Differentiable hierarchical and surrogate gradient search for spiking neural networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Spiking neural network (SNN) has been viewed as a potential candidate for the next generation of artificial intelligence with appealing characteristics such as sparse computation and inherent temporal dynamics. By adopting architectures of deep artificial neural networks (ANNs), SNNs are achieving competitive performances in benchmark tasks such as image classification. However, successful architectures of ANNs are not necessary ideal for SNN and when tasks become more diverse effective architectural variations could be critical. To this end, we develop a spike-based differentiable hierarchical search (SpikeDHS) framework, where spike-based computation is realized on both the cell and the layer level search space. Based on this framework, we find effective SNN architectures under limited computation cost. During the training of SNN, a suboptimal surrogate gradient function could lead to poor approximations of true gradients, making the network enter certain local minima. To address this problem, we extend the differential approach to surrogate gradient search where the SG function is efficiently optimized locally. Our models achieve state-of-the-art performances on classification of CIFAR10/100 and ImageNet with accuracy of 95.50%, 76.25% and 68.64%. On event-based deep stereo, our method finds optimal layer variation and surpasses the accuracy of specially designed ANNs meanwhile with 26$\times$ lower energy cost ($6.7\mathrm{mJ}$), demonstrating the advantage of SNN in processing highly sparse and dynamic signals. Code will be available soon.

1 Introduction

Inspired from biological neural networks, spiking neural network (SNN) [42] has been viewed as a potential candidate for the next generation of artificial intelligence, with appealing characteristics such as asynchronous computation, sparse activation and inherent temporal dynamics. However, the training of deep SNNs is challenging due to the binary spike which is incompatible with gradient-based backpropagation. To solve this problem, various surrogate gradient (SG) methods were proposed [5, 67, 49] where soft relaxed functions were used to approximate the original discontinuous gradient. Based on these methods, SNNs have achieved high level performances on benchmark image classification tasks such as CIFAR and ImageNet [59, 68, 76, 55, 78]. However, the accuracy of SNN often drops when directly adopting ANN architectures , such as ResNet [22] and VGG networks [60]. With recent improved SNN training methods [36, 16, 14] the performance gap is decreasing but remains. This gap is more evident in tasks where the network architecture requires more variation, such as dense image prediction [83, 21, 26].

Directly inheriting sophisticated ANN architectures might not be ideal for SNN and there is *a lack of study* of optimal architectures for spiking neurons given a particular task. Intuitively, topology in cortex should offer us some inspiration. Neuroscience has studied macro structures such as

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

columnar organization [45] and micro dynamics such as interaction between lateral and feedback connections [37]. However, on the network level the relation between connections of neural circuits and functionalities remains largely unknown. In this work, we develop a spike-based differentiable hierarchical search (SpikeDHS) framework in order to find optimal task-specific network architectures under limited computation cost. Based on traditional differentiable architecture search (DARTS) framework [41], we redesign the search space and information flow in the principle of spike-based computation. For both the cell and the layer search space, we study how to realize this principle towards a fully spiking network.SG provides an approximation of the non-existing gradient in SNN and its selection is not unique. To explore optimal SG functions for the training of SNN, we propose a differentiable SG search method to efficiently adapt the function locally, and demonstrate its efficiency on both static and event-based benchmark machine learning tasks. In summary, our contributions are following:

- We develop a differentiable hierarchical search framework for spiking neurons, realizing spike-based computation on both the cell and the layer level search space, based on which optimal SNN architectures can found under limited computation cost.
- To improve gradient approximation of deep SNNs, we propose a differentiable SG search method to efficiently optimize SG functions locally, which is easy to scale and also effective for binary networks.
- Extensive experiments show that our methods outperform SNNs based on sophisticated ANN architectures on image classification of CIFAR10, CIFAR100 and ImageNet datasets.
- On event-based deep stereo task, to the best of our knowledge we show the first time SNN surpasses specially designed ANNs on the Multi Vehicle Stereo Event Camera (MVSEC) [82] dataset in terms of accuracy, network sparsity and computation cost, demonstrating its advantage in processing highly sparse and dynamic signals with extremely low power and latency.

2 Related work

2.1 Architecture search

Designing high performance network architectures for specific tasks often requires expert experience and trial-and-error experiments. Neural architecture search (NAS) [15] aims to automate this manual process and has recently achieved highly competitive performance in tasks such as image classification [84, 85, 39, 56, 53], object detection [85, 9, 65, 20] and semantic segmentation [40, 77, 50, 38], etc. However, searching over a discrete set of candidate architectures often results in a massive number of potential combinations, leading to explosive computation cost. The recently proposed differentiable architecture search (DARTS) method [41] and its variations [69, 8, 11] address this problem using a continuous relaxation of the search space which enables learning a set of architecture coefficients by gradient descent, and has achieved competitive performances with the state-of-the-art using orders of magnitude fewer computation resources [41, 40, 10]. Recently, [47] studied pooling operations for downsampling in SNNs and applied NAS to reduce the the overall number of spikes. [27] applied NAS to improve SNN initialization and explore backward connections. However, both works only searched for different SNN cells or combinations of them under fixed network backbone and their application is limited to image classification.

2.2 Training of SNN

The success of deep ANNs in solving benchmark machine learning tasks has motivated efforts to make SNNs realize similar capabilities, either based on bio-inspired mechanisms [52, 48, 34, 46, 33, 28] or approximating ANN learning algorithms [5, 67, 49, 3]. Currently, two approaches have demonstrated their efficiency, showing the ability to solve hard problems at similar levels as their artificial counterparts, namely, ANN-to-SNN conversion [58, 6, 35] and directly training SNNs with SG [59, 68, 76, 36, 16, 14]. The theoretical soundness of the SG approach for training binary activation networks has been studied and justified [4, 71]. Experiments show that the training of SNN is robust to the shape of SG function as long as it meets certain criteria, such as the overall scale [73]. [21] shows that a suitable SG function is critical when the SNN goes deeper. [36] further improved the performance of SNN by optimizing the width (or temperature) of a continuous SG function,

through the guidance of an approximated gradient measured with finite difference gradient (FDG). However, a suitable hyperparameter setting of FDG is largely empirical. In terms of computation, it iterates over each element of the weight and is sequential across layers, making a global application of this method computationally impractical.

2.3 Event-based task with SNN

Inspired by biological retina, event camera [19] captures instantaneous changes of pixel intensity at microsecond resolution. Compared to traditional frame-based cameras, it covers a higher dynamical range (120dB) and offers a low power solution for vision tasks in high-speed scenario. Further combining it with neuromorphic processors [51, 57, 12, 43, 18, 31] can create a fully neuromorphic system, realizing extremely low power and low latency sensing. However, learning from highly sparse and asynchronous events is challenging. Given its inherited asynchronous dynamics, SNN is an ideal candidate for such task and recently a number of works have applied it for event-based problems such as classification [30, 35], tracking [70], detection [25], semantic segmentation [26] and optical flow estimation [32, 21], etc. Multi-view event-based deep stereo solves the problem of 3D scene reconstruction based on pixel differences of the same physical point from event streams obtained by multiple views. Given the problem's complexity, it has been addressed by specially designed deep ANNs. Several works use additional information such as camera motion to produce sparse depth maps [81, 80]. Estimating dense disparity images from sparse event inputs is more challenging. A recent work [63] addresses this problem by an event queue method which encodes events into event images through 3D convolution, followed by an hourglass network to estimate disparity. Based on [63], [1] further enhances local contours of the estimated disparity using image reconstruction.[44] creates feature pyramids with multi-scale correlation learned from a cycle of gray-scale images and event inputs. Inspired from biological neuron models, [75] developed discrete time convolution to encode events with temporal dynamic feature maps and proposed a dual-path encoder with spatially adaptive modulation to strengthen events representation. A very recent work [54] applies SNN to this problem with a handcrafted U-net structure. Nevertheless, compared with ANN models using geometric volumes its performance is suboptimal.

3 Method

3.1 Preliminary

We adopt the iterative leaky integrate-and-fire (LIF) neuron model [68] described by

$$u^{t,n} = \tau u^{t-1,n}(1 - y^{t-1,n}) + I^{t,n} \tag{1}$$

where superscripts n and t denote layer index and time step, respectively. τ is the membrane time constant, u is the membrane potential, y denotes the spike output and I denotes the synaptic input with $I^{t,n} = \sum_j w_j y_j^{t,n-1}$ where w is the weight. The neuron will fire a spike $y^{t,n} = 1$ when $u^{t,n}$ exceeds a threshold V_{th}, otherwise $y^{t,n} = 0$. In this work, we set $\tau = 0.2$ and $V_{th} = 0.5$. Given loss L, using chain rule the weight update of SNN can be expressed as:

$$\frac{\partial L}{\partial w} = \sum_t \frac{\partial L}{\partial y^t}\frac{\partial y^t}{\partial u^t}\frac{\partial u^t}{\partial I^t}\frac{\partial I^t}{\partial w} \tag{2}$$

where $\frac{\partial y^t}{\partial u^t}$ is the gradient of the spiking function, which is zero everywhere except at $u = V_{th}$. The SG approach uses continuous functions to approximate the real gradients, such as rectangular [78], triangular [2], Superspike [72], ArcTan [17] and exponential curves [59]. We adopt Dspike function from [36]:$\mathrm{Dspike}(x) = a \cdot \tanh(b(x-c)) + d$, which can cover a large range of smoothness by changing the temperature parameter b, with $\mathrm{Dspike}(x) = 1$ or 0 for $x > 1$ or $x < 0$. We set $c = V_{th} = 0.5$ and determine a and d by setting $\mathrm{Dspike}(0) = 0$, $\mathrm{Dspike}(1) = 1$.

3.2 Differentiable hierarchical search for SNN

3.2.1 Cell level search

Similar to traditional DARTS [41], a cell is defined as a repeated and searchable unit, which is a directed acyclic graph with N nodes, $\{x_i\}_N$, as depicted in Fig. 1. Each cell receives input from two

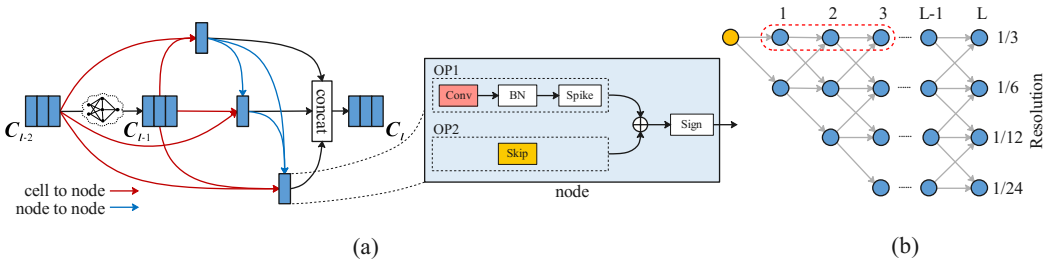


Figure 1: Hierarchical search space of SpikeDHS. (a) Cell level search space, with 3 nodes as an example. Within one node, operations are summed (2 operations here) and filtered by a sign function. BN denotes batch normalization. (b) Layer search space for event-based deep stereo.

previous cells and forms its output by concatenating all outputs of its nodes. In SpikeDHS, each node is a spiking neuron described by:

$$x_j = f(\sum_{i<j} o^{(i,j)}(x_i)) \tag{3}$$

where f is a sign function (or a spiking neuron) taking the sum of all operations as input, $o^{(i,j)}$ is the operation associated with the directed edge connecting node i and j. During search, each edge is represented by a weighted average of candidate operations, the information flow connecting node i and node j becomes:

$$\bar{o}^{(i,j)}(x) = \sum_{o \in O^{(i,j)}} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o \in O^{(i,j)}} \exp(\alpha_o^{(i,j)})} o(x) \tag{4}$$

where $O^{(i,j)}$ denotes the operation space on edge (i,j) and $\alpha_o^{(i,j)}$ is the weight of operation o, which is a trainable continuous variable. At the end of search, a discrete architecture is selected by replacing each mixed operation $\bar{o}^{(i,j)}$ with the most likely operation $o^{(i,j)} = \max_{o \in O^{(i,j)}} \alpha_o^{(i,j)}$. For spiking neurons, we can either mix the operation at the spike activation or at the input of the membrane potential, i.e. $y = \bar{o}$ or $I = \bar{o}$. The former allows a search over operations with different SG functions, while the latter transfers more accurate learning signals for α and leads to a concise node with less spiking filters (see supplement material for deduction). We choose the former for search phase and apply the latter particularly for SG search during retraining when operation type freeze (Section 3.4). Specifically, in the former choice, we distinguish between operations with the same forward computation but with different SG functions in candidate operation space. In addition, in original DARTS, a prepossessing step is required on operations between nodes and cells of previous layers in order to align the dimension of feature maps. We merge this step with subsequent candidate operations to reduce model complexity and improve inference speed.

3.2.2 Layer level search

Task specific knowledge has been proved to be helpful in speeding up the search process and improving network performance. For classification task, we adopt a fixed downsampling structure (Fig. 2) as in [41], with normal cells and reduction cells searched separately. For dense image prediction where high resolution output is needed and network architecture requires more variation, we implement differentiable search on the layer level as proposed in [40]. A set of scalars $\{\beta\}$ are trained to weight different potential layer resolutions and they are updated together with α. By the end of search, an optimal structure is decoded from a pre-defined L-layer trellis, as shown in Fig. 1. For upsampling layers, we use nearest interpolation to maintain binary feature maps. For the stereo matching task, following volumetric approaches [24, 74, 79] we construct a feature volume which embeds geometric knowledge of the binocular input, and search cell structure separately for the feature and matching subnetworks, similar to [10]. At the end of both subnetworks, the output of the last cell is upsampled to the initial resolution of the trellis using spike-based activation. We use batch normalization and ANN stem layers in the search phase. In the retraining phase, the ANN stem layers are converted to SNN by replacing the Relu function with spike function and retraining the weights from scratch. The reason for using Relu activation during search is to ensure more stable updating of the supernet, since deep SNNs may suffer from gradient vanish problem when the SG

functions is not chosen appropriately. In extended experiments we replace the Relu function in stem layers with Dspike function under appropriate hyperparameters, e.g $b = 3$, and the search process is also stable. Parameters of batch normalization are converted into convolution weights and biases after retraining [23, 58], leading to full spike-based network for inference. We provide more details for spike-based layer search space in the supplement.

3.3 Loss function, estimator and optimization

For classification task, we use an auxiliary loss as in [41], with weight 0.4. For event-based stereo matching, our network produces a matching cost tensor C of size $\frac{d_{max}}{2} \times h \times w$, based on which we estimate a disparity map $\hat{D}$ using a sub-pixel estimator [62].

$$\hat{D} = \sum_{d} D(d) \underset{d:|\hat{d}-d|<\delta}{\operatorname{softmin}} (C_{d,y,x}) \text{ , with } \hat{d} = \underset{d}{\operatorname{argmin}}(C_{d,y,x}) \tag{5}$$

where $\delta = 2$ is an estimator support and $D(d) = 2d$ is a disparity corresponding to index d in the matching cost tensor. We use a sub-pixel cross entropy loss [62] to train the network, which is described by:

$$L(\Theta) = \frac{1}{wh} \sum_{y,x} \sum_{d} \text{Laplace}(D(d)|\mu = D^{GT}_{y,x}, b) \cdot \log(\underset{d}{\operatorname{softmin}}(C_{d,y,x})) \tag{6}$$

where $\text{Laplace}(D|\mu = D^{GT}_{y,x}, b)$ is a discretized Laplace distribution with the mean equal to the ground truth disparity $\mu = D^{GT}_{y,x}$ and diversity $b = 2$. Following bi-level optimization [41], we update weight and architecture parameters $\{\alpha, \beta\}$ alternately based on two disjoint training sets *A* and *B*:

- Update network weights **w** by $\nabla_{\mathbf{w}} L(\mathbf{w}, \alpha, \beta)$ on *A*
- Update architecture parameters α and β by $\nabla_{\alpha,\beta} L(\mathbf{w}, \alpha, \beta)$ on *B*

We use first-order approximation to speed up the search process. After search, we decode the discrete cell structure by retaining two strongest afferent edges for each node. As to network structure, we decode it by finding the maximum probability path between different layers.

3.4 Differentiable surrogate gradient search

A recent work [21] shows that a suitable SG function is critical when the SNN goes deeper, and [36] demonstrates that by optimizing the width (or temperature) of the SG function the performance of SNN can be improved. Continuous relaxation through gradient descent is an efficient approach to explore diverse operations on the same path, inspired from this idea, we propose a differentiable surrogate gradient search (DGS) method to parallelly optimize local SGs for SNN.

In the retraining phase, with certain epoch intervals, we associate each operation path with N candidate SG functions, $\{g_i\}_N$, based on which we update the weight (or N copies of the weight) separately, leading to $\{w_{g_i}\}_N$. These weights are then combined to form a mixed operation weighted by a set of factors $\{\alpha_{g_i}\}_N$ through a softmax function, described as:

$$\hat{I} = \sum_{i=1}^{N} \frac{\exp(\alpha_{g_i})}{\sum_{j=1}^{N} \exp(\alpha_{g_j})} I_i \text{ , with } I_i = w_{g_i} x \tag{7}$$

We then update $\{\alpha_{g_i}\}_N$ through the loss of the mixed operation output. This process is repeated for multiple batches and finally we update the original SG to $\{g_i | i = \text{argmax}_i \langle \alpha_i \rangle\}$, with $\langle \cdot \rangle$ denoting the average over batches. Note that w_{g_i} can be obtained either by repeatedly calculate for each g_i, or directly estimating from the gradient of the original SG, $\nabla_{g,w} L$, if $\{g_i\}_N$ are linear to g. The psuedo code of the algorithm is summarized in Algorithm. 1. The intuition behind DGS is that the updated value of α_{g_i} indicates the contribution of w_{g_i} in decreasing the loss. So $\{g_i | i = \text{argmax}_{g_i} \langle \alpha_i \rangle\}$, which leads to the best updated weight, could be the most suitable SG function for the original local weight. Note that the difference between DGS and the SG search in Section 3.2.1 is that the former aims to optimize SG function for the local weight, while the latter is essentially a search of different operation types.

Algorithm 1: Differentiable surrogate gradient search (DGS)

Input: Training dataset, training epoch E, training iteration I, DGS iteration I_g, SG function g, candidate SG functions $\{g_i\}_N$, SG weighting factors $\{\alpha_{g_i}\}_N$ and their initializing value ϵ, epoch interval for DGS e_D

for *all* e *= 1, 2, ..., E-th epoch* **do**
  for *all* i *= 1, 2, ..., I-iteration* **do**
    Collect training data and labels, update weights w based on SG function g;
  if $e/e_D = \text{Int.}$ **then**
    for *all* j *=1, 2, ..., I_g-iteration* **do**
      Initialize $\{\alpha_{g_i}\}_N$ to ϵ, update weights w with $\nabla_{g_i,w}L$ based on $\{g_i\}_N$, obtain associated weights $\{w_{g_i}\}_N$;
      Combine $\{w_{g_i}\}_N$ to form mixed operation, update $\{\alpha_{g_i}\}_N$ with $\nabla_{\alpha_{g_i},\{w_{g_i}\}_N}L$.
    Update g to $\{g_i | i = \text{argmax}_{g_i}\ \langle\alpha_i\rangle\}$
return trained network.

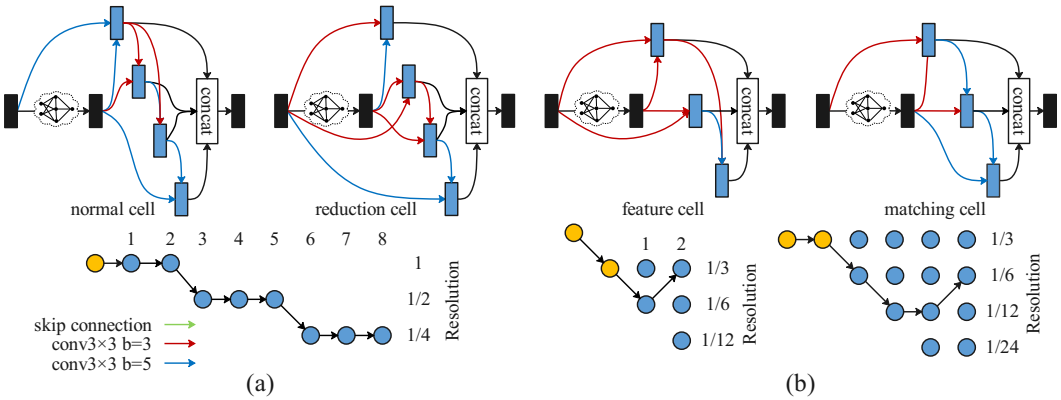


Figure 2: (a) Architecture for classification: SpikeDHS-CLA and (b) event-based stereo: SpikeDHS-Stereo.

4 Experiments

4.1 Classification

The CIFAR10 and CIFAR100 datasets [29] have 50K/10K training/testing RGB images with a spatial resolution of 32×32. The ImageNet dataset [13] contains more than 1250k training images and 50k test images. We apply SpikeDHS on CIFAR10 and then retrain on target datasets including CIFAR10, CIFAR100 and ImageNet. For ImageNet, we use a larger variant of the searched network with one more stem layer and two more cells. We use standard pre-processing and augmentation for training as in [22]. The test image is directly centered cropped to 224×224. More details about the model architecture and training are provided in the supplement material.

4.1.1 Architecture search and retrain

In the search phase, the training set is equally split into two subsets for bi-level optimazation. For retraining, the standard training/testing split is used. We use 4 nodes (n4) within one cell and a limited number of candidate operations to reduce search time, which are {conv3×3 with $g(b = 3)$, conv3×3 with $g(b = 5)$, skip connection}, with $g = \text{Dspike}$. For the sign function we use $g(b = 3)$. For the network architecture, we adopt an 8 layer fixed downsampling architecture proposed in [41]. The search phase takes 50 epochs with mini-batch size 50, the first 15 epochs are used to warm up convolution weights. We use SGD optimizer with momentum 0.9 and a learning rate of 0.025. The architecture search takes about 1.4 GPU day on a single NVIDIA Tesla V100 (32G) GPU. We term the searched architecture SpikeDHS-CLA and plot it in Fig. 2. After search, we retrain the model on target datasets with channel expansion for 100 epochs with mini-batch size

50 for CIFAR and 160 for ImageNet, with cosine learning rate 0.025. We use SGD optimizer with weight decay $3e^{-4}$ and momentum 0.9. The DGS method is applied to the first stem layer (s1), we set $I_g = 100$, $\epsilon = 0.001$, $e_D = 5$, $g = \text{Dspike}$ with $\{g_i\}_N$ having equal temperature interval: $\{g(b - \frac{(N-1)}{2}\Delta b), ..., g(b + \frac{(N-1)}{2}\Delta b)\}$ where $\Delta b = 0.2$ and $N = 5$. In extending experiments, we slightly increase the output channel of the first stem layer and use 3 nodes (n3) within a cell. In addition, we apply DGS to the first node of the 5th cell (c5).

4.1.2 Results

The results are summarized in Table 1 and the values of other models are obtained from literature. On CIFAR datasets, SpikeDHS-CLA with DGS achieves the highest accuracy comparing with other directly trained SNNs under similar model capacity. Note that both Dspike [36] and TET [14] use advanced training algorithms while plain SpikeDHS-CLA models are trained with fixed SG function. Our model also outperforms the recent SNN works with NAS in terms of accuracy, inference steps and model size[1]. As a reference, we run the original ANN DARTS network with the same architecture and it achieves 95.88% accuracy on CIFAR10. On ImageNet, our models surpass the ResNet-34-large model with much smaller model capacity and rivals VGG-16 with less than half of its size. We conjecture that the large number of jump connections both within and across cells potentially help gradient propagation through deep layers, which improves training of the model.

Table 1: Comparison on image classification. $^{\text{D}}$: DGS method. NoP: Number of parameters.

Dataset	Methods	Architecture	NoP	T	Accuracy[%]
CIFAR10	[76]TSSL-BP	CIFARNet	-	5	91.41
	[55]Diet-SNN	ResNet-20	-	10	92.54
	[78]STBP-tdBN	ResNet-19	13M	6	93.16
	[36]Dspike	ResNet-18	11M	6	94.25 ± 0.07
	[14]TET	ResNet-19	13M	6	94.50 ± 0.07
	[36]Dspike	ResNet-18	11M	6	94.25 ± 0.07
	[27]SNASNet	SNASNet-Bw	-	8	94.12 ± 0.25
	[47]AutoSNN	AutoSNN (C=128)	21M	8	93.15
	SpikeDHS	SpikeDHS-CLA (n4)	12M	6	**94.34 ± 0.06**
		SpikeDHS-CLA (n3)	14M	6	**95.35 ± 0.05**
	SpikeDHS$^{\text{D}}$	SpikeDHS-CLA (n4s1)	12M	6	**94.68 ± 0.05**
		SpikeDHS-CLA (n3s1)	14M	6	**95.36 ± 0.01**
		SpikeDHS-CLA (n3c5)	14M	6	**95.50 ± 0.03**
CIFAR100	[55]Diet-SNN	ResNet-20	-	5	64.07
	[78]STBP-tdBN	ResNet-19	13M	6	71.12 ± 0.57
	[36]Dspike	ResNet-18	11M	6	74.24 ± 0.10
	[14]TET	ResNet-19	13M	6	74.72 ± 0.28
	[27]SNASNet	SNASNet-Bw	-	5	73.04 ± 0.36
	[47]AutoSNN	AutoSNN (C=64)	5M	8	69.16
	SpikeDHS	SpikeDHS-CLA (n4)	12M	6	**75.70 ± 0.14**
		SpikeDHS-CLA (n3)	14M	6	**76.15 ± 0.20**
	SpikeDHS$^{\text{D}}$	SpikeDHS-CLA (n4s1)	12M	6	**76.03 ± 0.20**
		SpikeDHS-CLA (n3s1)	14M	6	**76.25 ± 0.10**
ImageNet	[68]STBP-tdBN	ResNet-34	22M	6	63.72
	[68]STBP-tdBN	ResNet-34-large	86M	6	67.05
	[55]Diet-SNN	VGG-16	138M	5	69.00
	SpikeDHS	SpikeDHS-CLA-large	58M	6	**67.96**
	SpikeDHS$^{\text{D}}$	SpikeDHS-CLA-large	58M	6	**68.64**

4.2 Event-based deep stereo

We further apply our method to event-based deep stereo matching on the widely used benchmark MVSEC dataset [82]. The dataset contains depth information recorded by LIDAR sensors and event streams collected from a pair of *Davis346* cameras, with synchronized 20 Hz gray scale images at

[1]The size of SNASNet is not given.

346×260 resolution. We split and preprocess the Indoor Flying dataset from the MVSEC following the same setting as [63, 1, 81]. We use the mean depth error (MDE), one-pixel-accuracy (1PA), median depth error, and mean disparity error as evaluation metrics. Learning from highly sparse raw events is challenging and a prepossessing step is often required to encode events. We use stacking based on time (SBT) [64] which merges events into temporally neighboring frames. During training, we use multiple consecutive stacks as one input, with an equal number of consecutive ground truth disparities as one label. More details can be found in the supplement material.

4.2.1 Architecture search and retrain

We use 3 nodes within one cell. For candidate operations, we use: {conv3×3 with $g(b=3)$, conv3×3 with $g(b=5)$, skip connection} for the feature net and {conv3×3×3 with $g(b=3)$, conv3×3×3 with $g(b=5)$, skip connection} for the matching net, with $g=\mathrm{Dspike}$. For the layer search space, we adopt a four-level trellis with downsampling rates of {3,2,2,2}. The number of layers is set to 2 and 4 for the feature and matching net, respectively. In addition, two stem layers are applied in front of both subnetworks to reduce input spatial resolution and increase channel number. We search the architecture for 12 epochs with batch size 1. The first 3 epochs are used to initiate the weight of the supernet. The rest 9 epochs are applied with bi-level optimization. We use SGD optimizer with momentum 0.9 and a learning rate of 0.002. The architecture search takes about 0.4 GPU day on a single NVIDIA Tesla V100 (32G) GPU. We term the searched architecture as SpikeDHS-Stereo and plot it in Fig. 2. Extensive random seed experiments show that the architectures are gradually optimized during the search phase. After search, we retrain the model with channel expansion for 200 epochs with mini-batch size 2. We use Adam optimizer with initial learning rate 0.001 and momentum $(0.9, 0.999)$, with learning rate decaying half at $[50, 100, 150]$ epochs. The DGS method is applied to the first stem layer of the feature net with the same setup as in classification. In extending experiments, we also retrain the network with operation mixed at the membrane potential (MM).

4.2.2 Results

We compare our method with other event-based stereo matching approaches (Section 2.3) on dense disparity estimation. The results are summarized in Table 8, values of other models are obtained from literature. Among SNN methods, SpikeDHS significantly outperforms StereoSpike, exhibiting the effectiveness of the searched architecture. Among events-only approaches, SpikeDHS even surpasses ANN-based specially designed DDES network in all criteria with only one-third of its number of parameters. The performance of SpikeDHS-Stereo is further improved with DGS. As shown in Fig. 4a-b, the temperature of the SG function is constantly optimized, which avoids vanishing gradient compared to training with fixed SG function, leading to more stable training of SNN. The MM approach also improves the performance of the network. A qualitative comparison of estimated disparities is shown in Fig. 3. It can be seen that SpikeDHS with DGS predicts better disparities compares to other models, especially in local edges.

Streaming inference: In real-world scenario, events are generated consecutively by the sensor with flexible lengths. To test the real-time applicability of our model, we fed the entire test split continuously into the model, which evolves an equal length of steps and estimates sequential disparities. The inference speed of our model achieves 44 FPS (26 FPS for the DDES model, see supplement material for measurement details) while achieving similar accuracy as in standard testing (Fig. 3), where the model always receives a fixed length of events.

4.2.3 Ablation study

The inherent temporal dynamics of SNN is assumed to be helpful for it to learn temporal correlation of the data. To investigate this property, we fix the membrane time constant τ to 0 to create binary neuron (BN) and perform architecture search and retrain. As shown in Table 8, the SpikeDHS-BN model performs worse than SpikeDHS, which shows the benefit of temporal dynamics in this task. In addition, we evaluate different SG functions including Superspike [72], Triangle [2], Arctan [17] SG functions with fixed hyperparameters during training and varying hyperparameters with DGS. The results demonstrate the robustness of DGS for different SG functions as it consistently improves network performance, as shown in Table 3. More details are provided in the supplement.

Table 2: We denote the best and second best results in bold and underscore. EO denotes events-only method. EITNet requires gray scale images for training but not for inference. Symbols meaning: -, unavailability of the value; (·), streaming tests; ^{D}, training with DGS; *, estimated value; NoP, Number of parameters.

Method	EO	NoP	MDE [cm] ↓		Median depth error [cm] ↓		Mean disparity error [pix] ↓		1PA [%] ↑	
			split1	split3	split1	split3	split1	split3	split1	split3
EIS [44]	✗	-	**13.7**	22.4	-	-	-	-	89.0	88.1
EITNet [1]	✓	>16M*	14.2	19.4	**5.9**	10.4	0.55	0.75	**92.1**	**89.6**
DDES [63]	✓	2.33M	16.7	27.8	6.8	14.7	0.59	0.94	89.4	74.8
StereoSpike [54]	✓	-	18.5	25.4	-	-	-	-	-	-
SpikeDHS-BN	✓	0.87M	17.5	20.3	7.1	10.8	0.59	0.74	88.7	88.3
SpikeDHS-BN^{D}	✓	0.87M	17.0	19.8	6.8	**10.2**	0.58	0.73	89.3	88.5
SpikeDHS	✓	0.87M	16.5(16.5)	19.4(19.5)	6.5(6.5)	10.6(10.6)	0.57(0.57)	0.73(0.74)	90.1(90.2)	88.5(88.4)
SpikeDHS^{D}	✓	0.87M	15.9(15.8)	**19.1**(19.3)	6.3(6.3)	10.4(10.5)	**0.54**(0.54)	**0.72**(0.74)	90.7(90.8)	88.9(88.8)
SpikeDHS (MM)	✓	0.87M	15.7(15.7)	-	6.3(6.3)	-	0.55(0.54)	-	91.0(91.1)	-
SpikeDHS^{D} (MM)	✓	0.87M	15.4(15.4)	-	6.0(6.0)	-	**0.54**(0.54)	-	91.3(91.4)	-

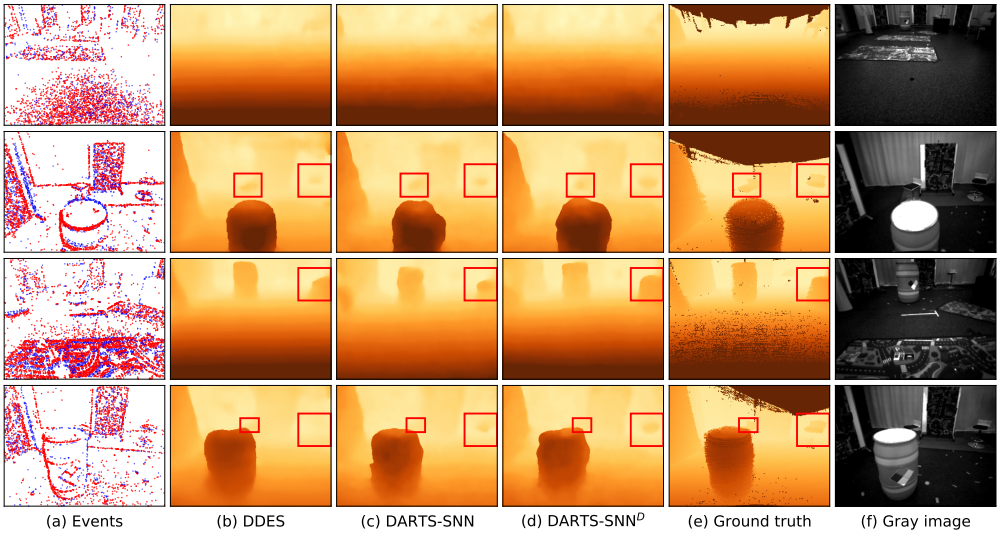


Figure 3: Qualitative comparison on MVSEC. Disparity maps of different methods are on same frames, see supplement material for details.

Table 3: Different SG functions for event-based stereo task on split 1 w/ and w/o DGS.

SG function	1PA w/ DGS (w/o DGS) [%] ↑
Triangle [2]	91.3 (90.9)
Arctan [17]	90.1 (89.3)
Superspike [72]	89.7 (89.6)

5 Sparsity and energy cost

The dense computation of deep ANNs came at a significant energy cost. In contrast, SNN performs sparse computing and multiplication-free inference. As Fig. 5 shows, our models show an overall sparse activity with different degrees across layers. For event-based stereo, we plot network activity along with the density of events for a short duration (Fig. 4c-d). The activity of the first stem layer in the feature net highly correlates with the density of the event streams. This relation is weakened with the increase of layer depth. As activity propagates, the sparsity of matching layers drops to create dense disparities. The operation number of ANN is collected by a public available PyTorch package [61]. Following [36], the addition count of SNN is calculated by $s * T * A$, where s is the mean sparsity, T is the time step and A is the addition number. The operation number and

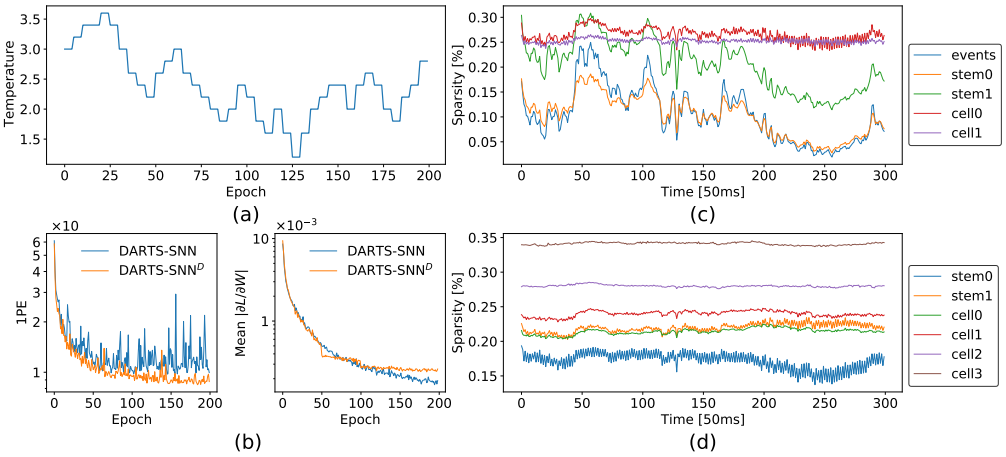
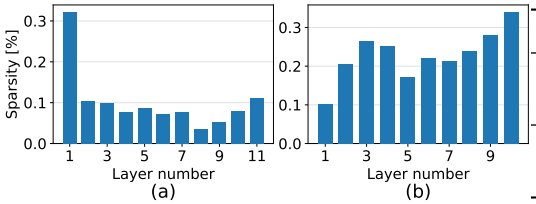


Figure 4: DGS training process and sparsity of SpikeDHS-Stereo. (a): Evolution of SG temperatures during DGS training. (b): 1PE (left) and first stem weight gradient (right) of DGS training and normal training. (c) and (d): Sparsity of the feature net and the matching net.

energy cost is in Table 3. We can see that SpikeDHS-Stereo has much lower operation number than DDES. Note that due to streaming inference, SNN realizes a natural usage of its temporal dynamics with $T = 1$. We measure the energy consumption following [55]. In 45nm CMOS technology, the addition operation in SNN costs 0.9pJ while the multiply-accumulate (MAC) operation in ANN consumes 4.6pJ. SpikeDHS-Stereo costs only 6.7mJ for a single forward, with $26\times$ lower energy than DDES.



Method	#Add.	#Mult.	Energy
DARTS CIFAR10	29654M	29654M	136.4 mJ
SpikeDHS-CLA (n4)	**5973M**	**19M**	**5.5** mJ
DDES (ANN)	37917M	37917M	174.4 mJ
SpikeDHS-Stereo	**7414M**	**0M**	**6.7** mJ

Figure 5 & Table 3: Left: Network sparsity for (a) classification and (b) event-based stereo. Right: The operation number and energy cost.

6 Discussion

In this paper, we develop the SpikeDHS framework which finds optimal architectures for SNNs. On image classification and event-based deep stereo tasks, our models outperform previous SNNs with ANN or handcrafted architectures. The latter task demonstrates the advantage of SNNs in processing highly sparse and dynamic event streams with extremely low power. One potential reason of the high performance of our architecture could be the large number of jump connections both within and across cells, which helps gradient propagation. Our method can be applied to other tasks where network requires more architectural variations such as object detection and semantic segmentation. The DGS algorithm efficiently optimizes SGs locally, which is of general usage in improving training of deep SNNs or binary networks. Currently we only apply DGS on a fixed layer, future works could develop a dynamical algorithm which applies it on optimal layers. In addition, we only use limited candidate operations, other types of biological plausible connections such as recurrent and feedback connections, as well as excitatory/inhibitory synapses could be considered in the future.

7 Cell search space

7.1 Mixed operation

For spiking neurons, we can either mix the operation at the spike activation or at the input of the membrane potential. The former allows a search over operations with different SG functions, while the latter transfers more accurate learning signals for α and leads to a concise node with fewer spiking filters, as we show in the following.

For mixed operation at the spike function, $\bar{y} = \bar{o}$ and $y = o$. The mixed operation is then:

$$\bar{y} = \sum_{k}^{K} \frac{e^{\alpha_k}}{Z} y_k \tag{8}$$

where $Z = \sum_{k}^{K} e^{\alpha_k}$ is the normalizing constant, K denotes the number of candidate operations on one edge and k is the index of the operation.

Given loss L, assuming $T = 1$ for simplification, the gradient of α_k is derived as:

$$\begin{aligned}
\frac{\partial L}{\partial \alpha_k} &= \sum_{t}^{T} \frac{\partial L}{\partial f^t} \frac{\partial f^t}{\partial \bar{y}^t} \frac{\partial \bar{y}^t}{\partial \alpha_k} \\
&= \frac{\partial L}{\partial f} \frac{\partial f}{\partial \bar{y}} \frac{e^{\alpha_k}}{Z} \left[y_k - \frac{\sum_{k}^{K} (e^{\alpha_k} y_k)}{Z} \right] \\
&= \frac{\partial L}{\partial f} \frac{\partial f}{\partial \bar{y}} \frac{e^{\alpha_k}}{Z} \Big(y_k - \bar{y} \Big)
\end{aligned} \tag{9}$$

where f is the sign function of the node and for $\frac{\partial f}{\partial \bar{y}}$ we apply a fixed SG function.

For mixed operation at the input of the membrane potential, $\bar{I} = \bar{o}$ and $I = o$. The mixed operation is then:

$$\bar{I} = \sum_{k}^{K} \frac{e^{\alpha_k}}{Z} I_k \tag{10}$$

Similarly, the gradient of α_k is then derived as:

$$\begin{aligned}
\frac{\partial L}{\partial \alpha_k} &= \sum_{t}^{T} \frac{\partial L}{\partial y^t} \frac{\partial y^t}{\partial u^t} \frac{\partial u^t}{\partial \bar{I}^t} \frac{\partial \bar{I}^t}{\partial \alpha_k} \\
&= \frac{\partial L}{\partial y} \cdot g(u) \cdot 1 \cdot \frac{e^{\alpha_k}}{Z} \Big(I_k - \bar{I}_k \Big)
\end{aligned} \tag{11}$$

For mixed operation at the spike activation, as Eq. 9 shows, the gradient of α_k stops at $y_k - \bar{y}$, so for candidate operations we can use different SG functions for the corresponding spike activation. However, the learning signal of α is also filtered by an additional SG function $\frac{\partial f}{\partial \bar{y}}$ at the node, which could cause additional noise. In addition, when applied for DGS method, the contribution of w_{g_i} will be filtered by the consecutive spike activation. If the difference between the original weight and w_{g_i} is minor, they could lead to the same spike activation and in the extreme case $y_{g_i} - \bar{y}$ could be 0. For mixed operation at the input of membrane potential, as Eq. 11 shows, the gradient of α_k applies a unified SG function $g(u)$ for different candidate operations which limits the exploration of diverse SG functions. However, since I_k directly depends on the weight, when applied for DGS method it strictly reflects different changes of the original weight, thus giving more accurate learning signals for α_k. Also, in forward path this leads to a more concise node without the sign filter, as shown in Fig. 6.

8 Layer search space

As stated in the main text, to ensure spike-based computation for the network, we apply the nearest interpolation for upsampling in order to maintain a binary feature map. For event-based deep stereo, at the end of each subnetwork, the feature map of the last cell is upsampled to the original resolution. In the feature subnetwork, this is done by upsampling (double scaled) followed by convolution with

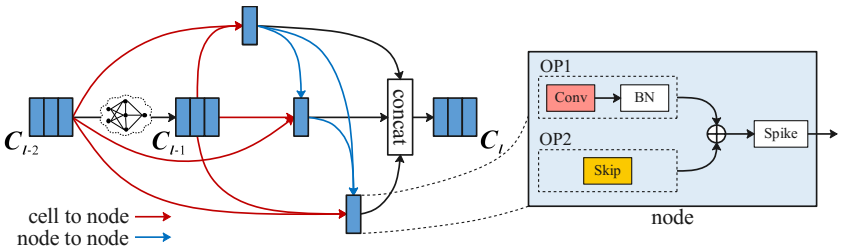


Figure 6: Mixed operation at the input of membrane potential. 2 operations are shown here as an example.

spike activation, this process is repeated for multiple times until the feature map is recovered to the original resolution. In the matching subnetwork, we cancel the last spike activation and directly output the membrane potential in order to increase the representation ability.

9 Classification

For classification task, we inherit the searching structure from DARTS [41] and make some adjustments. Same as most NAS structures, we adopt stem layer as our first layer to extract features. Meanwhile it's also a spike emitter to transmit floating numbers to spikes. Then we employ cell structure as our basic unit to construct our whole searching structure. The stem layer takes 3-channel images as input and outputs 48-channel spiking feature maps. For the rest of the network, we employ 6 normal cells and 2 reduction cells. Normal cells keep the dimension of feature maps unchanged, while reduction cells (Cell2 and Cell5) halve the spatial size and double the channel number of former feature maps. Finally, we use a global pooling and a fully-connected layer to end the whole classification process. The auxiliary loss is applied at Cell5. In retraining stage, we use 8 cells as well to avoid some intrinsic problem of DARTS to some extent, like depth-gap between searching structure and retraining structure as described in [8]. In this stage, stem layer uses 108 output channels rather than 48, to extract more features from row image data. The detailed network structure for retraining is shown in Table 4. In extending experiments, we slightly increase the output channel of the first stem layer to 144 and use 3 nodes within a cell, the size of the output feature map of the cell remains the same. Detailed architectures of SpikeDHS-CLA (node=4), SpikeDHS-CLA (node=3) and SpikeDHS-CLA-large (ImageNet) are listed in Table 4, 5 and 6.

Table 4: Network architecture of SpikeDHS-CLA for CIFAR with 4 nodes in a cell.

Layer	**Feature map size** $c \times h \times w$
Stem	$108 \times 32 \times 32$
Cell0	$144 \times 32 \times 32$
Cell1	$144 \times 32 \times 32$
Cell2	$288 \times 16 \times 16$
Cell3	$288 \times 16 \times 16$
Cell4	$288 \times 16 \times 16$
Cell5	$576 \times 8 \times 8$
Cell6	$576 \times 8 \times 8$
Cell7	$576 \times 8 \times 8$
Pooling	$576 \times 1 \times 1$
FC	10

10 Event-based deep stereo

We split and preprocess the Indoor Flying dataset from the MVSEC dataset following the same setting as [63, 1, 81]. In split one, 3110 samples from the Indoor Flying 2 and 3 are used as the training set while 861 and 200 samples from the Indoor Flying 1 are used as the test set and validation set. In split three, 2600 samples from the Indoor Flying 2 and 3 are used as the training set while 1343 and 200 samples from the Indoor Flying 1 are used as the test set and validation set.

Table 5: Network architecture of SpikeDHS-CLA for CIFAR with 3 nodes in a cell.

Layer	**Feature map size** $c \times h \times w$
Stem	$144 \times 32 \times 32$
Cell0	$144 \times 32 \times 32$
Cell1	$144 \times 32 \times 32$
Cell2	$288 \times 16 \times 16$
Cell3	$288 \times 16 \times 16$
Cell4	$288 \times 16 \times 16$
Cell5	$576 \times 8 \times 8$
Cell6	$576 \times 8 \times 8$
Cell7	$576 \times 8 \times 8$
Pooling	$576 \times 1 \times 1$
FC	10

Table 6: Network architecture of SpikeDHS-CLA-large for ImageNet with 3 nodes in a cell.

Layer	**Feature map size** $c \times h \times w$
Stem0	$72 \times 112 \times 112$
Stem1	$144 \times 56 \times 56$
Cell0	$144 \times 56 \times 56$
Cell1	$144 \times 56 \times 56$
Cell2	$288 \times 28 \times 28$
Cell3	$288 \times 28 \times 28$
Cell4	$576 \times 14 \times 14$
Cell5	$576 \times 14 \times 14$
Cell6	$576 \times 8 \times 8$
Cell7	$1152 \times 7 \times 7$
Cell8	$1152 \times 7 \times 7$
Cell9	$1152 \times 7 \times 7$
Pooling	$1152 \times 1 \times 1$
FC	1000

10.1 Event encoding

We use stacking based on time (SBT) [64] which merges events into temporally neighboring frames. Specifically, within one stack, a duration of Δt event stream is compressed into n frames. The value of each pixel in the ith frame is defined as the accumulated polarity of events:

$$P(x, y) = \text{sign}(\sum_{t \in T} p(x, y, t)) \tag{12}$$

where P is the value of the pixel at (x, y), t is the timestamp, p is the polarity of the event and $T \in [\frac{(i-1)\Delta t}{n}, \frac{i\Delta t}{n}]$ is the duration of events merged into one frame. The advantage of SBT is that it is simple and has low computation cost. Besides, event cameras are often embedded with accumulator modules that directly output events in SBT. However, when too few events happening during the interval, SBT may produce a very sparse event image. This limitation can be alleviated for SNN by its intrinsic temporal accumulation effect of the membrane potential. During training, we set $\Delta t = 50\text{ms}$, $n = 5$ and $T = 10\text{ms}$ for each stack and use $6 + 2$ consecutive stacks as one input (with the first 2 stacks as burn in time), corresponding to 6 consecutive ground truth disparities as one label. A training epoch on split one thus contains 518 batches (no time overlap between label batches).

10.2 Architecture search and retrain

We use random cropping for data augmentation as well as memory saving. The input images are cropped from 260×346 to 200×280 with 50% probability at random positions. Our feature subnetwork produces a pair of left and right feature maps F^L and F^R. Following volumetric approaches in deep stereo matching [24, 74, 79], we construct a feature volume by concatenating the left feature map and disparity shifted right feature map in channel dimension. With a given disparity

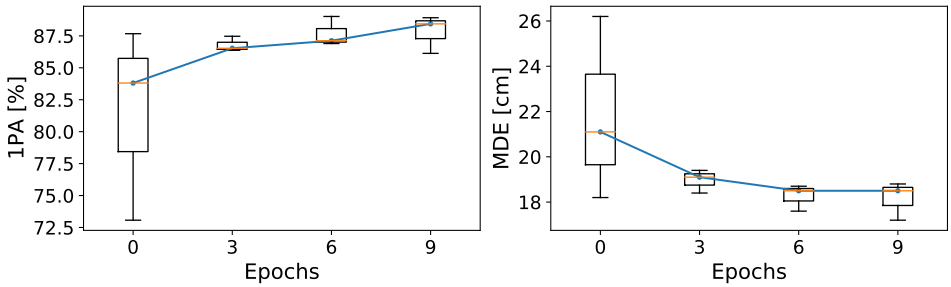


Figure 7: Search progress of SpikeDHS on MVSEC. We record the searched architecture for every 3 epochs. Each recorded architecture is retrained from scratch for 100 epochs and then evaluated on the validation set. We repeat the experiment for 3 times with different random seeds and report the best 1PA and MDE for each recorded architecture.

shift $d_s \in \{0, 1, 2, \ldots, D\}$, the feature volume for pixel $\mathbf{x}$ can be expressed as:

$$C(\mathbf{x}, d_s) = \text{concat}(F^L(\mathbf{x}), F^R(x - d_s)), C \in \mathbb{R}^{D \times H \times W} \tag{13}$$

For architecture search, we set the channel number of the last stem feature map to 12. In the retraining phase, we expand this channel number from 12 to 24 to improve network performance, the rest of the network also have channel number doubled accordingly. We set $D = 33$, close to the maximum disparity of the MVSEC dataset, which is 37. The detailed network architecture for retraining is shown in Table 7.

Table 7: Detailed architecture of SpikeDHS-Stereo. $C = 12, H = 260, W = 346, D = 33$.

Module	Layer	Feature map size
Feature net	Stem0	$C \times H \times W$
	Stem1	$2C \times \frac{1}{3}H \times \frac{1}{3}W$
	Cell0	$4C \times \frac{1}{6}H \times \frac{1}{6}W$
	Cell1	$2C \times \frac{1}{3}H \times \frac{1}{3}W$
	Upsampling	$2C \times \frac{1}{3}H \times \frac{1}{3}W$
Feature Volume	Concat	$4C \times D \times \frac{1}{3}H \times \frac{1}{3}W$
Matching net	Stem0	$2C \times D \times \frac{1}{3}H \times \frac{1}{3}W$
	Stem1	$2C \times D \times \frac{1}{3}H \times \frac{1}{3}W$
	Cell0	$4C \times \frac{1}{2}D \times \frac{1}{6}H \times \frac{1}{6}W$
	Cell1	$8C \times \frac{1}{4}D \times \frac{1}{12}H \times \frac{1}{12}W$
	Cell2	$8C \times \frac{1}{4}D \times \frac{1}{12}H \times \frac{1}{12}W$
	Cell3	$4C \times \frac{1}{2}D \times \frac{1}{6}H \times \frac{1}{6}W$
	Upsampling	$1 \times D \times \frac{1}{3}H \times \frac{1}{3}W$
Estimator	Estimator	$H \times W$

10.3 Random seed experiments

To determine the final architecture, we repeat the experiment for 3 times with different random seeds and select the best architecture based on the validation performance obtained by training from scratch for a short period. Fig. 7 shows that while the architecture exhibits certain sensitivity for initialization, it can be gradually optimized during the search phase. This gradual improvement in dense image prediction may in a degree owing to the layer level optimization. In the classification task, we didn't observe a smooth improvement of the architecture during the search phase. This could be due to the intrinsic optimization problem of the original DARTS [41]. This issue is discussed in [69, 11, 7, 66] where various solutions have proposed. Future works can potentially be improved using these methods.

Table 8: We denote the best and second best results in bold and underscore. Symbols meaning: (·), streaming tests;$^{\mathrm{D}}$, training with DGS; MM, mixed operation at the membrane potential

Method	EO	No. param.	MDE [cm] ↓		Median depth error [cm] ↓		Mean disparity error [pix] ↓		1PA [%] ↑	
			split1	split3	split1	split3	split1	split3	split1	split3
DDES [63]	✓	2.33M	16.7	27.8	6.8	14.7	0.59	0.94	89.4	74.8
SpikeDHS	✓	0.87M	16.5(16.5)	19.4(19.5)	6.5(6.5)	10.6(10.6)	0.57(0.57)	0.73(0.74)	90.1(90.2)	88.5(88.4)
SpikeDHS$^{\mathrm{D}}$	✓	0.87M	15.9(15.8)	**19.1**(19.3)	6.3(6.3)	**10.4**(10.5)	**0.54**(0.54)	**0.72**(0.74)	90.7(90.8)	**88.9**(88.8)
DDES-SNN	✓	2.33M	53.4	-	37.8	-	2.08	-	37.1	-
SpikeDHS (MM)	✓	0.87M	15.7(15.7)	-	6.3(6.3)	-	0.55(0.54)	-	91.0(91.1)	-
SpikeDHS$^{\mathrm{D}}$ (MM)	✓	0.87M	**15.4**(15.4)	-	**6.0**(6.0)	-	**0.54**(0.54)	-	**91.3**(91.4)	-

10.4 Further ablation study

In the main text, we reported results based on mixed operation at the spike function for the cell (for DGS we applied mixed operation at the membrane potential since there is no multiple spike functions at the stem layer) and results from extending experiments with mixed operation at the membrane potential (MM). To save computation time, we directly convert the searched architecture from mixed at spike to mixed at membrane potential. Specifically, we reinitialize all SG function to $b = 3$ and retrain the model on split one. We also apply DGS in the first stem layer as before. In addition, to compare with ANN converted SNN network, we convert the DDES model to spike-based model, i.e replace Relu units with LIF neurons, to create DDES-SNN model and directly train it. The results are summarized in Table 8. Models with mixed operation at the membrane potential achieves significantly better results, probably due to a more concise node which uses fewer SG function. The performance of the model is further improved with DGS method, demonstrating the effectiveness of the algorithm. The DDES-SNN model performs much worse than SpikeDHS-Stereo, though additional fine tuning could potentially improve its performance, it shows the difficulty of directly applying ANN architecture for spiking neurons in dense prediction task.

10.5 FPS calculation

For DARTS-SNN, We use 8 consecutive stacks as one input and calculate corresponding inference time of the network. Then we divide the time by 8 to obtain the inference time of producing one disparity map. For DDES we run the public available code with one input sample and produce one disparity map. Both experiments are repeated multiple times to obtain the average duration. These experiments are performed on a single NVIDIA Tesla V100 (32G) GPU.

10.6 Different SGs for DGS

We evaluate four different SG functions including Dspike [36] (Eq. 14), Triangle [2] (Eq. 15), Superspike [72] (Eq. 16) and Arctan [17] (Eq. 17) functions with fixed hyperparameters during training and varying hyperparameters with DGS. These functions are described as following

$$\delta'_D(x) = \frac{b}{2} \cdot \frac{1 - \tanh^2(b(x - \frac{1}{2}))}{\tanh(\frac{b}{2})} \text{ if } 0 \leq x \leq 1 \tag{14}$$

$$\delta'_T(x) = b \cdot \max\{0, 1 - |x - \frac{1}{2}|\} \tag{15}$$

$$\delta'_S(x) = \frac{1}{(b \cdot |x| + 1)^2} \tag{16}$$

$$\delta'_A(x) = \frac{b/3}{1 + (\pi x)^2} \tag{17}$$

Each SG function has a temperature factor b to control its shape through DGS. A visualization is shown in Fig. 8.

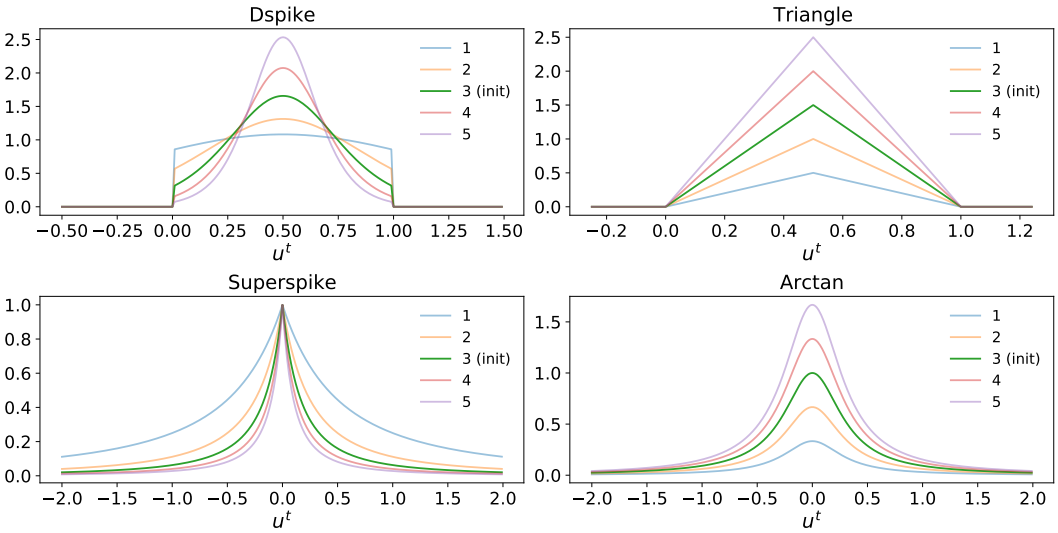


Figure 8: The shape of different SG functions with different temperature factor b. We set the initial value of b to 3.

References

[1] Soikat Hasan Ahmed, Hae Woong Jang, SM Nadim Uddin, and Yong Ju Jung. Deep event stereo leveraged by event-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 882–890, 2021.

[2] Guillaume Bellec, Darjan Salaj, Anand Subramoney, Robert Legenstein, and Wolfgang Maass. Long short-term memory and learning-to-learn in networks of spiking neurons. *Advances in neural information processing systems*, 31, 2018.

[3] Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 11(1):1–15, 2020.

[4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[5] Sander M Bohte, Joost N Kok, and Johannes A La Poutré. Spikeprop: backpropagation for networks of spiking neurons. In *ESANN*, volume 48, pages 419–424. Bruges, 2000.

[6] Tong Bu, Wei Fang, Jianhao Ding, PengLin Dai, Zhaofei Yu, and Tiejun Huang. Optimal ann-snn conversion for high-accuracy and ultra-low-latency spiking neural networks. In *International Conference on Learning Representations*, 2021.

[7] Xiangning Chen and Cho-Jui Hsieh. Stabilizing differentiable architecture search via perturbation-based regularization. In *International conference on machine learning*, pages 1554–1565. PMLR, 2020.

[8] Xin Chen, Lingxi Xie, Jun Wu, and Qi Tian. Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1294–1303, 2019.

[9] Yukang Chen, Tong Yang, Xiangyu Zhang, Gaofeng Meng, Xinyu Xiao, and Jian Sun. Detnas: Backbone search for object detection. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33:22158–22169, 2020.

[11] Xiangxiang Chu, Xiaoxing Wang, Bo Zhang, Shun Lu, Xiaolin Wei, and Junchi Yan. Darts-: robustly stepping out of performance collapse without indicators. *arXiv preprint arXiv:2009.01027*, 2020.

[12] 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*, 109(5):911–934, 2021.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. *arXiv preprint arXiv:2202.11946*, 2022.

[15] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

[16] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[17] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2661–2671, 2021.

[18] Steve B Furber, Francesco Galluppi, Steve Temple, and Luis A Plana. The spinnaker project. *Proceedings of the IEEE*, 102(5):652–665, 2014.

[19] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 154–180, 2020.

[20] Jianyuan Guo, Kai Han, Yunhe Wang, Chao Zhang, Zhaohui Yang, Han Wu, Xinghao Chen, and Chang Xu. Hit-detector: Hierarchical trinity architecture search for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11405–11414, 2020.

[21] Jesse Hagenaars, Federico Paredes-Vallés, and Guido De Croon. Self-supervised learning of event-based optical flow with spiking neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[24] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[25] Seijoon Kim, Seongsik Park, Byunggook Na, and Sungroh Yoon. Spiking-yolo: Spiking neural network for energy-efficient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11270–11277, 2020.

[26] Youngeun Kim, Joshua Chough, and Priyadarshini Panda. Beyond classification: Directly training spiking neural networks for semantic segmentation. *arXiv preprint arXiv:2110.07742*, 2021.

[27] Youngeun Kim, Yuhang Li, Hyoungseob Park, Yeshwanth Venkatesha, and Priyadarshini Panda. Neural architecture search for spiking neural networks. *arXiv preprint arXiv:2201.10355*, 2022.

[28] Agnes Korcsak-Gorzo, Michael G Müller, Andreas Baumbach, Luziwei Leng, Oliver J Breitwieser, Sacha J van Albada, Walter Senn, Karlheinz Meier, Robert Legenstein, and Mihai A Petrovici. Cortical oscillations support sampling-based computations in spiking neural networks. *PLoS computational biology*, 18(3):e1009753, 2022.

[29] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

[30] Alexander Kugele, Thomas Pfeil, Michael Pfeiffer, and Elisabetta Chicca. Efficient processing of spatio-temporal data streams with spiking neural networks. *Frontiers in neuroscience*, 14: 439, 2020.

[31] Akos F Kungl, Sebastian Schmitt, Johann Klähn, Paul Müller, Andreas Baumbach, Dominik Dold, Alexander Kugele, Eric Müller, Christoph Koke, Mitja Kleider, et al. Accelerated physical emulation of bayesian inference in spiking neural networks. *Frontiers in neuroscience*, page 1201, 2019.

[32] Chankyu Lee, Adarsh Kumar 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.

[33] Luziwei Leng. *Solving Machine Learning Problems with Biological Principles*. PhD thesis, 2020.

[34] Luziwei Leng, Roman Martel, Oliver Breitwieser, Ilja Bytschok, Walter Senn, Johannes Schemmel, Karlheinz Meier, and Mihai A Petrovici. Spiking neurons with short-term synaptic plasticity form superior generative networks. *Scientific reports*, 8(1):1–11, 2018.

[35] Yuhang Li, Shikuang Deng, Xin Dong, Ruihao Gong, and Shi Gu. A free lunch from ann: Towards efficient, accurate spiking neural networks calibration. In *International Conference on Machine Learning*, pages 6316–6325. PMLR, 2021.

[36] Yuhang Li, Yufei Guo, Shanghang Zhang, Shikuang Deng, Yongqing Hai, and Shi Gu. Differentiable spike: Rethinking gradient-descent for training spiking neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[37] Hualou Liang, Xiajing Gong, Minggui Chen, Yin Yan, Wu Li, and Charles D Gilbert. Interactions between feedback and lateral connections in the primary visual cortex. *Proceedings of the National Academy of Sciences*, 114(32):8637–8642, 2017.

[38] Peiwen Lin, Peng Sun, Guangliang Cheng, Sirui Xie, Xi Li, and Jianping Shi. Graph-guided architecture search for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4203–4212, 2020.

[39] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[40] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 82–92, 2019.

[41] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[42] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

[43] Paul A Merolla, John V Arthur, Rodrigo Alvarez-Icaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345 (6197):668–673, 2014.

[44] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4258–4267, 2021.

[45] Vernon B Mountcastle. The columnar organization of the neocortex. *Brain: a journal of neurology*, 120(4):701–722, 1997.

[46] Milad Mozafari, Saeed Reza Kheradpisheh, Timothée Masquelier, Abbas Nowzari-Dalini, and Mohammad Ganjtabesh. First-spike-based visual categorization using reward-modulated stdp. *IEEE transactions on neural networks and learning systems*, 29(12):6178–6190, 2018.

[47] Byunggook Na, Jisoo Mok, Seongsik Park, Dongjin Lee, Hyeokjun Choe, and Sungroh Yoon. Autosnn: Towards energy-efficient spiking neural networks. *arXiv preprint arXiv:2201.12738*, 2022.

[48] Emre Neftci, Srinjoy Das, Bruno Pedroni, Kenneth Kreutz-Delgado, and Gert Cauwenberghs. Event-driven contrastive divergence for spiking neuromorphic systems. *Frontiers in neuroscience*, 7:272, 2014.

[49] 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.

[50] Vladimir Nekrasov, Hao Chen, Chunhua Shen, and Ian Reid. Fast neural architecture search of compact semantic segmentation models via auxiliary cells. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9126–9135, 2019.

[51] J. Pei, L. Deng, S. Song, M. Zhao, and L. Shi. Towards artificial general intelligence with hybrid tianjic chip architecture. *Nature*, 572(7767):106, 2019.

[52] Mihai A Petrovici, Johannes Bill, Ilja Bytschok, Johannes Schemmel, and Karlheinz Meier. Stochastic inference with spiking neurons in the high-conductance state. *Physical Review E*, 94 (4):042312, 2016.

[53] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International conference on machine learning*, pages 4095–4104. PMLR, 2018.

[54] Ulysse Rançon, Javier Cuadrado-Anibarro, Benoit R Cottereau, and Timothée Masquelier. Stereospike: Depth learning with a spiking neural network. *arXiv preprint arXiv:2109.13751*, 2021.

[55] Nitin Rathi and Kaushik Roy. Diet-snn: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[56] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 4780–4789, 2019.

[57] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019.

[58] Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017.

[59] Sumit Bam Shrestha and Garrick Orchard. Slayer: Spike layer error reassignment in time. *arXiv preprint arXiv:1810.08646*, 2018.

[60] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[61] Vladislav Sovrasov. Flops counter for convolutional networks in pytorch framework, 2019. URL `https://github.com/sovrasov/flops-counter.pytorch/`.

[62] Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. *arXiv preprint arXiv:1806.01677*, 2018.

[63] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1527–1537, 2019.

[64] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019.

[65] Ning Wang, Yang Gao, Hao Chen, Peng Wang, Zhi Tian, Chunhua Shen, and Yanning Zhang. Nas-fcos: Fast neural architecture search for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11943–11951, 2020.

[66] Ruochen Wang, Minhao Cheng, Xiangning Chen, Xiaocheng Tang, and Cho-Jui Hsieh. Rethinking architecture selection in differentiable nas. *arXiv preprint arXiv:2108.04392*, 2021.

[67] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018.

[68] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, Yuan Xie, and Luping Shi. Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1311–1318, 2019.

[69] Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, and Hongkai Xiong. Pc-darts: Partial channel connections for memory-efficient architecture search. *arXiv preprint arXiv:1907.05737*, 2019.

[70] Zheyu Yang, Yujie Wu, Guanrui Wang, Yukuan Yang, Guoqi Li, Lei Deng, Jun Zhu, and Luping Shi. Dashnet: A hybrid artificial and spiking neural network for high-speed object tracking. *arXiv preprint arXiv:1909.12942*, 2019.

[71] Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662*, 2019.

[72] Friedemann Zenke and Surya Ganguli. Superspike: Supervised learning in multilayer spiking neural networks. *Neural computation*, 30(6):1514–1541, 2018.

[73] Friedemann Zenke and Tim P Vogels. The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. *Neural Computation*, 33(4):899–925, 2021.

[74] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019.

[75] Kaixuan Zhang, Kaiwei Che, Jianguo Zhang, Jie Cheng, Ziyang Zhang, Qinghai Guo, and Luziwei Leng. Discrete time convolution for fast event-based stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8676–8686, 2022.

[76] Wenrui Zhang and Peng Li. Temporal spike sequence learning via backpropagation for deep spiking neural networks. *Advances in Neural Information Processing Systems*, 33:12022–12033, 2020.

[77] Yiheng Zhang, Zhaofan Qiu, Jingen Liu, Ting Yao, Dong Liu, and Tao Mei. Customizable architecture search for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11641–11650, 2019.

[78] Hanle Zheng, Yujie Wu, Lei Deng, Yifan Hu, and Guoqi Li. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11062–11070, 2021.

[79] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017.

[80] Yi Zhou, Guillermo Gallego, Henri Rebecq, Laurent Kneip, Hongdong Li, and Davide Scaramuzza. Semi-dense 3d reconstruction with a stereo event camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018.

[81] Alex Zihao Zhu, Yibo Chen, and Kostas Daniilidis. Realtime time synchronized event-based stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 433–447, 2018.

[82] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, 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, 2018.

[83] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. *arXiv preprint arXiv:2201.10943*, 2022.

[84] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

[85] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.