Smart Wide-field Fluorescence Lifetime Imaging System with CMOS Single-photon Avalanche Diode Arrays

Dong Xiao, Zhenya Zang, Quan Wang, Ziao Jiao, Francescopaolo Mattioli Della Rocca, Yu Chen, and David Day Uei Li

Abstract—**Wide-field fluorescence lifetime imaging (FLIM) is a promising technique for biomedical and clinic applications. Integrating with CMOS single-photon avalanche diode (SPAD) sensor arrays can lead to cheaper and portable real-time FLIM systems. However, the FLIM data obtained by such sensor systems often have sophisticated noise features. There is still a lack of fast tools to efficiently recover lifetime parameters from highly noise-corrupted fluorescence signals. This paper proposes a smart wide-field FLIM system containing a 192 × 128 COMS SPAD sensor, and a field-programmable gate array (FPGA) embedded deep learning (DL) FLIM processor. The processor adopts a hardware-friendly and light-weighted neural network for fluorescence lifetime analysis, showing the advantages of high accuracy against noise, fast speed, and low power consumption. Experimental results demonstrate the proposed system's superior and robust performances, promising for many FLIM applications such as FLIM-guided clinical surgeries, cancer diagnosis, and biomedical imaging.**

I. Introduction

Fluorescence lifetime imaging microscopy (FLIM) is a widely applied imaging technique for biology, chemistry, and pharmacy applications. It provides a unique way to quantitively investigate cellular metabolisms, molecular biophysical microenvironments (including pH, Ca^{2+}, O_2), protein-protein interactions, and Förster resonance energy transfer (FRET) behaviors [1,2]. With the ability to monitor various molecular processes in living cells and tissues, FLIM is promising for disease diagnosis, drug delivery, and drug developments [3-5].

Conventional laser-scanning FLIM systems for cell imaging usually use discrete components, including an optical module, a scanning system, photomultipliers (PMT), a time-correlated single-photon counting (TCSPC) card, and a PC, which are expensive, bulky, and only suitable for laboratory environments. PMTs are fragile and need a high voltage supply (>1KV). Meanwhile, FLIM data is usually analyzed by iterative curve-fitting software tools based on the least square fitting (LSF) method, maximum likelihood estimation (MLE), or Bayesian methods. These methods are usually computationally intensive and, therefore, slow [6].

A single-photon avalanche diode (SPAD) is a p-n junction working at the Geiger mode, offering a sensitivity down to the single-photon level with excellent timing precision [7, 8]. Thanks to rapid advances in complementary metal-oxide-semiconductor (CMOS) manufacturing technologies, various CMOS SPAD array sensors have been developed for cheap, portable, and fast time-resolved imaging systems [9 - 12]. CMOS SPAD sensors are integrated with front-end electronics, timing electronics, and other functional blocks in a system-on-chip. Therefore, CMOS SPAD sensors show unparalleled advantages such as a small footprint, lower power consumption, and low cost. Moreover, CMOS SPAD array sensors offer parallel detection capacity, suitable for fast wide-field FLIM imaging. Compared with laser-scanning FLIM, wide-field FLIM techniques are simpler and faster.

However, SPAD array devices can be noisier than traditional single-channel SPAD or PMT sensors [7], even though new low-noise sensor technologies have been introduced [12 - 14]. Apart from dark count and shot noise, a bigger sensor array also has clock switching noise, ground bounce noise, mismatch problems (sensor gain), internal clock tree routings, and front-end circuits [12]. FLIM images obtained by SPAD array sensors can be therefore distorted. In addition, wide-field FLIM systems usually have sophisticated noise features because the signals are vulnerable to the scattering light from the out-of-focus plane and pixel crosstalk. Existing FLIM analysis methods cannot tackle this problem robustly. This study reports an intelligent wide-field SPAD FLIM system with an embedded deep learning processor on an FPGA board. DL techniques provide an alternative route to fast and high-precision FLIM analysis [15 -18]. In this work, a light-weighted neural network (NN) algorithm was designed and implemented on the DL processor for fast and accurate analysis of noise-contaminated FLIM data. The proposed system shows superior and robust analysis performances without additional software. Our study can facilitate the development of portable and real-time wide-field FLIM systems and related applications.

II. System Design and Evaluation

A. System Design

Figure 1 shows the overall view of the wide-field FLIM system. It comprises three subsystems: the optics, the SPAD system, and the DL processor. The optics subsystem is shown in Fig. 1 (a). The 485 nm pulsed diode laser (DD-485L Delta Diode, HORIBA Scientific) with a pulse width of 80 ps and a

* This work was supported in part by Medical Research Scotland (MRS-1179-2017), Photon Force, Ltd., BBSRC (BB/V019643/1 and BB/K013416/1), EPSRC (EP/T00097X/1), and nVIDIA. (*Corresponding author: David Day-Uei Li.*)

Dong Xiao, Zhenya Zang, Quan Wang, Ziao Jiao, and David Day-Uei Li are with the Department of Biomedical Engineering, Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, G4 0RE Glasgow, U.K. (dong.xiao@strath.ac.uk; zhenya.zang@strath.ac.uk; quan.wang.100@strath.ac.uk; ziao.jiao@strath.ac.uk ; david.li@strath.ac.uk).

Francescopaolo Mattioli Della Rocca is with School of Engineering, University of Edinburgh, Edinburgh, EH9 3JL UK (francesco.mdrocca@gmail.com)

Yu Chen is with the Department of Physics, University of Strathclyde, Glasgow, G4 0RE, Scotland, UK. (y.chen@strath.ac.uk).

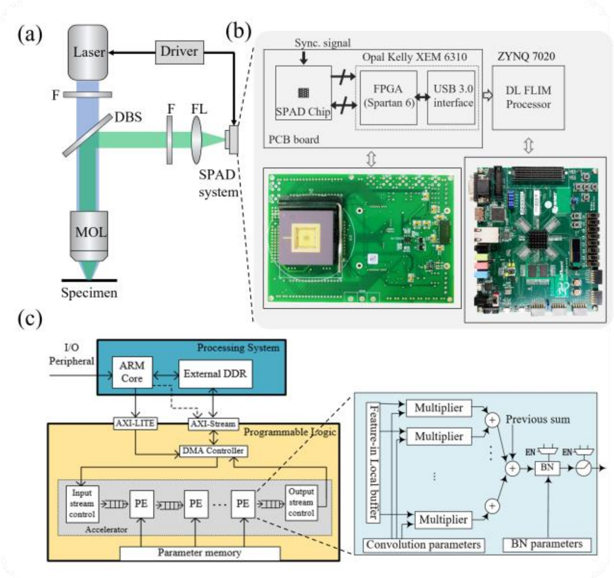

Fig. 1. Overview of the smart wide-field FLIM system. (a) The optics for fluorescence signal excitation and collection. (b) The block diagram of SPAD system with DL processor (upper panel) and top view of experimental devices (low panel). (c) The hardware block diagram of the FPGA-based DL processor. The insert shows the details of the processing element (PE).

repletion rate of 100 MHz illuminated the specimen slide through a filter (F), a dichroism beam splitter (DBS), and a microscope objective lens (MOL). The fluorescent photons emitted by the specimen passed through the DBS and the filter and were focused by a focus lens (FL). The SPAD system was packaged into a QuantiCam camera module and mounted on the microscope system to collect the signal on the focus plane of the FL. The laser driver generated the synchronized signal to drive the laser and provided a reference signal for timing the detected photon in the SPAD system. Figure 1(b) shows the detailed configuration of the SPAD system. The SPAD array fabricated in STMicroelectronics' 40 nm CMOS process was integrated into one 3.15 mm × 2.37 mm chip. This sensor chip comprises a 196×128 SPAD array, 64 parallel-to-serial converters, and corresponding addressing circuitry. This SPAD chip's detailed design and fabrication were reported in [12]. In our experiment, each pixel's time-to-digital converter (TDC) resolution was set to 39 ps. The SPAD chip was controlled by an Opal Kelly FPGA board (XEM6310-LX150, Xilinx) that contains a Xilinx Spartan 6 FPGA chip and a USB3 serial link. A custom printed circuit board (PCB) provides an interface for the SPAD chip and FPGA board. It also provides an input interface for the synchronized signal from the laser source. Fig.1(c) shows the hardware block diagram of the DL processor with an NN algorithm for FLIM data analysis. The DL processor was developed on the ZYNQ 7020 (XC7Z020-CLG484-1, Xilinx, USA) board for the proof-of-concept study. The FPGA device contains the programmable logic (PL) blocks and the processing system (PS). The NN architecture is implemented in PL blocks on multiple processing elements (PEs) for highly parallel computing. The PS with two ARM cores configures status registers of the Direct-Memory-Access (DMA) controller and peripherals. The AXI-Stream and AXI-LITE buses are for data transfer and configurations of pre-trained NN parameters, respectively. Raw FLIM data from the SPAD sensor will be fetched through I/Os and processed by PEs. The footprint of the whole SPAD system with DL FLIM processor has a compact size with only a half of A4 paper.

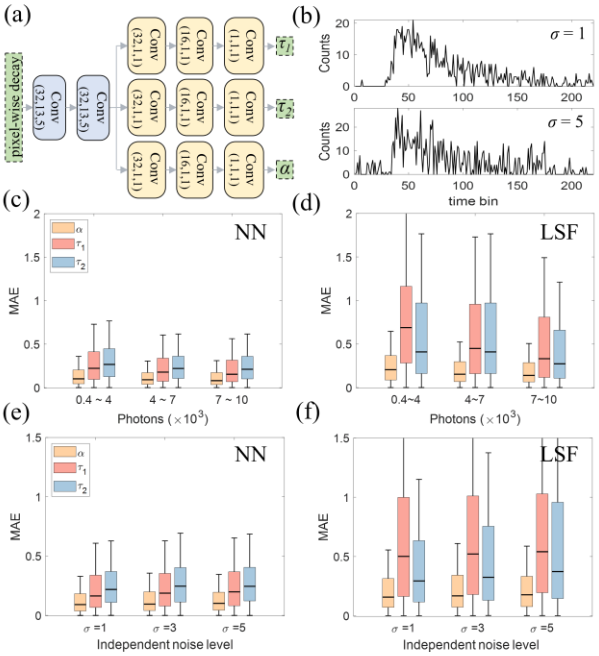

Fig. 2. (a) The architecture of the NN. The dash boxes indicate the inputs and outputs. The parameters of each convolutional block are the filter number × the kernel size × the stride. (b) Simulated decay profiles with different background noise levels. Both decays have same lifetime parameters with $\tau_1 = 1$ ns, $\tau_2 = 3$ ns, and $\alpha = 0.5$. Their total counts are 1000. The background noise level is quantized by the standard deviation σ of Gaussian noise. (c) – (f) Evaluation of the DL processor's performance using testing dataset. Each group in boxplots contains 2000 decay samples.

B. Neural network algorithm

One-dimensional convolutional blocks were applied to construct the NN backbone to design a hardware friendly NN algorithm with high throughput, low latency, and low energy consumption [17]. Each convolutional block contains three subsequential layers: a 1-D convolutional layer, a batch normalization (BN) layer, and a rectifier linear unit (ReLU) activation layer. The NN's topological structure is shown in Fig. 2(a). Our study focuses on bi-exponential decays, and the lifetime parameters (including the shorter lifetime τ_1, the longer lifetime τ_2, and the fraction ratio α) will be estimated. The first two layers with large kernel sizes and strides are for feature extraction, and three branches reconstruct lifetime parameters. Each branch contains three pointwise convolutional blocks for down pooling the information and obtaining final parameters.

The designed NN was trained by simulated synthetic data since both the fluorescence decay and noise mathematical models are well developed. The theoretical bi-exponential fluorescence decays measured by a FLIM system can be described as:

$$y(t) = N_T \cdot I(t) * [\alpha e^{-t/\tau_1} + (1-\alpha) e^{-t/\tau_2}], \quad (1)$$

where $I(t)$ is the FLIM system's instrument response function (IRF). The asterisk (*) refers to a convolution operator. The integral $\int I(t) * [\alpha e^{-t/\tau_1} + (1-\alpha) e^{-t/\tau_2}]\, dt$ is normalized to 1, and N_T is the total photon count of the decay. With noise included, the synthetic decay is:

$$Y(t) = y(t) + \sqrt{y(t)}\mathcal{N}(0,1) + \mathcal{N}(0,\sigma), \qquad (2)$$

where $\mathcal{N}(\mu, \sigma)$ is the Gaussian distribution with mean μ and a standard deviation σ. The last two terms of Eq. (2) refer to the shot and signal-independent noise, respectively. The shot noise originating from the discrete nature of photons follows a Poisson distribution, which is approximated using a normal distribution. The signal-independent noise has complex origins, including surrounding scattered light, TDC nonlinearity, circuit clock noise, and quantization noise. As the background noise can be subtracted in the preprocessing phase, the signal-independent noise can be described by an added Gaussian noise with zero mean and a standard deviation σ. Fig. 2 (b) shows two examples of decays with different background noise levels. Larger σ corresponds pixels with a high noise level. For network training, 40000 different decays as described in Eq. (2) were generated with lifetime parameters $\tau_1 \in$ [0.5, 2] ns, $\tau_2 \in$ [2, 3.5] ns, and $\alpha \in$ [0, 1]. The total N_T ranges from 400 to 1e4, and the standard deviation σ of the Gaussian noise empirically varies from 1 to 5 to emulate different noise levels of SPAD array pixels. We used Gaussian functions to fit the IRFs of the pixels in the SPAD array sensor, and the FWHM is 324 ps. The simulated data were generated using MATLAB, and the NN was implemented with *Pytorch* in the Python environment [19]. The loss function is

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} \left\| F_j(Y^i, \Theta) - \hat{Y}_j^i \right\|_2^2, \qquad (3)$$

where F is j^{th} (j=1, …, M) end-to-end mapping function and $M = 3$ is the number of the output branches. Y is the input signal and $\hat{Y}_j$ is the corresponding ground-truth target of the j^{th} output branch. N = 128 is the batch size number. Θ is the hyperparameter of the network. The optimizer is the Adam algorithm with a learning rate of 1e-4. Once the NN was well-trained, to reduce computational complexity and off-chip data transfer on the FPGA board, the weights and activations of layers were quantized to 1- and 4-bit, respectively [20]. The NN parameters were fetched via a Python script and stored in DL processor's on-chip memory.

Figs 2 (c) – (f) show the evaluation of the NN algorithm on a new testing dataset. As a comparison, the widely used LSF with the Levenberg-Marquardt algorithm was also used for data analysis. In Figs 2(c) and (d), the mean absolute errors (MAEs) of τ_1, τ_2, and α were evaluated by NN and LSF using samples with different total photon counts. The background noise levels randomly vary from σ = 1 to 5. In Figs 2(e) and (f), the MAEs of lifetime parameters using both algorithms were investigated under different background noise levels while the total photons of all samples vary from 400 to 1e4. Results show that our NN significantly outperforms LSF for resolving three different parameters. The NN algorithm is also more stable against significant background noise. In addition, the calculation speed for the NN algorithm is 300-fold faster than LSF. The DL FLIM processor consumes a small amount of hardware resource. The on-chip memory is about 3 Mb and the power for lifetime analysis is only around 4.5 W.

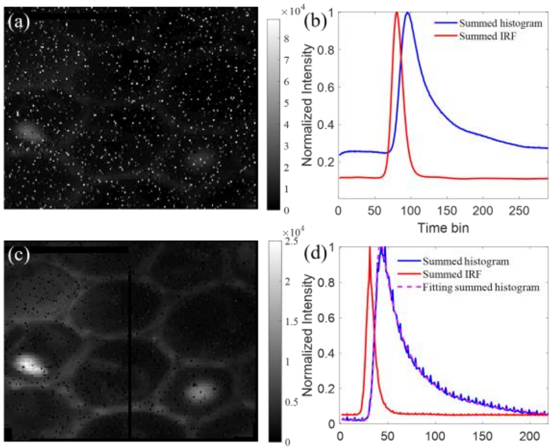

Fig. 3. The raw (a) and calibrated (c) intensity image obtained from the SPAD array sensor. The corresponding global histogram and IRF before (b) and after (d) calibration. The patterned noise in (d) is due to TDC nonlinearity.

III. Experimental Analysis

As a proof-of-concept demonstration, the specimen is the acridine-orange-staining *convallaria majlis* rhizome sample. Before testing the sample, the IRFs of the SPAD array sensor were measured by replacing the sample with the solution of Ludox. A neutral density filter was placed in front of the laser source to attenuate the laser intensity. The IRF map of each pixel was used for calibration of the sensor. For the acquisition speed, a single frame for our SPAD sensor only takes 2 ms. In comparison, conventional laser-scanning FLIM systems usually need several seconds. The raw intensity image of the specimen is shown in Fig. 3 (a). The image's content is hardly seen from background noise due to sensor pixels' varying noise levels and sensitivities. The bright pixels are hot pixels (with a high dark count rate). Figure 3 (b) shows the global IRF and fluorescence histogram by summing all pixels together. A high background noise level and a broader IRF can be observed. Therefore, pixel alignments are needed to calibrate the distortion. The IRF peak positions of all pixels are aligned at a pre-defined position to ensure that the rising edges of all fluorescence decays are also aligned. Besides, all decays subtract background noise to ensure the mean noise level is nearly zero. Hot and silent pixels can be identified and masked through their corresponding IRF profiles. The calibration procedure can be programmed in the FPGA for further automatic pre-processing. The calibrated image is shown in Fig. 3(c). The image becomes apparent as the vascular bundles' structure can be easily identified. The two bright spots in the images are photodamaged areas caused by long exposure to the high-intensity laser. The photodamage changes the FLIM intensity and leads to an ultrashort lifetime component [21]. Fig. 3(d) shows the corresponding calibrated global histogram and IRF. The global analysis was conducted using LSF. The histogram was fitting by a bi-exponential model and the lifetime parameters are τ_1 = 0.76 ns, τ_2 = 3.04 ns, and α = 0.58. The amplitude-weighted average lifetime defined by $\tau_A = \alpha e^{-t/\tau_1} + (1-\alpha)e^{-t/\tau_2}$ was also investigated. τ_A = 1.71 ns in global analysis.

The DL processor was used to analyze FLIM data. A bi-exponential model interpreted the experimental FLIM image.

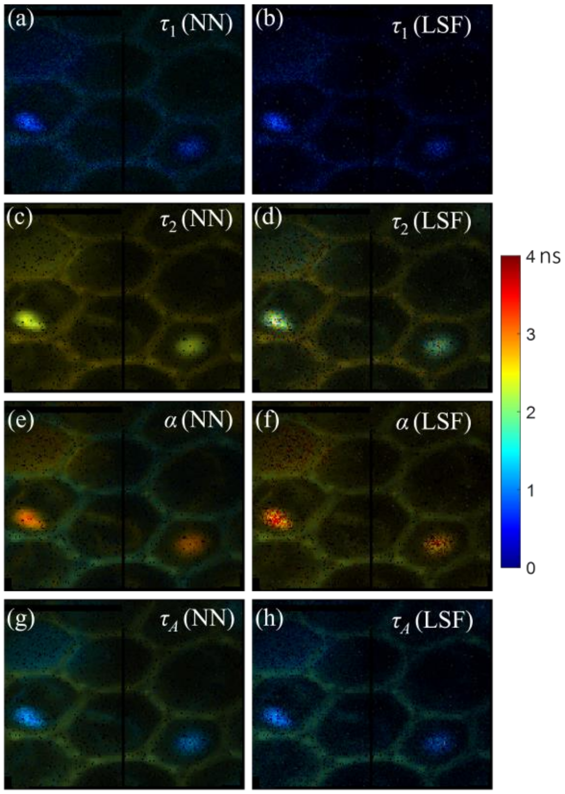

Fig. 4. Lifetime analysis using DL processor with NN algorithm and LSM, respectively.

The predicted results by the NN algorithm and LSF are shown in Fig.4. The NN algorithm delivers clear lifetime images for all parameters. The mean values of τ_1, τ_2, τ_A and α images are 0.9 ns and 2.84 ns, 1.78 ns, and 0.55, respectively. The results correspond well with the global analysis above. As a comparison, the τ_1, τ_2, τ_A and α images calculated by LSF are noisier, less stable with a significant variation. In addition, due to the considerable noise, LSF failed to converge in a substantial portion of pixels. The corresponding mean values of τ_1, τ_2, τ_A and α images are 0.32 ns, 2.3 ns, 0.82 ns, and 0.64, which show a significant discrepancy with previous results. The analysis of the experimental image using the NN algorithm in the DL processor only takes several a few hundred milliseconds. In contrast, traditional LSF needs several minutes on a desktop computer. The results show that the NN algorithm is faster, more robust, and more accurate to analyze fluorescence lifetimes. The intelligent wide-field FLIM system has a significant advantage in various applications such as lifetime-guided diagnosis, cell imaging, FLIM-FERT analysis.

IV. Conclusion

In conclusion, we designed a wide-field FLIM system with COMS SPAD sensors and a DL processor. The SPAD array can detect fluorescence signals parallel, leading to fast imaging speed. The processor was designed for intelligent data analysis, showing excellent performance in FLIM analysis from highly corrupted fluorescence signals. A light-weighted neural network algorithm was developed, trained, evaluated using simulated synthetic data, and implemented on the FPGA-based processor. The wide-field FLIM system was experimentally validated with cell samples to demonstrate its effectiveness. The system can be further developed into portable devices by designing compact optics and optimizing firmware of SPAD sensor and DL processor.

References

[1] Lakowicz, J. R. Principles of Fluorescence Spectroscopy. (Springer, 2006).

[2] Suhling, K. et al. Fluorescence lifetime imaging (FLIM): Basic concepts and some recent developments. Medical Photonics 27, 3-40, (2015).

[3] Sun, Y. et al. Fluorescence lifetime imaging microscopy for brain tumor image-guided surgery. J Biomed Opt 15, 056022, (2010).

[4] Yaseen, M. A. et al. Fluorescence lifetime microscopy of NADH distinguishes alterations in cerebral metabolism in vivo. Biomed Opt Express 8, 2368-2385, (2017).

[5] Renfrew, A. K., Bryce, N. S. & Hambley, T. W. Delivery and release of curcumin by a hypoxia-activated cobalt chaperone: a XANES and FLIM study. Chemical Science 4, (2013).

[6] R. Datta, T. M. Heaster, J. T. Sharick, A. A. Gillette, M. C. Skala. "Fluorescence lifetime imaging microscopy: fundamentals and advances in instrumentation, analysis, and applications," J Biomed Opt 25(7), 1-43, (2020).

[7] Caccia, M., Nardo, L., Santoro, R. & Schaffhauser, D. Silicon Photomultipliers and SPAD imagers in biophotonics: Advances and perspectives. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 926, 101-117, (2019).

[8] Bruschini, C., Homulle, H., Antolovic, I. M., Burri, S. & Charbon, E. Single-photon avalanche diode imagers in biophotonics: review and outlook. Light Sci Appl 8, 87, (2019).

[9] Li, D. D. Real-time fluorescence lifetime imaging system with a 32 × 32 0.13μm CMOS low dark-count single-photon avalanche diode array. Opt Express 18, 10257-10269, (2010).

[10] Li, D. D. et al. Video-rate fluorescence lifetime imaging camera with CMOS single-photon avalanche diode arrays and high-speed imaging algorithm. J Biomed Opt 16, 096012, (2011).

[11] Field, R. M., Realov, S. & Shepard, K. L. A 100 fps, Time-Correlated Single-Photon-Counting-Based Fluorescence-Lifetime Imager in 130 nm CMOS. IEEE Journal of Solid-State Circuits 49, 867-880, (2014).

[12] Henderson, R. K. et al. A 192 x 128 Time Correlated SPAD Image Sensor in 40-nm CMOS Technology. IEEE Journal of Solid-State Circuits, 1-10, (2019).

[13] Xu, H., Pancheri, L., Betta, G. D. & Stoppa, D. Design and characterization of a p+/n-well SPAD array in 150nm CMOS process. Opt Express 25, 12765-12778, (2017).

[14] Ulku, A. C. et al. A 512x512 SPAD Image Sensor with Integrated Gating for Widefield FLIM. IEEE J Sel Top Quantum Electron 25, (2019).

[15] Wu, G., Nowotny, T., Zhang, Y., Yu, H. Q. & Li, D. D. Artificial neural network approaches for fluorescence lifetime imaging techniques. Opt Lett 41, 2561-2564, (2016).

[16] Smith, J. T. et al. Fast fit-free analysis of fluorescence lifetime imaging via deep learning. Proc Natl Acad Sci U S A 116, 24019-24030, (2019).

[17] D. Xiao, Y. Chen and D. D. -U. Li, "One-Dimensional Deep Learning Architecture for Fast Fluorescence Lifetime Imaging," IEEE J. Sel. Top. Quantum Electron., vol. 27, no. 4, pp. 1-10, Aug. 2021.

[18] D. Xiao, Z. Zang, N. Sapermsap, Q. Wang, W. Xie, Y. Chen and D. D. -U. Li, "Dynamic fluorescence lifetime sensing with CMOS single-photon avalanche diode arrays and deep learning processors," Biomed. *Opt. Express*, vol. 12, no. 6, pp. 3450-3462, June 2021.

[19] Pytorch, https://pytorch.org/.

[20] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients," arXiv:1606.06160, (2018).

[21] W. Becker, *The bh TCSPC Handbook,* 9th edition (2021) available on www.becker-hickl.com.