Neural Modulation Fields for Conditional Cone Beam Neural Tomography

Samuele Papa [1 2] **David M. Knigge** [1 2] **Riccardo Valperga** [1] **Nikita Moriakov** [2] **Miltiadis Kofinas** [1]
Jan-Jakob Sonke [2] **Efstratios Gavves** [1]

Abstract

Conventional Computed Tomography (CT) methods require large numbers of noise-free projections for accurate density reconstructions, limiting their applicability to the more complex class of Cone Beam Geometry CT (CBCT) reconstruction. Recently, deep learning methods have been proposed to overcome these limitations, with methods based on neural fields (NF) showing strong performance, by approximating the reconstructed density through a continuous-in-space coordinate based neural network. Our focus is on improving such methods, however, unlike previous work, which requires training an NF from scratch for each new set of projections, we instead propose to leverage anatomical consistencies over different scans by training a single *conditional* NF on a dataset of projections. We propose a novel conditioning method where *local* modulations are modelled per patient as a field over the input domain through a Neural Modulation Field (NMF). The resulting Conditional Cone Beam Neural Tomography (CondCBNT) shows improved performance for both high and low numbers of available projections on noise-free and noisy data.

1. Introduction

In inverse problems, the goal is to infer a certain quantity of interest from indirect observations. They arise in many scientific fields, medical imaging (Louis, 1992), biology (Karwowski, 2009; Sridharan et al., 2022), and physics (Romanov, 2018; Collaboration, 2019). Unfortunately, many inverse problems are inherently *ill-posed*, i.e., there exist multiple solutions that agree with the measurements and these do not necessarily depend continuously on the data (Kabanikhin, 2008). These issues warrant further study, and tools from machine learning and deep learning in particular have attracted a lot of attention recently.

In this work, we focus on Computed Tomography (CT) (Oldendorf, 1978), a medical imaging technique for reconstructing material density[1] inside a patient, using the mathematical and physical properties of X-ray scanners. In CT, several X-ray scans–or *projections*–of the patient are acquired from various angles using a *detector*. An important variant of CT is Cone Beam CT (CBCT), which uses flat panel detectors to scan a large fraction of the volume in a single rotation. Unfortunately, CBCT reconstruction is harder in comparison to classical (helical) CT. This is caused by the inherent mathematical difficulty of Radon Transform inversion in the three-dimensional setting (Tuy, 1983), physical limits of the detector, and characteristics of the measurement process such as noise. Traditional reconstruction methods include FDK (Feldkamp et al., 1984), and iterative reconstruction (Kaipio & Somersalo, 2005). FDK filters the projections and applies other simple corrections to properly account for the physical geometry of the acquisition system. Iterative methods use optimization to find the density that most closely resemble the measurements once projected using a forward operator. In addition, deep learning has seen increasing use in the field, with algorithms such as learned primal-dual (Adler & Öktem, 2018), invertible learned primal-dual (Rudzusika et al., 2021) and LIRE (Moriakov et al., 2022).

Recently, reconstruction methods that employ Neural Fields (NFs) have been proposed. *NFs are a class of neural architectures that parameterize a field $f : \mathbb{R}^d \to \mathbb{R}^n$, i.e. a quantity defined over spatial and/or temporal coordinates, using a neural network f_θ* (see Xie et al. (2022) for a survey on NFs). In CT reconstruction, these architectures have been used to approximate the density directly over the volume space $\mathbb{R}^3$ (Zang et al., 2021; Zha et al., 2022; Lin et al., 2023). Zha et al. (2022) proposed Neural Attenuation Fields (NAF), an approach to supervise NFs using only the

[*] Equal contribution [1] Institute for Informatics, University of Amsterdam, Amsterdam, the Netherlands [2] Netherlands Cancer Institute, Amsterdam, the Netherlands. Correspondence to: Samuele Papa <s.papa@uva.nl>, David M. Knigge <d.m.knigge@uva.nl>.

Accepted after peer-review at the 1st workshop on Synergy of Scientific and Machine Learning Modeling, SynS & ML ICML, Honolulu, Hawaii, USA. July, 2023. Copyright 2023 by the author(s).

[1] To be precise, we try to find the *attenuation coefficients*, but we may use density interchangeably, as they are strongly related under assumptions that hold in our setting.

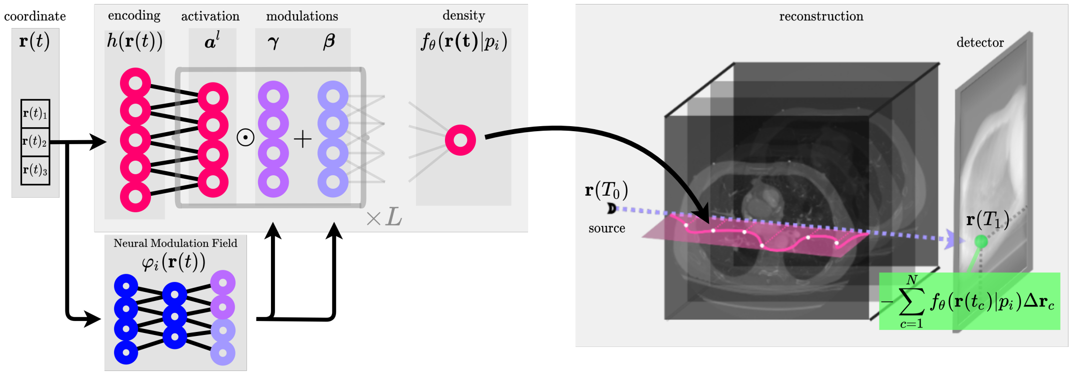

Figure 1. We propose *Conditional Cone Beam Neural Tomography* (CondCBNT), a framework for reconstructing Cone Beam Computed Tomography volumes using neural fields. An integral is taken over values sampled from a neural field f_θ at coordinates $\mathbf{r}(t)$ along a ray cast from source to sensor. The coordinates are encoded into a multiresolution hash-encoding $h(\mathbf{r}(t))$ (Müller et al., 2022), and passed through L linear layers. To leverage consistencies over anatomies of different patients, we propose to model the density for a specific patient p_i using a shared neural field f_θ, whose activations $\boldsymbol{a}^l$ are modulated by a patient-specific *Neural Modulation Field* (NMF) φ_i. This conditioning function learns a field of $\boldsymbol{\gamma}, \boldsymbol{\beta}$ FiLM modulations (Dumoulin et al., 2018) over the input space $\mathbb{R}^3$ for a patient p_i. The integral $-\sum_{c=1}^{N} f_\theta(\mathbf{r}(t_c))\Delta\mathbf{r}_c$ is supervised at the sensor using the corresponding observed projection value.

measured attenuated photon counts at the detector. Despite showing promising results, this method requires training a NF from scratch for each volume, prohibiting transfer of learned features across volumes through weight sharing. Instead, Lin et al. (2023) propose encoding a set of projections into a latent space shared over all training volumes, and decoding this into a density modelled as a NF. However, encoding of all available projections is only feasible when a small number of them is used, as it would otherwise result in prohibitive compute and memory requirements.

In this work, we instead aim to remove the need for an explicit decoder. We leverage the work of Park et al. (2019), who propose to learn latent codes for a dataset of 3D shapes using *auto-decoding*, where randomly initialized latent codes are optimized during training. Dupont et al. (2022) expand on by using these learned latent codes as modulations for a shared NF. Bauer et al. (2023) show that the use of a single global code per signal limits reconstruction quality, and instead use a spatially structured grid of codes. Their approach greatly increases reconstruction quality, but requires interpolating a grid of modulations, increasing computational requirements for signals over higher-dimensional domains. We introduce the **Neural Modulation Field** (NMF) which models a continuous field of modulations over the signal domain. We propose the **Conditional Cone Beam Neural Tomography** (CondCBNT) framework, which incorporates this *local conditioning function* to speed up reconstruction, while still processing all available projections, relieving restrictions on projection counts used in the reconstruction process. In doing so, we show considerable improvements in scenarios with both sufficient or limited projections, as well as in the presence of both noisy and noise-free data.

2. Method

Beer-Lambert's law relates the attenuation of electromagnetic radiation such as visible light or X-rays to the properties of the material it is traveling through (Swinehart, 1962). Let $\mathbf{r} : [T_0, T_1] \longrightarrow \mathbb{R}^3$ be the straight path taken by radiation through the medium. The radiation intensity $I(\mathbf{r}(T_1))$ at position $\mathbf{r}(T_1)$ is the line integral

$$I(\mathbf{r}(T_1)) = I_0 \exp\left[-\int_{T_0}^{T_1} \mu(\mathbf{r}(t)) \left|\mathbf{r}'(t)\right| dt\right], \quad (1)$$

where $\mu : \mathbb{R}^3 \longrightarrow \mathbb{R}^+$ is the attenuation coefficient of the medium and $I_0 = I(\mathbf{r}(T_0))$ is the initial intensity. The integral in (1) can be approximated by the sum

$$I(\mathbf{r}(T_1)) \approx I_0 \exp\left[-\sum_{c=1}^{N} \mu(\mathbf{r}(t_c)) \left|\mathbf{r}'(t_c)\right| \Delta t\right], \quad (2)$$

where $t_c \in [T_0, T_1]$ and $\left|\mathbf{r}'(t_c)\right| \Delta t = \Delta\mathbf{r}_c = |\mathbf{r}(t_{c+1}) - \mathbf{r}(t_c)|$. Given a set of 2D CBCT projections $v_\alpha \in \mathbb{R}^{H \times W}$ with H, W the height and width of the sensor and α the angle under which the projection was taken, we are trying to estimate density values along rays cast from source to sensor. Each ray is the straight path $\mathbf{r}$ which connects the source to pixels in the detector. For simplicity, we bound the patient volume with a box and assume zero attenuation outside the box. Therefore, for every path, we compute the sum in (2) with only those $\mathbf{r}(t_c)$ that are contained in the bounding box. By taking the logarithm we can avoid the computationally tedious exponential and use $\log I(\mathbf{r}(T_1)) \approx -\sum_{c=1}^{N} \mu(\mathbf{r}(t_c))\Delta\mathbf{r}_c + \log I_0$ and discard the constant that depends on the initial intensity, which we assume is the same for all projections. We use a neural field

$f_\theta : \mathbb{R}^3 \longrightarrow \mathbb{R}^+$ to approximate the density μ such that the intensity $I(\mathbf{r}(T_1))$ coincides with the intensity recorded by the detector at the position $\mathbf{r}(T_1)$:

$$\log I(\mathbf{r}(T_1)) \approx -\sum_{c=1}^{N} f_\theta(\mathbf{r}(t_c))\Delta \mathbf{r}_c. \quad (3)$$

Coordinate embedding. Tancik et al. (2020) showed that ReLU MLPs suffer from spectral bias, limiting their capacity to model high frequency functions on low-dimensional domains. As a solution, they note that it is possible to embed coordinates $\mathbf{r}(t_c) \in \mathbb{R}^3$ into a higher-dimensional space $\mathbb{R}^e$ with $e \gg 3$ before passing them through the MLP. We choose to follow Müller et al. (2022) and use the *multiresolution hash-encoding*, denoted $h(\mathbf{r}(t_i))$, as it empirically shows fastest convergence in our experiments. See Appx. A for a full description of this embedding.

Conditioning with Neural Modulation Fields. Conditioning in neural fields consists of modulating the weights θ or activations $\boldsymbol{a}$ of a NF f_θ with a conditioning variable $\boldsymbol{z}$ to vary the NF's output (Xie et al., 2022), a method often used to encode different samples x_i from a single dataset X through a set of latents $\{\boldsymbol{z}_i | x_i \in X\}$. Intuitively, in the setting of CT reconstruction, we could fairly assume the densities for patients $p_i \in P$ share a lot of anatomical structure. A conditional NF that is tasked with reconstructing a dataset of multiple volumes would be able to leverage this consistency in anatomical information in its reconstruction (e.g. inferring from noisy or missing data), with patient-specific characteristics being refined with the conditioning variable $\boldsymbol{z}_i$. To this end, we could in principle use the aforementioned auto-decoding approach with a *global* conditioning latent $\boldsymbol{z}_i$. However, global conditioning has been shown to result in reconstructions with limited detail (Dupont et al., 2022; Bauer et al., 2023). This limitation is significant because patient-specific fine-grained details in scans contain information crucial for medical purposes. We instead opt for *local* conditioning, where the conditioning variable $\boldsymbol{z}_i$ depends on the input coordinate $\boldsymbol{r}(t)$. In previous works, this is done through interpolation of a trainable discrete data structure, e.g. a grid of latent codes (Shaham et al., 2021; Yu et al., 2021; Bauer et al., 2023). Instead, to further increase expressivity of the resulting modulation and forego modelling choices such as code grid resolution and interpolation method, we propose to abstract the learning of modulations away from a discrete data structure and model the modulations themselves as a continuous field through a patient-specific *Neural Modulation Field* (NMF) we denote φ_i. During training, parameters θ_i of the patient-specific NMFs φ_{θ_i} are optimized alongside the weights of the shared NF f_θ, during inference - for a novel set of projections - only the parameters for θ_i are optimized.

For the activation modulation, we use feature-wise linear modulations (FiLM) (Dumoulin et al., 2018), such that activations $\boldsymbol{a}^l$ at a layer l with weights $\boldsymbol{W}^l$ and bias $\boldsymbol{b}^l$ are transformed with patient-specific *local* scaling and shifting modulations $\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i$, as follows:

$$\boldsymbol{a}_i^l = \mathrm{ReLU}((\boldsymbol{W}^l \boldsymbol{a}_i^{l-1} + \boldsymbol{b}^l) \odot \boldsymbol{\gamma}_i + \boldsymbol{\beta}_i), \quad (4)$$

where $\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i$ are obtained from the NMF $\varphi_{\theta_i} : \mathbb{R}^3 \to \mathbb{R}^{\dim(\boldsymbol{\gamma})+\dim(\boldsymbol{\beta})}$. For specific architectural choices of the NMF and shared NF, see Appx. C. We term the resulting model *Conditional Cone Beam Neural Tomography* (CondCBNT). See Fig. 1 for an overview of the framework.

Dataset. The dataset used is derived from the LIDC-IDRI (Armato III et al., 2015). This is a collection of diagnostic lung cancer screening thoracic CT scans. A random selection of 250 cases was chosen and the CT scan resampled to 2mm resolution. Then, each volume is projected using 256×256 pixel, 2mm resolution detectors. Angles equally spaced between $0°$ and $205°$ are used. 400 projections are created, first without any noise, then with Poisson noise, used to simulate measurement noise with 5×10^5 photons. A subset of 50 equally-spaced projections is obtained from both. The 250 volumes are split into $200/25/25$ for training, validation, and testing. The resulting dataset will be made publicly available upon acceptance.

Metrics. For quantitve evaluation we rely on the *Peak Signal to Noise Ratio* (**PSNR**), a classical measure of signal quality, and the *Structural Similarity Index Measure* (**SSIM**), which captures the perceptive similarity between two images by analyzing small local chunks (Wang et al., 2004). Historically, both metrics have been defined for images, but we compute them over full volumes. Finally, we track the GPU memory used and the time required to reconstruct a volume.

Baselines. FDK reconstruction (Feldkamp et al., 1984) was performed using Operator Discretization Library (Adler et al., 2017). As an iterative reconstruction baseline, we implemented Landweber iteration with Total Variation regularization (Kaipio & Somersalo, 2005), where parameters such as step size, iteration count and the amount of regularization were chosen via grid search on the validation set. As a deep learning reconstruction baseline, we use LIRE-32(L) architecture from Moriakov et al. (2022), which is a dedicated lightweight, memory-efficient variant of learned primal-dual method from Adler & Öktem (2018) for CBCT reconstruction. From the NF class of models, we compare with Zha et al. (2022); we do not compare with Lin et al. (2023) due to their prohibitive computational costs.

Table 1. Mean ± standard deviation of metrics over test set for FDK (Feldkamp et al., 1984), Iterative (Kaipio & Somersalo, 2005), LIRE-L (Moriakov et al., 2022), NAF (Zha et al., 2022), and CondCBNT (ours). LIRE-L slightly outperforms CondCBNT but requires more GPU memory. Our method excels with less memory and comparable runtime.

		Noisy			Noise-free			
P.	Method	PSNR (↑)	SSIM (↑)	Time (s/vol)	PSNR (↑)	SSIM (↑)	Time (s/vol)	Mem. (MiB)
50	FDK	14.54 ± 2.90	$.20 \pm .07$	0.8	16.09 ± 3.22	$.43 \pm .09$	0.8	100
	Iterative	26.36 ± 2.11	$.70 \pm .08$	7.7	27.13 ± 2.80	$.71 \pm .08$	30.8	300
	LIRE-L	29.48 ± 2.07	$.83 \pm .05$	3.9	-	-	-	2.1k
	NAF	22.83 ± 2.24	$.58 \pm .10$	161	24.26 ± 2.52	$.72 \pm .08$	582	18
	CondCBNT	28.31 ± 1.22	$.80 \pm .05$	124	30.21 ± 1.42	$.86 \pm .05$	647	96
400	FDK	16.43 ± 3.38	$.45 \pm .12$	7	16.71 ± 3.47	$.65 \pm .09$	7	100
	Iterative	28.38 ± 3.27	$.78 \pm .11$	87.4	31.40 ± 6.22	$.91 \pm .07$	174	600
	LIRE-L	30.70 ± 2.25	$.88 \pm .05$	12.8	-	-	-	4k
	NAF	25.93 ± 2.45	$.75 \pm .08$	275	25.04 ± 2.91	$.77 \pm .08$	580	205
	CondCBNT	29.89 ± 1.39	$.86 \pm .05$	763	30.63 ± 1.43	$.88 \pm .04$	595	96

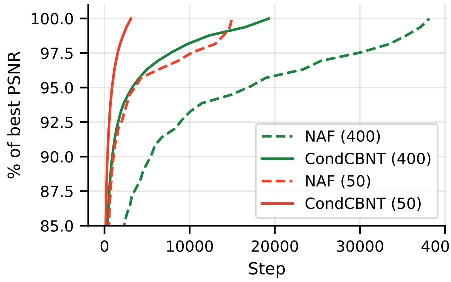

Figure 2. Using noisy projections, the percentage of the best PSNR (↑) that a model can reach over the number of steps required to achieve it. CondCBNT converges significantly faster.

3. Experiments

Hyperparameter search for NAF, CondCBNT, and the Iterative method was carried out on the validation set. With noisy projections, early stopping was used to avoid overfitting the noise. Instead, with noise-free projections, we decided to stop after about 10 minutes of training. Although more time would have improved performance further, it would not have provided any additional insights. It is worth noting that individual volume optimization was not conducted to reflect the constraints of a realistic scenario.

During training, we followed Lin et al. (2023) and directly supervised the neural field with density values, as we observed this greatly improved stability. During inference on validation and test sets, we kept the shared NF fixed and only optimized the randomly initialized NMF weights for each unseen scan (see Appx. C). We first evaluated the model on the test set using 50 and 400 noise-free projections respectively, results shown in Tab. 1 right. CondCBNT greatly improves reconstruction quality both in terms of PSNR and SSIM, compared to classical methods and NAF. Next, we validated the model on 50 and 400 noisy projections, results for which are shown in Tab. 1 left.

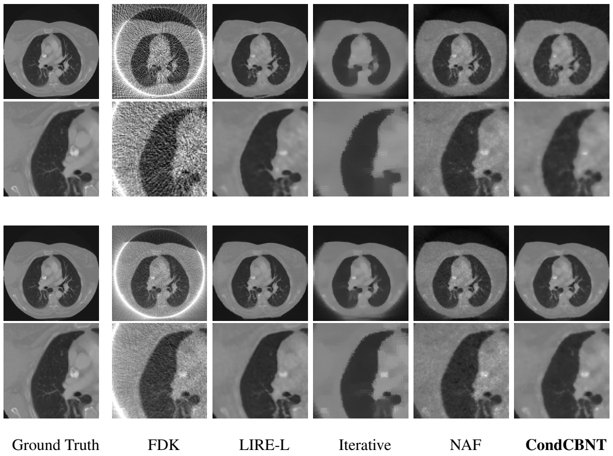

Figure 3. Ground truth and reconstructions using all the methods applied to noisy projections. Top 50, bottom 400 projections. Grayscale with density in $[0 - 0.04]$. Our method does not overfit the noise and maintains tissue contrast. High-res in Appx. D.

Again, we see considerable improvements in our method over all baseline approaches. LIRE-L is the exception, achieving a performance slightly better than CondCBNT with significantly faster reconstruction speed at the cost of an increased memory footprint.

Qualitative assessment in the noisy case is possible from Fig. 3, where it is evident that NAF overfits the noise. The iterative method over-smooths the reconstruction and exhibits blocky artifacts. The FDK reconstruction suffers from artifacts caused by the detector size, noise, and the low number of projections. LIRE-L and CondCBNT both reconstruct the volume with better soft-tissue contrast and without overfitting the noise.

Comparing convergence speed from Tab. 1 is hard because of diverging implementation choices and final performance reached. Therefore, we normalized performance by maximum PSNR reached after optimization. Additionally, given that dataset and batch size were the same, we decided to compare using the number of iterations instead of wall-clock time (Fig 2). This shows how CondCBNT quickly reaches a satisfying performance with both noisy and noise-free projections. Especially interesting is that, in the 400 projection case, CondCBNT was optimized for only half of a full epoch and still managed to outperform NAF and be within 1 standard deviation of LIRE-L. Since our method does not require training the whole model from scratch for a newly obtained set of projections, the model converges considerably faster.

4. Conclusion

We improve noise resistance of neural field (NF)-based CBCT reconstruction methods by sharing a conditional

NF over scans taken from different patients. We propose learning a continuous, local conditioning function expressed through a sample-specific *Neural Modulation Field* which modulates activations in the conditional NF to express volume-specific details. *Conditional Cone-Beam Neural Tomography* (CondCBNT) represents an efficient improvement over previous approaches, in terms of GPU memory scalability and reconstruction quality on both noise-free and noisy data and with varying numbers of available projections.

Broader impact

This work yields significant implications for the Medical Imaging field. By utilizing our method, it becomes possible to diminish radiation exposure and scan duration, thereby increasing the number of patients who can access treatment. Additionally, the superior quality of the reconstruction obtained opens up avenues for enhanced performance in subsequent tasks. This benefit extends to various imaging modalities like MRI or PET, not solely limited to CT scans.

While the proposed method exhibits minimal susceptibility to this concern, it is essential to acknowledge that complete interpretability remains a challenge for all deep learning models. Consequently, accurately identifying artifacts when they arise could prove difficult.

References

Adler, J., Kohr, H., and Öktem, O. Odl 0.6.0, April 2017. URL https://doi.org/10.5281/zenodo.556409.

Adler, J. and Öktem, O. Learned Primal-dual Reconstruction. 37(6):1322–1332, 2018. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2018.2799231. URL http://arxiv.org/abs/1707.06474.

Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E. A., MacMahon, H., Van Beek, E. J. R., Yankelevitz, D., Biancardi, A. M., Bland, P. H., Brown, M. S., Engelmann, R. M., Laderach, G. E., Max, D., Pais, R. C., Qing, D. P. Y., Roberts, R. Y., Smith, A. R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G. W., Jude, C. M., Munden, R. F., Petkovska, I., Quint, L. E., Schwartz, L. H., Sundaram, B., Dodd, L. E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, A. V., Gupte, S., Sallam, M., Heath, M. D., Kuhn, M. H., Dharaiya, E., Burns, R., Fryd, D. S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B. Y., and Clarke, L. P. Data from LIDC-IDRI. https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX, 2015.

Bauer, M., Dupont, E., Brock, A., Rosenbaum, D., Schwarz, J., and Kim, H. Spatial functa: Scaling functa to imagenet classification and generation. *arXiv preprint arXiv:2302.03130*, 2023.

Collaboration, T. E. H. T. First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole, 2019. URL http://arxiv.org/abs/1906.11238.

Dumoulin, V., Perez, E., Schucher, N., Strub, F., Vries, H. d., Courville, A., and Bengio, Y. Feature-wise transformations. *Distill*, 3(7):e11, 2018.

Dupont, E., Kim, H., Eslami, S., Rezende, D., and Rosenbaum, D. From data to functa: Your data point is a function and you can treat it like one. *arXiv preprint arXiv:2201.12204*, 2022.

Feldkamp, L. A., Davis, L. C., and Kress, J. W. Practical cone-beam algorithm. *Josa a*, 1(6):612–619, 1984.

Kabanikhin, S. I. Definitions and examples of inverse and ill-posed problems. 2008.

Kaipio, J. and Somersalo, E. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005. ISBN 0-387-22073-9. doi: 10.1007/b138659. URL http://link.springer.com/10.1007/b138659.

Karwowski, J. Inverse problems in quantum chemistry. 109(11):2456–2463, 2009. ISSN 1097-461X. doi: 10.1002/qua.22048. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.22048.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Lin, Y., Luo, Z., Zhao, W., and Li, X. Learning deep intensity field for extremely sparse-view cbct reconstruction. *arXiv preprint arXiv:2303.06681*, 2023.

Louis, A. K. Medical imaging: State of the art and future development. 8(5):709, 1992. ISSN 0266-5611. doi: 10.1088/0266-5611/8/5/003. URL https://dx.doi.org/10.1088/0266-5611/8/5/003.

Moriakov, N., Sonke, J.-J., and Teuwen, J. Lire: Learned invertible reconstruction for cone beam ct, 2022.

Müller, T., Evans, A., Schied, C., and Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

Oldendorf, W. H. The quest for an image of brain. 28(6):517–517, 1978. ISSN 0028-3878. doi: 10.1212/WNL.28.6.517. URL https://n.neurology.org/content/28/6/517.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Romanov, V. G. Inverse Problems of Mathematical Physics. In *Inverse Problems of Mathematical Physics*. De Gruyter, 2018. ISBN 978-3-11-092601-9. doi: 10.1515/9783110926019. URL https://www.degruyter.com/document/doi/10.1515/9783110926019/html.

Rudzusika, J., Bajic, B., Öktem, O., Schönlieb, C.-B., and Etmann, C. Invertible learned primal-dual. In *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*, 2021. URL https://openreview.net/forum?id=DhgpsRWHl4Z.

Shaham, T. R., Gharbi, M., Zhang, R., Shechtman, E., and Michaeli, T. Spatially-adaptive pixelwise networks for fast image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14882–14891, 2021.

Sridharan, B., Goel, M., and Priyakumar, U. D. Modern machine learning for tackling inverse problems in chemistry: Molecular design to realization. 58(35):5316–5331, 2022. ISSN 1364-548X. doi: 10.1039/D1CC07035E. URL https://pubs.rsc.org/en/content/articlelanding/2022/cc/d1cc07035e.

Swinehart, D. F. The Beer-Lambert Law. 39(7):333, 1962. ISSN 0021-9584. doi: 10.1021/ed039p333. URL https://doi.org/10.1021/ed039p333.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

Tuy, H. K. An inversion formula for cone-beam reconstruction. *SIAM Journal on Applied Mathematics*, 43(3):546–552, 1983. ISSN 00361399. URL http://www.jstor.org/stable/2101324.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., and Sridhar, S. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, volume 41, pp. 641–676. Wiley Online Library, 2022.

Yu, A., Ye, V., Tancik, M., and Kanazawa, A. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4578–4587, 2021.

Zang, G., Idoughi, R., Li, R., Wonka, P., and Heidrich, W. Intratomo: Self-supervised learning-based tomography via sinogram synthesis and prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1940–1950, 2021. doi: 10.1109/ICCV48922.2021.00197.

Zha, R., Zhang, Y., and Li, H. Naf: Neural attenuation fields for sparse-view cbct reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pp. 442–452. Springer, 2022.

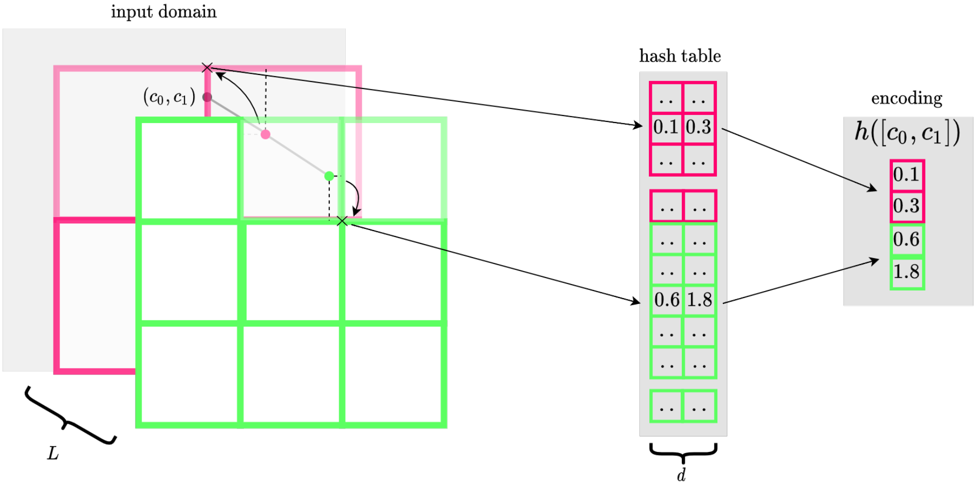

Figure 4. Multi-resolution hash encoding. L grids of multiple resolutions (green, red) are defined over the input domain. Each grid-point corresponds to an entry in a hash table. Each entry in this hash table consists of a $d-$dim freely trainable weight vector. To encode a coordinate (c_0, c_1), it is mapped to its closest grid-points on every grid (in practice a linear interpolation of the n nearest grid points is taken), and the encoding for this coordinate is given by concatenating the grid points corresponding feature vectors, to end up with a $(L \cdot d)$- dimensional embedding.

A. Multiresolution Hash Encoding

Müller et al. (2022) propose a multi-resolution hash encoding as coordinate embedding for neural fields. Here, we briefly describe the method in more detail.

Multi-resolution hash encoding is a parametric embedding, meaning the embedding function itself contains additional trainable parameters. In multi-resolution hash encoding this is done through assigning freely trainable weights to grid points from a set of multi-resolution grids defined over the input space. These parameters are then looked up and interpolated for a specific input coordinate $\mathbf{x}$. Formally, the embedding consists of a number of levels L, which correspond to the multiple grid resolutions, a feature dimensionality d denoting the dimensionality of each trainable vector attached at a grid point, a base resolution denoting the number of grid points for the lowest resolution grid, a per-level resolution increase factor r and a maximum hash-table size.

To encode a coordinate (c_0, c_1), it is mapped to its closest grid-points on every grid (in practice a linear interpolation of the n nearest grid points is taken), and the encoding for this coordinate is given by concatenating the grid points corresponding feature vectors, to end up with a $(L \cdot d)$- dimensional embedding. See figure 4 for an illustration.

B. Metrics

The metrics used to quantitatively evaluate the reconstruction quality are the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). Both metrics can be adapted in the 3D setting in a straightforward way. Given two volumes $x, y \in \mathbb{R}^{H \times W \times D}$ where H, W, and D are the height, width, and depth of the volume respectively, y is the ground truth and x is the reconstruction, the PSNR is the following

$$\begin{aligned} \text{PSNR}(x, y) &= 10 \cdot \log_{10} \frac{(\max y)^2}{\text{MSE}(x, y)} && (5) \\ &= 20 \cdot \log_{10} \max y - 10 \cdot \log_{10} \text{MSE}(x, y), && (6) \end{aligned}$$

where the second step improves numerically stability and the MSE is the voxel-wise Mean Squared Error:

$$\text{MSE}(x, y) = \frac{1}{NML} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \sum_{k=0}^{L-1} (x(i, j, k) - y(i, j, k))^2. \quad (7)$$

The SSIM is computed over a small $K \times K \times K$ cube within the volume. This is repeated for all pixels, padding when necessary with zeros. Here we show the formula for the entire volume, although the original definition is for a single region:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)},$$

where μ indicates the mean, σ the covariance, $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$ with $k_1 = 0.01$, $k_2 = 0.03$, and L the difference between the maximum value and minimum value in y.

C. Experimental details

When training NAF and CondCBNT, rays are sampled at random to form a batch. Then, a number of samples are selected along the ray to form the inputs of the model. While in NAF the batch is created using rays sampled at random from a single projection, for ConCBNT we sampled rays from any projection.

Projection noise was added using the Poisson distribution, to simulate the effect of measurement noise. This is also called *shot noise*, and it happens in all devices which measures the amount of photons that hit them. The probability of detecting photons can be modelled using a Poisson distribution. Intuitively, a thicker and denser substance in the path of the ray will result in a lower probability of detection and more noise in the projection. To be specific, assuming a projected value of p and a fixed photon count π (set at 5×10^5 in our experiments), the Poisson distribution's rate is defined as $\lambda = \pi e^{-p}$. Thus, the probability of detecting a specific number of photons, q, can be expressed as:

$$P(q; \lambda) = \frac{e^{-\lambda}\lambda^q}{q!} = \frac{(\pi e^{-p})^q e^{-\pi e^{-p}}}{q!} \tag{8}$$

By sampling a value q from this distribution, the resulting projected value is then calculated as:

$$\tilde{p} = -\log\left(\frac{q}{\pi}\right) \tag{9}$$

C.1. Architectural details

Here, we describe the architectural specifications for the shared neural field and the patient-specific modulation neural fields.

The shared neural field f_θ consists of a multi-resolution hash encoding, as described in A, with 16 levels of feature dimensionality 2, a base resolution of $16 \times 16 \times 16$, a per-level resolution increase factor of 2, and a hash-table with maximum size of 2^{19} parameters per level. This results in a 32-dimensional embedding, which is passed through 2 linear layers with hidden size 128, each followed by patient-specific FiLM modulation - as described in Sec. 2 - and ReLU activations. Each modulation neural field φ_{θ_i} also uses multi-resolution hash encoding to embed the input coordinate, followed by 2 linear layers of hidden dimensionality 128 with ReLU activations, which outputs into a $2 \cdot 128$ dimensional code $\boldsymbol{z}$ split into $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{128}$.

C.2. Hyperparameters

In this section, we describe the hyperparameters used in the experiments. for all experiments, the code was implemented in PyTorch (Paszke et al., 2019), optimized using Adam (Kingma & Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$.

CondCBNT was trained for 15 hours on an A100 GPU using all 200 volumes from the training set. The learning rate used for the MNF was 10^{-4} while 10^{-3} for the shared NF. During training the batch size was $16,384$. During validation and testing, the MNFs are optimized individually for each patient, with a batch size of 1024 rays and 300 samples along the ray. We sample only points within the bounding box of the patient, defined by the original CT scan.

NAF was optimized on each volume individually, with a learning rate of 5×10^{-4}, optimized through hyperparameter search on the validation set. For the noise-free projection settings, the model used reflected the specifications from the original paper. The hash encoding used a base resolution of 16, the maximum size of the hash table was 2^{21}, the number of

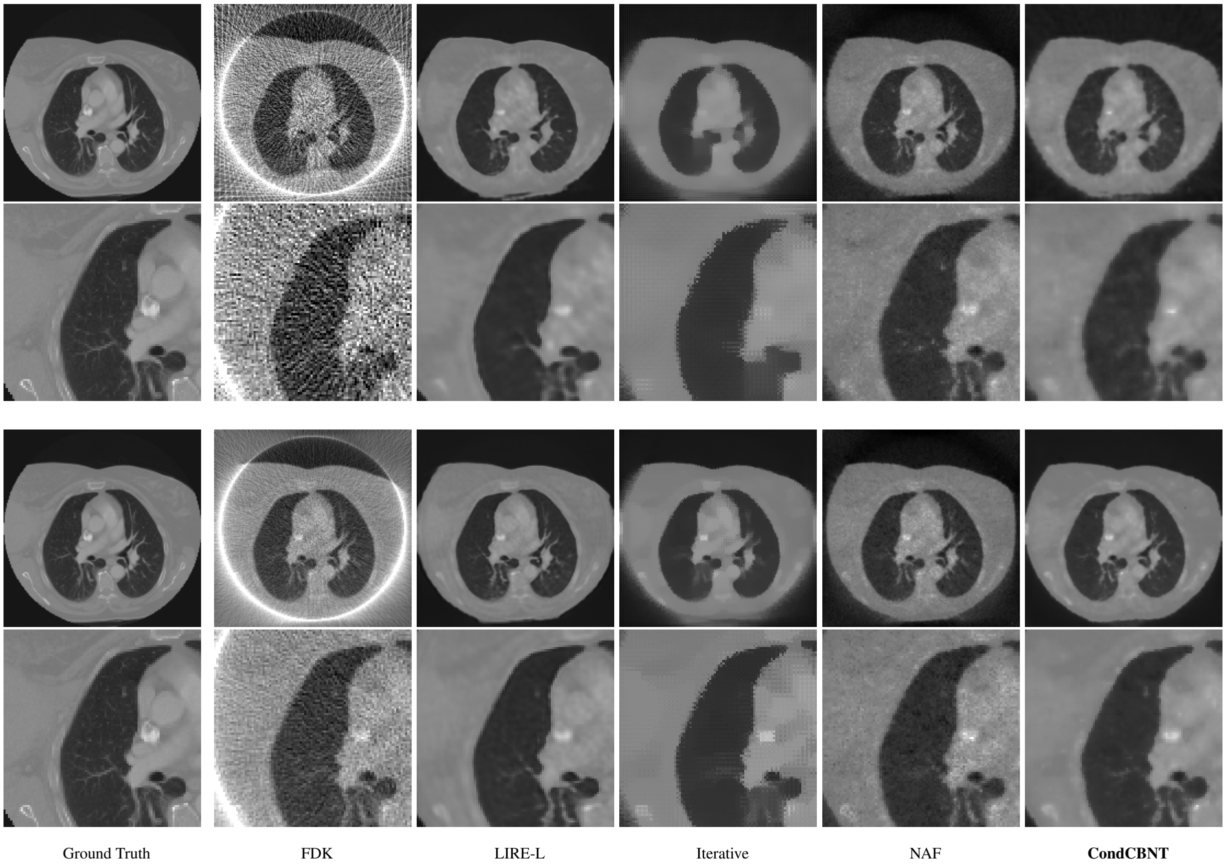

Figure 5. Ground truth and reconstructions using all the methods applied to noisy projections. Top 50 projections, bottom 400 projections. Grayscale colormap with density in $[0 - 0.04]$. The detector size causes a dense ring to appear in the FDK reconstruction. NAF overfits the noise with both 50 and 400 projections. Iterative over-smooths the soft tissues and removes bones. LIRE-L succeeds in keeping soft-tissue contrast and reconstructing bones. Our method succeeds in not overfitting the noise and maintaining higher tissue constrast.

levels was 16 and the size of the feature vector for each level was 2. Instead, validation revealed that a base resolution of 8, with 8 levels and a hash table size of 2^{19} resulted in better reconstruction, as it avoided overfitting to the noise more often. For both settings, an MLP with LeakyReLU activations, 4 layers, and 32 neurons per layer was used. The batch size is also 1024 rays, with 300 points sampled per ray.

D. Additional experimental results

D.1. Larger-scale images of reconstructed volumes

For an improved viewing experience, we include larger-scale versions of our experimental results in Fig. 5 and Fig. 6. The latter shows a volume with less noise in the projections.

E. Acknowledgments

The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

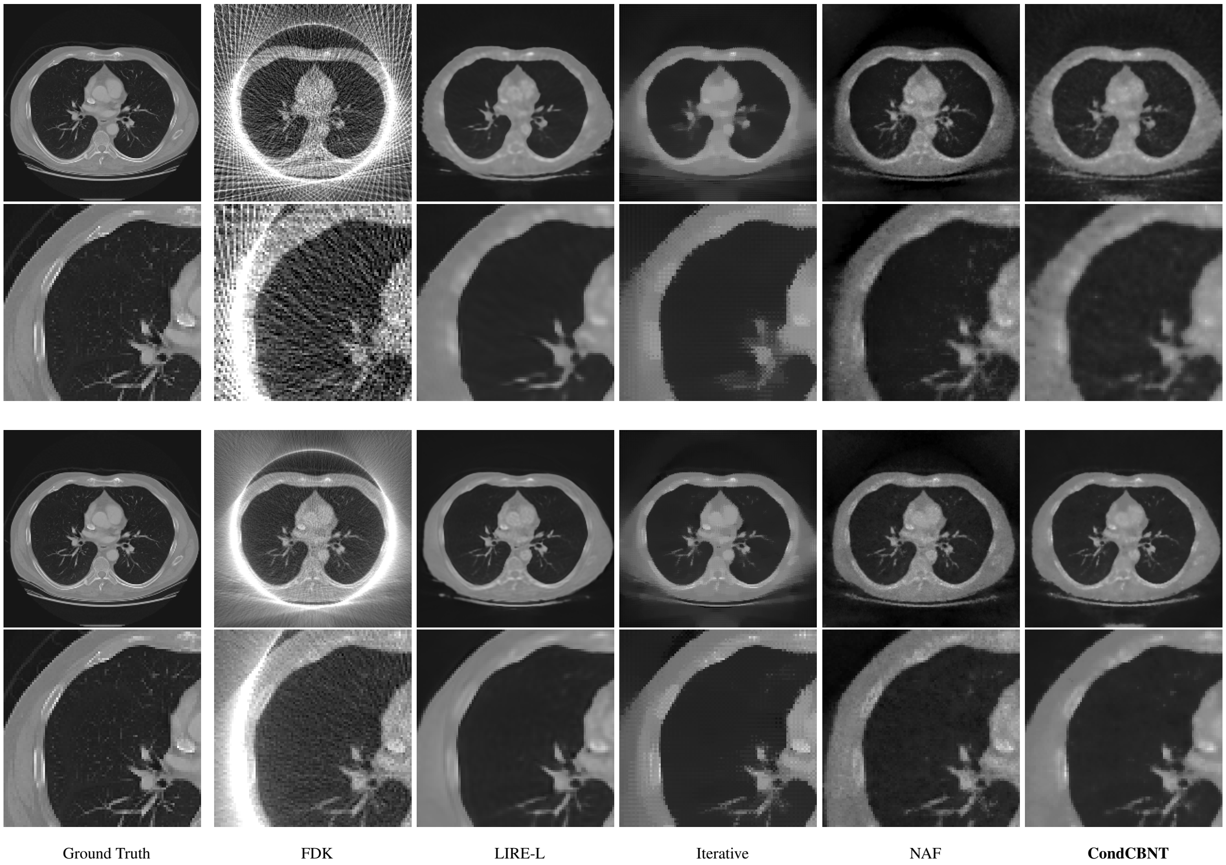

Figure 6. Ground truth and reconstructions using all the methods applied to noisy projections. Top 50 projections, bottom 400 projections. Grayscale colormap with density in $[0 - 0.04]$. Similar behavior to the one shown in Fig. 5. Soft-tissue contrast resolution very clear for CondCBNT and LIRE-L, thanks to less noise in the projections. NAF still overfits the noise. Less over-smoothing by the iterative method.