

StereoSpike: Depth Learning with a Spiking Neural Network

Ulysse Rançon, Javier Cuadrado-Anibarro, Benoit R. Cottureau, Timothée Masquelier
CerCo CNRS UMR 5549, Université Toulouse III
Toulouse, France

ulysse.rancon@u-bordeaux.fr, {javier.cuadrado, benoit.cottureau, timothee.masquelier}@cnrs.fr

Abstract

Depth estimation is an important computer vision task, useful in particular for navigation in autonomous vehicles, or for object manipulation in robotics. Here, we propose to solve it using StereoSpike, an end-to-end neuromorphic approach, combining two event-based cameras and a Spiking Neural Network (SNN) with a slightly modified U-Net-like encoder-decoder architecture. More specifically, we used the Multi Vehicle Stereo Event Camera Dataset (MVSEC). It provides a depth ground-truth, which was used to train StereoSpike in a supervised manner, using surrogate gradient descent. We propose a novel readout paradigm to obtain a dense analog prediction—the depth of each pixel—from the spikes of the decoder. We demonstrate that this architecture generalizes very well, even better than its non-spiking counterparts, leading to state-of-the-art test accuracy. To the best of our knowledge, it is the first time that such a large-scale regression problem is solved by a fully spiking neural network. Finally, we show that low firing rates ($<10\%$) can be obtained via regularization, with a minimal cost in accuracy. This means that StereoSpike could be efficiently implemented on neuromorphic chips, opening the door for low power and real time embedded systems.

1. Introduction

Depth is an important feature of the surrounding space whose estimation finds its place in various tasks across many different fields. Potential applications can be as diverse as object manipulation in robotics or collision avoidance for autonomous vehicles during navigation. In humans, depth processing is extremely well developed and relies on monocular (e.g., occlusions, perspectives or motion parallax) and binocular (retinal disparities) visual cues [2]. This processing consumes very little energy as the visual system encodes retinal information under the form of action potentials, or *spikes* and it is believed that the brain only requires about 20 Watts to function [24]. Over the last years, it has motivated the development of numerous bio-

inspired approaches based on neuromorphic sensors and spiking neural networks to process depth in embedded systems.

Dynamic Vision Sensors (DVS) have recently gathered the interest of scientists and industrial actors, thanks to a growing number of research papers explaining how to process their output [9]. Reasons for this recent popularity are the very high dynamic range, excellent temporal resolution and unrivaled energy efficiency of these event cameras, making them especially suitable in automotive scenarios with strong energy and latency constraints. Similarly, as retinal ganglion cells in the animal retina, their pixels asynchronously emit an action potential (i.e. a *spike*, or *event*) whenever the change in log-luminance at this location since the last event, reaches a threshold.

Spiking Neural Networks (SNNs) are a good fit for DVSSs, as they can leverage the sparsity of their output event streams. Implemented on dedicated chips such as Intel Loihi [3], IBM TrueNorth [1], Brainchip Akida [35] or Tianjic [29], these models could become a new paradigm for ultra-low power computation in the coming years. In addition, SNNs maintain the same level of biological plausibility as silicon retinæ, making them new models of choice among computational neuroscientists. SNNs have recently attracted the deep learning community since the breakthrough of Surrogate Gradient (SG) learning [25] [36], which enabled the training of networks with back-propagation despite the non-differentiable condition for spike emission. While SNNs generally remain less accurate than their analog counterpart (i.e., Analog Neural Networks or ANNs), the gap in accuracy is decreasing, even on challenging problems like ImageNet [8].

In this context, we propose an ultra-low power spiking neural network for depth estimation, capable of dense depth predictions even at places without events and performing on par with the current state-of-the-art. Despite the dynamic nature of DVS data, we show that the problem of depth estimation from such neuromorphic event streams can be treated as a non-temporal task; therefore, our model is purely stateless in the sense that we reset all neurons to

a membrane potential of zero at each timesteps. If this feature might not fully take advantage of the temporal processing abilities of SNNs, it drastically decreases the computational and energy footprint of our model.

In section 2, we introduce a few related works that inspired our approach. We explain our methodology in section 3, including the data pre-processing, network architecture, and training details. In section 4, we compare our method with prior studies in terms on performances. We also show that StereoSpike surpasses equivalent analog Neural Networks (ANNs) with similar architectures, both in terms of performances and energy-efficiency.

2. Related Work

Deep learning approaches for depth estimation have had a long tradition on the timescale of modern deep learning techniques. First methods were based on luminance-field data from traditional frame-based cameras, either in mono- or binocular setups. The model in [5] was the first successful multi-scale architecture designed for depth estimation from RGB images, and was consequently followed by advances based on similar approaches [22] [21] [12].

Consistently with the recent interest of the scientific community in event-based cameras, a few works have successfully tackled the problem with neuromorphic data. Historically, several groups have used bio-inspired approaches such as Spiking Neural Networks (SNNs), in a very hardware-oriented direction, but not in a “deep learning” setting. For instance, the authors of [32] implemented a spike-based algorithm on a FPGA to regress low-resolution depth maps on a small size dataset. Furthermore, [13] proposed a SNN for processing depth from defocus (DFD); this work targeted neuromorphic chips and was able of recovering depth at full resolution, but the reconstruction was not dense and did not include learning.

Following the latter, [40] presented a neural network that jointly predicted camera pose and per-pixel disparity from stereo inputs. However, their reconstructed depth maps remained sparse as they restricted their analyses to pixels where events occurred. [34] addressed this problem and pushed even further the state-of-the-art on indoor scenarios thanks to 3D convolutions exploiting a specific event embedding. In [17], dense metric depth was recovered from only one camera, and showed very good performances with a recurrent, monocular encoder-decoder architecture on outdoor sequences.

Another inspiration for our work has been the task of large-scale optical flow regression from neuromorphic data, which is similar to depth reconstruction because it is also a large-scale image regression task. Similarly, large-scale optical flow regression from neuromorphic data has been an active area of research over the past few years, and has

seen the emergence of methods with growing hardware-friendliness. EV-FlowNet [37], arguably considered as the precursor of encoder-decoder models for optical flow reconstruction from event data, consisted in a feedforward analog encoder-decoder architecture. As a direct sequel, the hybrid model Spike-FlowNet [20] used spiking neurons in the encoder of a similar backbone, while maintaining the same levels of performances. In this approach, spiking neurons were shown to be able of encoding abilities close to analog ones and with a reduced computational cost. On the other hand, authors kept the remaining part of their network analog, to counteract the lack of expressivity in SNNs. More recently, the model proposed in [28] showed very good performances but it cannot be considered as a fully spiking network because real-value intermediate predictions of the outputs were reinjected within the network and mixed with binary spike tensors. In addition, they upsampled low-scale representations with the bilinear upsampling method, which breaks the binary spike constraint necessary for an implementation on neuromorphic hardware. Nevertheless, it is the first success in a large-scale regression task with a network that is spiking for its vast majority.

So far, SNNs have been used for classification tasks like image recognition [8], object detection [18], or motion segmentation [27]. Only a few works employed them for regression tasks. A notable exception is [10], but they only regressed 3 variables, while we propose here to regress the values of $260 \times 346 = 89960$ pixels.

3. Method

We used PyTorch and SpikingJelly [6] as our main development libraries. PyTorch is currently one of the most popular tools for deep learning and automatic differentiation, while SpikingJelly is an open-source framework for spiking neural networks, based on PyTorch and with rising popularity. Our codes will be released on GitHub upon publication of the paper. For the time being, a tentative anonymous repository can be accessed at the following address: <https://github.com/urancon/StereoSpike>

3.1. Neuron model

We use the McCulloch and Pitts model [23], outputting a binary activation whether the amount of weighted spikes integrated from lower layers reaches a threshold:

$$S^l = \Theta(V_{reset}^l + \sum w^{l-1} * S^{l-1}) \quad (1)$$

where Θ is the Heaviside step function, l denotes the layer number, and w synapse weights. V_{reset} corresponds to the potential of neurons at rest, and acts as an offset – a bias – to facilitate or hinder neurons from spiking.

This model is equivalent to the Integrate-and-Fire (IF) model deprived of the implicit recurrence in the membrane potential; all neuron potentials are reset to a value of V_{reset} at every timestep. As a result this model is stateless, contrarily to the traditional IF model. We use SpikingJelly’s *IFNode* class for its implementation.

Such neurons are inexpensive to simulate and can be deployed in large models, whereas more complex ones such as Hodgkin-Huxley [14], Izhikevich [16] or even SRM [11] are still too computationally expensive on modern hardware to train them.

A problem that has long prevented researchers from using simple Integrate-and-Fire models with standard deep learning techniques, is the gradient of the activation function –the Heaviside function– being zero everywhere (except in 0 where it is not defined). A recent solution to this is the replacement of the true gradient by a surrogate [25], which lets more room for the gradient to flow. We use the derivative of the arctan function as our surrogate gradient in this paper, as suggested in [8].

3.2. Dataset

We trained and tested our network on the Multi Vehicle Stereo Event Camera (MVSEC) dataset [39]. Because of its large size and variability, it has become one of the most popular benchmarks for depth reconstruction from neuromorphic events. It was collected from two DAVIS346 cameras with a resolution of 346×260 pixels, mounted on several vehicles such as a car, a motorbike or a drone. The depth groundtruth was provided by a Velodyne Puck Lite LIDAR mounted on the top of the two event cameras and with a sampling frequency of 20 Hz, hence providing a ground truth depth map every 50 ms.

We applied our method on the *indoor-flying* sequences of MVSEC, which was recorded on a quadricopter flying inside a large room. We used the data splits that were defined in [38] and [34]. We followed these previous works and removed take-off and landing parts of the sequence, because they contained very noisy event streams and inaccurate groundtruths.

3.3. Event Representation

We adopted a rather common representation of data: we binned all incoming spikes on each pixel on a time window of 50 ms (see Figure 1). Furthermore, we accumulated spikes for each polarity in a different channel. Because there are two polarities, the resulting tensor had a shape of $(2, Height, Width)$ and contained positive integers, corresponding to the number of spikes of each polarity that showed up at each position of the scene during the time window. We further refer to this format as *spike histograms* or *spike frames* interchangeably.

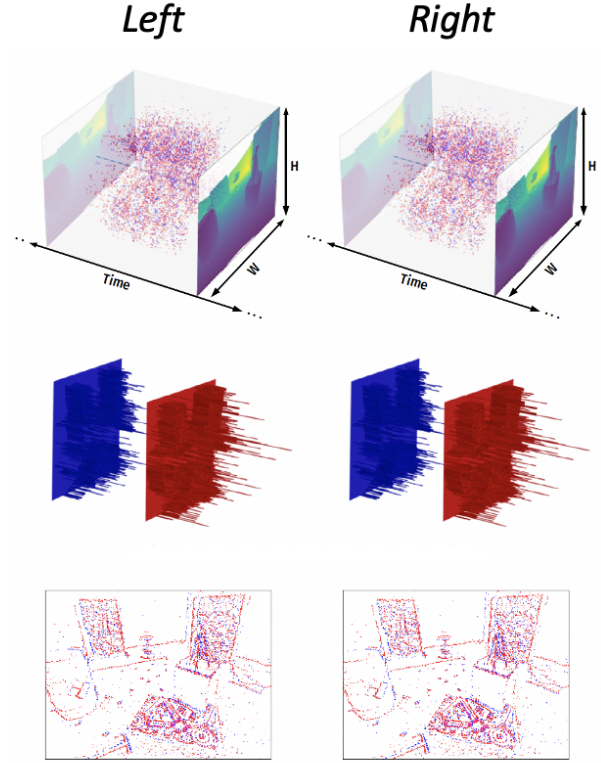


Figure 1. (top): ON and OFF events are binned, per-pixel, within time windows of 50 ms. Frames displayed on the temporal axis are the ground-truth depth maps, provided by the LIDAR at 20 Hz. (middle): This operation results in a 2-channel histogram of spikes, containing integer spike counts at every pixel and for each polarity. (bottom): Such spike frames are commonly visualized according to the following convention: pixels reporting at least one ON event are colored in red, those reporting OFF events in blue, and those reporting both types in pink.

The duration of 50 ms for binning was motivated by empirical results, as this value lead to better model performances than durations of 25 ms or 100 ms.

The final input tensor is obtained by concatenating the spike frames from left and right cameras together channel-wise, hence resulting in a $(4, Height, Width)$ -shaped volume.

Many ANN approaches normalize this kind of input tensor (e.g. divide by the maximum number of spike count) for an easier-to-learn distribution of data and better generalization [17]. We believe that this operation is not adapted to neuromorphic hardware, and can lead to non-integer number of spikes in the normalized input tensor. For this reason, we prefer feeding raw spike frames directly to our network.

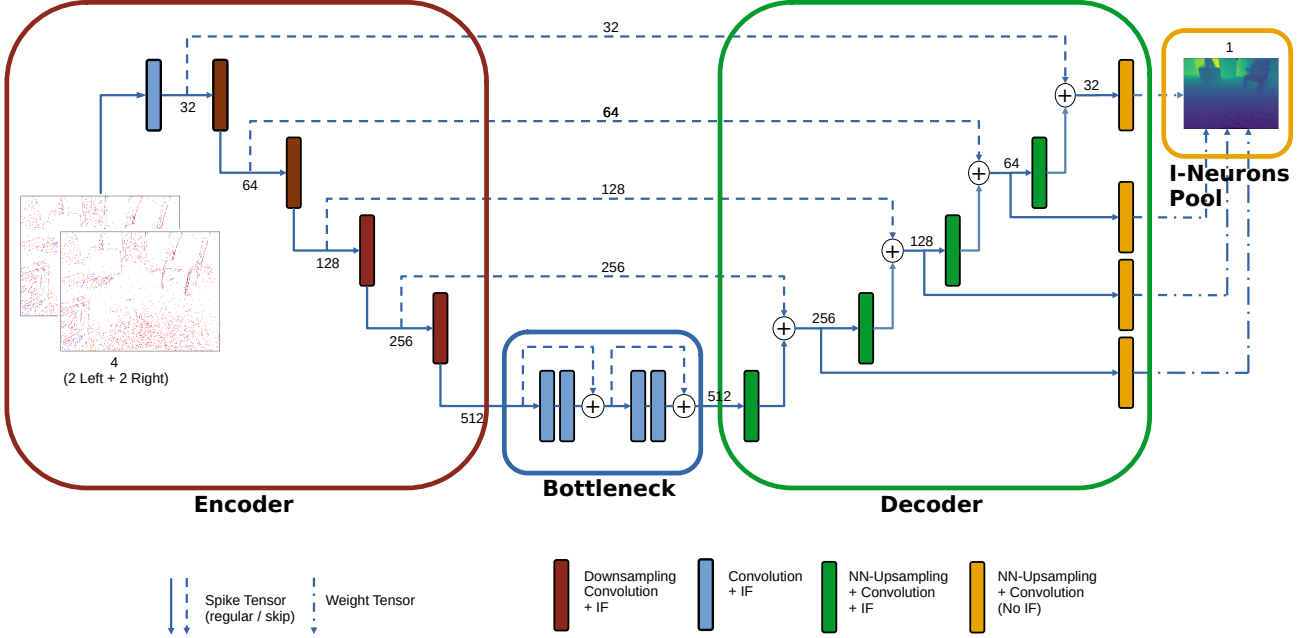


Figure 2. Detailed architecture of StereoSpike. Its encoder contains one or two branches (depending on the modality), each for a different DVS camera. The output of both branches are combined with a sum function, and further processed through a bottleneck consisting of 2 chained SEW-Resblocks. As a result, the tensor out of these residual layers is composed of integers in range $[0, 4]$. This latent representation is progressively upsampled by decoder layers and whose outputs are summed with same-level encoder spike tensors, leading to integer values in range $[0, 3]$. In parallel, prediction synapses from different scales directly project to output I-neurons whose potentials bear the final prediction. The numbers indicate the size of channel dimension for each spike volume. Best viewed in color.

3.4. Architecture

Our model was based on a U-Net backbone [33] consisting of an encoder, a bottleneck and a decoder whose non-linearities were achieved by spiking neurons (see Figure 2). We used the McCulloch and Pitts model presented in (1) with a reset potential of $V_{reset} = 0$, except for output neurons where this parameter is learned and shared.

Downsampling in the encoder was performed by 2-strided convolutions, which divided the spatial resolution by 2 while doubling the channel resolution. The bottleneck consisted in 2 SEWResBlocks [7] following each other and with *ADD* connect function.

Because transposed convolutions are known to generate checkerboard artifacts [26], we rather used nearest neighbor (NN) upsampling followed by a convolution. Contrarily to bilinear upsampling, we believe NN to be neuromorphic-hardware friendly, because it essentially keeps integer spike counts in the upsampled volume. In terms of biological plausibility, it can be viewed as a single neuron at low scale projecting synapses towards several higher-scale neurons.

In addition, none of the layers in our network used a bias

term nor Batch Normalization (BatchNorm) because adding constant biases is costly on neuromorphic hardware and is not biologically plausible.

The output of the network was carried by the potentials of a pool of non-leaky neurons with an infinite threshold. Before each inference and like all neurons in the network, their membrane potentials were set to a shared and learnable V_{reset} parameter. One critical problem of SNNs is their lack of expressivity, since they can only update the membrane potentials of output neurons with discrete weighted spikes. With synapses coming from the full scale level of the encoder only as in a standard architecture, the model would have too few spikes and too few different parameters (i.e., synaptic weights) to achieve top performances at rendering a large-scale and diverse depth scene. To counteract this effect, we increased the number of spikes to update the readout neurons' potential by linking them to lower levels of the network *via* intermediary prediction layers. These layers essentially consisted in nearest neighbor upsampling followed by convolution; they were the same as the upsampling layers in the decoder, except that they upsampled spike tensors directly up to the full, original scale.

3.5. Loss Function

Similarly as in [17], we used a combination of a regression loss with a regularization loss. Noting $R = \hat{D} - D$ the residual between the groundtruth and predicted depth maps, the first term can be written as:

$$L_{regression} = \frac{1}{n} \left(\sum_u R(u)^2 \right) - \frac{1}{n^2} \left(\sum_u R(u) \right)^2 \quad (2)$$

where n is the number of valid groundtruth pixels u . With the same notations, the regularization loss is computed with:

$$L_{smooth} = \frac{1}{n} \sum_u |\nabla_x R_s(u)| + |\nabla_y R_s(u)| \quad (3)$$

According to [17], the minimization of this term encourages smooth depth changes as well as sharp depth discontinuities in the depth map prediction, hence helping the network to represent objects that stand out of the background of the scene (e.g., because they are closer), while respecting its overall topology. Finally, we weighted both terms with a factor λ in the total loss:

$$L_{tot} = L_{regression} + \lambda L_{smooth} \quad (4)$$

We used a value of 0.5 for λ , which was determined empirically.

3.6. Training Procedure

Learning of parameter values in our model was done using surrogate gradient learning [25], as implemented in Spikingjelly Python library [6].

Because our network is feed-forward and only processes one step for inference, we trained it with regular back-propagation, not Back-Propagation Through Time (BPTT) and on the shuffled dataset. We adopted the following procedure: 1) initialize all potentials (including output) to zero, 2) forward pass a sample or a batch of them, 3) calculate loss and back-propagate it, 4) update the weights and 5) start again with a new sample.

We used Adam optimizer [19] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We trained the network for 30 epochs with an initial learning rate set to 2.10^{-4} and divided by 10 at epoch 10, 25 and 40. Batch size was set to 1 and we did not use any weight decay.

4. Experiments

4.1. Performances

Our model performances are equivalent, or even slightly better, to those obtained from DDES [34] (the current state-of-the-art), which is a fully-fledged ANN using 3D convolutions within a less general framework. In addition, StereoSpike’s architecture is not specific to the estimation of

depth and could be used for any other large-scale regression task. StereoSpike also outperforms TSES [38] and CopNet [31], which only predict depth at event positions. Please refer to Figure 3 for qualitative visualizations of depth reconstructions obtained with our model and to Table 1 for a quantitative comparison with previous works on the Mean Depth Error (MDE), the most common metric used for characterizing depth estimation on MVSEC.

Table 1. Test Mean Depth Error (MDE) in centimeter on several dataset splits of *indoor_flying* sequence. From three randomized training trials, the best model on the validation set is selected and then evaluated on the test set, from which these metrics are calculated. * indicates that the evaluation is done on the sparse groundtruth, i.e., only at pixels where events occurred.

Model	MDE [cm]		
	Split 1	Split 2	Split 3
Tulyakov et al. [34]	16.7	29.4	27.8
Ours (Binocular)	18.5	29.4	25.4
Ours (Monocular)	21.7	30.8	27.7
TSES* [38]	36	44	36
CopNet* [31]	61	100	64

In terms of MDE, monocular models (i.e., receiving data from only one camera) also lead to good depth estimates, but with a consistent drop in accuracy across data splits. This suggests that in addition to being –for the most part– a non-temporal task, depth reconstruction from DVS data can be efficiently tackled on a monocular setting, at a reasonable cost in performance. Therefore, fusing left and right data as early as in the first convolution layer reveals itself to be a good strategy for exploiting binocular disparity and reaching state-of-the-art performances.

4.2. Ablation studies

4.2.1 Intermediary predictions

Contrarily to a standard U-Net [33], StereoSpike makes a coarse-to-fine reconstruction of the depth scene *via* four prediction layers, which can be seen as synapses projecting from different levels of the network body (i.e., decoder) to the pool of readout neurons. This technique is new, and to determine its added value, we conducted an ablation study on the architecture by “cutting” the latter and observing the performances of models partially deprived of their means of expressivity. Results are compiled in Table 2.

We observe that performances gradually decay as we remove intermediary prediction layers. Best performances are obtained by the full model with all four layers as presented in Figure 2, while the model equipped with top-level prediction layer only (i.e., classical encoder-decoder configu-

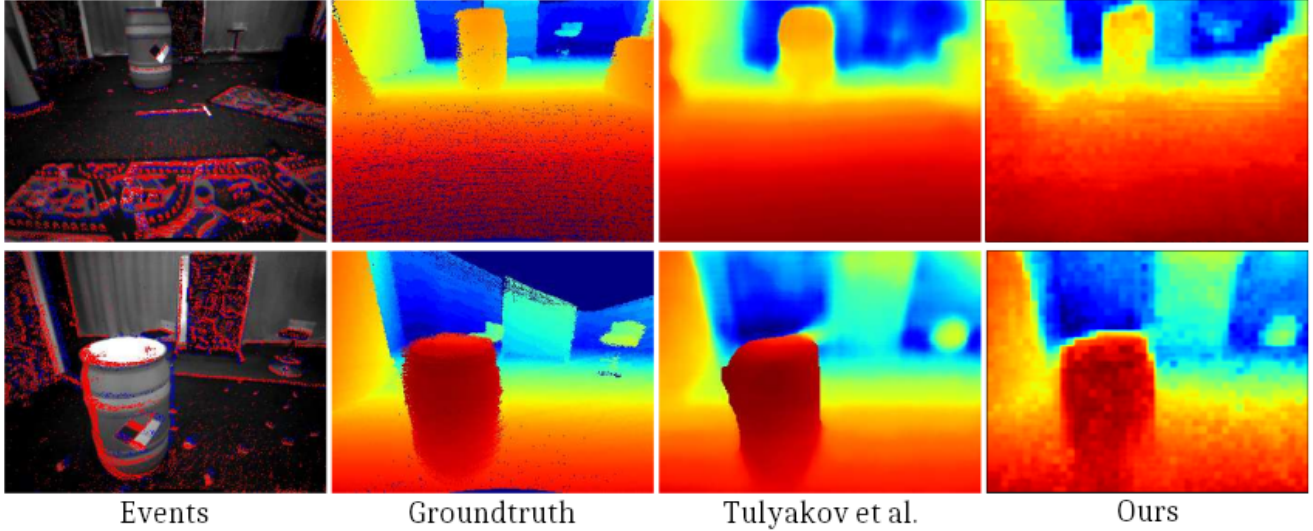


Figure 3. Qualitative comparison of our method with the current state-of-the-art detailed by [34]. Similarly to what they did in their paper, we selected the same input frames and run our model to infer depth from this data; Event groundtruth and other prediction images were borrowed from their article. The top row corresponds to frame #1700 of *indoor_flying3* and frame #980 of *indoor_flying1* sequences. The pixelated aspect of our predictions comes from the Nearest-Neighbor interpolation in the prediction layer from very-low to full scale. Even so, it can be seen that our fully spiking network captures the scene just as well as a cutting edge ANN using a heavy framework and 3D convolutions. Adapted from [34].

Table 2. Test Mean Depth Error (MDE) on split 1 of *indoor_flying* sequence. Reported errors are averaged and provided with standard deviations over three randomized training trials. Prediction layers are depicted by a number $\in \{1, 2, 3, 4\}$, where 1 and 4 are the top- and bottom- level prediction layers, respectively.

Model prediction layer number	MDE [cm]
$\{1, 2, 3, 4\}$	19.07 ± 0.48
$\{1, 2, 3\}$	22.40 ± 0.63
$\{1, 2\}$	24.13 ± 0.31
$\{1\}$	25.43 ± 0.21

ration) reports the worst metrics. Thus, multiplying regression layers constitutes a efficient strategy to enable more precise predictions. More synapses means more parameters but also more possible spikes to update the potentials of output neurons, and therefore more degrees of freedom. This strategy is translated into an improved accuracy compared to standard auto-encoder architectures.

4.2.2 Skip connections

Skip connections are a standard feature of encoder-decoder neural network architectures in the field of computer vision. However, they can be difficult to implement on some neuromorphic hardware. With the concern of staying close

to low-level designers, we studied the effect of entirely removing them from StereoSpike.

Table 3. Test Mean Depth Error (MDE) on split 1 of *indoor_flying* sequence. Reported errors are averaged and provided with standard deviations over three randomized training trials.

Model	MDE [cm]	
Skip connections	train	test
With	15.06	19.07
Without	9.52	22.41

Models trained without skip connections turned out to overfit more than standard StereoSpike on training data, to the detriment of test accuracy. However, this performance drop is low, and a model without skip connections could still be reliably used in real situations. Therefore, skip connections seem to act like a regularizer here. In the future, we believe that neuromorphic chips should be able to implement this typical architectural feature.

4.3. StereoSpike - ANN comparison

A lesson learnt from our study and [20] is that SNNs can encode information very optimally, even with binary values. While Spike-FlowNet used ANNs to decode the latent space representation, we only use spiking neurons. In addition, we do not mix real-valued intermediary predic-

tions with integer spikes as in [28]. SNNs can therefore efficiently encode information as well as decode it, even for large scale regression tasks.

In an attempt to compare our model with fully-fledged ANNs, we trained equivalent ANN models. These models had a exactly the same architecture and output paradigm consisting in a pool of I-neurons; however, and with the idea of using the full power of analog models, we replaced IFN-odes by common activation functions, used Batch Normalization (BN) [15] and trainable biases in convolution layers. As can be seen in Table 4, the ANNs outperform the SNN on the training set, but not on the testing set. In other words, the ANNs overfit more than the SNN. This suggests that spikes, in addition to increasing hardware-friendliness, constitute an efficient regularization mechanism, causing the SNN to generalize better. To the best of our knowledge, it is the first time that this desirable regularization effect is reported - but of course, other regularization methods could be tried to limit overfitting in the ANNs (e.g. drop out).

Table 4. Comparative evaluation of our SNN vs equivalent ANN models on split 1 of *indoor_flying* sequence. For a given model, the train MDE, train loss and test loss are all sampled from the epoch yielding the best test MDE. Our fully spiking network surpasses by a large margin all of its analog relatives.

Model	MDE [cm]		Loss [au]	
	train	test	train	test
StereoSpike	13.3	18.5	0.91	1.32
ANN (Sigmoid + BN)	9.5	24.5	0.72	1.48
ANN (Tanh + BN)	9.8	25.6	0.77	1.76
ANN (LeakyReLU + BN)	16.8	28.1	0.80	1.55

The LeakyReLU ANN achieves the lowest training loss and among the lowest training MDE, but has the worse test metrics by a large margin, hinting at a strong overfitting. Networks with step-like activations (Sigmoid, Tanh) show better results on test sets and higher error on training sets, therefore suggesting a reduced overfitting phenomenon. Finally, the StereoSpike network –equivalent to an ANN with Heaviside step function as activation– achieves the best test loss and MDE but also the worst training metrics, therefore generalizing best, despite the absence of BN.

4.4. Computational Efficiency

A promising feature of SNNs is their ability to solve complex tasks at performances comparable to conventional networks (ANNs), but with sparse activations. In order to quantify the computational efficiency of our network, we measure the firing rates of its layers, i.e., the density of intermediary tensors computed during inference on the test set. Sparse volumes can be leveraged by dedicated hardware capable of sparse computation, hence diminishing in-

ference time as well as energy consumption.

It appears that the firing rates of our best model grow as layers become closer to the output I-neuron pool. That is, convolution layers in the decoder report an average firing rate of 23.5% compared to 8.1% in the encoder. We suggest that in this large-scale regression task, a minimum number of pre-synaptic spikes preceding output neurons is necessary to faithfully render the visual scene. Similarly, a certain amount of spikes could be necessary to encode the information contained in the input histograms.

To encompass this trade-off and estimate these minimal firing rates, we apply a regularization loss term explained in [30] and train a new binocular model on split 1. This secondary loss, which we also call *quadratic spike penalization loss*, penalizes the mean of the squared spike tensor. Therefore, for a given layer containing K spiking units whose output at time-step n is $S_k[n] \in \{0 \dots 4\}$, it can be defined as:

$$\begin{aligned} L_{spikes} &= \frac{1}{2NK} \sum_n \sum_k S_k[n]^2 \\ &= \frac{1}{2K} \sum_k S_k[n]^2 \end{aligned} \quad (5)$$

Because the number of time-steps to do one prediction is $N = 1$, as our model is purely stateless/feedforward. We apply this loss on the tensor out of the bottleneck and on the resulting tensors of the skip connections, that are used by predictions layers at different scales. Penalizing these tensors also indirectly affects the activity of encoder layers, as their output conditions the density of same-level encoder volumes because of the skip connections. Therefore, this regularization is less aggressive than penalizing all intermediary tensors and performances are less negatively affected.

We then evaluate the network trained with spike penalization on the test set and compare obtained firing rates with our unconstrained model. More specifically, we plot the average test accuracy as a function of the average firing rate in Figure 4. Regularized models show a drastic decrease in activity, at a very reasonable accuracy loss for a penalization weight if 0.5 (< 3 cm MDE). With these results, we can imagine our model implemented efficiently on dedicated hardware.

4.4.1 Target Hardware

Our model has resolutely been developed in the philosophy of spiking neural networks. As a result, it is essentially implementable on dedicated neuromorphic hardware,

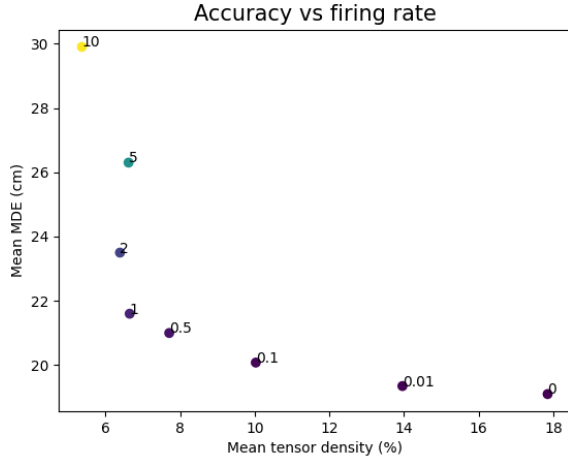


Figure 4. Test accuracy as a function of the mean firing rate. Mean firing rate is calculated as the average of the density of all activation (spike) tensors calculated within the network during inference. Labels correspond to the weight of the spike penalization loss in comparison to the objective loss. Unconstrained models generally perform better than models trained while encouraging sparsity. High weight values for the spike penalization loss do not always result, on average, in higher sparsity. We reckon this is due to a significant change in the landscape of the total loss. Gradient descent is not a perfect optimization algorithm and due to local minima, a penalization weight of 5 result in both higher error and density than with a weight of 2.

such as Intel Loihi [3], IBM TrueNorth [1]. These chips can leverage the binarity and sparsity of spike tensors navigating through the network. In addition, we believe that our model being feedforward and requiring a reset on all of its neurons at each timestep is not a problem, because resetting membrane potentials is actually less costly than applying a leak. Therefore, statelessness can be seen as an advantage over recurrence in spiking models with similar performances. However, we are aware that current neuromorphic chips are initially designed for the implementation of stateful units, and acknowledge that we do not leverage this feature. Consequently, we believe that it rather fits to dedicated hardware for stateless models with sparse activations quantized on 1 bit. We therefore consider that Brainchip’s Akida chip [35] is a good fit. Finally, we would like to emphasize that our class of model with sparse binary activations and less constrained weights provides a good compromise between Spiking Neural Networks (SNNs) and Binary Neural Networks (BNNs).

4.4.2 Handling integer (non-binary) spike counts

Because of the sum operations present in residual layers of our architecture and at skip connections, bottleneck and decoder tensors can contain integer (non-binary) numbers of spikes. We explain here why we do not consider it as a problem. On most digital neuromorphic chips, spikes are represented by multi-bit messages containing destination and/or source addressing, and a few bits for a graded-value payloads; this is the case for Loihi, see [4]. In our case, the spike counts are included in $[0, 3]$ and thus can be coded on 2 bits. For the chips that can only handle binary spikes, a spike count of N could be handled by N serial binary spike operations. Another possibility to make our model fit in a small-size neuromorphic chip, could be to cut the skip connections as indicated in 4.2.2, or to remove intermediate prediction layers. The latter are indeed necessary to reach near state-of-the-art performances, but are not absolutely required to show reasonable performances.

5. Conclusion

We have proposed StereoSpike, the first fully spiking, deep neural network architecture for large scale regression task with sparse activity. The key is expressivity and we tackle this problem by increasing the number of spikes generated by deeper layers, with a pool of perfect integrator neurons bearing the final prediction with their membrane potential. The same strategy could presumably be used for other dense regression problems with SNNs, for example optic flow prediction.

We have shown that depth estimation from DVS data can be brought back to a stateless, static, non-temporal task. As a result, hardware implementations of StereoSpike could consume substantially less than recurrent versions. However, we are aware that our model neither takes advantage of this implicit recurrence of SNNs to capture temporal dependencies, nor of the very high temporal resolution of DVSs; these two aspects deserve investigation for further improvement of our algorithm.

Furthermore, our experiments hint towards the fact that because of the constraint on their output (i.e. binarity) SNNs might generalize better than their ANN counterparts. Consequently, our work is yet another evidence that binary encoding in latent space can represent input data as efficiently as real-valued projections. This encourages more research in the field of SNNs and neuromorphic computing in general, or Binarized Neural Networks (BNNs).

Finally, the combination of event cameras and spiking neural networks within the same framework is a more biologically plausible approximation of the visual nervous system, and could allow researchers to understand the processing of depth in the brain with large-scale models.

6. Acknowledgements

We would like to thank Amirreza Yousefzadeh for his help and expertise on digital neuromorphic hardware. Our thanks also go to Wei Fang for his wonderful SpikingJelly library, without which this work would not have been possible. This research was supported by a grant from the ‘Agence Nationale de la Recherche’ (ANR-16-CE37-0002-01, ANR JCJC 3D3M) awarded to Benoit R. Cottureau.

References

- [1] Filipp Akopyan, Jun Sawada, Andrew Cassidy, Rodrigo Alvarez-Icaza, John Arthur, Paul Merolla, Nabil Imam, Yutaka Nakamura, Pallab Datta, Gi-Joon Nam, Brian Taba, Michael Beakes, Bernard Brezzo, Jente B. Kuang, Rajit Manohar, William P. Risk, Bryan Jackson, and Dharmendra S. Modha. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 34(10):1537–1557, 2015. 1, 8
- [2] James Cutting and Peter Vishton. *Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth*. Academic Press, 1995. 1
- [3] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Andrew Lines, Andreas Wild, Hong Wang, and Deepak Mathaikutty. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, PP:1–1, 01 2018. 1, 8
- [4] Mike Davies, Andreas Wild, Garrick Orchard, Yulia Sandamirskaya, Gabriel A Fonseca Guerra, Prasad Joshi, Philipp Plank, and Sumedh R. Risbud. Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook. *Proceedings of the IEEE*, X:1–24, 2021. 8
- [5] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2650–2658, 2015. 2
- [6] Wei Fang, Yanqi Chen, Jianhao Ding, Ding Chen, Zhaofei Yu, Huihui Zhou, Yonghong Tian, and other contributors. Spikingjelly. <https://github.com/fangwei123456/spikingjelly>, 2020. Accessed: 2021-08-01. 2, 5
- [7] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks, 2021. 4
- [8] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating Learnable Membrane Time Constant to Enhance Learning of Spiking Neural Networks. In *IEEE/CVF ICCV*, 2021. 1, 2, 3
- [9] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1
- [10] Mathias Gehrig, Sumit Bam Shrestha, Daniel Mouritzen, and Davide Scaramuzza. Event-based angular velocity regression with spiking networks. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4195–4202, 2020. 2
- [11] W Gerstner, R Ritz, and J L van Hemmen. Why spikes? Hebbian learning and retrieval of time-resolved excitation patterns. *Biol Cybern*, 69(5-6):503–515, 1993. 3
- [12] Clément Godard, Oisín Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611, 2017. 2
- [13] Germain Haessig, Xavier Berthelon, Sio-Hoi Ieng, and Ryad Benosman. A spiking neural network model of depth from defocus for event-based neuromorphic vision. *Scientific Reports*, 9:3744, 03 2019. 2
- [14] A L Hodgkin and A F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol*, 117(4):500–544, 1952. 3
- [15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 7
- [16] E M Izhikevich. Simple model of spiking neurons. *IEEE Trans Neural Netw*, 14(6):1569–1572, 2003. 3
- [17] Daniel Gehrig Javier Hidalgo-Carrio and Davide Scaramuzza. Learning monocular dense depth from events. *IEEE International Conference on 3D Vision (3DV)*, 2020. 2, 3, 5
- [18] Seijoon Kim, Seongsik Park, Byunggook Na, and Sungroh Yoon. Spiking-yolo: Spiking neural network for energy-efficient object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11270–11277, Apr. 2020. 2
- [19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 5
- [20] Chankyu Lee, Adarsh Kosta, Alex Zihao Zhu, Kenneth Chaney, Kostas Daniilidis, and Kaushik Roy. Spike-flownet: Event-based optical flow estimation with energy-efficient hybrid neural networks. In *European Conference on Computer Vision*, pages 366–382. Springer, 2020. 2, 6
- [21] Z. Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 2
- [22] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 2
- [23] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943. 2
- [24] J W Mink, R J Blumenshine, and D B Adams. Ratio of central nervous system to body metabolism in vertebrates: its constancy and functional basis. *The American journal of physiology*, 241(3):R203–R212, 1981. 1

- [25] Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019. 1, 3, 5
- [26] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 4
- [27] Chethan M. Parameshwara, Simin Li, Cornelia Fermüller, Nitin J. Sanket, Matthew Evanusa, and Yiannis Aloimonos. Spikems: Deep spiking neural network for motion segmentation. *CoRR*, abs/2105.06562, 2021. 2
- [28] Federico Paredes-Valles, Jesse Hagenaars, and Guido de Croon. Self-supervised learning of event-based optical flow with spiking neural networks, 2021. 2, 7
- [29] Jing Pei, Lei Deng, Sen Song, Mingguo Zhao, Youhui Zhang, Shuang Wu, Guanrui Wang, Zhe Zou, Zhenzhi Wu, Wei He, Feng Chen, Ning Deng, Si Wu, Yu Wang, Yujie Wu, Zheyu Yang, Cheng Ma, Guoqi Li, Wentao Han, Huan-glong Li, Huaqiang Wu, Rong Zhao, Yuan Xie, and Luping Shi. Towards artificial general intelligence with hybrid Tian-jie chip architecture. *Nature*, 2019. 1
- [30] Thomas Pellegrini, Romain Zimmer, and Timothee Masque-lier. Low-Activity Supervised Convolutional Spiking Neural Networks Applied to Speech Commands Recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 97–103. IEEE, jan 2021. 7
- [31] Ewa Piatkowska, Jurgen Kogler, Nabil Belbachir, and Margrit Gelautz. Improved cooperative stereo matching for dynamic vision sensors with ground truth evaluation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 370–377, July 2017. 5
- [32] Nicoletta Risi, Enrico Calabrese, and Giacomo Indiveri. Instantaneous stereo depth estimation of real-world stimuli with a neuromorphic stereo-vision setup. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2021. 2
- [33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 4, 5
- [34] S. Tulyakov, F. Fleuret, M. Kiefel, P. Gehler, and M. Hirsch. Learning an event sequence embedding for event-based deep stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 5, 6
- [35] Anup Vanarse, Adam Osseiran, Alexander Rassau, and Peter van der Made. A hardware-deployable neuromorphic solution for encoding and classification of electronic nose data. *Sensors*, 19(22), 2019. 1, 8
- [36] Friedemann Zenke, Sander M. Bohtë, Claudia Clopath, Iulia M. Comşa, Julian Göltz, Wolfgang Maass, Timothée Masquelier, Richard Naud, Emre O. Neftci, Mihai A. Petrovici, Franz Scherr, and Dan F.M. Goodman. Visualizing a joint future of neuroscience and neuromorphic engineering. *Neuron*, 109(4):571–575, feb 2021. 1
- [37] Alex Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. 2
- [38] Alex Zihao Zhu, Yibo Chen, and Kostas Daniilidis. Realtime time synchronized event-based stereo. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 438–452, Cham, 2018. Springer International Publishing. 3, 5
- [39] Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, Jul 2018. 3
- [40] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 989–997, 2019. 2