Synthesis of Spectral CT Tissue Maps from Single-Energy CT Using Conditional Diffusion Models

Jerry Sarubbe

Abstract—Spectral computed tomography (CT) is an advanced imaging technique that allows for the decomposition of tissues into constituent materials, offering significant advantages in medical diagnostics. However, it typically requires specialized dual-energy systems, which can increase cost, radiation dose, and procedure complexity. We propose using a conditional denoising diffusion probabilistic model (C-DDPM), an approach that allows for the synthesis of three distinct tissue decomposition maps (adipose, fibroglandular, and calcification) directly from single-energy 50 kVp CT images, potentially overcoming the limitations of traditional dual-energy systems. Our approach conditions a diffusion model on the input CT image via channel concatenation and trains with a weighted noise prediction loss. Evaluated using the AAPM DL-Spectral-CT Challenge dataset (1,000 breast phantom slices at 512×512 resolution), our method achieves an average RMSE of 0.042, SSIM of 0.867, and PSNR of 30.1 dB across all tissue types after applying post-processing. We additionally compare against a U-Net baseline, which was found to produce sharper tissue maps with higher overall empirical performance, though it exhibits errors at the tissue boundaries.

Index Terms—Spectral CT, material decomposition, conditional diffusion models, medical imaging.

1 Introduction

Dual-energy computed tomography (DECT) is a spectral CT imaging modality that acquires images at two different X-ray energy levels. This leverages the fact that different materials have different X-ray attenuation at various energy levels, allowing for the decomposition of tissue into its constituent materials based on their attenuation properties [1]. There are several motivations for using DECT in clinical practice; It can enable differentiation between contrast media and bone or calcified vessels, allow the generation of virtual non-contrast images via material decomposition (eliminating the need for both contrasted and non-contrasted scans), and improve detection and characterization of tumors—particularly for establishing tumor boundaries, which is important for preventing damage to healthy tissue during radiation therapy [2].

However, DECT requires specialized systems which are significantly more expensive than SECT systems, which prevents their ubiquity. Additionally, traditional dual-energy CT imaging allows for the decomposition of tissue into only two basis materials due to capturing images at two different energies, whereas learning-based approaches may be able to expand upon this limitation.

This project aims to investigate the viability of using diffusion-based approaches for estimating material maps from a single-energy CT scan. We address this problem using a conditional denoising diffusion probabilistic model (C-DDPM), a class of generative models that has recently shown strong results in image-to-image translation tasks [3], [4]. Our model takes a single 50 kVp CT image as input and generates three tissue fraction maps (adipose, fibroglandular, and calcification) through an iterative denoising process conditioned on the input image. A significant challenge is the extreme sparsity of calcification tissue, where signal is not present in a vast majority of pixels. This sparsity can lead the model to predict near-zero values for this channel unless addressed with specific loss weighting and post-processing techniques.

• *Department of Electrical Engineering, Stanford University. E-mail: jsarubbe@stanford.edu*

2 Related Work

Material decomposition from CT. Traditional material decomposition approaches rely on physics-based methods that exploit the known energy dependence of X-ray attenuation coefficients [5]. Lyu *et al.* [6] proposed a CNN-based approach to estimate high-energy CT images from corresponding low-energy CT input images and a single high-energy image, enabling material decomposition without utilizing a full dual-energy scan. Their work demonstrated that deep learning could learn the mapping between single-energy images and material-specific information.

Conditional DDPMs for CT synthesis. Denoising diffusion probabilistic models have been shown to be a useful class of generative models that can produce high-quality samples through iterative denoising. Ho *et al.* [3] notably outlined the DDPM framework, with subsequent work by others such as Nichol and Dhariwal introducing improvements [7]. Other have taken this approach as conditioned the DDPM model for medical applications. Gao *et al.* [8] proposed a C-DDPM for generating contrast-enhanced DECT from non-contrast SECT by concatenating the conditioning image channel-wise with the noisy target at each reverse step. Their approach, which focused on head and neck data, outperformed Pix2PixGAN and CNN baselines. A prior paper from the same group demonstrated that the same framework can be utilized for iodine map synthesis [9].

Diffusion posterior sampling. Jiang *et al.* used spectral diffusion posterior sampling (DPS) as a framework for solving multi-material decomposition. This approach performed on par with, if not better than, conventional material decomposition methods. It had faster compute times, and as the authors state, it does not need to be retrained for specific scenario of the task, as this method is more generalizable. [10]

3 METHODS

3.1 Problem Statement

Given a single-energy CT image $y \in \mathbb{R}^{1\times H\times W}$, our goal is to predict the corresponding material decomposition maps $x_0 \in \mathbb{R}^{3\times H\times W}$, where the three channels represent adipose, fibroglandular, and calcification tissue volume fractions. Each tissue map takes values in $[0, 1]$, representing the fractional contribution of that material.

The material decomposition satisfies the physical constraint:

$$\mu(E) = \alpha_{\text{adip}}\,\mu_{\text{adip}}(E) + \alpha_{\text{fibro}}\,\mu_{\text{fibro}}(E) + \alpha_{\text{calc}}\,\mu_{\text{calc}}(E), \quad (1)$$

where $\mu(E)$ is the measured linear attenuation coefficient at energy E, and α_i denotes the volume fraction of each material.

We note that the original AAPM Challenge was designed for dual-energy decomposition using both low- and high-energy CT images as input; our single-energy formulation using only the 50 kVp image is a strictly harder problem, as it discards the complementary energy information that conventional methods rely on.

3.2 Dataset

We used the AAPM DL-Spectral-CT Grand Challenge dataset [11], which consists of 1,000 simulated breast CT slices at 512×512 resolution. Each slice includes a 50 kVp single-energy CT image and three ground-truth tissue fraction maps (adipose, fibroglandular, calcification) derived from a known breast phantom composition. This dataset is one of the few publicly available dual-energy CT datasets with paired ground truth, which is generally unavailable in clinical settings.

An example of a slice from the dataset is shown in Fig. 1, with each image including the high and low energy images, the tissue maps for that given slice, and the material sinograms. For the purposes of this project, we chose to operate in the pixel domain and thus are synthesizing the reconstruction of the tissue maps.

We partitioned the data into 850 training, 100 validation, and 50 test slices (85/10/5 split). Input CT images (raw range approximately $[-0.07, 1.80]$) were clamped and then linearly scaled to $[-1, 1]$. Target tissue maps were linearly scaled from $[0, 1]$ to $[-1, 1]$ for the diffusion process and rescaled back for evaluation. Data augmentation consisted of random horizontal and vertical flips, applied identically to both the input CT image and target tissue maps.

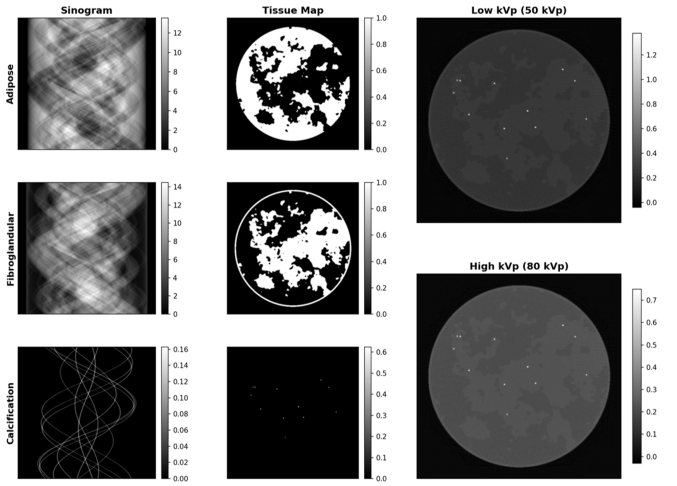

Fig. 1. Example slice from the AAPM DL-Spectral-CT dataset showing sinograms (left column), ground-truth tissue maps (center column), and the corresponding low kVp (50 kVp) and high kVp (80 kVp) CT images (right column). Only the low kVp image and tissue maps are used in our approach.

3.3 Conditional DDPM

In this project, our primary model is a conditional denoising diffusion probabilistic model. A DDPM is a generative model that predicts the Gaussian noise added at a specific timestep, enabling iterative image reconstruction by reversing the forward diffusion process to recover an image from pure noise [3]. In our conditional formulation, the input CT image is provided to the network at every step of the reverse process, allowing it to leverage the anatomical structure present in the conditioning image throughout generation.

The forward diffusion process gradually adds Gaussian noise to the clean tissue maps x_0 over $T = 1000$ timesteps:

$$x_t = \sqrt{\bar{\alpha}_t}\,x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad (2)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the cumulative product of the noise schedule.

Conditioning on the input CT image y is achieved through channel concatenation: at each denoising step, the noisy sample $x_t \in \mathbb{R}^{3\times H\times W}$ is concatenated with y along the channel dimension, producing a 4-channel input to the U-Net, which predicts the added noise. The model is trained using the following objective:

$$\mathcal{L} = \mathbb{E}_{x_0,y,t,\epsilon}\left[\sum_{c=1}^{3} w_c \left\|\epsilon_c - \epsilon_\theta(x_t, t, y)_c\right\|_2^2\right], \qquad (3)$$

where w_c are per-channel weights set to 1.0 for adipose and fibroglandular and 5.0 for calcification. The elevated calcification weight was necessary to prevent the case where the model trivially minimizes the loss by predicting zero calcification everywhere due to the extreme sparsity of this tissue type. An graphical representation of the DDPM process is shown in Fig. 2.

3.4 Network Architecture

Our U-Net backbone is based on the architecture proposed by Nichol and Dhariwal [7]. It uses four resolution levels.

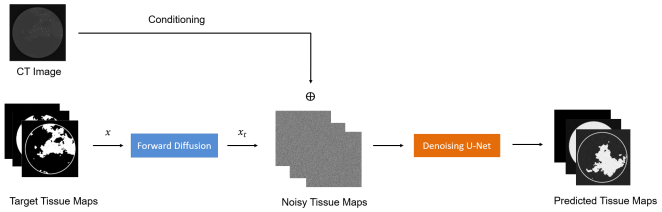

Fig. 2. Overview of the C-DDPM framework. The input 50 kVp CT image is concatenated channel-wise with the noisy tissue maps x_t at each reverse diffusion step. The denoising U-Net predicts the noise ϵ, conditioned on both x_t and y via channel concatenation and on timestep t.

Each resolution level contains residual blocks with GroupNorm normalization and SiLU activations. Skip connections concatenate the encoder features with decoder features at each level, and downsampling is performed via strided convolutions.

The integer timestep t is encoded via a sinusoidal positional embedding, projected through a two-layer MLP, and used to modulate each residual block's features via learned scale and shift parameters. A self-attention layer is applied in the middle block at the bottleneck resolution to capture long-range dependencies. The model takes a 4-channel input (three noisy tissue channels plus one conditioning CT channel) and produces a 3-channel output (predicted noise for each tissue type), with approximately 58.5 million trainable parameters.

3.5 Training Details

We trained the model using AdamW with a learning rate of 10^{-4}. We used a batch size of 8. Training proceeded for approximately 185,000 iterations. Training time took approximately 14 hours of run time.

3.6 Post-Processing

Additionally, after training, to recover less noisy image results, we applied post-processing techniques to the C-DDPM tissue maps to better align them with the ground truth. For the adipose and fibroglandular channels, we apply a sigmoid function which pushes the values that are near 0 and 1 to their extremes while preserving intermediary values that exist at the tissue boundaries. For the calcification maps, histogram matching was applied, where the distribution of predicted calcification values is remapped to match the distribution of ground-truth calcification values from the training set. This is effective because the ground-truth calcification distribution consists of virtually all zero values. The histogram matching suppresses background noise while preserving the actual calcification signals. Values below 0.05 are then thresholded to zero.

4 Results

We evaluated performance using three standard metrics computed per tissue and averaged: root mean square error (RMSE), structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR). All metrics were computed on the 50-sample test set. Given the sparsity of calcification within the data, RMSE serves as the primary evaluation metric, though SSIM and PSNR are also reported for reference.

TABLE 1
U-Net baseline per-tissue performance metrics.

Tissue	RMSE	SSIM	PSNR (dB)
Adipose	0.0269	0.9849	31.48
Fibroglandular	0.0271	0.9846	31.41
Calcification	0.0003	0.9999	69.73
Average	**0.0181**	**0.9890**	**44.21**

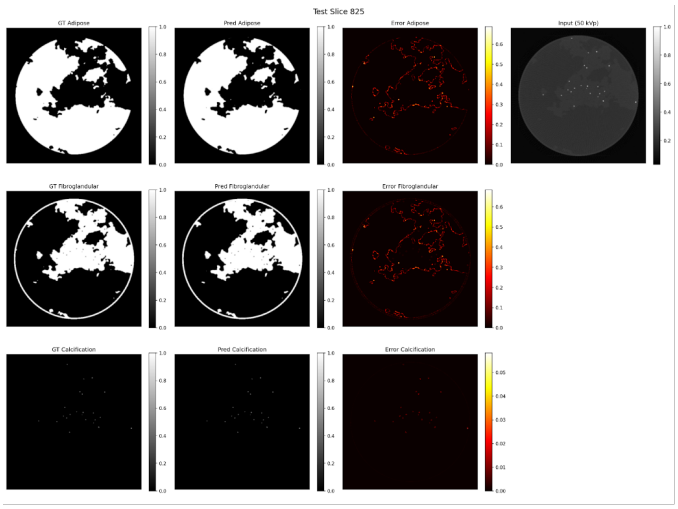

Fig. 3. Baseline U-Net results for test slice 825, showing ground truth, predicted tissue maps, and absolute error maps for each tissue type. Errors are concentrated along tissue boundaries where the model struggles to predict sharp transitions.

4.1 Baseline Comparison

For a baseline comparison, we trained a deterministic U-Net with the same backbone architecture as the denoising U-Net, but without timestep conditioning and with BatchNorm in place of GroupNorm. The same objective function of minimizing the weighted MSE was used for training. After 100 epochs the U-Net was fairly performant, achieving an average RMSE of 0.0181 and PSNR of 44.21 dB, as shown in Table 1. Example outputs from the baseline are shown in Fig. 3. It can be observed in the error maps that the model generally fails along the borders between tissues—which clinically is the most crucial area to predict accurately.

4.2 DDIM Sampling Steps

At inference time, we use DDIM sampling. Initially we used 50 denoising steps rather than the full 1,000-step DDPM reverse process; however, we found that increasing to 100 steps had a non-trivial improvement on image quality across all tissue types, as shown in Table 2, with a notable 12% improvement in average RMSE. All subsequent C-DDPM results use 100 DDIM steps.

4.3 Training Progression

Fig. 4 shows the training loss over approximately 185k iterations. The loss decreases rapidly early in training and continues to converge gradually thereafter. Fig. 5 shows how the generated tissue maps for an example slice in

TABLE 2
Effect of DDIM sampling steps on raw C-DDPM performance (no post-processing).

Metric	DDIM-50	DDIM-100	Change
Adipose RMSE	0.113	0.093	18% improved
Fibro RMSE	0.186	0.131	29% improved
Calc RMSE	0.464	0.446	4% improved
Avg RMSE	0.254	0.223	12% improved
Avg SSIM	0.119	0.138	16% improved
Avg PSNR (dB)	13.45	15.45	+2.0 dB

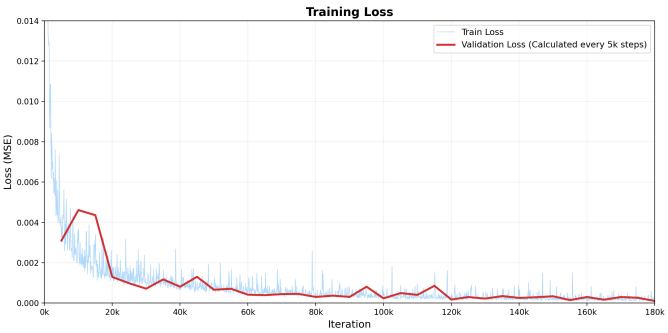

Fig. 4. Training and validation loss over approximately 185k iterations. Validation loss is computed every 5k steps.

TABLE 3
C-DDPM per-tissue metrics before and after post-processing (PP) (100 DDIM steps).

	RMSE		SSIM		PSNR (dB)	
Tissue	Raw	PP	Raw	PP	Raw	PP
Adipose	0.093	0.053	0.229	0.907	20.9	25.6
Fibrogland.	0.131	0.064	0.183	0.725	18.0	24.0
Calcification	0.446	0.010	0.001	0.971	7.4	40.6
Average	**0.223**	**0.042**	**0.138**	**0.867**	**15.5**	**30.1**

the test set evolve at various training iterations, showing progressive refinement of tissue structure and boundary detail and also show that the differences between the later iterations, at least visually, appears minimal.

4.4 Quantitative Results

Table 3 presents the per-tissue C-DDPM results before and after post-processing. The raw model outputs have poor SSIM due to residual diffusion noise, but the underlying tissue structure is visible. The post-processing methods empirically improve all metrics with a change in average RMSE going from 0.223 to 0.042 and PSNR from 15.5 to 30.1 a substantial increase.

Calcification achieves the highest SSIM and PSNR after post-processing despite being the hardest tissue to predict, which we attribute to the effectiveness of histogram matching the sparse signal. Fibroglandular tissue has the lowest SSIM likely because it has the most complex boundary morphology and intermediate-valued transition regions.

4.5 Qualitative Results

Fig. 6 shows representative error maps for a test case, comparing the raw predicted C-DDPM outputs against the post-processed outputs. The predicted tissue maps show the overall morphology but contain noise, particularly in the calcification map, which appears as uniform noise with sparse bright dots that are not as noticeable from the background. After post-processing, the adipose and fibroglandular maps become significantly cleaner, with the interior regions approaching ground truth values. The remaining errors are concentrated along tissue boundaries—similar to what was observed in the U-Net baseline.

5 Discussion

This C-DDPM approach demonstrates that single-energy CT material decomposition is feasible with learned generative models, achieving an average RMSE of 0.042 and PSNR of 30.1 dB across all three tissue types after post-processing. These results empirically seem strong, particularly for adipose tissue (SSIM 0.907) and calcification detection (SSIM 0.971 after post-processing). However, generally post-processing data in this manner is not ideal and a sufficiently performant model would eliminate the need for this. Additionally, it can be visually seen in the error maps that the pre-processed predicted tissue maps actually have less error around the boundary compared to the post-processing which result in most error occurring near the tissue boundary.

Compared to the U-Net, the C-DDPM presents a different set of trade-offs. The baseline U-Net produces clearer outputs with a lower overall RMSE (0.018 vs. 0.042 post-processed). However, both the U-Net and post-processed images share the same issue where errors are concentrated at tissue boundaries, which are clinically the most important regions. The C-DDPM predicted tissue error map outputs show that the model captures the tissue structure details well, but the primary deficiency in the raw outputs is a residual diffusion haze rather than structural inaccuracy.

5.1 Limitations

The main constraints of this work are data scale, the use of synthetic data, and tissue boundary accuracy. The model was trained on only 1,000 slices from a single simulated breast phantom. The AAPM Challenge dataset is simulated rather than clinical, so performance on real acquisitions remains an open question. DDIM inference with 100 steps is also substantially slower than a single-pass U-Net, which could theoretically be a limiting factor considering time-sensitive clinical workflows.

5.2 Future Work

Several directions could improve upon these results. First, there are additional enhancements and approaches to DDPM that can be used such as using a different noise scheduler like cosine scheduling or incorporating more robust attention mechanisms. Other approaches like fine-tuning a pretrained diffusion model rather than training from scratch could address the data-scale limitation, as pretrained models have likely already learned key image

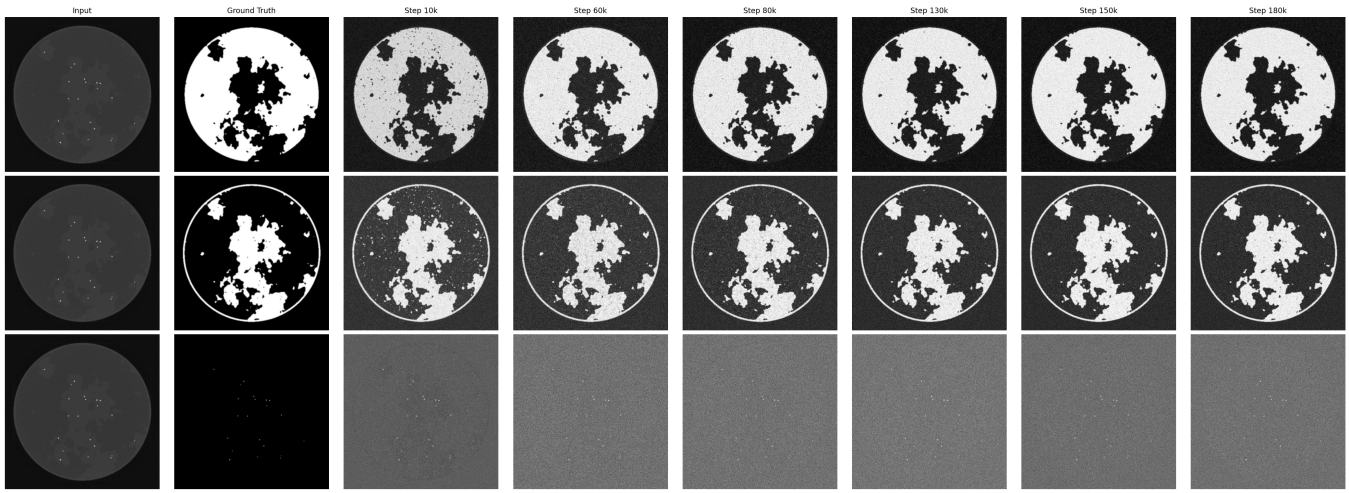

Fig. 5. Training progression for a test case across checkpoints (10k to 180k iterations). Each column shows the model output at a different training stage for adipose (top), fibroglandular (middle), and calcification (bottom), with the input CT image and ground truth in the leftmost columns.

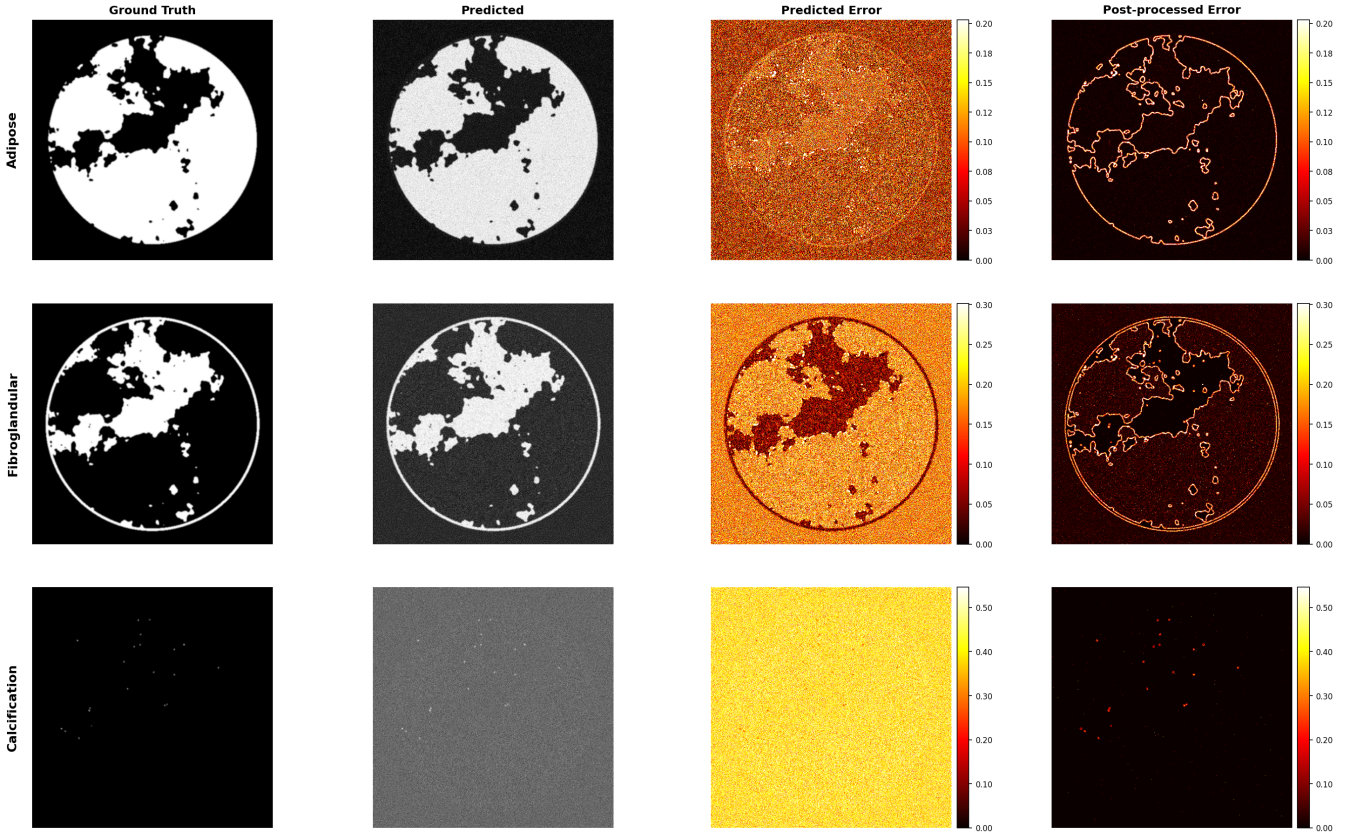

Fig. 6. Error map comparison for a test slice. Columns show ground truth, raw C-DDPM prediction, raw prediction error, and post-processed prediction error. Rows correspond to adipose, fibroglandular, and calcification.

semantic information that can be applicable to alternative image sets. Diffusion posterior sampling (DPS) offers an alternative approach where an unconditional diffusion prior is guided at inference time via a known forward model, which could eliminate the need for paired training data. Additionally, incorporating sinograms as the conditioning input could be useful as it would allow the model to generate the material decomposition maps while simultaneously learning the filtered back projection.

6 Conclusion

We presented a conditional DDPM framework for synthesizing tissue decomposition maps from single-energy CT images. Our approach generates three material decomposition maps (adipose, fibroglandular, calcification) with fair accuracy, particularly considering the tissue boundary. Conditional diffusion models show promise in estimating the tissue decomposition maps from single-energy CT images

and demonstrate the potential to approximate spectral CT information without dual-energy hardware, though residual noise and data availability remain a challenge - this can likely be improved by the use of a more robust clinical dataset and by leveraging alterative approaches and improvements.

REFERENCES

[1] K. Taguchi, I. Blevis, and K. Iniewski, Eds., *Spectral, Photon Counting Computed Tomography: Technology and Applications*. Boca Raton, FL: CRC Press, 2020.

[2] T. R. C. Johnson, "Dual-energy ct: General principles," *American Journal of Roentgenology*, vol. 199, no. 5 Supplement, pp. S3–S8, 2012.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[4] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022.

[5] R. E. Alvarez and A. Macovski, "Energy-selective reconstructions in x-ray computerized tomography," *Physics in Medicine & Biology*, vol. 21, no. 5, pp. 733–744, 1976.

[6] T. Lyu, W. Zhao, Y. Zhu, Z. Wu, Y. Zhang, Y. Chen, L. Luo, S. Li, and L. Xing, "Estimating dual-energy ct imaging from single-energy ct data with material decomposition convolutional neural network," *Medical Image Analysis*, vol. 70, p. 102001, 2021.

[7] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 8162–8171.

[8] Y. Gao, R. L. J. Qiu, H. Xie, C.-W. Chang, T. Wang, B. Ghavidel, J. Roper, J. Zhou, and X. Yang, "Ct-based synthetic contrast-enhanced dual-energy ct generation using conditional denoising diffusion probabilistic model," *Physics in Medicine & Biology*, vol. 69, no. 16, p. 165015, 2024.

[9] Y. Gao, H. Xie, C.-W. Chang, J. Peng, R. L. J. Qiu, T. Wang, J. Roper, B. Ghavidel, J. Zhou, and X. Yang, "Iodine map synthesis from non-contrast ct using diffusion model," in *Medical Imaging 2024: Physics of Medical Imaging*, ser. Proceedings of SPIE, vol. 12925, 2024, p. 129254P.

[10] X. Jiang, G. J. Gang, and J. W. Stayman, "Multi-material decomposition using spectral diffusion posterior sampling," *IEEE Transactions on Biomedical Engineering*, vol. 72, no. 8, pp. 2447–2461, 2025.

[11] E. Y. Sidky and X. Pan, "Report on the aapm deep-learning spectral ct grand challenge," *Medical Physics*, vol. 51, no. 2, pp. 772–785, 2024.