Diffusion-based Generative Image Outpainting for Recovery of FOV-Truncated CT Images

Michelle Espranita Liman[1,2], Daniel Rueckert[1,3], Florian J. Fintelmann[†,2,4], and Philip Müller[†,1](✉)

[1] Technical University of Munich, Germany
{michelle.liman,daniel.rueckert,philip.j.mueller}@tum.de
[2] Massachusetts General Hospital, USA
fintelmann@mgh.harvard.edu
[3] Imperial College London, UK
[4] Harvard Medical School, USA

Abstract. Field-of-view (FOV) recovery of truncated chest CT scans is crucial for accurate body composition analysis, which involves quantifying skeletal muscle and subcutaneous adipose tissue (SAT) on CT slices. This, in turn, enables disease prognostication. Here, we present a method for recovering truncated CT slices using generative image outpainting. We train a diffusion model and apply it to truncated CT slices generated by simulating a small FOV. Our model reliably recovers the truncated anatomy and outperforms the previous state-of-the-art despite being trained on 87% less data. Our code is available at https://github.com/michelleespranita/ct_palette.

Keywords: Computed tomography · Diffusion models · Field-of-view recovery · Generative image outpainting

1 Introduction

Body composition analysis is the analysis of the quantity, quality, and distribution of skeletal muscle and adipose tissue in the body [18]. It adds prognostic value to routine CT scans, including for patients with lung cancer [23] and pancreatic cancer [3]. However, chest CT scans are frequently subject to field-of-view (FOV) truncation [23], meaning that part of the chest wall muscle and adipose tissue is intentionally excluded to increase the image quality of the lung tissue [13]. Consequently, body composition analysis cannot be accurately conducted on truncated scans, and valuable information is lost.

Recovering FOV-truncated portions of chest CT scans presents a notable challenge with two key issues. Firstly, while recovering the truncated information from the raw data prior to image reconstruction would preserve data fidelity [8, 12, 14, 17], the raw data are periodically purged from CT scanners, making this approach impractical in clinical practice. Therefore, recovering the

[†]Shared last authorship

truncated part of the anatomy following image reconstruction is more realistic, as performed by the current state-of-the-art method known as S-EFOV [24]. Secondly, the FOV recovery problem is an ill-posed problem, meaning there exists a distribution of possible outputs for a given input. S-EFOV only generates a single recovered slice given a truncated CT slice. This feature does not allow its users to explore other possible recovered slices in the distribution.

To overcome these limitations, we introduce a novel approach for recovering truncated CT slices using generative image outpainting called *CT-Palette*. Inspired by the recent advancements of generative AI models, CT-Palette is built upon diffusion models [11, 21], which have been proven to outperform generative adversarial networks (GANs) for image-to-image tasks [7]. Leveraging diffusion models enables CT-Palette to generate multiple recovered slices from a truncated CT slice, making it ideal for tackling the ill-posed FOV recovery problem.

Our contributions are as follows:

- We propose *CT-Palette*, a diffusion-based CT field-of-view (FOV) recovery method capable of generating multiple recovered slices from a truncated CT slice. Subsequently, we utilize the muscle and SAT areas from each slice to select the most representative recovered slice as the final output.
- We train and evaluate CT-Palette using truncated CT slices generated by simulating real-world FOV truncation.
- We design a novel method for mask generation, which reduces the area of the unknown region to be outpainted in the CT slice. This improvement makes training more feasible, particularly on a small dataset.
- Despite being trained only on a small dataset, our experiments show that CT-Palette outperforms previous state-of-the-art models.

2 Method

As shown in Figure 1, CT-Palette consists of two components: 1) a body bounding box detector which estimates the bounding box representing the untruncated body on the truncated CT slice; 2) an image outpainting model which recovers the tissues of the truncated CT slice. Between the body bounding box detector and the image outpainting model, we "zoom out" the slice to ensure the bounding box of the untruncated body fits within the image borders. This ensures sufficient space for the outpainting model to recover the truncated tissues.

In addition to the zoomed-out slice, a binary mask is required as input to the outpainting model. This mask specifies the (unknown) region where the outpainting model should recover the tissues. Typically, this is the FOV mask detected from the truncated CT slice (the pixel values [Hounsfield Unit/HU] located outside the FOV are usually marked by a pre-defined value). However, this approach would cause a large part of the unknown region to include regions outside the body, which is unnecessary for the model to learn.

Therefore, we propose a novel and more effective method to generate masks for the outpainting model, aiming to reduce the size of the unknown region. We refer to the resulting mask as the "small mask". To create the small mask, we

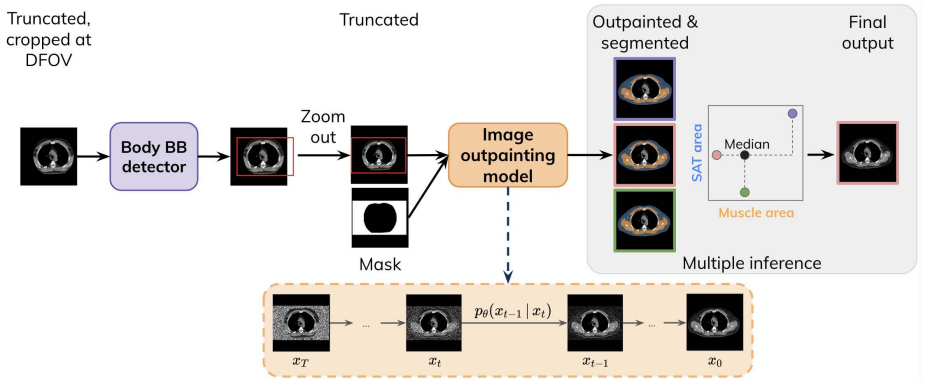

Fig. 1. An overview of CT-Palette. The body bounding box detector estimates the bounding box representing the complete/untruncated body, and the image outpainting model recovers the tissues of the truncated CT slice. The white region in the mask indicates the region to be outpainted. CT-Palette generates different slices at each run by drawing samples from the distribution it has learned during training. Using a body composition segmentation model, we extract the muscle and SAT areas from each slice. The final output is the slice with muscle and SAT areas closest to the median.

use the body bounding box scaled during the zoom-out. Then, we subtract the region occupied by the body (identified using [22]) from the region covered by the bounding box since the former is a known region that does not need to be learned by the model.

2.1 Synthetic data generation

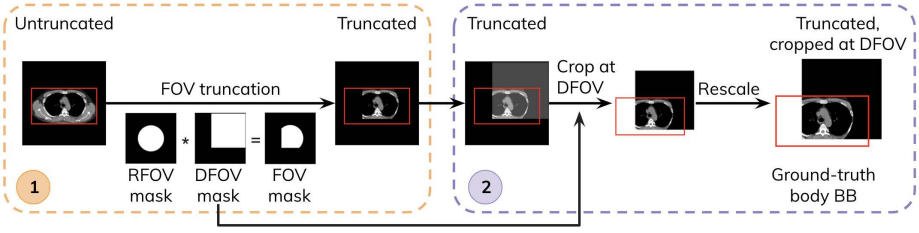

Fig. 2. Synthetic data generation process: 1) To train the image outpainting model, we generate truncated slices and their corresponding small masks (explained above); 2) To train the body bounding box detector, we generate truncated slices cropped at the display FOV (DFOV) and bounding boxes of the untruncated bodies.

We train the body bounding box detector and the image outpainting model separately (Figure 2). To train the body bounding box detector, we need trun-

cated slices cropped at the display FOV (DFOV) and bounding boxes of the untruncated bodies. To train the image outpainting model, we need truncated slices, their corresponding small masks, and untruncated slices.

To simulate the FOV truncation that occurs in CT scans, we generate synthetic FOV masks by combining the following FOV types [24]:

1. Reconstruction FOV (RFOV): the region of the scanned area used for image reconstruction. We generate a circular mask at the center, defined by the ratio r_{RFOV} relative to the slice resolution.
2. Display FOV (DFOV): the region displayed as the final output to the radiologist. We generate a square mask with a length defined by the ratio r_{DFOV} relative to the slice resolution. It is then offset from the center of the slice by $(x_{\text{DFOV}}, y_{\text{DFOV}})$, chosen from the range of $[-\frac{D}{2}, \frac{D}{2}]$, where $D = d * (1 - r_{\text{DFOV}})$ and d is the slice resolution.

We use the generated FOV masks to truncate the original untruncated slices, resulting in truncated slices for the outpainting model. We extract the bounding boxes of the untruncated bodies by identifying the body using the body mask identification algorithm developed by [22] and crop the truncated slices at the DFOV for the bounding box detector.

2.2 Training

Body bounding box detector This is a ResNet-18 [9] model pre-trained on ImageNet [6], with the last fully-connected layer replaced with a fully-connected layer which has 4 output nodes representing the axis-aligned bounding box coordinates of the untruncated body $\hat{B} = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$.

Image outpainting model Inspired by [19], this is a conditional diffusion model [5, 20] that learns the distribution of possible untruncated slices given a truncated CT slice. At the beginning of the diffusion process, we inject random Gaussian noise exclusively into the unknown region to be outpainted and keep the known region fixed. Over iterations, the model denoises the unknown region, resulting in an outpainted, untruncated slice.

We train for 36 hours on 8 NVIDIA Quadro RTX 8000 GPUs with up to 48 GB of memory assigned to every GPU, resulting in around 340 epochs. We use a total batch size of 128 and optimize the model using the Adam [15] optimizer with a learning rate of 5×10^{-5} and no weight decay. All the other hyperparameters are set to the same values as in [19].

2.3 Inference

As the diffusion process begins with random noise in the unknown region, we obtain different outpainted slices with each run. Based on this, we propose a novel inference method called *multiple inference*: We generate $n > 1$ (default:

$n = 5$) outpainted slices given a truncated slice and select the most representative slice based on statistics. This prevents an outlier from being the final output.

To select the most representative slice, we extract the body composition metrics, i.e. muscle and SAT areas, from each outpainted slice using a segmentation model [4] and calculate the median muscle area $\bar{x}_{\text{muscle}}$ and the median SAT area $\bar{x}_{\text{SAT}}$. Then, we calculate the L_1 distance from the muscle and SAT areas of each outpainted slice to their corresponding medians:

$$L_1(x_i, \bar{x}) = |x_{i,\text{muscle}} - \bar{x}_{\text{muscle}}| + |x_{i,\text{SAT}} - \bar{x}_{\text{SAT}}| \tag{1}$$

where $x_{i,\{\text{muscle/SAT}\}}$ is the muscle/SAT area of the outpainted slice x_i. Finally, the outpainted slice with the smallest distance to the median is selected.

3 Experimental Setup

3.1 Dataset and Pre-processing

We use chest CT scans from the Framingham Heart Study [1] for training and evaluation. The scans were obtained between 2002 and 2011. After an image quality review by a radiologist, the included patients are 54% female, 91% white, with a mean age of 56.1 ± 10.4 years, and a mean BMI of 29.9 ± 5.4 kg/m^2.

We utilize only slices representing the T5, T8, and T10 vertebral levels of the CT scans, commonly used for body composition analysis on chest CTs [4]. We process the slices by applying a soft-tissue window of range [-160, 240] HU and linear transformation to ensure pixel values range between [-1, 1]. Afterwards, we use the body mask corresponding to the CT slice to remove extraneous objects outside the body, such as the scan table. Finally, the CT slices are downsampled from a resolution of 512 to 256 using bilinear interpolation.

To prevent data leakage, we split the dataset into train (9,061 slices from 3,152 patients), val (899 slices from 177 patients), and test (998 slices from 210 patients) sets by patient ID. There is no significant difference in patient characteristics between the sets.

3.2 Baselines

As baselines, we use S-EFOV [24], the current state-of-the-art; and RFR-Net [16], the underlying architecture of S-EFOV's outpainting model. We only compare the outpainting models, as all methods use the same body bounding box detector.

S-EFOV was originally trained on 71,319 slices from the Vanderbilt Lung Screening Program (VLSP) [2] dataset, covering vertebral levels slightly above T5 and below T10. To ensure a fair comparison, we use the published weights for S-EFOV with and without fine-tuning on our dataset. For fine-tuning, we adopt the same training settings outlined in [24], incorporating data augmentation and utilizing FOV masks as input to the outpainting model.

To highlight the difference in performance between our method and RFR-Net, we train RFR-Net from scratch on our dataset using small masks and the

same values for the synthetic data generation parameters ($r_{\text{RFOV}} \sim \mathcal{U}[0.5, 0.7]$ and $r_{\text{DFOV}} \sim \mathcal{U}[0.65, 0.9]$). We optimize RFR-Net using Adam with a learning rate of 2×10^{-4} for a maximum of 200 epochs. Early stopping is implemented with a patience threshold of 60 epochs.

3.3 Evaluation Strategy

Quantitative Evaluation We evaluate the outpainting models based on their ability to recover the truncated slice. This involves ensuring that the muscle and SAT areas of the recovered slice closely align with those of the corresponding ground-truth untruncated slice, measured by RMSE. We extract the muscle and SAT areas using the segmentation model developed by [4] and calculate the segmented areas (cm^2) using the pixel spacing obtained from the corresponding DICOM files.

Qualitative Evaluation We qualitatively evaluate the slices recovered by CT-Palette against those by S-EFOV using real-world truncated slices without ground truth. To that end, we randomly select 40 samples for each vertebral level (T5, T8, and T10). Two trained radiologists (A, a radiology resident with 3 years of experience; B, a radiology attending with over 10 years of experience) independently perform the evaluation. Unaware of which recovered slice corresponds to which model, they are tasked with selecting the more realistic recovered slice or choosing "no difference" if neither is more realistic than the other.

4 Results

4.1 Quantitative Evaluation

Figure 3 compares examples of outpainted slices by various models from different vertebral levels. CT-Palette consistently produces realistic slices, while RFR-Net struggles to generate coherent outpaintings despite being trained on the same dataset. Notably, CT-Palette successfully restores the shoulder blade (see red arrow) in the T5 example, which is missed by other models. Additionally, CT-Palette achieves the lowest overall FID [10] (26% reduction from S-EFOV, see supp. material), indicating that the distribution of its outpainted slices matches that of the ground-truth untruncated slices the most.

In recovering muscle and SAT areas (Table 1), CT-Palette achieves a significant overall reduction in RMSE compared to the best-performing baseline S-EFOV (56.36% for muscle, 36.66% for SAT when using multiple inference; 44.70% for muscle, 20.38% for SAT when using single inference), despite S-EFOV being pre-trained on a substantially larger dataset.

The box plots in Figure 4 illustrate that the differences in muscle and SAT areas between the untruncated and outpainted slices by CT-Palette follow a Gaussian-like distribution, with an equal tendency for over- or underestimation.

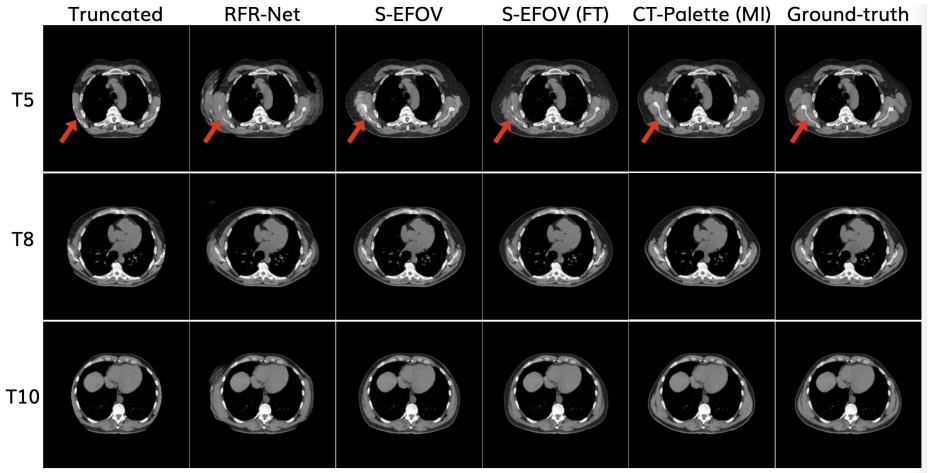

Fig. 3. Visual comparison of outpainted slices from different vertebral levels by various models, given a truncated CT slice. (FT): Fine-tuned. (MI): Multiple inference. CT-Palette recovers the truncated slice most realistically. Particularly in the T5 example, CT-Palette successfully restores the shoulder blade, as indicated by the red arrow.

Table 1. RMSE for muscle and SAT areas between ground-truth untruncated slices and outpainted slices by various models, stratified by vertebral levels. The topmost row indicates the RMSE between untruncated and truncated slices for reference. CT-Palette achieves the biggest reduction in RMSE for both muscle and SAT. (FT): Fine-tuned. (SI): Single inference. (MI): Multiple inference.

Method	Overall (n=997)		T5 (n=339)		T8 (n=348)		T10 (n=310)	
	Muscle	SAT	Muscle	SAT	Muscle	SAT	Muscle	SAT
Truncated	26.654	76.876	38.397	83.630	21.276	80.062	12.824	64.521
RFR-Net	9.022	35.258	10.778	32.388	9.134	40.776	6.412	31.376
S-EFOV	9.091	20.480	11.983	18.096	8.833	24.354	4.606	18.030
S-EFOV (FT)	8.202	24.974	9.014	24.252	9.188	26.653	5.723	23.773
CT-Palette (SI)	5.027	16.307	5.999	16.541	5.224	17.430	3.361	14.663
CT-Palette (MI)	**3.967**	**12.973**	**4.381**	**12.471**	**4.231**	**13.467**	**3.088**	**12.946**

CT-Palette's box plots exhibit a narrow interquartile range and a median close to zero, indicating no significant difference in area between untruncated and outpainted slices. In contrast, S-EFOV consistently underestimates both muscle and SAT areas, likely due to differences between our dataset and the dataset on which S-EFOV was originally trained. Fine-tuning S-EFOV shifts the distribution of SAT area differences toward a greater inclination for overestimation. The Wilcoxon signed-rank test shows no significant difference in the muscle/SAT area distribution between ground-truth images and those outpainted by our CT-Palette, while other models show a significant difference (see supp. material).

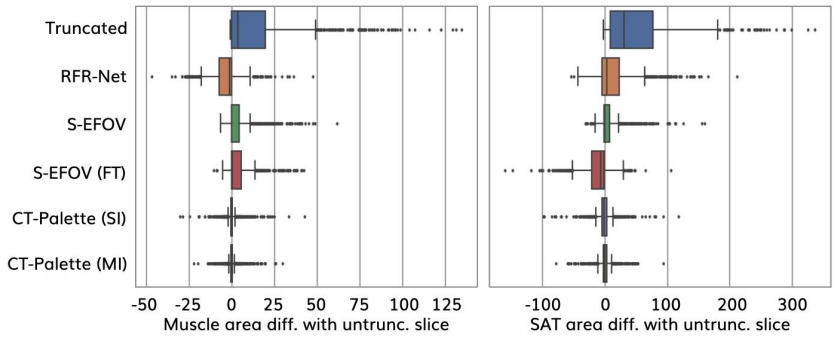

Fig. 4. Box plots of the differences in muscle and SAT areas between ground-truth untruncated and outpainted slices by various models. The topmost box plots illustrate the difference in area between the untruncated and truncated slices for reference. Positive values indicate underestimation, negative values indicate overestimation, and 0 indicates no difference. CT-Palette has a median close to zero and very narrow IQR, suggesting little to no deviation from ground-truth in muscle and SAT areas.

Lastly, our method achieves the highest Dice score overall (0.979 for muscle, 0.954 for SAT when using multiple inference), significantly outperforming S-EFOV by 0.31% in muscle and 0.63% in SAT (see supp. material).

We hypothesize that CT-Palette excels even with training only on a small dataset due to its nature as a diffusion model that directly learns the underlying data distribution. In contrast, RFR-Net is designed to minimize a task-specific supervised loss for image inpainting/outpainting. This loss function is composed of style loss, perceptual loss, L1 loss, and TV loss, all of which require extensive experimentation with their coefficients. This inflexibility may hinder RFR-Net's ability to generalize effectively when trained on a limited dataset.

4.2 Qualitative evaluation

Radiologist A perceives no significant difference between the CT slices recovered by CT-Palette and S-EFOV in 76.67% of the samples, while radiologist B finds CT-Palette's slices more realistic than S-EFOV's in 58.33%. The Cohen's Kappa Coefficient of 0.01 indicates low agreement. Overall, radiologists would tend to find slices from both methods equally realistic, with a slight preference towards those recovered by CT-Palette. The distribution of choices of both radiologists is presented in the supp. material.

5 Discussion and Conclusion

We presented CT-Palette, a novel approach for recovering truncated chest CT slices via generative image outpainting. It enhances the accuracy of body composition analysis (muscle and SAT areas), thereby enabling disease prognostication.

CT-Palette improves on the previous state-of-the-art by generating multiple possible recovered slices given an input. To increase training efficiency on a small dataset, we designed a novel mask generation method that substantially reduces the size of the unknown region for outpainting. CT-Palette demands considerably less data than the previous state-of-the-art while achieving superior performance.

The application of CT-Palette extends beyond chest CT scans to any scan obtained for routine care, including the neck and abdomen. Currently, CT-Palette's main drawback is its slow inference speed due to the large number of steps involved in image generation by the outpainting model. We leave the exploration of techniques to enhance CT-Palette's inference speed for future work.

Acknowledgments. The Framingham Heart Study is supported by Contract No. HHSN268201500001I from the National Heart, Lung, and Blood Institute (NHLBI) with additional support from other sources. This work was supported by the FHS Core Contract (NHLBI award #75N92019D00031). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or conclusions of the Framingham Heart Study or the NHLBI.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Framingham heart study. https://www.framinghamheartstudy.org/
2. Vlsp. https://www.vumc.org/radiology/lung
3. Babic, A., Rosenthal, M.H., Sundaresan, T.K., Khalaf, N., Lee, V., Brais, L.K., Loftus, M., Caplan, L., Denning, S., Gurung, A., Harrod, J., Schawkat, K., Yuan, C., Wang, Q.L., Lee, A.A., Biller, L.H., Yurgelun, M.B., Ng, K., Nowak, J.A., Aguirre, A.J., Bhatia, S.N., Vander Heiden, M.G., Van Den Eeden, S.K., Caan, B.J., Wolpin, B.M.: Adipose tissue and skeletal muscle wasting precede clinical diagnosis of pancreatic cancer. Nature Communications **14**(1), 4317 (Jul 2023). https://doi.org/10.1038/s41467-023-40024-3, https://doi.org/10.1038/s41467-023-40024-3
4. Bridge, C.P., Best, T.D., Wrobel, M.M., Marquardt, J.P., Magudia, K., Javidan, C., Chung, J.H., Kalpathy-Cramer, J., Andriole, K.P., Fintelmann, F.J.: A fully automated deep learning pipeline for multi–vertebral level quantification and characterization of muscle and adipose tissue on chest ct scans. Radiology: Artificial Intelligence **4**(1), e210080 (2022). https://doi.org/10.1148/ryai.210080, https://doi.org/10.1148/ryai.210080
5. Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., Chan, W.: Wavegrad: Estimating gradients for waveform generation (2020)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis (2021)
8. Éric Fournié, Baer-Beck, M., Stierstorfer, K.: Ct field of view extension using combined channels extension and deep learning methods (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)

10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
12. Huang, Y., Gao, L., Preuhs, A., Maier, A.: Field of view extension in computed tomography using deep learning prior (2019)
13. Kazerooni, E.A., Austin, J.H., Black, W.C., Dyer, D.S., Hazelton, T.R., Leung, A.N., McNitt-Gray, M.F., Munden, R.F., Pipavath, S.: Acr–str practice parameter for the performance and reporting of lung cancer screening thoracic computed tomography (ct): 2014 (resolution 4)*. Journal of Thoracic Imaging **29**(5) (2014), https://journals.lww.com/thoracicimaging/fulltext/2014/09000/acr_str_practice_parameter_for_the_performance_and.12.aspx
14. Ketola, J.H.J., Heino, H., Juntunen, M.A.K., Nieminen, M.T., Siltanen, S., Inkinen, S.I.: Generative adversarial networks improve interior computed tomography angiography reconstruction. Biomed. Phys. Eng. Express **7**(6), 065041 (Oct 2021)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
16. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting (2020)
17. Li, Z., Cai, A., Wang, L., Zhang, W., Tang, C., Li, L., Liang, N., Yan, B.: Promising generative adversarial network based sinogram inpainting method for ultra-limited-angle computed tomography imaging. Sensors (Basel) **19**(18), 3941 (Sep 2019)
18. Magudia, K., Bridge, C.P., Bay, C.P., Babic, A., Fintelmann, F.J., Troschel, F.M., Miskin, N., Wrobel, W.C., Brais, L.K., Andriole, K.P., Wolpin, B.M., Rosenthal, M.H.: Population-scale CT-based body composition analysis of a large outpatient population using deep learning to derive age-, sex-, and race-specific reference curves. Radiology **298**(2), 319–329 (Feb 2021)
19. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models (2022)
20. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement (2021)
21. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics (2015)
22. Tang, Y., Gao, R., Han, S., Chen, Y., Gao, D., Nath, V., Bermudez, C., Savona, M.R., Bao, S., Lyu, I., Huo, Y., Landman, B.A.: Body part regression with self-supervision. IEEE Transactions on Medical Imaging **40**(5), 1499–1507 (2021). https://doi.org/10.1109/TMI.2021.3058281
23. Troschel, A.S., Troschel, F.M., Best, T.D., Gaissert, H.A., Torriani, M., Muniappan, A., Van Seventer, E.E., Nipp, R.D., Roeland, E.J., Temel, J.S., Fintelmann, F.J.: Computed tomography-based body composition analysis and its role in lung cancer care. J. Thorac. Imaging **35**(2), 91–100 (Mar 2020)
24. Xu, K., Li, T., Khan, M.S., Gao, R., Antic, S.L., Huo, Y., Sandler, K.L., Maldonado, F., Landman, B.A.: Body composition assessment with limited field-of-view computed tomography: A semantic image extension perspective (2023)