

LETTER

Artifact Removal Using Attention Guided Local-Global Dual-Stream Network for Sparse-View CT Reconstruction

Chang SUN[†], Yitong LIU^{†a)}, and Hongwen YANG[†], *Nonmembers*

SUMMARY Sparse-view CT reconstruction has gained significant attention due to the growing concerns about radiation safety. Although recent deep learning-based image domain reconstruction methods have achieved encouraging performance over iterative methods, effectively capturing intricate details and organ structures while suppressing noise remains challenging. This study presents a novel dual-stream encoder-decoder-based reconstruction network that combines global path reconstruction from the entire image with local path reconstruction from image patches. These two branches interact through an attention module, which enhances visual quality and preserves image details by learning correlations between image features and patch features. Visual and numerical results show that the proposed method has superior reconstruction capabilities to state-of-the-art 180-, 90-, and 45-view CT reconstruction methods.

key words: artifact removal, sparse-view CT, dual-stream, attention

1. Introduction

Computed Tomography (CT) is a widely used medical imaging modality for diagnosing various injuries, diseases, and radiation therapy. However, concerns about the potentially harmful effects of radiation and the need for radiation protection have become crucial issues in CT scans, as high radiation doses can increase the risk of cellular DNA damage and cancer later in life. To address these concerns, the International Commission on Radiologic Protection (ICRP) introduced the “As Low As Reasonably Achievable” (ALARA) principle in 1977, which aims to strike a balance between the risks and benefits of radiation used for diagnostics. One technique used to reduce CT radiation doses is sparse-view CT scanning, which involves acquiring fewer measurements of projection data. However, this approach often leads to high levels of noise and streak artifacts in the reconstructed images. Consequently, researchers have devoted their efforts to developing advanced image reconstruction methods to mitigate these challenges and enable the practical implementation of sparse-view CT scanners.

Recently, different kinds of deep learning (DL)-based sparse view CT reconstruction algorithms have been proposed, demonstrating superior reconstructed image quality over traditional iterative methods. These DL methods can be classified into four categories: post-processing in the image domain [1]–[4], pre-processing in the projection domain [5],

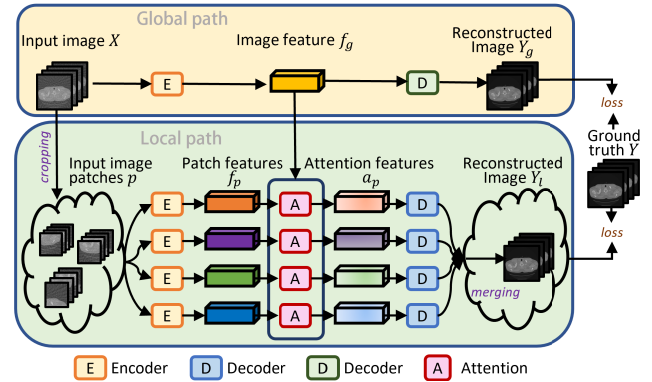


Fig. 1 Overview of the proposed network.

joint processing in the projection and image domains [6], and direct reconstruction in the projection domain [7]. In this study, we focus on artifact removal and structure refinement of sparse-view CT images in the image domain. The image-domain-based approach is not required for projection data, making it more advantageous in combination with commercial CT reconstruction software.

Image-domain DL-based reconstruction methods usually utilize convolutional neural networks (CNN) to improve initial reconstructions (often obtained from filtered back projection, FBP) by supervised learning. DD-Net [1], Improved GoogLeNet [2] and Framing U-Net [3] are popular end-to-end image-domain DL methods that improve image quality using extensive training datasets and advanced DL networks. Recently, influential ideas in deep learning, such as GAN [8], attention mechanisms [9] have been applied in sparse-view CT reconstruction, e.g., Lee et al. [4] designed a multi-level wavelet convolutional neural network combining a wavelet transform with modified U-Net, showing better streak artifact suppression capability than the standard U-Net. However, the previous methods did not consider the potential benefits of utilizing local details from image patches and structural information derived from the entire image to achieve a balance between noise reduction and the preservation of structural details. Therefore, this study aims to address this gap by proposing a solution that effectively combines artifact removal and structure refinement in sparse-view CT reconstruction.

This study presents an end-to-end dual-stream, sparse-view CT reconstruction network, as shown in Fig. 1. This network comprises two components: global path reconstruction and local path reconstruction. Both paths utilize the

Manuscript received August 5, 2023.

Manuscript revised February 28, 2024.

Manuscript publicized March 29, 2024.

[†]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876 China.

a) E-mail: liuyitong@bupt.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2023EDL8049

encoder-decoder architecture, consisting of an encoder and a decoder, to reconstruct the entire image and image patches, respectively. The shared encoder ensures that image features and patch features are embedded within the same feature context. However, different decoder facilities reconstruct image features and patch features at different levels of detail. Furthermore, the local path reconstruction incorporates an attention module designed to capture long-range dependencies among deep features of patches and the entire image. Unlike the attention modules used in FRD-Net [10], our attention module focuses on exploiting the self-similarity between patch-based and image-based features.

The proposed method has two advantages: Firstly, it learns from multiple spatial scale inputs, including image-scale and patch-scale, enabling the exploration of global and local features. Secondly, the model fully exploits long-range correlations between patch features and image features by the attention mechanisms, resulting in better performance on local details preservation and global structure reconstruction.

2. Proposed Method

2.1 Global Branch

The encoder extract the deep features of the image $f_g \in \mathbb{R}^{c \times h_g \times w_g}$ from the input image $X \in \mathbb{R}^{H \times W}$, and then the decoder predicts the denoised image Y_g from f_g . In this study, we designed two specific instantiations for the encoder (Table 1) and decoder (Table 2). It should be mentioned that each convolutional layer and transposed convolutional layer is followed by a BatchNormalization and a ReLu function. Besides, a sigmoid function is implied at the end of the decoder to scale the output into $[0, 1]$.

2.2 Local Branch

The input image is first divided into m un-overlapping patches, and then the encoder extracts the deep patch features $f_p \in \mathbb{R}^{c \times h_p \times w_p}$. After that, each patch feature f_p goes

Table 1 Our encoder with ResNet-34 [11] backbone. The input dimensions for an image and a patch are $1 \times 512 \times 512$ and $1 \times 256 \times 256$, respectively. Residual blocks are shown in brackets

Block	Layer	Output Size (image)	Output Size (patch)
$conv_1$	$7 \times 7, 64, s2, p3^a$	$64 \times 256 \times 256$	$64 \times 128 \times 128$
$pool_1$	$3 \times 3, s2, p1$	$64 \times 128 \times 128$	$64 \times 64 \times 64$
res_{l2}	$3 \times 3, 64$	$128 \times 64 \times 64$	$128 \times 32 \times 32$
	$3 \times 3, 64$		
res_{l3}	$3 \times 3, 256$	$256 \times 32 \times 32$	$256 \times 16 \times 16$
	$3 \times 3, 256$		
res_{l4}	$3 \times 3, 512$	$512 \times 16 \times 16$	$512 \times 8 \times 8$
	$3 \times 3, 512$		

^a2D kernels in 7×7 , stride=2, padding=3.

through an attention module to capture global contextual information from the image feature f_g .

As shown in Fig. 2, a multi-head self-attention mechanism [9] is adopted in the transformer block. The patch feature f_p is transformed into the query matrices Q . The image feature f_g is transformed into the key matrix K and the value matrix V . Then the scaled dot-product attention [9] $F_{HEAD}()$ is performed on the Q, K, V , followed by a linear transform $F_{LINEAR}()$. After that, the output feature is reshaped and concatenated with f_p and fed into the feed-forward layer $F_{FORWARD}()$. Finally, the weighted patch feature is calculated by adding the f_p and the output of $F_{FORWARD}()$. After going through n stacked transformer blocks, the decoder block reconstructs the patch features.

In this study, we divided the input image into $m = 4$ patches, set the number of transformer blocks $n = 8$, and the number of attention heads $N_h = 8$. The dimensions of queries, keys and values vectors are 64. The feed-forward block contains two convolutional layers with kernel size 1. Each convolutional layer is followed by batch normalization and the ReLu function.

2.3 Loss Function

Denote the ground-truth image, the output of the global branch and the output of the local branch as Y, Y_g and Y_p , respectively. The training loss is defined as the sum of the

Table 2 Network architecture of the decoder module.

Block	Layer	Output Size (image)	Output Size (patch)
$conv_1$	conv(2048,k3,s1,p1) ^a up(512,k3s2p1) ^b	$512 \times 32 \times 32$	$512 \times 16 \times 16$
$conv_2$	conv(1024,k3,s1,p1) up(256,k3s2p1)	$256 \times 64 \times 64$	$256 \times 32 \times 32$
$conv_3$	conv(512,k3,s1,p1) up(128,k3s2p1)	$128 \times 128 \times 128$	$128 \times 64 \times 64$
$conv_4$	conv(256,k3,s1,p1) up(64,k3s2p1)	$64 \times 256 \times 256$	$64 \times 128 \times 128$
$conv_5$	conv(256,k3,s1,p1) up(64,k3s2p1)	$64 \times 512 \times 512$	$64 \times 256 \times 256$
$conv_6$	conv(32,k3,s1,p1)	$32 \times 512 \times 512$	$32 \times 256 \times 256$
$conv_7$	conv(1,k3,s1,p0)	$1 \times 512 \times 512$	$1 \times 256 \times 256$

^a2D kernels in 3×3 , stride=1, padding=1.

^btransposed convolution operator(kernel size 3×3 , stride=2, padding=1).

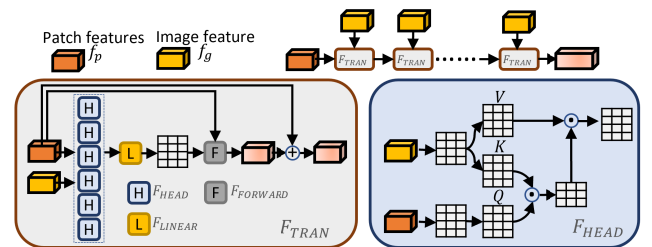


Fig. 2 Architecture of the attention module. The attention module is composed of a stack of n transformer blocks $F_{TRAN}()$. $F_{TRAN}()$ contains an attention layer $F_{ATTEN}()$ and a feed-forward block $F_{FORWARD}()$. $F_{ATTEN}()$ consists of N_h parallelly calculated attention heads $F_{HEAD}()$ and a linear block $F_{LINEAR}()$.

mean square error (MSE) between Y and Y_g , and the MSE between Y and Y_p .

$$L_{loss} = \alpha \|Y - Y_g\|_2^2 + \|Y - Y_p\|_2^2 \quad (1)$$

α is a tradeoff parameter.

3. Experiments and Results

3.1 Experiment Datasets

An open clinical dataset LungCT-Diagnosis [12], [13] is adopted in this study. 60 lung CT scans acquired between the years 2006 and 2009 are selected, including 4609 images with pixel size 512×512 . The CT scans are divided into three datasets: 43 scans for training (3198 images), 5 scans (408 images) for validation and 12 scans (1003 images) for testing.

To model degradation in a real CT scanner, the CT image pixel in terms of a Hounsfield Units (HU) value is first converted to an attenuation coefficient value of μ . After that, given a number of projection views, the Operator Discretization Library (ODL) [14] is used to generate tomographic projection data S in a commonly used fan beam geometry: the radius of the source and detector are 346 and 261, respectively. Then we simulated Poisson noise and Gaussian noise to the final intensities of the X-ray and recalculated S using the Beer-Lambert Law [15].

3.2 Experiment Setup

All the models are implemented by PyTorch and trained on a single GPU (NVIDIA GeForce RTX 2080 Ti) using the Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). We trained the models in 65 epochs (batch size = 2), starting with a learning rate of 1×10^{-4} and reducing it by half at the 50th epoch. We test the performance of the model per epoch on the validation dataset and select the best model with the lowest MSE.

PSNR and SSIM are used for quantitative evaluation. For calculation, we used a $[-1000, 1500]$ HU window for all images and normalized the CT values to $[0, 255]$.

3.3 Ablation Study of Training Loss

An ablation study was conducted to analyze the effect of α on the reconstruction performance. The results are summarized in Fig. 3 (a). It can be seen that $\alpha = 0.25$ has the worst PSNR on the validation dataset, $\alpha = 0.5$ has a slight improvement, and $\alpha = 1$ achieves the best reconstruction performance in terms of PSNR. Hence, $\alpha = 1$ is chosen for the rest of the experiments.

3.4 Ablation Study of Number of Transformer Blocks

We performed an ablation study to compare the performance of models with different numbers of transformer blocks n . In this experiment, we tested $n = 5, 6, 7, 8$ as shown in Fig. 3 (b).

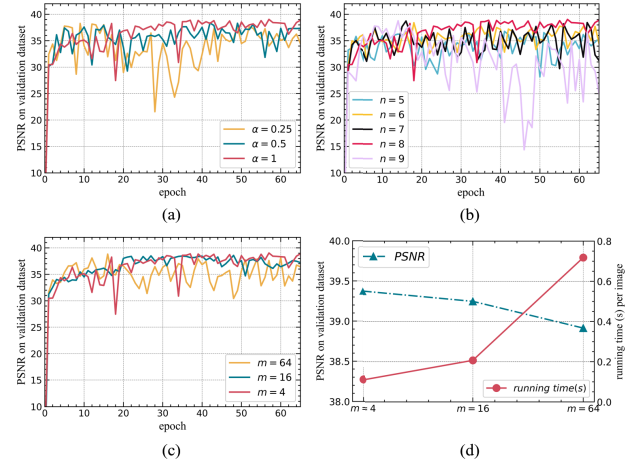


Fig. 3 Results of ablation study. (a) Comparison of PSNR of models trained with different α parameters on the validation dataset. (b) Comparison of PSNR of models trained with different number of transformer blocks n on the validation dataset. (c) Comparison of PSNR of models trained with different number of local patches m on the validation dataset. (d) Comparison of running time of models trained with different number of local patches m .

It can be noticed that the PSNR shows a slight improvement as n increases from 5 to 7, while $n = 8$ shows a relatively more considerable improvement compared to $n = 7$. However, as n increases to 9, the performance of the reconstructed model decreases severely, suggesting that a large number of transformer blocks may lead to overfitting the model to the training dataset. Hence, $n = 8$ is chosen for the rest of the experiments.

3.5 Ablation Study of Patch Size

In the local branch of the proposed network, we first divided the input image into m patches. We performed an ablation study to compare the performance of models with $m = 4$, $m = 16$ and $m = 32$. The results are summarized in Fig. 3 (c). It can be seen that the PSNR on the validation dataset of $m = 4$ and $m = 16$ are larger than that of $m = 32$. Besides, we compared the running time (s) per image of these three models. Specifically, we run each model ten times each (reconstructing one image each time) and calculate the running time for each run. Finally, the average of the ten running times is used as the metric. The results are shown in Fig. 3 (d). The running time for $m = 4$, $m = 16$ and $m = 32$ are 0.109s, 0.207s and 0.718s per image. Considering the computation load and the reconstruction image quality, $m = 4$ is chosen for the rest of the experiments.

3.6 Ablation Study of Attention Module

To understand the contribution of the attention module to the proposed network, we designed two additional baseline models, *Image_model* and *Patch_model* for evaluation. *Image_model* retains only the global branches of the proposed architecture, while *Patch_model* includes local branches

without attention modules. The numerical results summarized in Table 3 show that our model produces the highest PSNR and SSIM for all 180 views, 90 views, and 45 views-reconstruction, which validates the effectiveness of the attention module and dual-stream learning strategy. Fig-

Table 3 Comparison of average PSNR/SSIM in ablation study of attention module.

Method	180 views	90 views	45 views
FBP	26.2113/0.7598	19.7327/0.5917	15.0599/0.4597
<i>Image_model</i>	39.2138/0.9681	34.7704/0.9301	30.3258/0.8806
<i>Patch_model</i>	38.9136/0.9654	34.6407/0.9263	30.2567/0.8802
ours	39.3790/0.9681	34.9991/0.9314	30.4982/0.8902

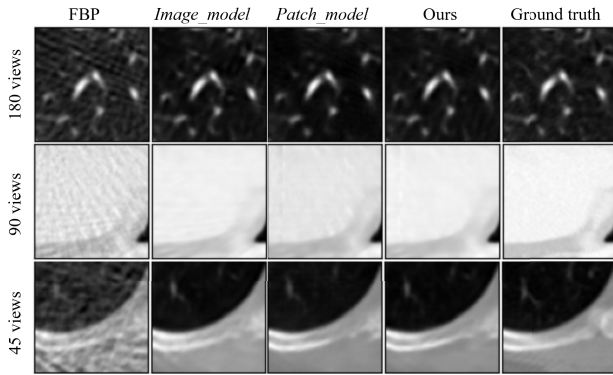


Fig. 4 Visual results of zoomed-images in the ablation study of attention modules.

Table 4 Comparison of average PSNR/RMSE between different reconstruction algorithms on the test dataset.

Method	180 views	90 views	45 views
FBP	26.2113/0.7598	19.7327/0.5917	15.0599/0.4597
Zhang [1]	37.8333/0.9428	33.1354/0.8922	28.9530/0.8177
Xie [2]	38.4415/0.9582	34.0091/0.8892	29.5440/0.8282
Han [3]	38.4478/0.9598	34.0644/0.8891	29.7970/0.8379
Lee [4]	38.9174/0.9636	34.1834/0.8902	30.0450/0.8414
ours	39.3790/0.9681	34.9991/0.9314	30.4982/0.8902

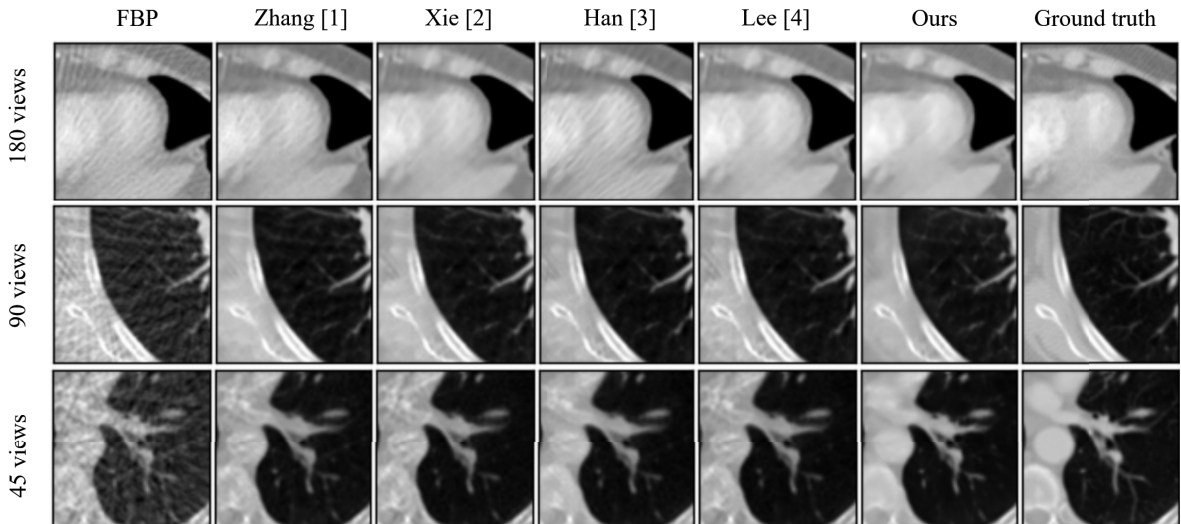


Fig. 5 Comparison of different reconstruction algorithms. Visual results of zoomed-in images.

ure 4 compares the de-artifacting performance of different methods visually, where our method better preserves delicate structures and removes streaking artifacts.

3.7 Comparison with Existing Methods

We compared the performance of the proposed method with the state-of-the-art image-domain-based DL methods Zhang [1], Xie [2], Han [3] and Lee [4]. Table 4 shows that our method performs best regarding PSNR and SSIM. We also compared the running speed of these methods. FBP, Zhang [1], Xie [2], Han [3], Lee [4] and the proposed method take about 0.038s, 0.045s, 0.042s, 0.142s, 0.077s and 0.109s to reconstruct a CT image on GPU. It shows that our model achieves comparable reconstruction speed. The zoomed visual results are shown in Fig. 5. It can be seen that the proposed method eliminates most of the artifacts and suppresses the noise compared to other state-of-the-art methods.

4. Conclusion

In this study, we propose an attention-guided local-global dual-stream encoder-decoder network to improve the image quality of sparse-view CT reconstruction. By exploiting the correlation between the patch features in the local path and the image features in the global path, the proposed method combines detailed information from the patches and the global structure of the whole image. The experimental results validate the effectiveness of the proposed model and show its potential to suppress streak artifacts and preserve detail structure, which is expected to provide practical guidance for improving the de-artifacting ability of image-domain methods in sparse-view CT reconstruction tasks.

Acknowledgments

C. Sun is supported by BUPT Excellent Ph.D. Students Foun-

dition (CX2022203).

References

- [1] Z. Zhang, X. Liang, X. Dong, Y. Xie, and G. Cao, "A sparse-view CT reconstruction method based on combination of DenseNet and deconvolution," *IEEE Trans. Med. Imag.*, vol.37, no.6, pp.1407–1417, 2018.
- [2] S. Xie, X. Zheng, Y. Chen, L. Xie, J. Liu, Y. Zhang, J. Yan, H. Zhu, and Y. Hu, "Artifact removal using improved GoogLeNet for sparse-view CT reconstruction," *Scientific reports*, vol.8, p.6700, 2018.
- [3] Y. Han and J.C. Ye, "Framing U-Net via deep convolutional framelets: Application to sparse-view CT," *IEEE Trans. Med. Imag.*, vol.37, no.6, pp.1418–1429, 2018.
- [4] M. Lee, H. Kim, and H.-J. Kim, "Sparse-view CT reconstruction based on multi-level wavelet convolution neural network," *Physica Medica*, vol.80, pp.352–362, 2020.
- [5] Z. Li, W. Zhang, L. Wang, A. Cai, N. Liang, B. Yan, and L. Li, "A sinogram inpainting method based on generative adversarial network for limited-angle computed tomography," *15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, vol.11072, SPIE, pp.345–349, 2019.
- [6] W. Wu, D. Hu, C. Niu, H. Yu, V. Vardhanabhuti, and G. Wang, "Drone: dual-domain residual-based optimization network for sparse-view ct reconstruction," *IEEE Trans. Med. Imag.*, vol.40, no.11, pp.3002–3014, 2021.
- [7] W. Wang, X.-G. Xia, C. He, Z. Ren, J. Lu, T. Wang, and B. Lei, "An end-to-end deep network for reconstructing ct images directly from sparse sinograms," *IEEE Transactions on Computational Imaging*, vol.6, pp.1548–1560, 2020.
- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol.63, no.11, pp.139–144, 2020.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol.30, 2017.
- [10] C. Du and Z. Qiao, "Epri sparse reconstruction method based on deep learning," *Magnetic Resonance Imaging*, vol.97, pp.24–30, 2023.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE conference on computer vision and pattern recognition*, pp.770–778, 2016.
- [12] O. Grove, A.E. Berglund, M.B. Schabath, H.J.W.L. Aerts, A. Dekker, H. Wang, E.R. Velazquez, P. Lambin, Y. Gu, Y. Balagurunathan, E. Eikman, R.A. Gatenby, S. Eschrich, and R.J. Gillies, "Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma," *PloS one*, vol.10, no.3, p.e0118261, 2015.
- [13] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol.26, pp.1045–1057, 2013.
- [14] J. Adler, H. Kohr, and O. Oktem, "Operator discretization library (odl)," *Zenodo*, 2017.
- [15] D.F. Swinehart, "The beer-lambert law," *Journal of chemical education*, vol.39, no.7, p.333, 1962.