This is a peer-reviewed, accepted author manuscript of the following paper:
Kirkland, P., Kapitany, V., Lyons, A., Soraghan, J., Turpin, A., Faccio, D., & Di Caterina, G. (2020).
Imaging from temporal data via spiking convolutional neural networks. In G. S. Buller, R. C. Hollins, R. A. Lamb, M. Laurenzis,
A. Camposseo, M. Farsari, L. Persano, & L. E. Busse (Eds.), Emerging Imaging and Sensing Technologies for Security and
Defence V; and Advanced Manufacturing Technologies for Micro- and Nanosystems in Security and Defence III (Vol. 11540).
SPIE. https://doi.org/10.1117/12.2573484

Imaging from Temporal Data via Spiking Convolutional Neural Networks

Paul Kirkland[1], Valentin Kapitany[2], Ashley Lyons[2], John Soraghan[1], Alex Turpin[2], Daniele Faccio[2], and Gaetano Di Caterina[1]

[1]Univ. of Strathclyde, Glasgow, UK
[2]Univ. of Glasgow, Glasgow, UK

ABSTRACT

A new approach for imaging that is solely based on the time of flight of photons coming from the entire imaged scene, combined with a novel machine learning algorithm for image reconstruction: a spiking convolutional neural network (SCNN) named Spike-SPI (Spiking - Single Pixel Imager). The approach uses a single point detector and the corresponding time-counting electronics, which provide the arrival time of photons in the form of spikes distributed over time. This data is transformed into a temporal histogram containing the number of photons per arrival time. A SCNN that converts the 1D temporal histograms into a 3D image (2D image with depth map) by exploiting the feature extraction capabilities of convolutional neural networks (CNNs), the high dimensional compressed latent space representations of a variational encoder-decoder network structure, and the asynchronous processing capabilities of a spiking neural network (SNN). The performance of the proposed SCNN is analysed to demonstrate the state-of-the-art feature extraction capabilities of CNNs and the low latency asynchronous processing of SNNs that offer both higher throughput and higher accuracy in image reconstruction from the ToF data, when compared to standard ANNs. The results of Spike-SPI show an increase in spatial accuracy of 15% over then ANN, using the Intersection of Union (IoU) for the objects in the scene. While also delivering a 100% increase over then ANN in object reconstruction signal to noise ratio (RSNR) from ∼3dB to ∼6dB. These results are also consistent across a range of IRF (Instrument Response Functions) values and photo counts, highlighting the robust nature of the new network structure. Moreover, the asynchronous processing nature of the spiking neurons allow for a faster throughput and less computational overhead, benefiting from the operational sparsity in the single point sensor.

Keywords: LiDAR, SPAD, SCNN, CNN, Neural Network, Imaging, Depth

1. INTRODUCTION

Most imaging methods can be divided into two categories. In the first, the scene is flood-illuminated with light, that is, all regions of the scene are illuminated simultaneously. The light reflected by the scene is then imaged onto many detector pixels via a lens. In the second, only a known sub-region of the scene is illuminated, and the light reflected from that sub-region is collected onto a single pixel. By dividing up the scene into many of these sub-regions, and measuring light from only one region at a time, one can scan over the scene. By combining the time of flight information from the sub-regions, the entire scene can be reconstructed. These processes extend also to three dimensional (3D) imaging, where distance from the sensor can be inferred from stereoscopic imaging, holographic, or time-of-flight (ToF) methods.[1–4]

The first approach has the advantage that the scene only needs to be illuminated once, giving a substantial advantage in speed over a point/structure scan. However, as the second approach relies on a single pixel only, it may be operated at a higher framerate and may be much less bulky than a whole array of pixels. Therefore, the next frontier for high speed, high framerate imaging would be to combine the speed benefit of flood illumination with the electronic/mechanical benefit of requiring only a single pixel for detection.

Further author information: Send correspondence to Paul Kirkland (Spiking Neural Networks), Valentin Kapitany (LiDAR).
Paul Kirkland: E-mail: paul.kirkland@strath.ac.uk Valentin Kapitany: E-mail: v.kapitany.1@research.gla.ac.uk

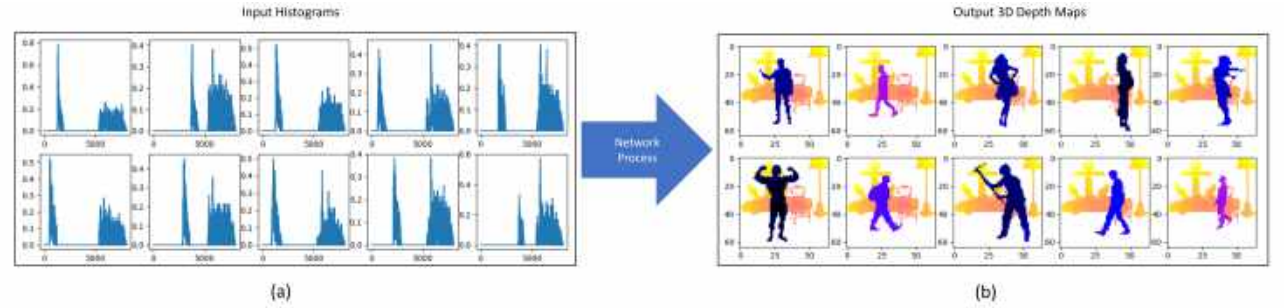

Figure 1. Illustration of the different histogram inputs (a), that are converted into 3D depth maps through the novel network proposed in this paper (b).

In a recent work, Turpin et al.[5] have demonstrated the feasibility of such an approach. They flood-illuminated a scene with a pulsed laser source, and focused the back-reflected light onto a single detector pixel. To achieve this, a single photon avalanche diode (SPAD) and Time Correlated Single Photon Counting (TCSPC) were used to measure the ToF of the photons between emission and back-reflection from the scene. Instead of measuring the spatial structure of the light, as in the aforementioned methods, the images were reconstructed from the ToF alone, with the temporal data being interpreted by an Artificial Neural Network (ANN) trained from both example ToF histograms and ground-truth 3D images. Figure 1 depicts these input histogram (a) to output 3D images (b) process. The difficulty arises from losing the spatial structure of the scene, resulting in the inverse image retrieval problem becoming heavily ill-posed. As the whole scene is illuminated simultaneously, and all photons are collected onto only 1 pixel, a photon measured at time t may originate from any point on the surface of a spheroid, or if the illumination source right next to a detector, a sphere (see Figure 2). The radius of the sphere of possible reflection point is given by $r = ct$.

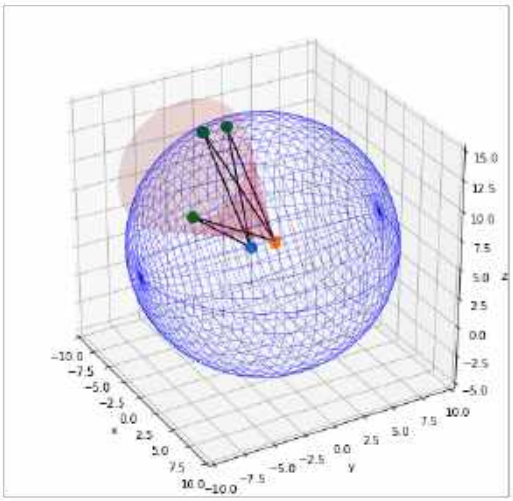

Figure 2. Geometry of the problem. For a given time of flight t, the distance d that the photon travelled is known ($d = ct$). The set of points of fixed travel distance d lie on the surface of a prolate spheroid (blue mesh), whose foci are the photon emitting laser (orange point) and the detector SPAD (blue point). The spheroid becomes a sphere in the limiting case when the SPAD and the laser are at the same point. Conversely, the spheroid becomes more ellipsoidal as the separation between SPAD and laser increases. The set of possible reflection points, then, is the intersection of this spheroid with the illumination beam (light cone shown in red). Three such possible reflection points are shown in green and their associated paths in black.

In other words, just from the arrival time of a photon, it is analytically impossible to determine which point

the light came from. However, that is not to say that the retrieved temporal information is fully uncorrelated to the spatial structure of the scene. Objects (such as people, chairs, cars, etc.) reflect photons with recognisable temporal traces, yet these traces are dependent on the object's orientation, reflectivity, distance from the camera, size, vicinity to other reflective objects and so on. To further complicate the issue, multiple objects can have the same temporal trace. As a result, it is practically infeasible to try to reconstruct objects by implementing a dictionary mapping temporal traces to potential sources. However, Turpin et al.[5] have shown that a machine learning algorithm can identify structures in the temporal signal which correspond to spatial structures in the scene, allowing them to recover a 3D scene from purely temporal data.

This paper presents a novel solution to this temporal imaging problem through the use of an asynchronous sparse processing method. The solution presented in Turpin et al[5] was a standard fully connected ANN with approximately 10 million parameters, which yielded good results. However, this particular problem has many traits in which a neuromorphic approach with Spiking Neural Networks (SNN) could be beneficial. Indeed the ToF sensor is akin to that of a spiking sensor, in that it records single instances of photons returning after a pulse has been transmitted, assigning a time-stamp to each return. This is then similar to a Neuromorphic Vision Sensor,[6] albeit without the other three values of x, y and polarity, but an importance on timing. Typically this sensor would capture thousands if not tens of thousands of photons to get a good distribution of the light reflecting in the scene, then process this histogram of time returns in order to retrieve a 3D depth map, as shown in Figure 1.

The inverse retrieval problem presented has two main areas that this paper looks to improve. First is the re-imaging of the network from a ANN to a variational fully convolutional encoder-decoder network, which is an accurate, high quality image creator,[7,8] that can also be adapted to spiking networks successfully.[9,10] In this case though, the variational term is referring to the translation from the 1D depth domain to the 3D depth map domain, where the transverse positional features are inferred by the network. The second improvement comes in the processing overhead and latency. The processing overhead is seen as the amount of calculations that need to be carried out for the network to return an output. The spiking nature of proposed network Spike-SPI (Spiking - Single Pixel Imager) allows a reduction in information being propagated through the network, thanks to the spiking neurons thresholding ability. In terms of latency, each processing stage can only be complete when the photon counter has reached the desired captured value; this results in a dead time in waiting for the sensor to return enough information before processing, and a similar constraint in waiting for the processing stage to finish before the next batch can run. The asynchronous characteristic of the spiking neuron could help to reduce this by allowing continual processing of direct or buffered inputs.

The remainder of this paper is organised as follows. Section 2 the methodology, cover the simulation and proposed network details. Section 3 details the results of the simulations comparing the proposed Spike-SPI network with the ANN model of Turpin et.al.[5] Section 4 details the broader impact of this work, while Section 5 contains the conclusion.

2. METHODOLOGY

The 3D imaging approach consists of three main elements: i) a pulsed light source, ii) a single-point time-resolving sensor, and iii) an image retrieval algorithm. The scene is flood-illuminated with the pulsed source and the resulting back-scattered photons are collected by the sensor. A single-point SPAD detector, operated together with time-correlated single-photon counting (TCSPC) electronics, forms a temporal histogram [Figure 1(a)] from the photons arrival time. Objects placed at different positions within the scene and objects with different shapes provide different distributions of arrival times at the sensor[11]

In this paper, we trained our neural networks on synthetic data (used by Turpin et al.[5]), which had been designed to simulate the imaging setup described above. The simulations contain humanoid silhouettes in various poses in front of a background consisting of some objects, as illustrated in Figure 3. The silhouettes and background objects form a 20m^3 3D environment, as seen by a simulated 3D camera, where the distance from the virtual camera is encoded in the colour of the scene.

We then find the photon arrival time probability density function (PDF) for the virtual scene. Then, the signal observed by a virtual single-point SPAD can be estimated by convolving the photon arrival time PDF with

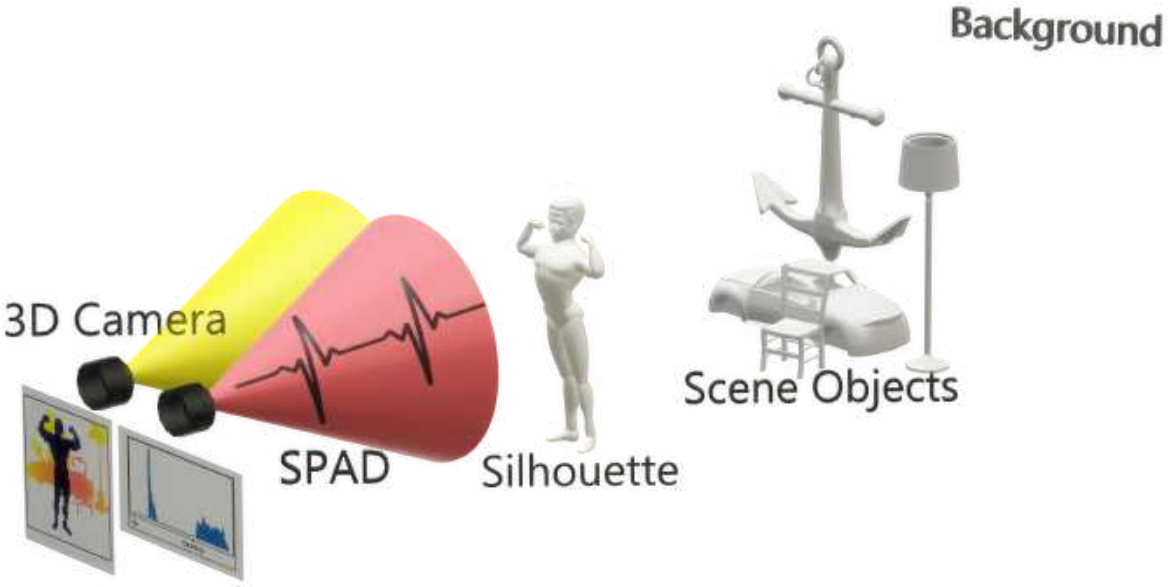

Figure 3. Mock-up of the synthetic scene used, with the 3D camera and SPAD capture the scene in depth image and histogram formats. The scene is made up of the human silhouette and the scene objects of a chair, car, lamp and anchor.

a Gaussian instrument response function (IRF). Finally, assuming that subsequent photon measurements are independent and identically distributed, we simulate n photons detected by the virtual SPAD with TCPSC by sampling the convolved photon arrival time PDF n times, assigning each photon into one of 8000 time bins, with the bin size being 2.3ps. This bin width is convolved with the two IRFs of 20 and 100ps, resulting in values more consistent with practical values of time counting electronics. Photon counts of 1000 and 9500 were also selected to represent a typical amount of photon returns selected with 9500 and a fraction of that to test the ability to batch process smaller amounts of information.

The new variational Encoder-Decoder Spike-SPI network is shown in Figure 4. It comprises 18 layers with 10 layers for the encoder and 8 layers for the decoder. The input comprising the 8000 bin histogram of the captured ToF data. These are processed though the network's convolutional layers as detailed in Table 1. The ANN network[5] used as comparison has 3 fully connect layers with 1024, 512, and 256 nodes respectively. Both networks were given the same training data which was split into 4 different experiments, with the two different photon counts and two different IRFs. This was done to allow a range of testing to mimic some best and worst case scenarios, from the high photon count of 9500 and small time windows of 20ps, through to the 1000 photon count and 100ps time window. This also helps to determine the robustness of the network to differing input conditions. The networks are both trained on the 11600 samples using 29 different silhouettes, with all training and validation testing instances ensuring images and their mirrored version are kept together. These mirrored pairs are kept due to the histogram of silhouettes mirrored on the center of the y axis are identical; however this does not mean the full histogram is identical as different sections of the scene are occluded in each case as illustrated in Figure 5. For final testing, the networks are shown histograms of a previously unseen silhouette in a range of depths and potions in which it has to reconstruct. In total there are 20 x positions, and 10 z positions, for each silhouette all with the exact same scene of objects.

Utilising the feature extraction capabilities of Convolutional Neural Networks (CNN), the high dimensional compressed latent space representations of an Encoder-Decoder network structure and the asynchronous processing capabilities of a SNN, we develop a novel spiking convolutional neural network (SCNN) structure that converts

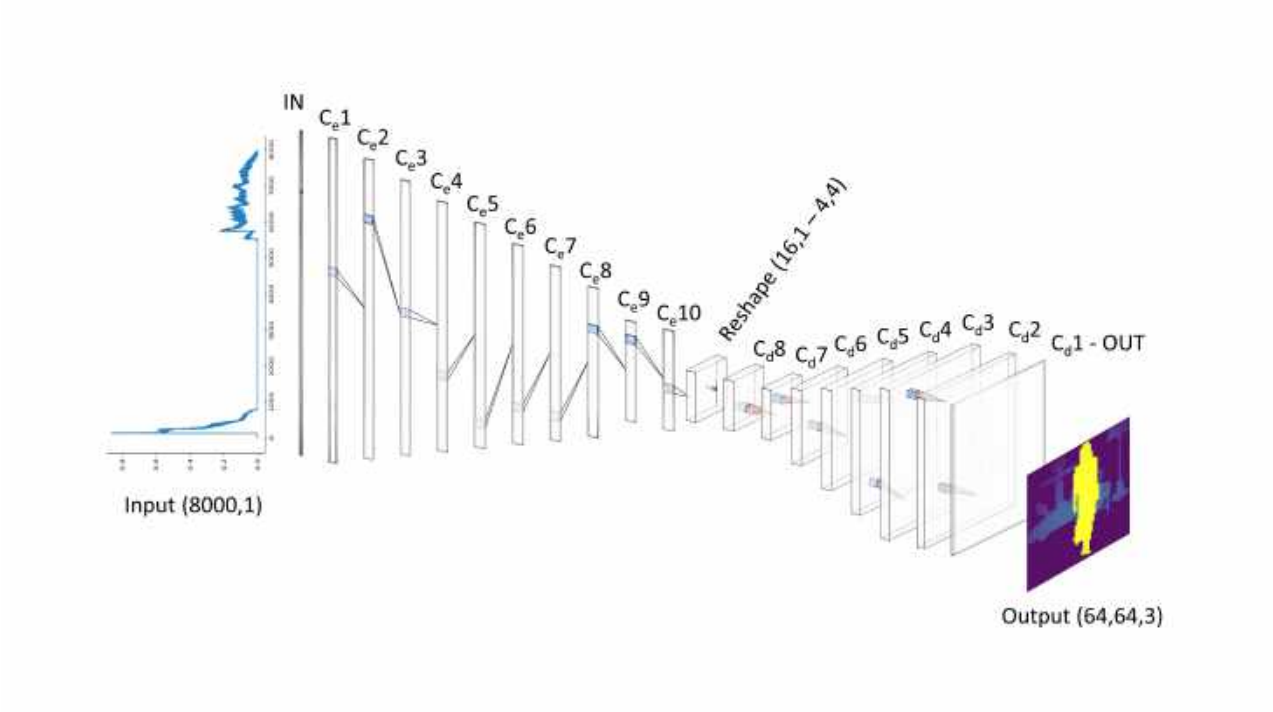

Figure 4. Spike-SPI Network structure with 1D convolutions reshaped into 2D convolutions. Input of 8000 long vector, reshaped at the transformation stage from 16 long vector to 4x4 matrix, before being output as a 64,64 matrix with 3 colour channels.

the 1D depth histograms into a Depth Map. Through the use of 1D convolutions the network is able to encode subtle differences in local spatial regions within the depth histograms into a high dimensional latent space. This allows the subtle differences in the histogram due to the silhouette placement in the scene and the area which it occludes to be captured. This latent space is then decoded by a 2D convolutional decoder, exploiting the strong spatially local correlation present in natural images. All this while the asynchronous processing nature of the spiking neurons allow for a faster throughput and less computational overhead, reinforcing the benefits of the operational sparsity in the single point sensor.

Table 1. Details of the Spike-SPI network

Network	**Spike-SPI**								**ANN**		
Layer	**Ce1**	**Ce2-3**	**Ce4-10**	**Cd8-5**	**Cd4-3**	**Cd2**	**Cd2**	**Up**	**FC1**	**FC2**	**FC3**
Kernel Size	7	7	7	5	5	5	5	2			
Feature Number	64	128	256	256	128	64	1		1024	512	256

The spiking neural network described, is trained as a traditional CNN then converted to a SCNN through the use of Nengo,[12] which is a tool for constructing and simulating neural networks that is similar to TensorFlow. Although Nengo can be used to create TensorFlow[13]-style networks, it has been primarily designed for a different style of modelling: neuromorphic networks. Such networks include features drawn from biological neural networks, in an effort to understand or recreate the functionality of biological brains. Note that these models fall on a spectrum with standard artificial neural networks, with different approaches incorporating different biological features. But in general the structure and parameterisation of these networks often differs significantly from standard deep network architectures. A common characteristic is the use of more complicated neuron models, in particular spiking neurons. In contrast to "rate" neurons (like relu) that output a continuous value, spiking neurons communicate via discrete bursts of output called spikes, which allows both the asynchronous features and lower computational overhead. Within this work NengoDL[14] is used to convert the trained CNN model into

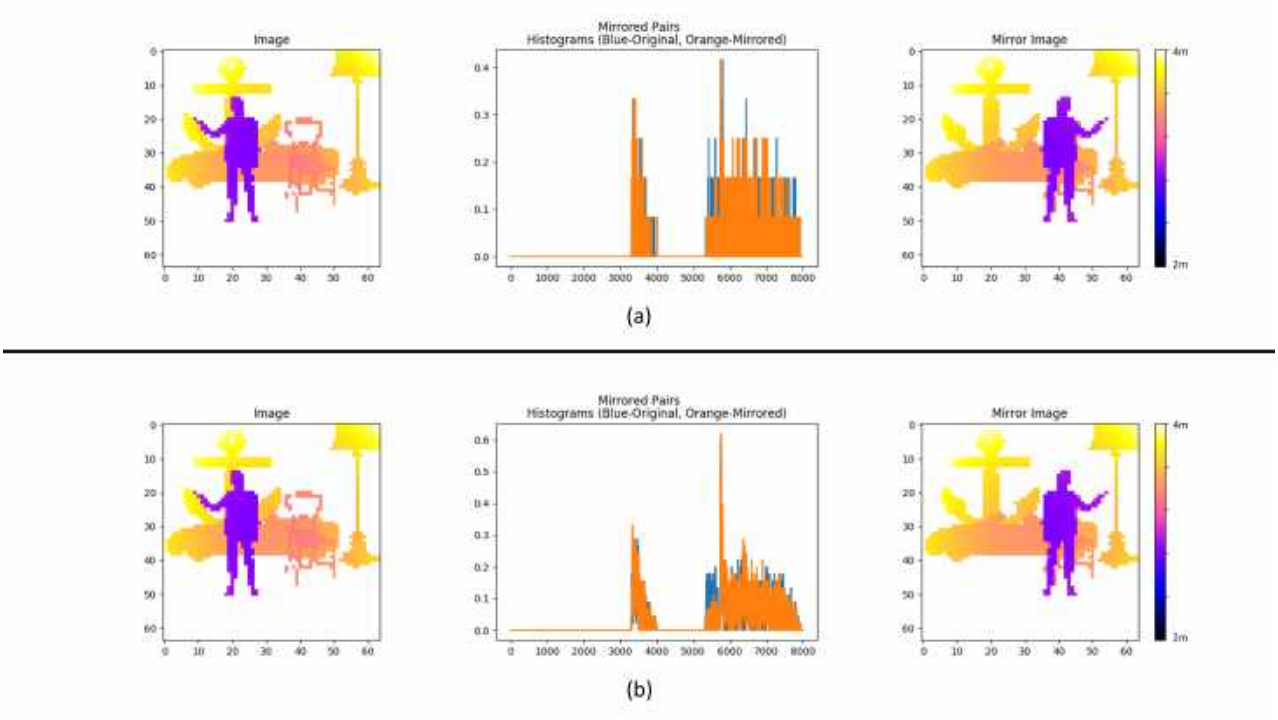

Figure 5. Histograms for both (a) 1000 photons and (b) 9500 Photons, showing the similar silhouette histogram but differing scene histograms.

an asynchronous spiking model, which allow lower computational throughput along with a continual processing approach.

Qualitative aspects from the reconstruction can be gauged visually by comparing the outputs of the two systems, namely the ANN and Spike-SPI. This allows for comparisons on not only the structural shapes of the object resolved, but the associated depth at which they reside, encoded within the colour data of the depth map. Along with qualitative measures, a series of quantitative measures are also proposed within the less common metrics equations given as follows. The restoration quality was evaluated using the reconstruction signal-to-noise ratio (RSNR),[15,16] which is a metric that has been used for depth image reconstruction on sparse single photon data.

$$\text{RSNR} = 10\,\log_{10}\left(\frac{\|y\|^2}{\|y - y^*\|^2}\right) \tag{1}$$

where y is the predicted depth map and y^* is the ground truth. $\|y\|^2$ is the ℓ_2 norm given by $y^T y$.

The other metrics are typical quantitative measures in depth estimation.[17] The equations for the other chosen metrics: SNR, AbsRel, SqRel and RMSE can be found in the appendix. Hoever the less common metrics are listed as follows.

The scale invariant log root mean squared error, si-logRSME[17] is calculated with

$$\text{si-logRMSE} = \frac{1}{2|T|}\sum_i \left(\log\, y_i - \log\, y_i^* + \frac{1}{|T|}\sum_i (\log\, y_i^* - \log\, y_i)\right)^2 \tag{2}$$

For any prediction y, $\frac{1}{n}\sum_i (\log\, y_i^* - \log\, y_i)$ is the scale that best aligns it to the ground truth. All scalar multiples of y have the same error, hence the scale invariance.

The accuracy score is set through the number of pixels that remain within a threshold such that

$$\% \text{ of } y_i \mid \max\left(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}\right) = \delta < thr \tag{3}$$

where δ is compared to three set threshold 1.25, 1.25^2, 1.25^3, with each value allowing a greater depth error to count in the accuracy.[17]

Lastly, the comparison metrics also include the Intersection over Union (IoU), that allows the comparison of the foreground objects of the ground truth y^* to be compared with the with the reconstructed foreground objects in y. This is done through first masking the foreground objects from the background. The background in these experiments are uniform and at maximal distance from the sensor. In order to threshold this and mask the foreground elements, the background is set at everything greater than 99% of the max distance. This means the constant scene objects and the silhouette mask are compared against the ground truth version to see how many pixels are correctly identified as belonging to an object. Meaning it is essentially looking for background pixels that have been incorrectly given a depth punishing a blurred edge smoothing or averaging approach. The IoU is calculated with the following[18]

$$IoU(y_{mask}, y^*_{mask}) = \frac{y_{mask} \wedge y^*_{mask}}{y_{mask} \vee y^*_{mask}} \tag{4}$$

with y_{mask} and y^*_{mask} representing the masked reconstruction and ground truth images respectively.

The chosen metrics include some of the typical full image comparison measures, however it is apparent that due to the averaging nature across some of those metrics, if the network learns to blur across the areas of interest, on average it will receive a good score. This results in overall depth estimation being compared and not the spatial reconstruction. This is why there are a number of relative metrics and a IoU scoring system to test the two models outputs. As the testing should indicate which model best recovers the shape and depth of the silhouette and scene objects.

3. RESULTS

This sections details the results from the 4 experimental setups with photon counts being 9500 and 1000, while the IRF is set to 100ps or 20ps. The results of both networks are first tested on the validation test data and then on the unseen test data, to compare their depth map reconstruction abilities. Results from the unseen testing scenario are found in Table 2 comparing the two models, Spike-SPI and the ANN from Turpin et al.,[5] against all the metrics detailed in the methodology.

Table 2. Results from the experimental testing split into 4 sections with 2 photon count values of 1000 and 9500 and two IRF for each of those at 100ps and 20ps.
Best (Bold), *best in test (Italic)*, higher is better (↑), lower is better (↓)

	Photon Count	**IRF** *ps*	**IoU** ↑	**SNR** *dB* ↑	**R-SNR** *dB* ↑	**absRel** ↓	**sqRel** ↓	**RMSE** ↓	**si-log RMSE** ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
ANN	1000	100	0.650	*14.844*	2.880	22.074	4.444	*0.189*	0.474	0.853	0.886	*0.908*
Spike-SPI			***0.783***	14.284	***6.502***	*3.597*	*1.323*	0.201	*0.456*	*0.871*	***0.890***	0.906
ANN		20	0.650	*14.708*	2.890	26.354	5.142	*0.192*	0.476	0.853	0.886	*0.908*
Spike-SPI			*0.760*	14.155	*5.842*	*6.571*	*1.502*	0.202	*0.456*	*0.868*	*0.889*	0.906
ANN	9500	100	0.637	***15.076***	2.614	20.558	3.846	***0.187***	0.468	0.856	0.889	*0.909*
Spike-SPI			*0.780*	14.391	*6.360*	***3.066***	***1.163***	0.198	*0.456*	***0.871***	*0.890*	0.905
ANN		20	0.631	*15.070*	2.500	15.124	2.559	*0.188*	0.470	0.856	0.888	***0.910***
Spike-SPI			*0.778*	14.424	*6.358*	*3.461*	*1.438*	0.198	***0.456***	*0.870*	*0.888*	0.904

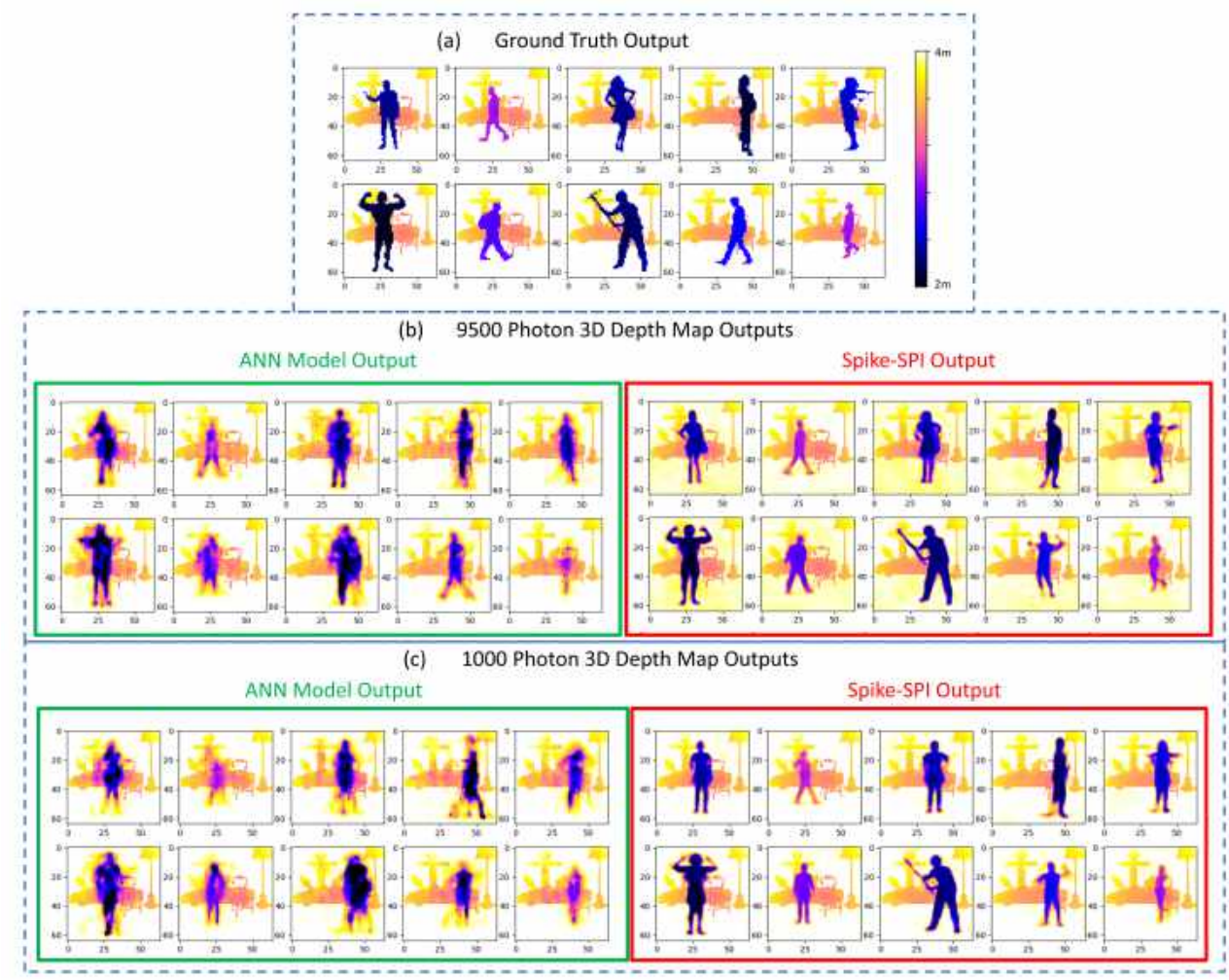

Figure 6. Outputs of the two networks, ANN and Spike-SPI for (b) 9500 Photon, 20 IRF and (c) 1000 Photon, 100 IRF validation data. With (a) the ground truth data displaying 10 examples of silhouettes in the scene at the top. ANN in green on the left and Spike-SPI in red on the right.

The experimental results are broken into two sections, with qualitative results for the validation testing and both qualitative and quantitative results for the unseen testing data, where as highlighted in Table 2 Spike-SPI scores better on all metrics other than the SNR and RMSE. However, within the qualitative results it is clear as to why this is the case, when looking at the precision of both reconstructions. For all the other metrics tested, Spike-SPI was able to achieve better reconstruction results especially in metrics that focus on the spatial reconstruction such as the IoU, where an increase of over 0.12 on average represents a considerable increase in accuracy. The accuracy metrics shown in terms of $\delta < 1.25^{1,2,3}$ as seen in (3) are a great indicator of the why some of the full image averaging metrics score the ANN higher than Spike-SPI. As the threshold for δ increases this is allowing a greater inaccuracy to count as correct on a per pixel basis, showing that with the tightest threshold of 1.25 (25% relative difference between reconstruction and ground truth) that Spike-SPI scores better across all tests. However, when increasing the threshold to 1.25^2, the accuracy only slightly favour Spike-SPI and, with the largest threshold 1.25^3, the ANN scores higher. Further insight into these values based on the accuracy metrics would appear to show that around 87% of the pixels in the image have a depth estimate that is very accurate, while the remaining are considerably off. It is these misclassifications of pixels that brings the average score of Spike-SPI down on the full image metrics. Where a high precision, without 100% accuracy leads to

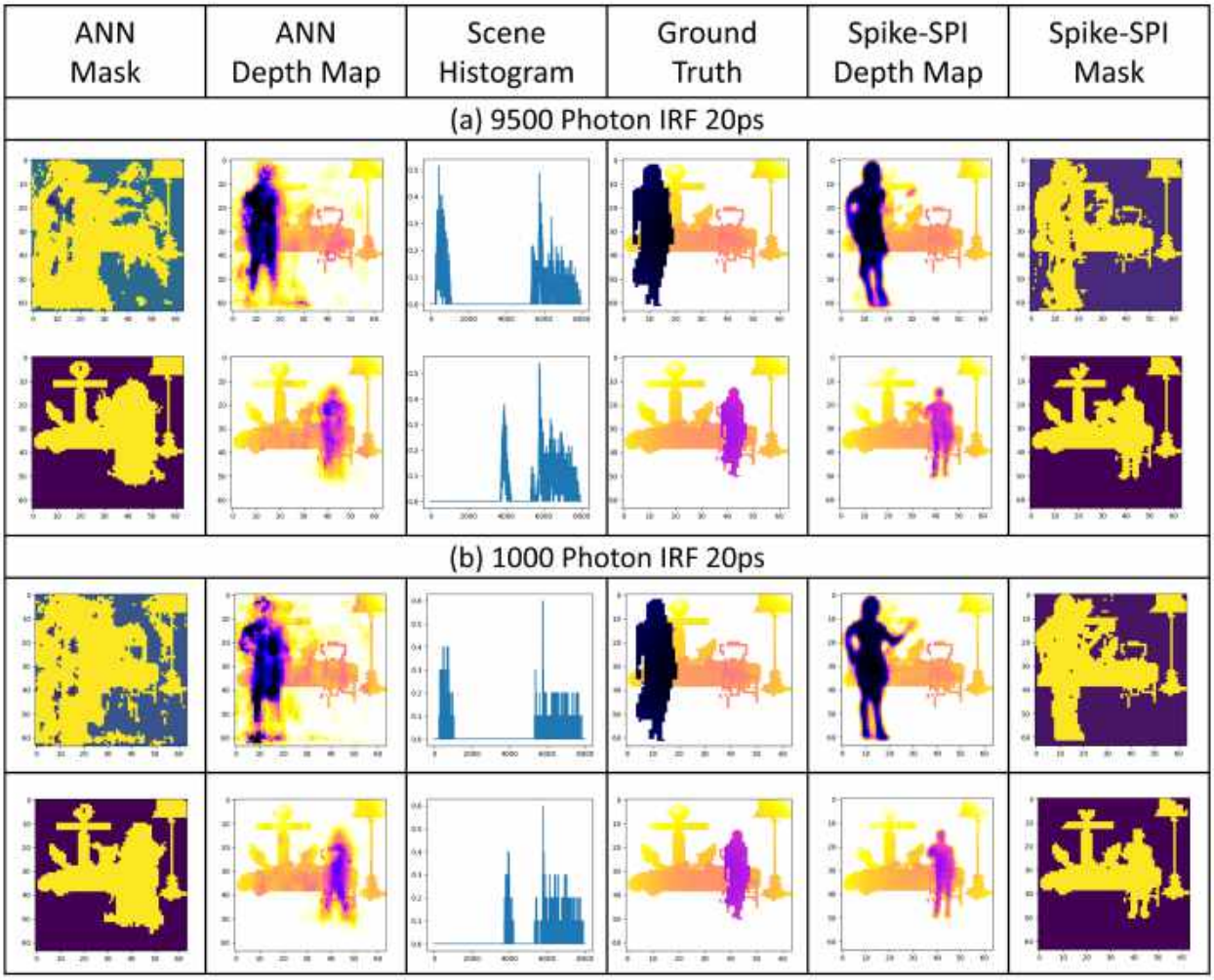

Figure 7. Results on unseen silhouette data with reconstructions shown within depth map columns. Results of IoU masking also shown for this test data within the Mask columns. Testing shown for both (a) the 9500 Photon, 20 IRF and (b) the 1000 Photon, 100 IRF data.

instances where the majority of the pixels may be correct but those that are incorrect have a large error value. This can be seen as the manifestation of Spike-SPIs convolution layers for decoding, guessing which silhouette is in the image from the compressed latent space representation. This silhouette guess then turns into a precise reconstruction which inevitably has errors.

This characteristic described previously are seen in Figures 6 (validation data) and 7 (unseen test data). Within both figures if the silhouette has a different outline from the ground truth, or is attempting to guess an unseen outline, the areas in which the reconstruction are incorrect happen to be a high magnitude depth error. This is due to the pixels normally belonging to the background which is set to max depth being misclassified. However, this resulting large depth error might actually be only 1 or 2 pixels off in terms of spatial error. Ultimately this results in metrics that favour depth accuracy over spatial accuracy, giving a better score to the ANN model over Spike-SPI. As a result close attention has to be paid to what the actual values of the quantitative results in Table 2 actually mean. The results of this spatial or depth focus is shown within the results illustrated in Figures 6 and 7, where the ANN reconstructions with the blurred edges of the estimated silhouettes favours depth, while the

convolution process of Spike-SPI reduces this blurring significantly favouring spatial errors.

Looking closer at the qualitative results in Figure 6, which depicts the ground truth images shown above the ANN (left) and Spike-SPI (right), it can be seen that Spike-SPI has a higher spatial acuity, highlighted within Figure 6 as the silhouette holding the pic axe, where both the arms and the handle of the axe have been well resolved. This acuity is also still captured within the 1000 photo 6 (a) data as well as the 9500 photo data 6 (b). It is noticeable the deterioration in the spatial outline of the silhouettes within the ANN data, when comparing the 1000 and 9500 photon results of Figure 6 (a) and (b) respectively. This consistency across a lower photon count is also reflected within the quantitative results in Table 2, where the results of Spike-SPI have less of a spread than the ANN across the 4 experiments.

Figure 7 illustrates the results of the unseen silhouette testing, while also illustrating the results of the masking process for the IoU measurement. The masking process not only highlights the spatial acuity of the Spike-SPI model compared to the ANN, but serves as a visual explanation of the accuracy results with the increasing thresholds. Within both photon counts shown in Figure 7 (a) and (b), Spike-SPIs results consistently mask a smaller percentage of the scene as foreground. This helps to visualise this acuity that penalises the method on some metrics, as it is apparent that is has misclassified some of the pixels when compared to the ground truth. This resulting high error value across a small number of pixels is in contrast to the ANN model, which has a lower spatial acuity as seen in the mask, but it also has a lower error value across a larger number of pixels. Since within this task our objective is to spatial resolve from depth measurements, Spike-SPI can be seen to quantitatively and qualitatively outperform the ANN.

3.1 Spiking Benefits

From the results discussed so far, the main benefits of the Spike-SPI network stem from the CNN approach of the SCNN and not the spiking elements. To see the results of the spiking neurons the histograms must be feed into a simulation with an asynchronous processing pipeline. From this simulation the results of a spiking approach are illustrated in Figure 8 (a), showing that not only is Spike-SPI better at image reconstruction, due to the CNN Encoder-Decoder structure, but it can achieve this with less processing power. Figure 8 (a) highlights this reduction in processing power with the displaying of the spiking activity in the networks over 60 time steps within the simulation. Figure 8 (a) shows an average neuron firing rate of 1Hz with a maximum neuron rate of 150Hz, while the overall activity rate is only 11%. This method also has the ability to scale the neurons firing rate, meaning more processing power being drawn, but as a result the simulation requires less steps to produce an image as shown in Figure 8 (b), where now the image is mostly formed by the 18th time step, and by the 24th it is resolved but with the background needing to settle. Figure 8 (b) also shows that some of the neurons are now firing twice during the activity, with similar higher activity in the mean, max and activity rate overall. Even with this increase only a quarter of the overall neurons available are active. This ability to process with less neurons is in fact a negative characteristic of the CNN structure, that the SCNN can exploit. Often within a CNN many of the neurons are not propagating useful information forward, in that they are only reporting a very small similarity of the kernel within a location of the image. As highlighted in Figure 8, the majority of neurons in our model at any given time are inactive. However, within the typical CNN approach there is no neuron threshold to stop the forward propagation of this information. This reduction in information propagation can lead to a reduction in reconstruction accuracy if the hyper-parameters of the spiking neurons are not correctly set. Although, throughout the testing in this paper no conceivable difference was recorded between the CNN and SCNN models in both qualitative and quantitative terms, and as such the CNN results are not compared.

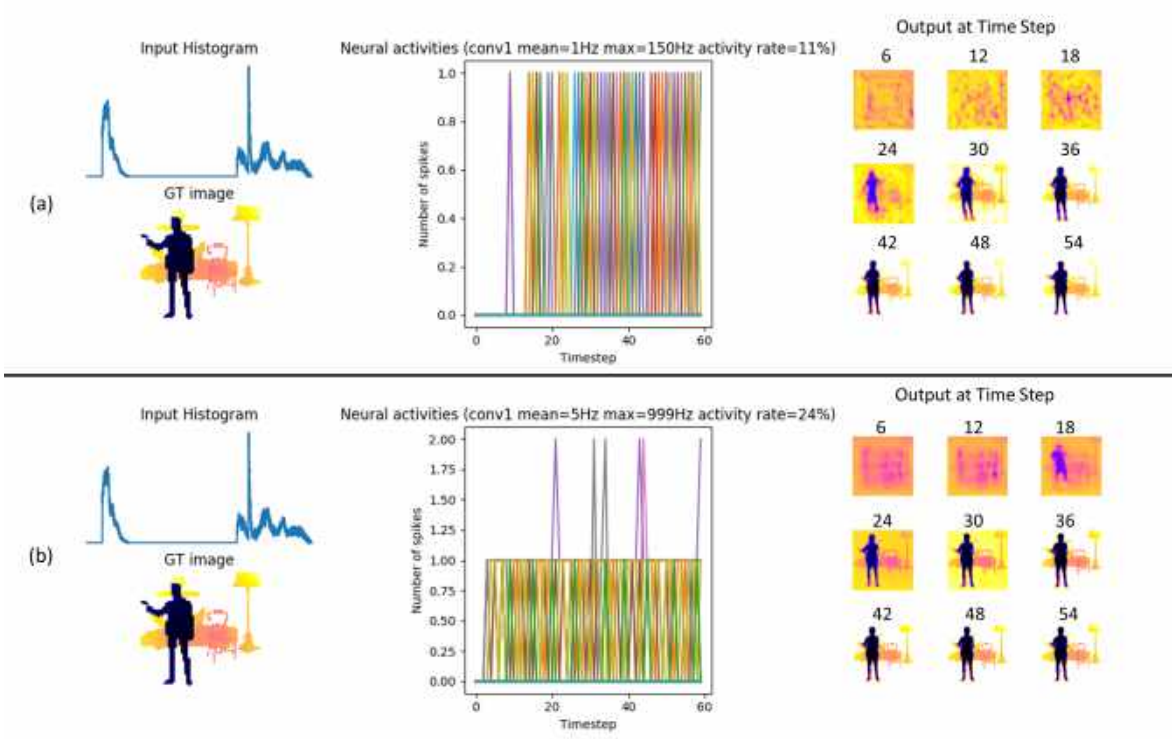

Figure 8. Visualised results of the spiking impact on the network, with the spike rates of a cross section of neurons shown within the middle section called Neural Activities, for the given input histogram with corresponding ground truth image. The figure also show the output of the spiking network at given time steps, as the information is processed in an asynchronous manner. (a) illustrates a lower firing rate while (b) illustrates a higher firing rate with faster processing but more neural activity.

3.2 Feature Visualisation

This sections is used to form insight and understanding into what the SCNN network is processing. That is what information does the network use to help understand the histogram and convert it into a depth map. Understanding of the network can be formed through weight and feature map visualisation. Figure 9 shows the weights and features maps associated with Spike-SPI from the Photon Count 9500, IRF 20, where weights and feature maps from 4 layers are illustrated. The first of these is the Conv 1 Encoding layer of the encoding, which looks at the histogram and tries to find useful spatial features that can help to describe the histogram within a windowed area covering 7 time bins. Figure 9 displays the corresponding weights from this process and the resulting feature map of each of the weights. The feature maps at this point look very similar as the kernel size of 7 compared to 8000 is relatively small. In the Conv 9 layer of the network the relative size of the kernel to the start histogram is considerably larger with each kernel now covering around 500 time bins. The resulting feature maps now depict a latent representation of the original histogram, with feature maps highlighting if the feature belonged to an object that was shallow of deep within the scene. It can also be seen within Figure 9 that some of the feature maps of Conv 9 have a horizontal line meaning no useful information from this weight is found. Considering 3 of the 24 maps shown contain this out of the 256 available it is apparent how the spiking neuron can reduce the neuron activity of the network. The second half of Figure 9 illustrates the 2D convolution weights and feature maps of the decoding process.These 2D weight decode both the spatial and depth information from the latent space, with the feature maps showing the output of the network for the according weight. Conv 4 has a much more subtle output compared to Conv 1 Decoding, which the resulting feature maps highlight the weight interest area. That being either the silhouette, scene objects or background. Similar to the encoding process the decoder also has some neurons that are essentially inactive depicted with all back feature maps. The image associated with these activation is the same as the top image in Figure 7, which is why the Conv 1 decoding

feature maps appear to show the silhouette of a woman on the left of the scene.

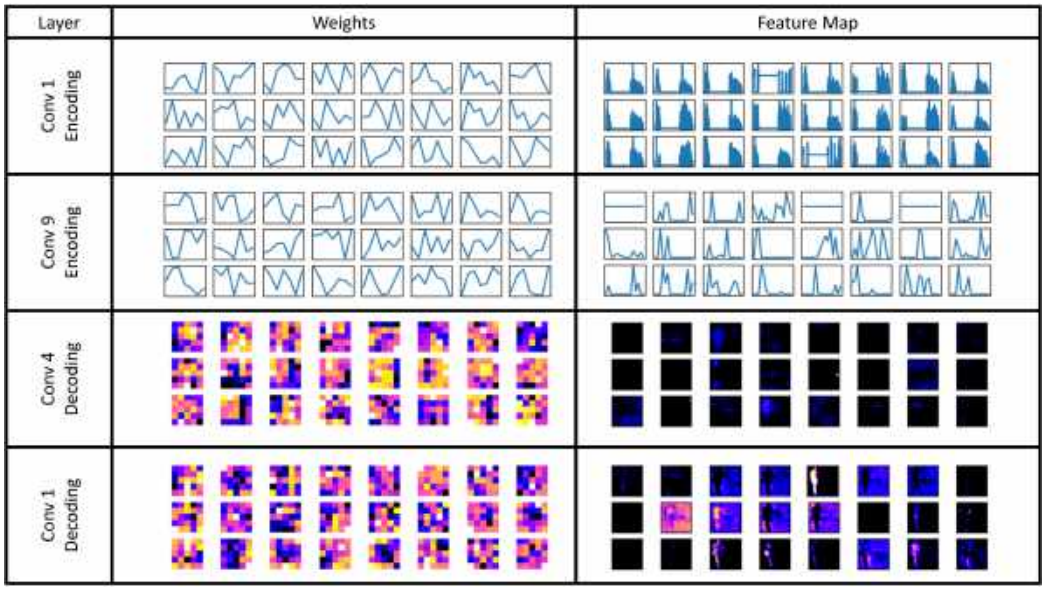

Figure 9. Showing selected weight and feature map results from the 9500 Photon, 20ps IRF data. From top to bottom depicts the encoding early and late stage layers weights and feature maps, then a late and early stage decoding layers weights and feature maps. Feature maps are the results of the activation of the weights to a given input, with the equivalent weight and feature map being collocated within respective columns.

4. CONCLUSION

Throughout this paper novel approach to depth imaging is shown to be able to outperform the previous state of the art. Spike-SPI not only delivers better spatial reconstruction and depth estimation within the depth maps but is able to do so in a asynchronous spiking manner. This allows not only a theoretical reduction in processing time, but an actual reduction in the amount of processing power required to produce an image, thanks to less neuron activity and the ability to produce depth map with less photons captured. These characteristic are illustrated with Spike-SPI being able to resolve the scene with less photons to a higher fidelity, while only using a fraction of the computations. Through utilising multiple aspects of a variety of machine and deep learning approaches, Spike-SPI is able to exploits the useful characteristic of these approaches while offsetting the drawbacks. This research highlights the benefits of a pragmatic approach to problem solving, utilising benefits of many system to deliver state of the art results.

ACKNOWLEDGMENTS

Thanks to Leonardo MW Ltd for helping bring about the collaboration. We acknowledge funding from Alexander von Humboldt-Stiftung, Engineering and Physical Sciences Research Council (EP/M01326X/1) and Amazon Web Services.

5. BROADER IMPACT

The experiments were carried out in scenes where objects were moving in front of a static scene. This makes our approach well suited for applications where the device needs to be placed at a fixed position during operation, i.e. with a fixed scene. There are multiple situations where operating in a fixed environment is useful. Examples are surveillance and security in public spaces, etc. These are examples where the scene and background (e.g. walls of the room, buildings) do not change at all and they are also very widespread scenarios. Currently, cities have spaces that are constantly monitored with CCTV cameras that also potentially record information from which it

is possible to extract information that breaches data protection policies. Our approach is therefore useful for cases where one requires human activity in a fixed area and in a data-compliant way. The approach shown here would be also valid in a slowly changing environment, where training could in principle be continuously updated. Indeed, background objects within the scene will appear static if they change at a slower rate (and/or are at a larger distance) with respect to the dynamic elements of the scene or slower than the acquisition rate of the sensor. An interesting route for future research is of course to also investigate methods that account for dynamic scenes, especially considering the new ability of asynchronous processing at lower photo return counts, would allow for a much shorter determination of what is relatively 'static' and 'dynamic'.

REFERENCES

[1] Barnard, S. T. and Fischler, M. A., "Computational stereo," ACM Comput. Surv. **14**, 553–572 (Dec. 1982).

[2] Frauel, Y., Naughton, T. J., Matoba, O., Tajahuerce, E., and Javidi, B., "Three-dimensional imaging and processing using computational holographic imaging," Proceedings of the IEEE **94**(3), 636–653 (2006).

[3] Sun, B., Edgar, M. P., Bowman, R., Vittert, L. E., Welsh, S., Bowman, A., and Padgett, M., "3d computational imaging with single-pixel detectors," Science **340**(6134), 844–847 (2013).

[4] Sun, M.-J., Edgar, M. P., Gibson, G. M., Sun, B., Radwell, N., Lamb, R., and Padgett, M. J., "Single-pixel three-dimensional imaging with time-based depth resolution," Nature communications **7**(1), 1–6 (2016).

[5] Turpin, A., Musarra, G., Kapitany, V., Tonolini, F., Lyons, A., Starshynov, I., Villa, F., Conca, E., Fioranelli, F., Murray-Smith, R., and Faccio, D., "Spatial images from temporal data," Optica **7**, 900–905 (Aug 2020).

[6] Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S. Davison, A. J., Conradt, J., Daniilidis, K., and Scaramuzza, D., "Event-based vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[7] Garg, R., Vijay Kumar, B. G., Carneiro, G., and Reid, I., "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in [Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)], **9912 LNCS**, 740–756, Springer Verlag (2016).

[8] Zheng, C., Cham, T. J., and Cai, J., "T2 Net: Synthetic-to-realistic translation for solving single-image depth estimation tasks," in [Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)], **11211 LNCS**, 798–814, Springer Verlag (sep 2018).

[9] Panda, P. and Roy, K., "Unsupervised regenerative learning of hierarchical features in Spiking Deep Networks for object recognition," in [Proceedings of the International Joint Conference on Neural Networks], **2016-Octob**, 299–306, Institute of Electrical and Electronics Engineers Inc. (oct 2016).

[10] Kirkland, P., Caterina, G. D., Soraghan, J., and Matich, G., "SpikeSEG: Spiking segmentation via STDP saliency mapping," in [Proceedings of the International Joint Conference on Neural Networks], (jul 2020).

[11] Caramazza, P., Boccolini, A., Buschek, D., Hullin, M., Higham, C. F., Henderson, R., Murray-Smith, R., and Faccio, D., "Neural network identification of people hidden from view with a single-pixel, single-photon detector," Scientific Reports **8**, 11945 (dec 2018).

[12] Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T., Rasmussen, D., Choo, X., Voelker, A., and Eliasmith, C., "Nengo: a Python tool for building large-scale functional brain models," Frontiers in Neuroinformatics **7**(48), 1–13 (2014).

[13] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X., "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015). Software available from tensorflow.org.

[14] Rasmussen, D., "NengoDL: Combining deep learning and neuromorphic modelling methods," arXiv **1805.11144**, 1–22 (2018).

[15] Kang, Y., Li, L., Liu, D., Li, D., Zhang, T., and Zhao, W., "Fast long-range photon counting depth imaging with sparse single-photon data," IEEE Photonics Journal **10** (jun 2018).

[16] Halimi, A., Altmann, Y., McCarthy, A., Ren, X., Tobin, R., Buller, G. S., and McLaughlin, S., "Restoration of intensity and depth images constructed using sparse single-photon data," in [European Signal Processing Conference], **2016-Novem**, 86–90, European Signal Processing Conference, EUSIPCO (nov 2016).
[17] Eigen, D., Puhrsch, C., and Fergus, R., "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," 2366–2374, Curran Associates, Inc. (2014).
[18] Jaccard, P., "Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines," Bull Soc Vaudoise Sci Nat **37**, 241–272 (1901).

APPENDIX A. FURTHER EQUATIONS FOR COMPARISON METRICS

The Signal to Noise ratio is used to determine the amount of noise exists in the reconstruction compared to the ground truth depth image.

$$\text{SNR} = 10 \log_{10}\left(\frac{y^*}{y^* - y}\right) \tag{5}$$

where y is the predicted depth map and y^* is the ground truth.

The absolute relative error is calculated with

$$\text{Abs Rel} = \frac{1}{|T|}\sum_{y \in T}\frac{|y_i - y_i^*|}{y^*} \tag{6}$$

where each n^{th} pixel is indexed by i to form y_i, y_i^*, giving a per pixel metric. T is the total number of pixels per image.

The squared relative error is calculated with

$$\text{Sq Rel} = \frac{1}{|T|}\sum_{y \in T}\frac{\|y_i - y_i^*\|^2}{y^*} \tag{7}$$

The root mean squared error is calculated with

$$\text{RMSE} = \sqrt{\frac{1}{|T|}\sum_{y \in T}\|y_i - y_i^*\|^2} \tag{8}$$

APPENDIX B. FURTHER IMAGES

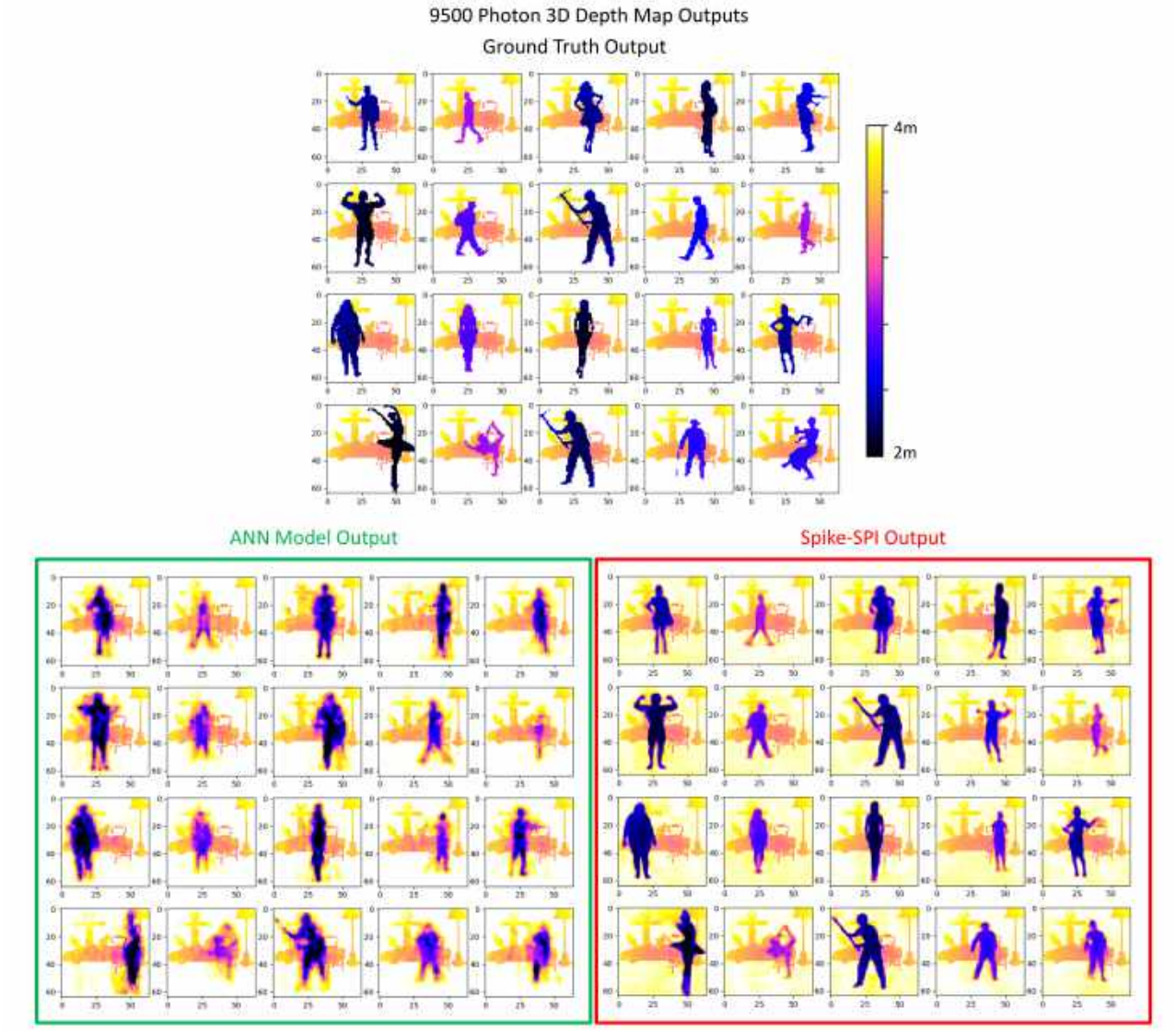

Figure 10. Full image of the result from the 9500 photon count 20ps IRF data

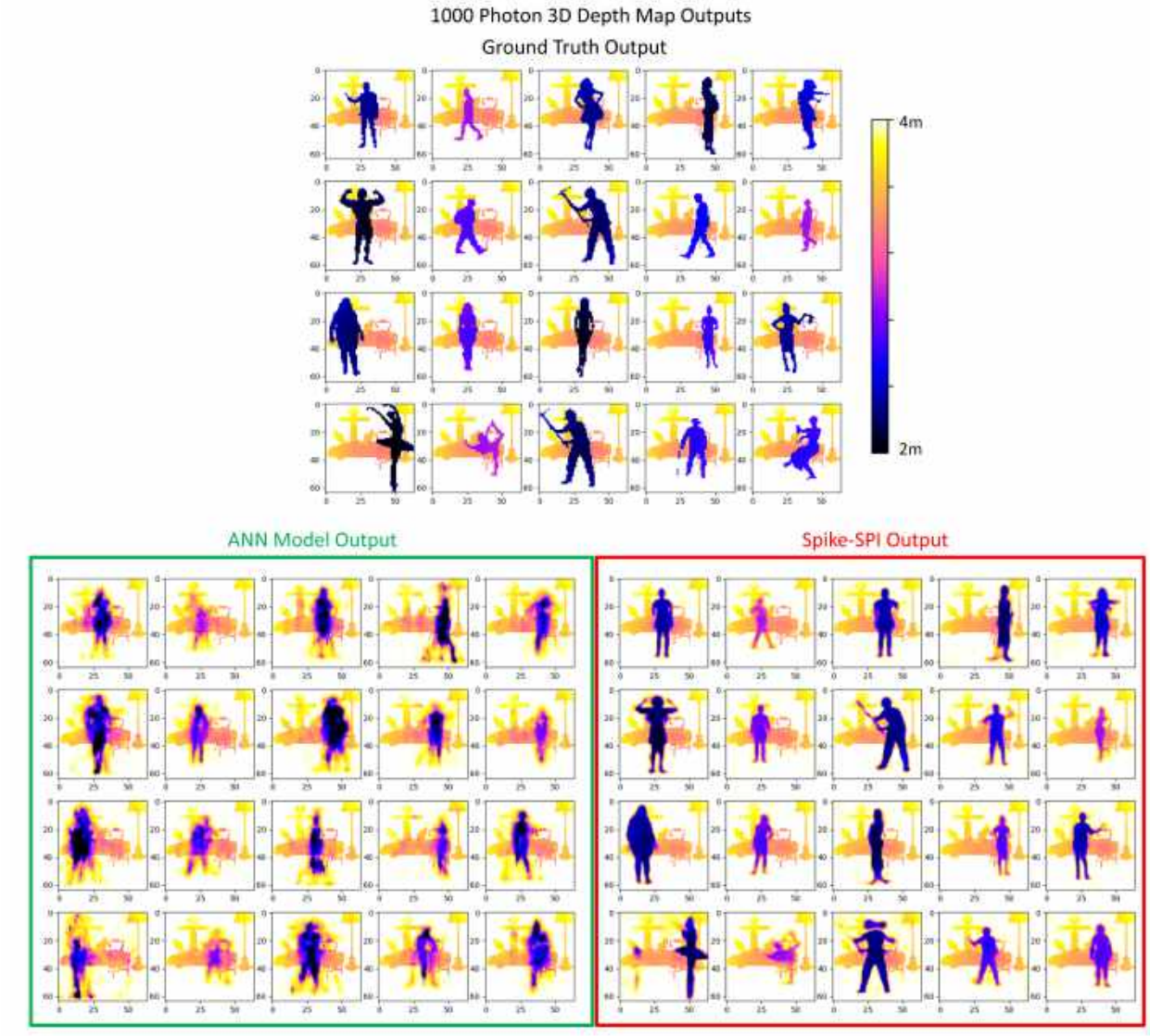

Figure 11. Full image of the result from the 1000 photon count 100ps IRF data

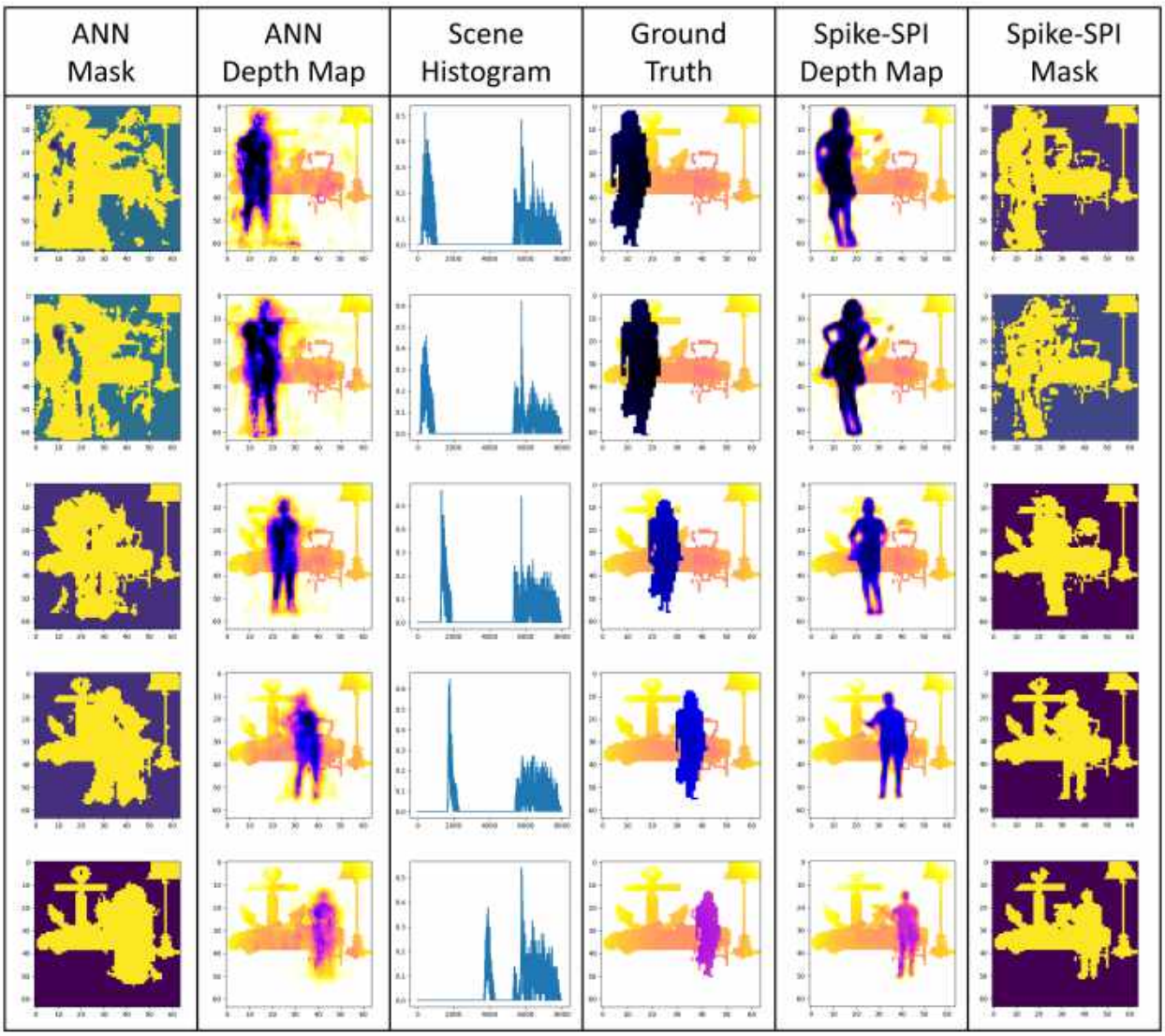

Figure 12. Full image of the result from the 9500 photon count 20ps IRF data for the unseen testing data including the masking for the IoU results

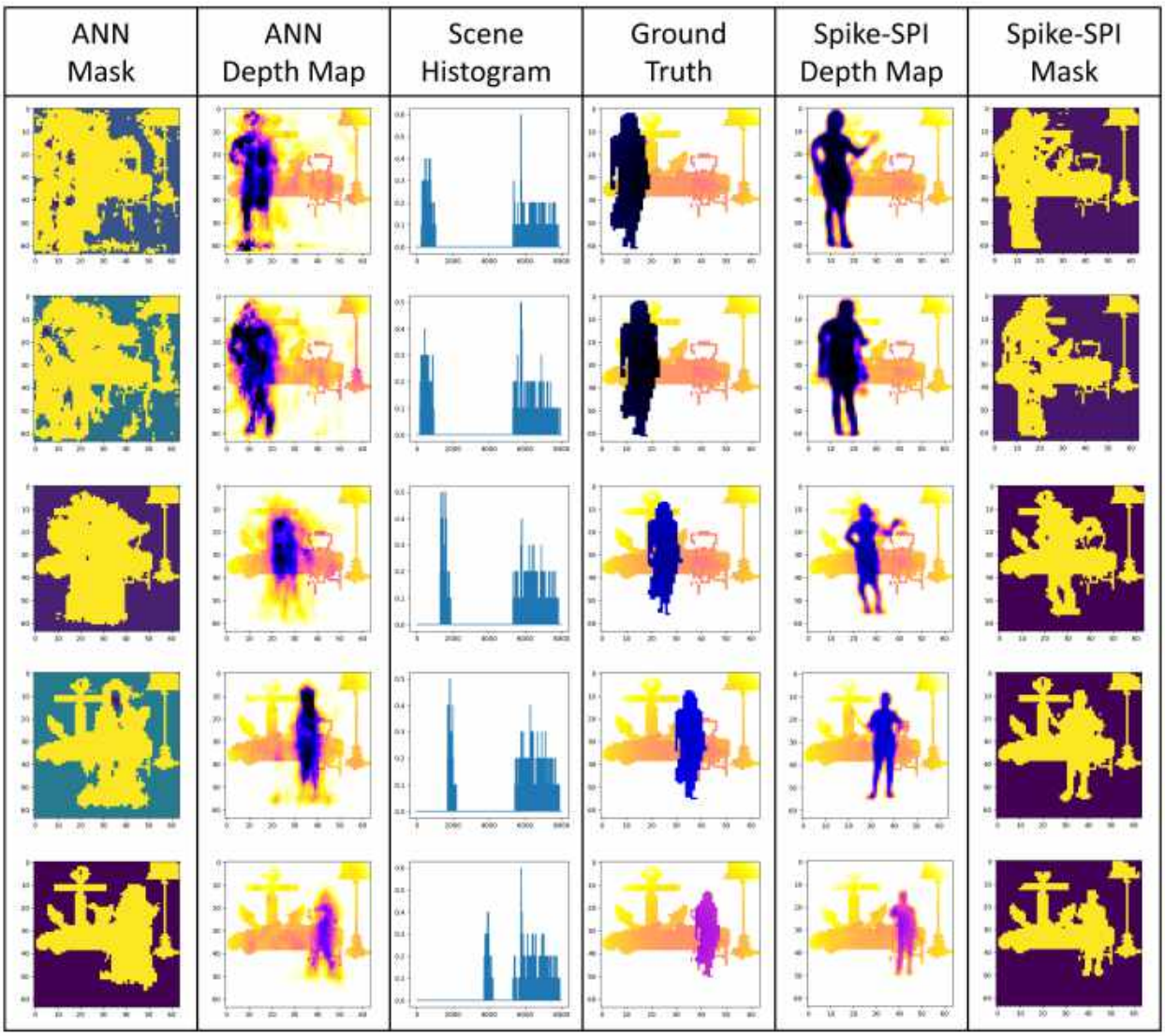

Figure 13. Full image of the result from the 1000 photon count 100ps IRF data for the unseen testing data including the masking for the IoU results

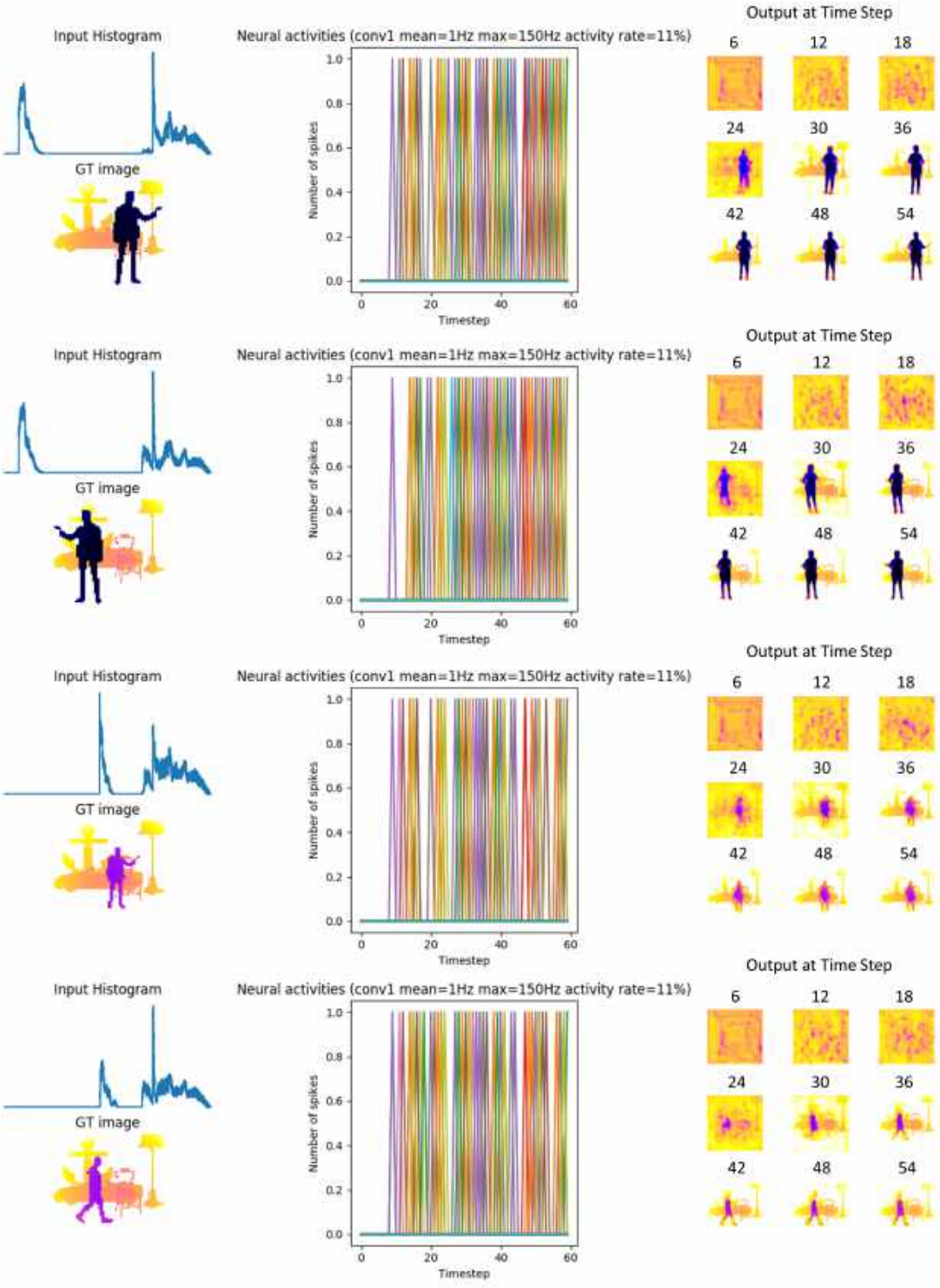

Figure 14. Spiking Influence on the Network for more examples showing similar results across all tests

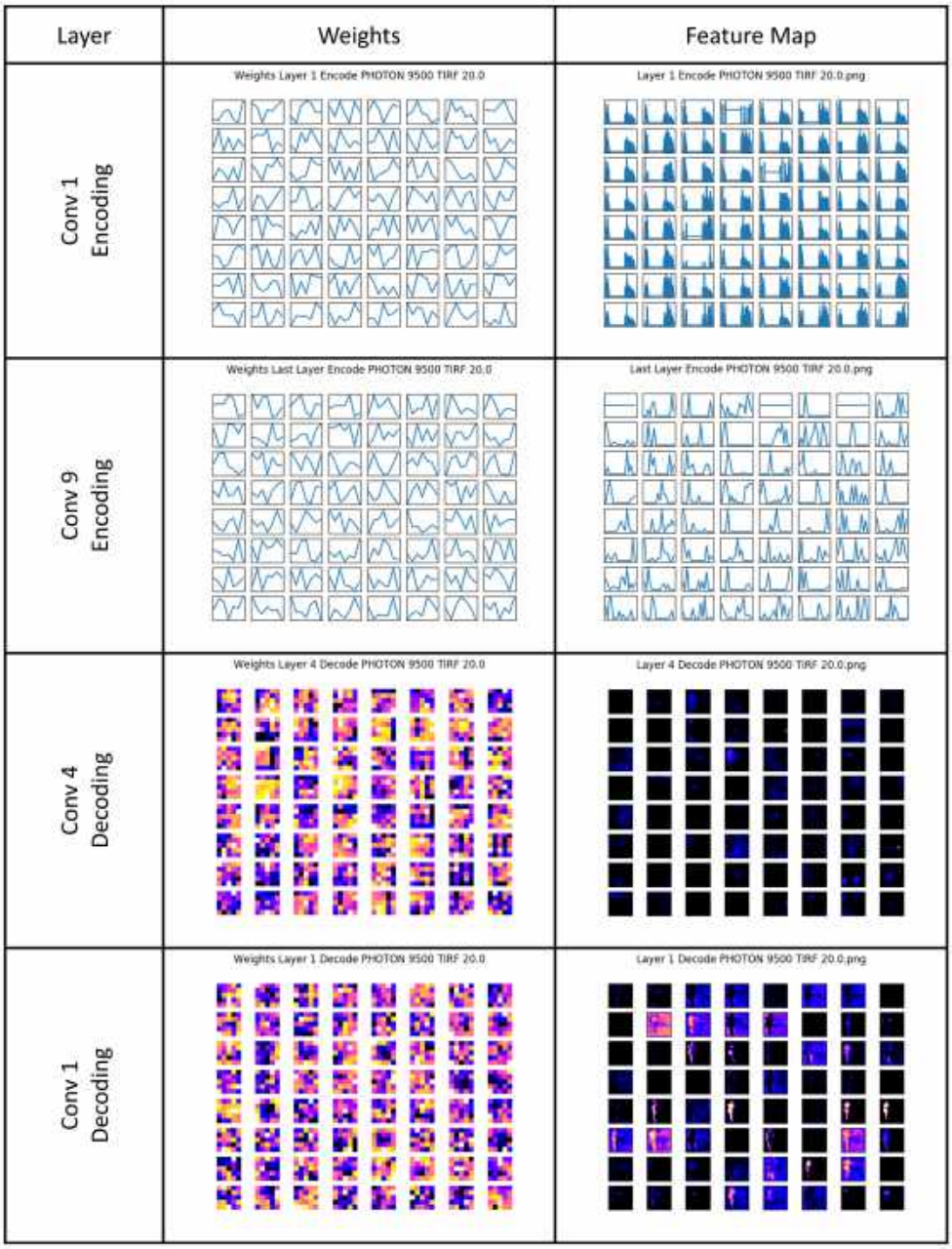

Figure 15. Weight and Feature Map visualisation for the best Spike-SPI network with data from 9500 photons and 20ps IRF. This image helps to show the similarities and difference between the learned features of each of the individual kernels and what areas they help to encode/decode