set, with the number of CT volumes used from each dataset.

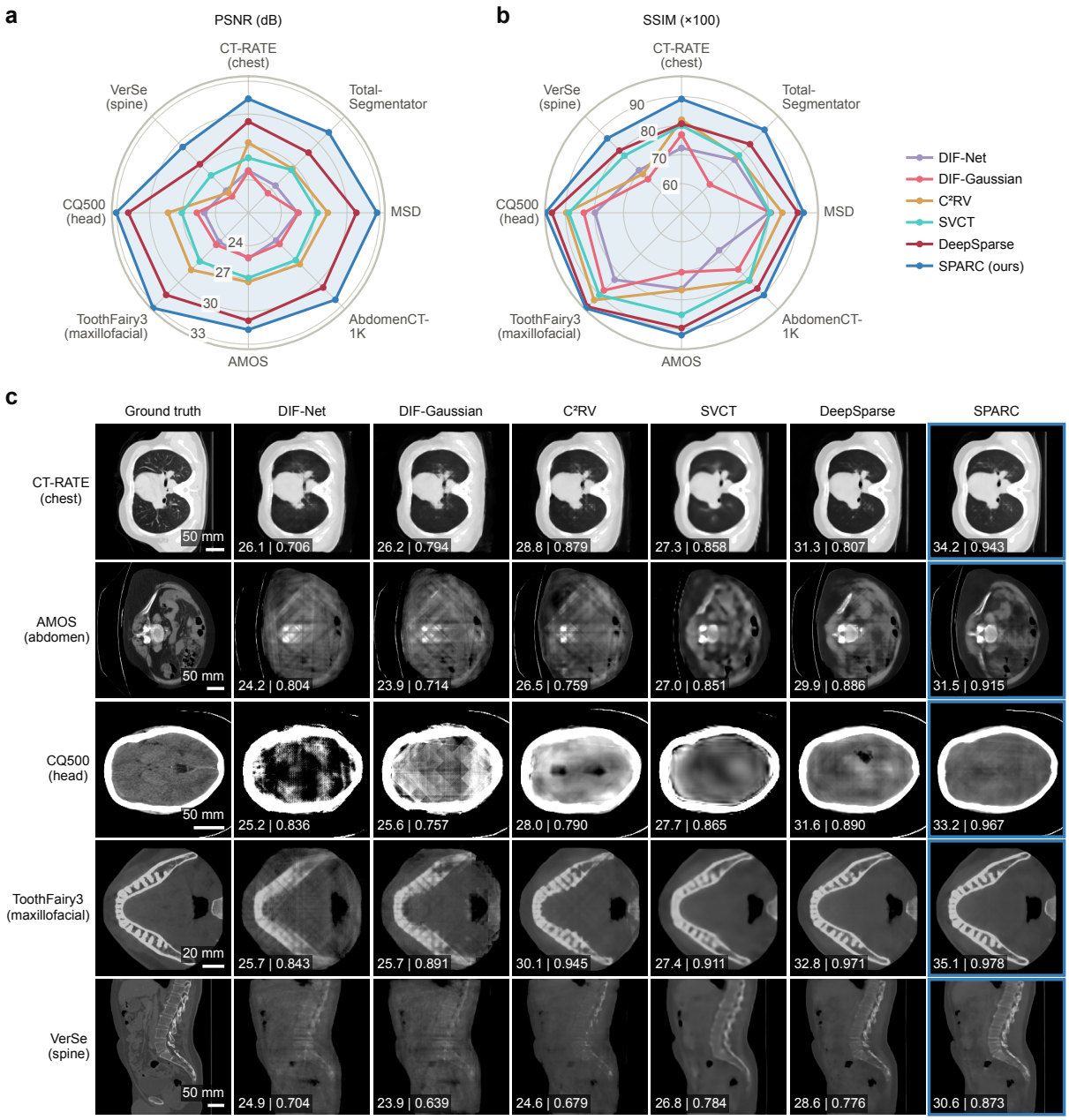


Fig. 2 | Reconstruction fidelity with eight projections. a, b. Mean PSNR (**a**) and SSIM (×100, **b**) of each method on the eight test sets. Radial axes start at 21 dB and 50. **c.** Reconstructions of one test volume from each of five datasets by every method, on the same cross-section, crop and display window within each row; labels give the volume's PSNR (dB) | SSIM, and SPARC is outlined in blue. Full results are reported in Supplementary Tab. 1–3. Supplementary Fig. 3 shows reconstructions of the other three datasets with eight projections. Supplementary Fig. 4 shows the mean PSNR and SSIM with four projections, and Supplementary Fig. 5 and 6 show reconstructions of all eight datasets with four projections.

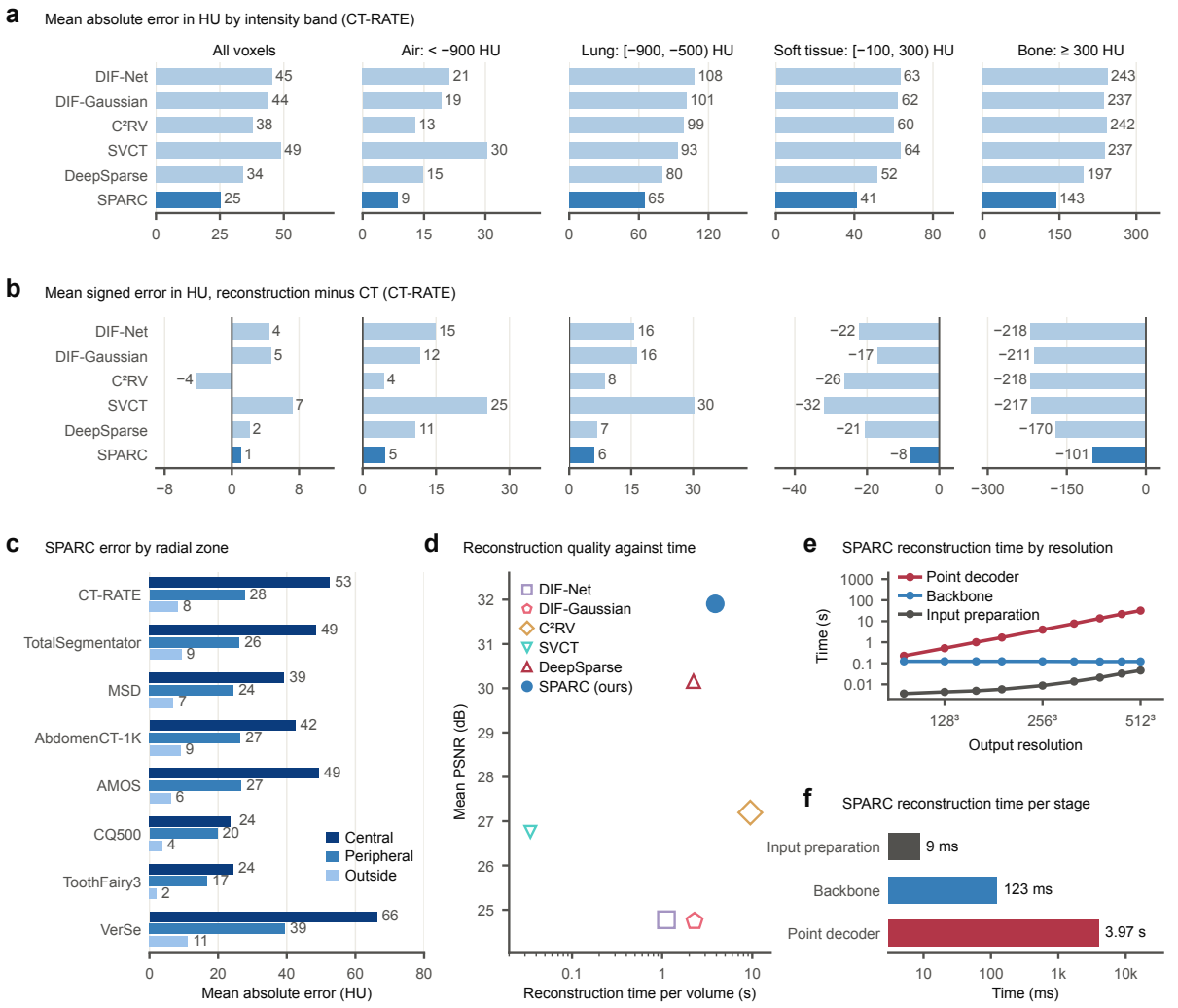


Fig. 3 | Hounsfield-unit accuracy, spatial error and computational cost. Eight projections. **a, b.** Mean absolute error (**a**) and mean signed error, reconstruction minus CT (**b**), in HU within intensity bands, for each method on the CT-RATE test set. **c.** SPARC's mean absolute error in three zones around the central axis of each volume, defined by its body radius R (Methods): central ($r < R/2$), peripheral ($R/2 \leq r < R$) and outside ($r \geq R$), on the eight test sets. **d.** Mean PSNR over the eight test sets against reconstruction time per volume; larger markers denote higher peak GPU memory. **e, f.** SPARC's time per volume for input preparation, the backbone and the point decoder, against the output resolution (**e**) and at a 256^3 output (**f**). Point estimates are shown. Full results are reported in Supplementary Tab. 4–9 and Supplementary Fig. 7 and 8.

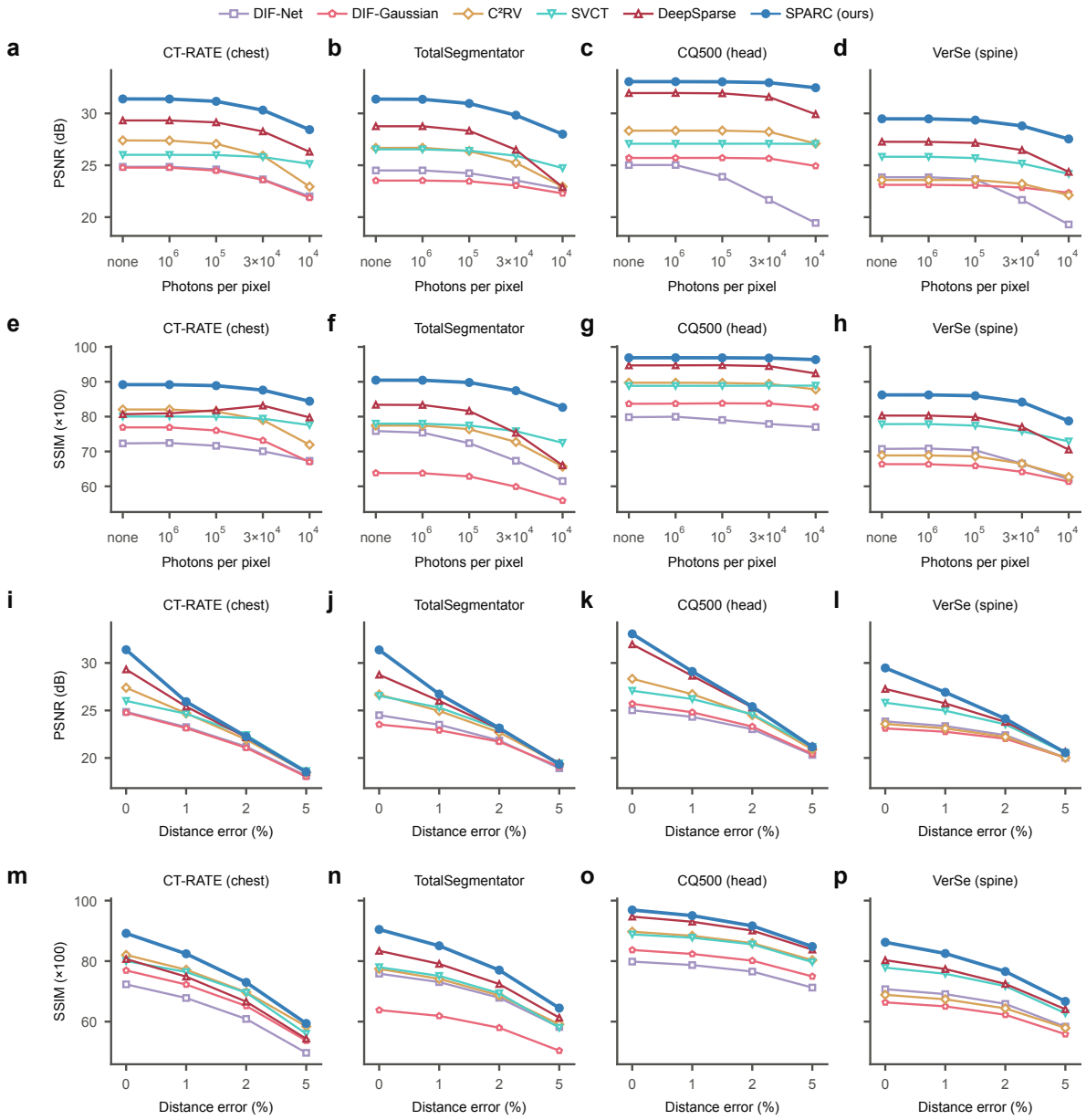


Fig. 4 | Robustness to acquisition errors. Every trained eight-projection model was evaluated again on perturbed projections of the test sets, without retraining (Methods). **a–h.** PSNR (**a–d**) and SSIM ($\times 100$, **e–h**) under photon noise on CT-RATE, TotalSegmentator, CQ500 and VerSe, against the number of incident photons per detector pixel. **i–p.** PSNR (**i–l**) and SSIM (**m–p**) under a source-to-detector distance error of 1, 2 or 5%. The leftmost point of each panel is the unperturbed evaluation. Point estimates are shown. Full results, including four projections and the other perturbations, are reported in Supplementary Tab. 11–17 and Supplementary Fig. 9–11.

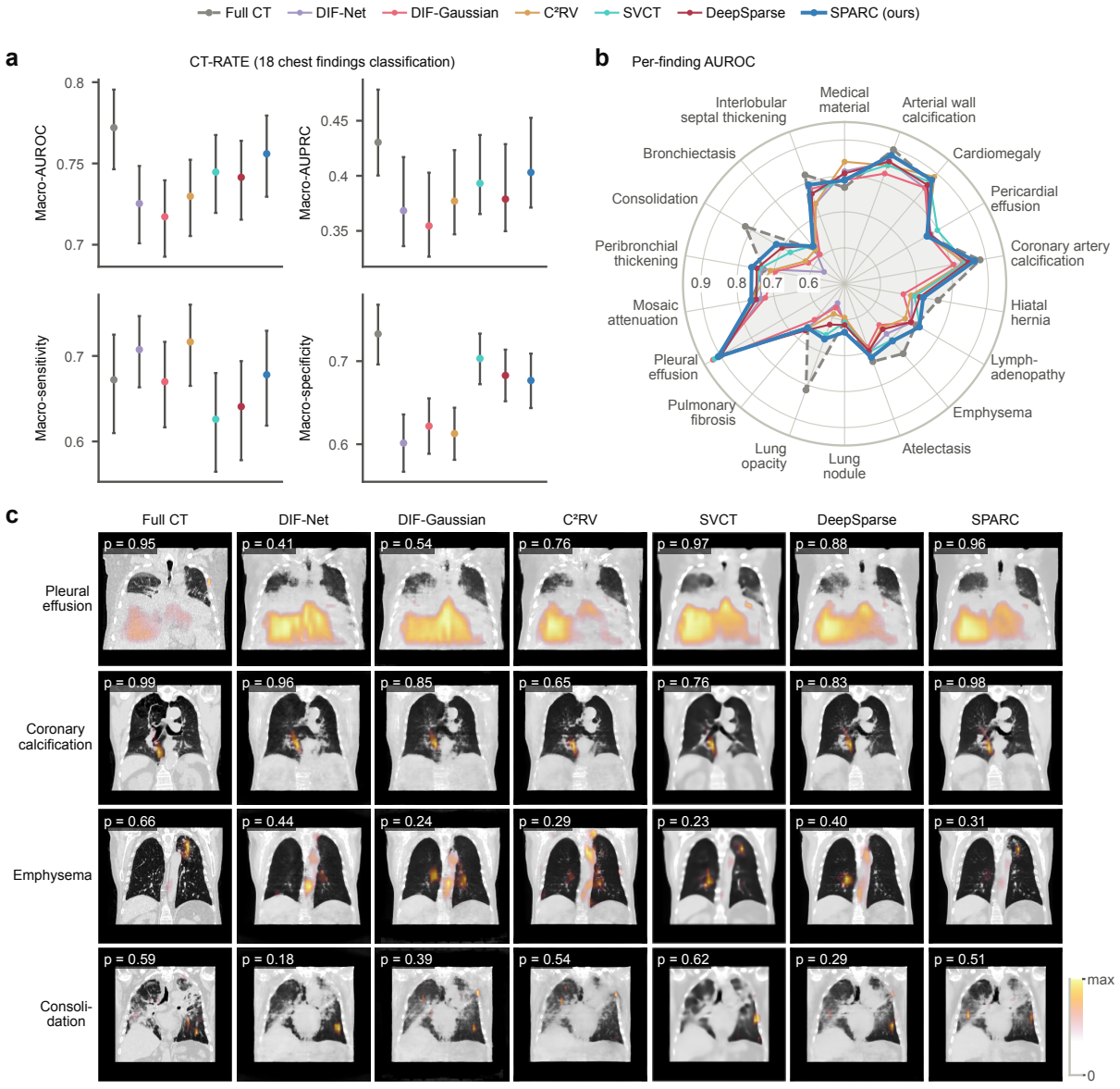


Fig. 5 | Abnormality classification by a classifier trained on full CT. Eight projections. Each method's reconstructions were given, unchanged, to the classifier. **a.** Macro-averaged AUROC, AUPRC, sensitivity and specificity over the 18 findings on the CT-RATE test set (602 volumes), with the classifier on full CT in grey. **b.** AUROC for each finding. **c.** Grad-CAM maps of one training seed of the classifier for four findings, on full CT and on each reconstruction; p, that classifier's probability for the finding. The findings are labelled from radiology reports without lesion outlines, so the maps are qualitative. Error bars denote 95% confidence intervals. Full results are reported in Supplementary Tab. 18–21; results with four projections are shown in Supplementary Fig. 12, and other ways of using SPARC for this task in Supplementary Fig. 13.

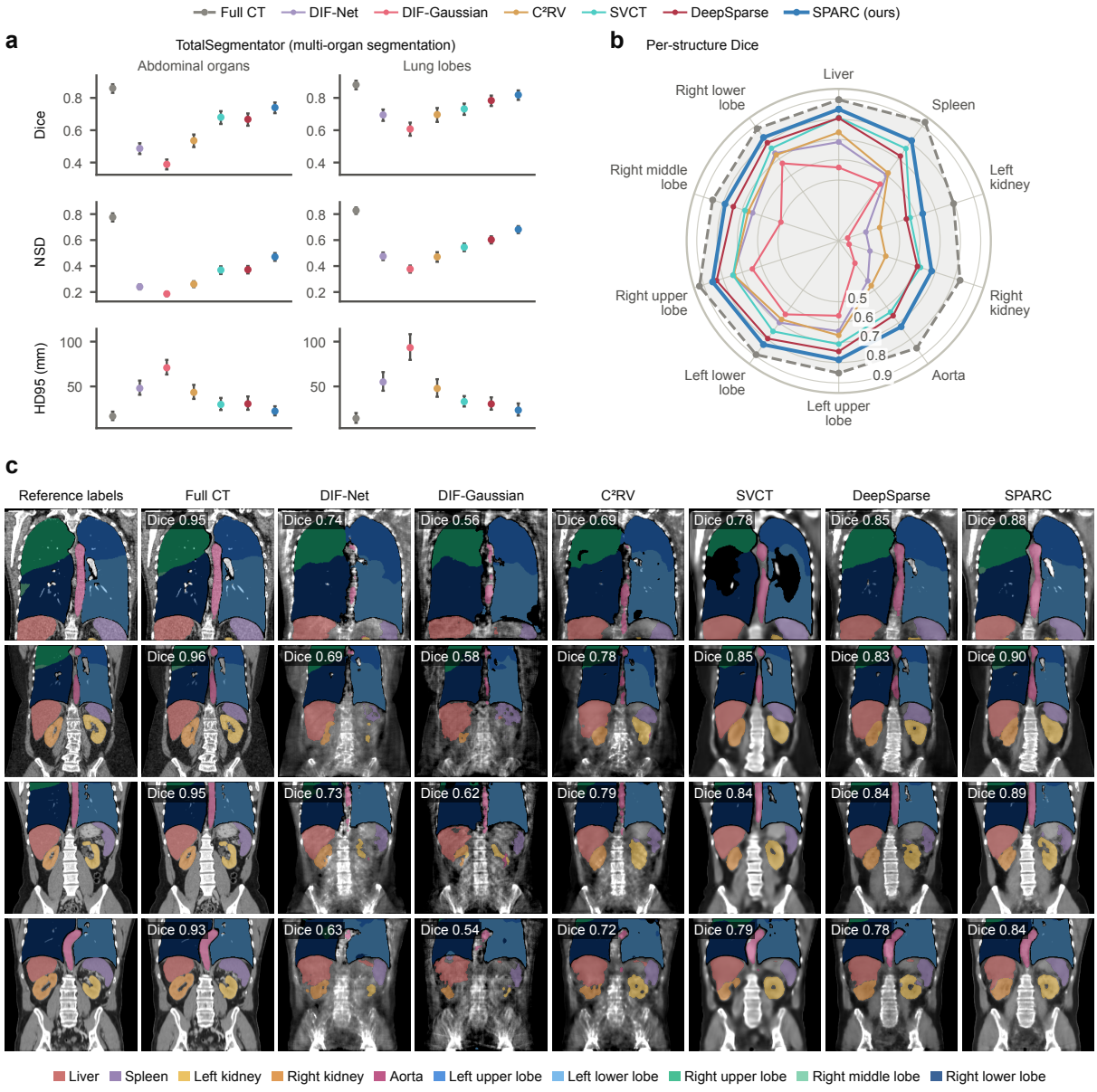


Fig. 6 | Multi-organ segmentation with a model trained on full CT. Eight projections. Each method's reconstructions were given, unchanged, to the segmentation model. **a.** Dice, normalised surface Dice at a 2-mm tolerance (NSD) and 95th-percentile Hausdorff distance (HD95) over the five abdominal organs (left) and the five lung lobes (right) on the TotalSegmentator test set (79 volumes), with the segmentation model on full CT in grey. **b.** Dice for each structure. **c.** Reference labels and the predictions of one training seed of the segmentation model on the CT and on each reconstruction for four test volumes, with each volume's mean Dice. Error bars denote 95% confidence intervals. Full results are reported in Supplementary Tab. 22–24; results with four projections are shown in Supplementary Fig. 14, and other ways of using SPARC for this task in Supplementary Fig. 13.

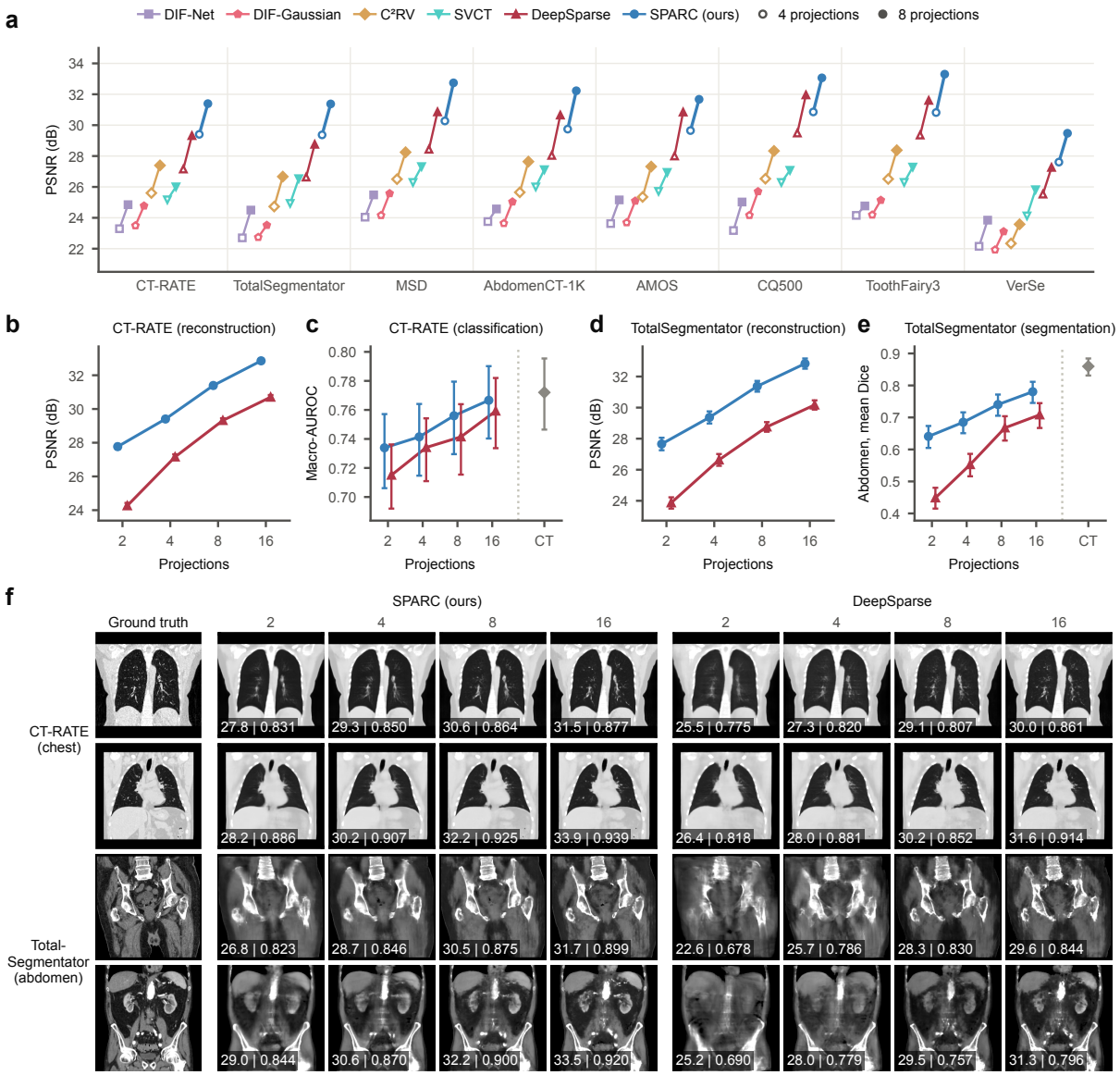


Fig. 7 | Effect of the number of projections. Each number of projections is a separately trained model. **a.** Mean PSNR of each method on each test set with four projections (open markers) and eight projections (filled markers). **b–e.** SPARC and DeepSparse with 2, 4, 8 and 16 projections. On CT-RATE, reconstruction PSNR (**b**) and macro-AUROC of the classifier trained on full CT (**c**). On TotalSegmentator, reconstruction PSNR (**d**) and abdominal Dice of the segmentation model trained on full CT (**e**). Grey diamonds denote the classifier and the segmentation model applied to full CT. **f.** Reconstructions by SPARC and DeepSparse from 2 to 16 projections for two test volumes from each of CT-RATE and TotalSegmentator; labels give the volume's PSNR (dB) | SSIM. Error bars denote 95% confidence intervals. Full results are reported in Supplementary Tab. 1, 2 and 26.

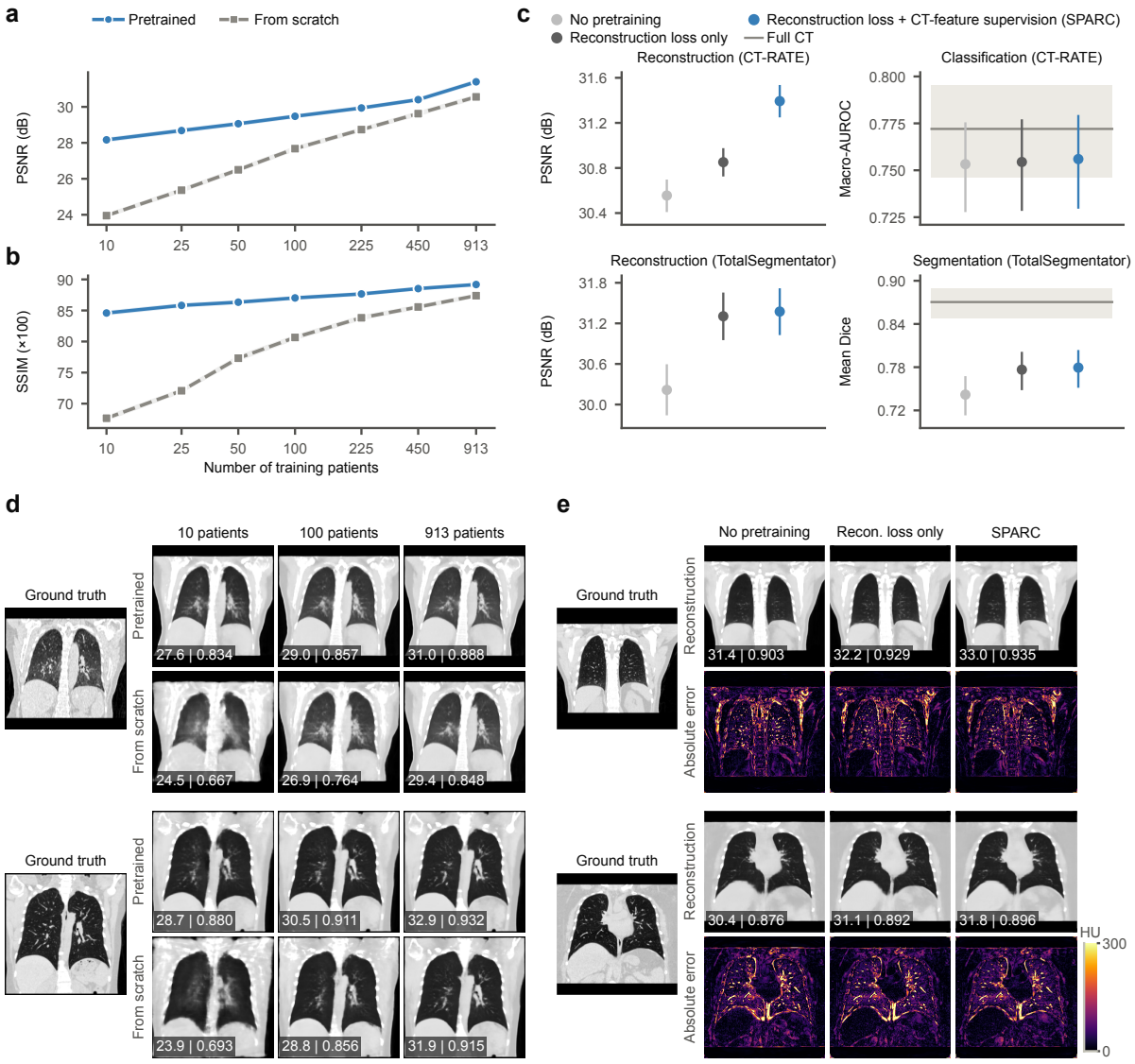


Fig. 8 | What pretraining contributes to SPARC. Eight projections; CT-RATE unless stated. **a, b.** Reconstruction PSNR (**a**) and SSIM (**b**) against the number of training patients, that is, patients with paired projections and CT used for adaptation, for SPARC (Pretrained) and the same architecture trained from random initialisation (From scratch). **c.** The same architecture without pretraining, pretrained with the reconstruction loss only, or pretrained with both losses (SPARC), each adapted in the same way. On CT-RATE, reconstruction PSNR and the macro-AUROC of the classifier trained on full CT. On TotalSegmentator, reconstruction PSNR and the mean Dice over the ten structures of the segmentation model trained on full CT. The classifier and the segmentation model were given the reconstructions of each of the three models; grey lines show their results on full CT. **d.** Reconstructions of two test volumes after adaptation with 10, 100 and 913 patients. **e.** Reconstructions of two test volumes by the three models of **c**, with their absolute error. Labels give the volume's PSNR (dB) | SSIM. Bands and error bars denote 95% confidence intervals. Full results are reported in Supplementary Tab. 27 and 28.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supptabs.pdf
- suppfigs.pdf