A MULTI-FILTER AND MULTI-SCALE U-NET FOR CONE-BEAM COMPUTED TOMOGRAPHY WITH HARDWARE CONSTRAINTS

Andreas Hauptmann[1,2], *Mustafa Al-Rubaye*[3], *Miika T. Nieminen*[3,4], *Mikael A.K. Brix*[3,4]

[1]Research Unit of Mathematical Sciences, University of Oulu, Finland
[2]Department of Computer Science, University College London, United Kingdom
[3]Research Unit of Health Sciences and Technology, University of Oulu, Finland
[4]Oulu University Hospital, Department of Diagnostic Radiology, Finland

ABSTRACT

Learned reconstructions for 3D cone-beam computed tomography (CBCT) require significant hardware resources for training as well as evaluation. In this challenge paper we aim to improve performance of the U-Net architecture for post-processing by creating multiple inputs to the network using varying frequency filters. The networks are able to be trained on a single GPU and achieved 3rd place in the ICASSP 2024 3D-CBCT grand challenge.

***Index Terms*—** Cone-beam computed tomography, high resolution, learned reconstructions, post-processing.

1. INTRODUCTION

We consider a learned post-processing approach to improve reconstructions of CBCT in 3D under hardware constraints. To be specific, we are given measured CBCT data y obtained from the ground-truth images x^* by the ray transform with a cone-beam geometry. From the measurements, we can obtain a reconstruction by the FDK (Feldkamp, Davis, Kress [1]) algorithm $\mathcal{F} : y \mapsto x$ as

$$x_{\mathrm{rec}} = \mathcal{F}y. \tag{1}$$

If y is undersampled or corrupted by noise the reconstructions $x_{\mathbf{rec}}$ will suffer from reconstruction artifacts, additionally due to the cone-beam geometry we suffer cone-beam artifacts in the image area where only a part of rays pass through the target. Thus, it is desirable to improve the reconstructions obtained by (1). A popular approach is to use a post-processing network Λ_θ, such as a convolutional neural network with parameters θ, that is trained to improve the initial reconstructions x_{rec} by removing artifacts and noise from the tomographic image. A very successful architecture for this task is the popular U-Net [2]. Unfortunately, training such a convolutional neural network in 3D can be very memory demanding and often only a limited amount of memory may be available for training and testing. Specifically, here we consider a limited memory setting, which means that we only have a single GPU workstation setup. Consequently, the network size one can fit on the GPU is limited and hence the expressivity and performance of the network are reduced. This scenario is specifically relevant in clinical settings, where access to a server may not be possible or is restricted due to data confidentiality.

In the following, we present an approach to improve network performance by producing a more informative input to the post-processing network, inspired by [3]. To achieve this, we consider a multi-filter FDK reconstruction by utilising several frequency cut-offs for the frequency filter. Additionally, we utilise discretisation invariance of the ray transform enabling initial reconstructions on multiple scales similar to [3]. This way we provide wider information content to the network and simplify the learning task. In an ablation study on downsampled resolution we will demonstrate that the increased filter size can lead to improvement in reconstruction quality. The proposed multi-filter and multi-scale U-Net (multi U-Net) and a baseline U-Net have been submitted to the reconstruction challenge and achieved 3rd place.

2. THE MULTI-FILTER AND MULTI-SCALE U-NET

The post-processing network is trained on the initial reconstructions. Instead of a single input we use the multi-filter and multi-scale input. To create this set we will modify the Ram-Lak filter in the frequency space of the FDK reconstruction. First, note that the Ram-Lak filter in frequency space is simply a ramp filter $|k|$. We consider here a simple normalised frequency cut-off, which means that the unmodified FDK uses all frequencies $|k| \in [0, 1]$ whereas a frequency cut-off uses frequencies up to η, i.e., $|k| \in [0, \eta]$ and 0 otherwise. In the following we use a set $\eta_j = j/m$ for $j = 1, \dots, m$. Additionally, we reconstruct images on multiple scales by FDK: the full-resolution reconstructions are $n_1 = 256^3$ and we reconstruct a set of reconstructions for the lower resolution of $n_2 = 128^3$. Now we create a set of reconstructions

This study was funded by the research council of Finland (project nos. 338408, 353093) and Business Finland (Project 8095/31/2022).

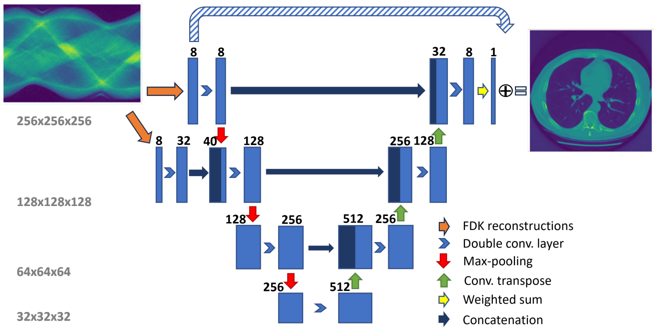

Fig. 1. The multi U-Net with $m = 8$ filters. The sinogram is reconstructed into a set of reconstructions and processed on two scales in the U-Net. A residual connection on the highest resolution reconstruction without frequency cut-off is used.

$$x_{i,j} = \mathcal{F}_{i,j} y,$$

where $\mathcal{F}_{i,j}$ denotes the FDK with frequency cut-off η_j on resolution n_i. For the input we compute the set $\{x\}_{i,j}$ with $i = 1, 2$ and $j = 1, \ldots, m$, which is fed into the network Λ_θ. For the challenge we chose $m = 8$, which results in 16 reconstructions for the U-Net. The architecture is shown in Fig. 1, which is largely a standard U-Net with the additional inputs on two scales. The convolutional layers used a filter size of $3 \times 3 \times 3$, ReLU activation, followed by group norms with group size 8. We limited the channel size on the finest layer due to limited GPU memory. The GPU used was a Nvidia Quadro RTX 6000 with 24GB memory.

3. NUMERICAL TESTS

Training procedure: The network is trained fully supervised to minimise the ℓ^2-loss over paired training data. That is, given training pair $\{x^*, y\}$ we first compute the multi-filter and multi-scale reconstructions $\{x\}_{i,j}$ and then we minimise the loss function $\|\Lambda_\theta(\{x\}_{i,j}) - x^*\|_2^2$, or if written in one step $\|\Lambda_\theta(\{\mathcal{F}_{i,j} y\}_{i,j}) - x^*\|_2^2$. The networks are trained for $5 \cdot 10^4$ iterations, using Adam optimiser, with an initial learning rate of 10^{-3} and a cosine decay. Training of the networks took 5 days for both the low-dose and clinical-dose cases. Additionally, we trained a standard U-Net for the low-dose case that only uses one FDK reconstruction without frequency cut-off as input to examine possible improvements of the multi U-Net. Training took slightly less than 4 days.

Ablation study: Time constraints of the challenge did not permit an ablation study at full resolution. Before training the full-scale networks for the challenge, we have performed initial tests that suggested improved performance of the multi U-Net. Here, we will shortly present an ablation study on the resolution 128^3 that has been performed after the challenge to confirm this observation. The training has been performed as above with one exception in the network architecture, where we chose the filter size of the finest full resolution scale as 32, instead of 8 as in Fig 1. This allows the network to use the provided information more efficiently. The multi U-Net can improve reconstructions (in the downsampled case) by up to 0.5dB (Table 1).

PSNR (dB)	mean	std	min	max
2 filters	37.51	1.93	32.9	43.58
4 filters	37.52	1.92	33.01	43.62
8 filters	37.46	1.9	33.12	43.53
baseline U-Net	37.00	1.84	32.85	42.64

Table 1. Ablation study on resolution 128^3 for low-dose validation set of 100 samples.

4. CHALLENGE RESULTS

Three trained networks have been submitted for the challenge: multi U-Net for the low-dose and clinical dose, as well as the baseline U-Net for the low-dose case. Mean squared errors have been evaluated by the challenge organisers on the test set. In the clinical dose we achieved an error of $9.66 \cdot 10^{-4}$ and 3rd place with the multi U-Net. For the low-dose we achieved $1.679 \cdot 10^{-3}$ for the baseline U-Net and $1.706 \cdot 10^{-3}$ with the multi U-Net, which correspond to 3rd and 4th place.

5. DISCUSSION AND CONCLUSIONS

The challenge has provided insights that a properly trained U-Net can provide competitive results, especially when limited GPU memory is available for training. While the ablation study on the downsampled case shows that the multi U-Net architecture can improve performance, this was not confirmed for the full resolution as the baseline U-Net performed slightly better than the multi U-Net in the challenge results. This is likely due to limited channel size on the full resolution scales (8 compared to 32 in the ablation study) and hence the network can not fully utilise the increased information. Timeliness of the challenge did not permit ablation studies at full resolution, but it will be interesting to investigate possible improved performance of the multi U-Net further.

6. REFERENCES

[1] L. Feldkamp, L. Davis, and J. Kress, "Practical cone-beam algorithm," *Josa a*, vol. 1, no. 6, pp. 612–619, 1984.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI 2015*, 2015.

[3] A. Hauptmann, J. Adler, S. Arridge, and O. Öktem, "Multi-scale learned iterative reconstruction," *IEEE Trans. Comp. Img.*, 2020.