Photon-Efficient 3D Imaging with A Non-Local Neural Network: Supplementary Material

Jiayong Peng, Zhiwei Xiong(✉), Xin Huang, Zheng-Ping Li,
Dong Liu, and Feihu Xu

University of Science and Technology of China
jiayong@mail.ustc.edu.cn, zwxiong@ustc.edu.cn
{hx9711,lizhp}@mail.ustc.edu.cn, {dongeliu,feihuxu}@ustc.edu.cn

In this supplementary material, we provide more details about our network architecture and long-distance coaxial single-photon imaging system.

1 Network architecture

The flowchart of our proposed network is shown in Fig. 1. The network consists of four parts: a feature extraction block, a non-local block, a feature integration block (consisting of a downsampling operator and several 3D dense dilated fusion sub-blocks), and a reconstruction block. The details of network layers as well as the parameter numbers are demonstrated in Sec. 1.1. The details of the non-local block are demonstrated in Sec. 1.2. The amount of GPU memory requirements during training phase is presented in Sec. 1.3.

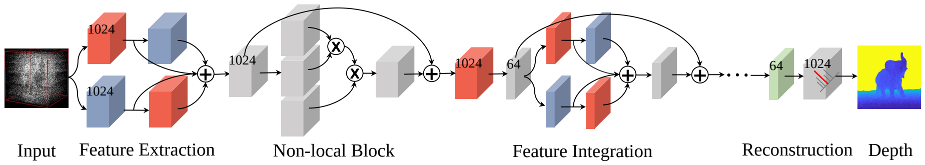

Fig. 1. The flowchart of our proposed network for depth reconstruction from the input raw photon-efficient measurements. The cuboids denote features which are in 3D volume (spatial and temporal). The temporal dimension of features is 1024 originally, and 64 after downsampling. Note that we only show one channel of features for simplification. The red, blue, and green colors denote the 3D convolution, dilated convolution and deconvolution with kernel size of 3×3×3, respectively. The gray color denotes the 3D convolution with kernel size of 1×1×1. "+" and "×" with circular blocks denote the concatenation and matrix multiplication, respectively. Each layer (except for the last one) adopts ReLU as the activation function, which is omitted here for simplification

1.1 Details of network layers and parameter numbers

The network layers and parameter numbers are shown in Table 1. C, D, and DC denote the 3D convolution, dilate convolution, and deconvolution, respectively.

Table 1. The layers and parameter numbers for each block in our network

Block	Layer	Parameter No.
Feature Extraction	$C-3,3,3-2; D-3,3,3-2$ $D-3,3,3-2; C-3,3,3-2$ $C-1,1,1-2$	350
Non-local	$C-1,1,1-2; C-1,1,1-2; C-1,1,1-2$ $C-1,1,1-2$	20
Feature Integration	$\begin{bmatrix} C-3,3,3-4 \\ C-3,3,3-8 \\ C-3,3,3-16 \\ C-3,3,3-32 \end{bmatrix}$ $\left\{\begin{matrix} C-1,1,1-16 \\ C-3,3,3-8; D-3,3,3-8 \\ D-3,3,3-4; C-3,3,3-4 \\ C-1,1,1-8 \end{matrix}\right\} \times 8$	97140
Reconstruction	$\begin{bmatrix} DC-3,3,3-48 \\ DC-3,3,3-24 \\ DC-3,3,3-12 \\ DC-3,3,3-6 \end{bmatrix}$ $C-1,1,1-1$ SoftArgMax	165247
Total		262757

The figures after them denote kernel size (in the order of depth, height, and width) and filter number. Layers in "{}" constitute the 3D dilated dense fusion sub-block, and the layers in "[]" constitute the downsampling and upsampling operations for feature tensors in our network.

1.2 Details of Non-local Block

We further clarify the details of non-local block with Fig. 2. Denote the three blocks in Fig. 1 from top to bottom in the non-local block as G(X), P(X), and T(X), they represent feature tensors by passing a common input feature tensor X to three 3D convolutions with kernel size of 1×1×1 (conv_G, conv_P, and conv_T). The size of X is H*W*D*C (Height=Width=32, Depth=1024, and Channel=2). G(X), P(X), and T(X) have the same size as X originally, and they are vectorized before the next step multiplication to be with size of HWDC*1. As shown in Fig. 2, we first compute z=P(X)'*G(X), which results in a scalar z; we then compute Y=T(X)*z, which results in a vector Y; the output Y is again reshaped to a feature tensor with size of H*W*D*C for the next 3D convolution with kernel size of 1×1×1. The final output of the non-local block is the concatenation of the feature tensor from the 1×1×1 3D convolution and the input feature tensor X. In this way, our NLB requires much less parameters than that would be required for a fully connected layer. Similar operations for tensor multiplication have been used in [4, 2].

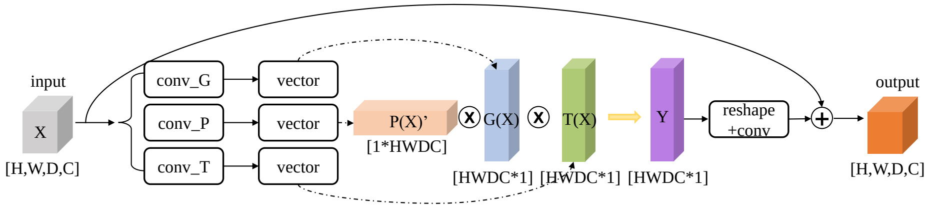

Fig. 2. The flowchart of the non-local block. The conv_G, conv_P, and conv_T are the three 3D convolutions with kernel size of 1×1×1 and "vector" denotes the vectorization on feature tensors. "+" and "×" with circular blocks denote the concatenation and matrix multiplication, respectively. "reshape+conv" denotes the reshape operation on feature tensors and the following 3D convolution with kernel size of 1×1×1

1.3 GPU memory requirements

Our network as well as the ones in [1] are trained on large-volume 3D photon-efficient measurements, which require relatively large GPU memory. We further compare the GPU memory requirements of our network with Lindell and Lindell_I in [1]. Training on the same dataset with a resolution of 32×32×1024, Lindell and Lindell_I use 6.9GB and 7.6GB GPU memory, while ours only uses 3.0GB.

2 Long-distance coaxial single-photon imaging system

For our imaging system, the active illumination source is a standard erbium-doped near-infrared fiber laser (1550 nm, 500 ps pulse width, 100 kHz repetition rate). The experiments are performed at near-infrared band to reduce the influence of solar background so that our imaging system has a capability of all-time working. In addition, operating in the near-infrared range ensures low atmospheric loss and guarantees our system eye-safe. The field of view (FoV) for our imaging system is 22.3 μrad. The laser is transmitted from the transceiver telescope that is a commercial Cassegrain telescope with a 280 mm aperture. After reaching the target, the reflected photons are collected by the same telescope, then delivered through the scanning mirror, perforated mirror and polarization beam splitter, and finally detected by the single photon detector. The single photon detector is a InGaAs/InP single-photon avalanche diode (SPAD) [3] with dark counts of 2200 counts/s, dead time of 1 us and efficiency of 15% (side note: the SPAD is operating in free-running mode). Photon detections are given timestamps by a time-to-digital converter (TDC) with about 50 ps time jitter. The whole system jitter is measured to be 1 ns which is equivalent to 15 cm depth uncertainty. In addition, a standard commercial camera with spatial resolution of 1280×960 is paraxially fixed on the telescope to provide a rough view of the scenes.

In order to capture a larger picture of the scene, we adopt a two-stage scanning scheme. Firstly, we utilize the scanning mirror to steer the light beam in both x and y axes to get a sub-picture. The mirror is mounted on a two-axis piezoelectric ceramic actuator to provide a raster scanning for both illumination and detection. Secondly, the telescope is rotated integrally by a rotary table to get the next sub-picture. This cycle repeats until the whole picture is taken.

We capture three different outdoor scenes with various distances and resolutions. The first scene is a *Hotel* that locates 1 km away from our imaging system. We scan 320×320 points at an acquisition time of 0.3 ms per point with a laser power of 10 mW. The scene is captured in the daylight with the background flux strength of about 300 counts/s within the 200 ns time window. The average photon per pixel (PPP) is about 1 and the signal-to-background (SBR) is 0.1. The second scene is a *Castle* that locates 4 km away from our imaging system. We scan 256×256 points at an acquisition time of 12.5 ms per point with a laser power of 9.5 mW. The scene is captured in the daylight with the background flux strength of about 300 counts/s within the 200 ns time window. The average PPP is about 1 and the SBR is 0.9. The third scene is a tall building named *K11* that locates 21 km away from our imaging system. We scan 128×128 points at an acquisition time of 40 ms per point with a laser power of 100 mW. The scene is captured in the daylight with the background flux strength of about 360 counts/s within the 200 ns time window. The average PPP is about 1 and the SBR is 0.1.

References

1. Lindell, D.B., O'Toole, M., Wetzstein, G.: Single-photon 3d imaging with deep sensor fusion. ACM Transactions on Graphics **37**(4), 113 (2018)
2. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
3. Yu, C., Shangguan, M., Xia, H., Zhang, J., Dou, X., Pan, J.W.: Fully integrated free-running ingaas/inp single-photon detector for accurate lidar applications. Optics Express **25**(13), 14611–14620 (2017)
4. Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., Xu, F.: Compact generalized non-local network. In: International Conference on Neural Information Processing Systems. pp. 6510–6519 (2018)