Pushing the Limits of Asynchronous Graph-based Object Detection with Event Cameras

Daniel Gehrig and Davide Scaramuzza
Robotics and Perception Group, University of Zurich

Abstract

State-of-the-art machine-learning methods for event cameras treat events as dense representations and process them with conventional deep neural networks. Thus, they fail to maintain the sparsity and asynchronous nature of event data, thereby imposing significant computation and latency constraints on downstream systems. A recent line of work tackles this issue by modeling events as spatiotemporally evolving graphs that can be efficiently and asynchronously processed using graph neural networks. These works showed impressive computation reductions, yet their accuracy is still limited by the small scale and shallow depth of their network, both of which are required to reduce computation. In this work, we break this glass ceiling by introducing several architecture choices which allow us to scale the depth and complexity of such models while maintaining low computation. On object detection tasks, our smallest model shows up to 3.7 times lower computation, while outperforming state-of-the-art asynchronous methods by 7.4 mAP. Even when scaling to larger model sizes, we are 13% more efficient than state-of-the-art while outperforming it by 11.5 mAP. As a result, our method runs 3.7 times faster than a dense graph neural network, taking only 8.4 ms per forward pass. This opens the door to efficient, and accurate object detection in edge-case scenarios.

1. Introduction

Humans can detect fast-moving objects in the blink of an eye thanks to their sophisticated visual cortex originally designed to hunt and spot prey. Today, computer vision researchers try to emulate these systems with data-driven object detection algorithms, which have found widespread application in robotic and automotive settings. However, state-of-the-art approaches operate on data from frame-based sensors like RGB cameras or LiDARs and, for this reason, suffer from a bandwidth-latency tradeoff: at high speeds, they require a high framerate to reduce perceptual latency, but this introduces a significant bandwidth overhead for downstream systems; reducing the framerate reduces the bandwidth requirements but at the cost of missing important scene dynamics, like the ones in Fig. 1, due to increased perceptual latency.

In recent years, event cameras have emerged as alternative sensors that do not suffer from this tradeoff: they are bio-inspired vision sensors that only measure changes in intensity that exceed a given threshold. These changes, called events, are recorded asynchronously and with microsecond

Figure 1. We introduce a new class of event-based object detectors, which we term Efficient Asynchronous Graph Neural Networks (EAGR). They process events (a) as spatiotemporally evolving event graphs (b) and can be deployed in an efficient, asynchronous mode where they only perform local updates for each new event, thus significantly reducing computational complexity. EAGRs have a two times deeper architecture compared with other asynchronous GNNs [27, 45] and are more efficient in event-by-event processing (d, blue box). This opens the door to efficient, and accurate object detection in edge-case scenarios (c) (from [16]), where methods like YOLOX-s [14] based on standard images fail.
Asynchronous methods hit a glass ceiling, since, as network depth increases, they are still behind in terms of expressiveness and accuracy. This is because current asynchronous GNNs are artificially kept shallow and lightweight to limit the per-event computation. Per-event computation still scales with the feature dimension and size of the subgraphs, which grow as the depth of the network increases. Thus, asynchronous methods hit a glass ceiling, since, as network depth increases their benefits over synchronous processing diminish.

In this work, we introduce several architecture design choices that allow us to scale the depth and capacity of GNNs significantly while maintaining highly efficient per-event processing. To maintain a low computational complexity, we first investigate effective ways to prune node updates caused by new events, so that they do not need to be propagated to lower layers. We find that, when combined with max pooling in early layers and node position rounding, we can skip up to 73% of the computation in lower layers with a minimal performance impact. As a next step, we show the significant benefit of performing early temporal node aggregation, which simultaneously boosts performance by 10.6 mAP and allows us to deploy our networks with novel and efficient Look-up-based Spline Convolutions (LUT-SCs). These LUT-SCs are trained as regular Spline Convolutions [12], but are later deployed as efficient look-up tables, which require 4.5 times less computation. Finally, we leverage directed event graphs (DEGs) at the input to boost our model’s performance by 1.8 mAP with a minimal computational cost. This is because, the k-hop subgraphs of DEGs maintain a constant size, thereby limiting their growth in lower layers. In summary:

- We introduce architecture designs to scale the depth and capacity of asynchronous graph neural networks while maintaining highly efficient processing. We achieve this by leveraging node update pruning, early temporal aggregation through max pooling, novel efficient Look-Up-Spline Convolutions, and directed event graph processing in early layers. Cumulatively, these factors reduce the computation by a factor of 33.1, while boosting the performance by 10.6 mAP.

- We introduce several event-based object detectors, which follow these design principles, and come in four sizes: nano, small, medium, and large. Our nano model uses 3.8 times less computation while outperforming the most accurate asynchronous methods by 7.4 mAP. Simultaneously, our medium model still has a 13% lower computational complexity than the most efficient method, while outperforming it by 11.5 mAP.

- We show that our small model performs on par with dense feed-forward neural networks while outperforming state-of-the-art asynchronous methods in terms of efficiency. Finally, our large model outperforms all feed-forward dense methods, paving the way for both accurate and efficient event-based object detection.

2. Related Work

Since the introduction of powerful object detectors in classical image-based computer vision, such as R-CNN [18, 19, 43], SSD [30] and the YOLO series [14, 41, 42], event-based object detection research has focused on leveraging the available models on dense, “image-like” event representations [5, 6, 21, 24, 26, 38]. This approach allows to use pre-training, and well-established architecture designs and loss functions, while maintaining the advantages of events, such as a high-dynamic range, and negligible motion blur.

Most recent examples of such methods include RED [38] and ASTMNet [26] which operate recurrently on events and have shown high performance on detection tasks in automotive settings. However, due to the nature of their method, these approaches necessarily need to convert events into dense frames. This invariably sacrifices the efficiency and high temporal resolution present in the events, which
are critical in many application scenarios such as low-power, always-on, surveillance [22, 33], and low-latency, low-power object detection and avoidance [10, 44].

As a result, a parallel line of research has emerged which tries to reintroduce sparsity into the present models by either adopting spiking neural network architectures [7] or geometric learning approaches [32, 45]. Of these, spiking neural networks are capable of processing raw events asynchronously and are thus closest in spirit to event-based data. However, they lack efficient learning rules and thus do not yet scale to complex tasks and datasets [1, 17, 25, 36, 37, 47]. Recently, geometric learning approaches have filled this gap. They treat events as spatio-temporal point-clouds [46], submanifolds [32] or graphs [2, 27, 34, 45], and process them with specialized neural networks. These methods retain the spatio-temporal sparsity in the events and can be implemented recursively, where single event insertions are highly efficient. Of these, processing events with graph neural networks has proven to be most scalable, achieving high performance on complex tasks like object recognition [2, 9, 27], object detection [45] and motion segmentation [35]. Simultaneously, they can be updated efficiently and asynchronously, for each new event, by only limiting computation to locally changed subgraphs, and propagating these changes to deeper layers in the network.

However, they are still far from achieving the same level of accuracy as dense methods. Due to efficiency requirements, current asynchronous graph neural networks are limited in terms of capacity and depth of the networks [9, 27, 45]. This is because asynchronous methods become less efficient as the depth of the network increases, and their complexity still scales with the number of network parameters. In this work, we address this limitation by introducing a class of graph neural networks, which simultaneously has deep and high-capacity networks, but only has low computational complexity.

3. Method

In Sec. 3.1 we start by reviewing the generation and data structure of events, followed by a description of how to create a graph from events. Next, in Sec. 3.2 we introduce our neural network, and finally describe the proposed asynchronous operation in Sec. 3.3.

3.1. Event Generation and Graph Construction

Event cameras have independent pixels which respond asynchronously to changes in logarithmic brightness $L$. Whenever the magnitude of this change exceeds the contrast threshold $C$, that pixel triggers an event $e_i = (x_i, t_i, p_i)$ characterized by the position $x_i$, timestamp $t_i$ with microsecond resolution and polarity (sign) $p_i \in \{-1, 1\}$ of the change. An event is triggered when

$$p_i[L(x_i, t_i) - L(x_i, t_i - \Delta t_i)] > C. \quad (1)$$

The event camera thus outputs a sparse stream of events $E = \{e_i\}_{i=0}^{N-1}$. As in [2, 9, 27, 35, 45], we interpret events as a 3D point cloud, connected via spatio-temporal edges.

Our event graph $G = \{V, E\}$ consists of nodes $V$ and edges $E$. Each event $e_i$ corresponds to a node. These nodes $n^i \in V$ are characterized by their position $n^i_p = (x_i, \beta t_i) \in$
\(\mathbb{R}^3\) and node features \(n^i_j = p_i \in \mathbb{R}\). Here \(x_i\) is the event pixel coordinate, normalized by the height and width, and \(t_i\) and \(p_i\) are taken from the corresponding event. To map \(t_i\) into the same range as \(x_i\), we rescale it by a factor of \(\beta = 10^{-6}\). These nodes are connected via edges, \((i,j) \in E\), connecting nodes \(n_i\) and \(n_j\), each with edge attributes \(e_{ij}\) \(\in \mathbb{R}^{d_e}\). We connect nodes that are within a spatio-temporal distance from each other and temporally ordered

\[
(i,j) \in E \text{ if } ||n^i_p - n^j_p||_{\infty} < R \text{ and } t_i < t_j
\]

(2)

Here \(||.||_{\infty}\) denotes the Manhattan distance, which returns the absolute value of the largest component. Constructing the graph in this way gives us several advantages: First, we can leverage the queue-based graph construction method in [27] to implement a highly parallel graph construction algorithm on GPU. Our implementation constructs full event graphs with 50,000 nodes in 1.75 ms, and inserts single events in 0.3 ms on a Quadro RTX 4000 laptop GPU. Secondly, the temporally ordering constraint above, makes the event graph directed [27, 35] which will enable high efficiency in early layers before pooling (see Sec. 3.3). In this work, we select \(R = 0.01\) and limit the number of neighbors of each node to 16.  

3.2. Efficient Asynchronous Graph Neural Network

An overview of our neural network architecture, which we term Efficient Asynchronous Graph Neural Network, is shown in Fig. 2. It processes the spatio-temporal graphs from Sec. 3.1 (Fig. 2, top left) and outputs object detection at multiple scales (top right). It consists of five alternating residual layers (Fig. 2 A) and max pooling blocks (Fig. 2 B), followed by a YOLOX-inspired detection head at two scales (Fig. 2 C). Crucially, our network has a total of 13 convolution layers. By contrast, the methods in [27] and [45] feature only 5 and 7 layers respectively, making our network almost twice as deep as previous methods. Before each residual layer, we concatenate the \(x\) and \(y\) coordinates of the node position onto the node feature, which is indicated by \(+2\) at the residual layer input. Residual layers and the detection head use the Look-Up-Table Spline Convolutions (LUT-SC) as the basic building block (Fig. 2 B). These LUT-SC are trained as a standard Spline Convolution [12, 45] and later deployed as an efficient look-up-table (see Sec. 3.3).  

Spline Convolutions Spline Convolutions update the node features by aggregating messages from neighboring nodes:

\[
n^i_j = W n^i_j + \sum_{(j,i) \in E} W(e_{ij}) n^j_i, \quad \text{and} \quad n^i_p = n^i_p
\]

(3)

Here \(n^i_j\) is the updated feature at node \(n_j\), \(W \in \mathbb{R}^{d_n \times d_n}\) is a matrix that maps the current node feature \(n^i_j\) to the output, and \(W(e_{ij}) \in \mathbb{R}^{d_n \times d_n}\) a matrix that maps neighboring node features \(n^j_i\) to the output. In [12], \(W(e_{ij})\) is a matrix-valued smooth function of the edge feature \(e_{ij}\). We set the edge feature to be \(e_{ij} = |n^i_y - n^j_y|/2r + 1/2\), where \(n^i_y\) denotes only the \(x\) and \(y\) component of the node position, and \(r\) is chosen such that \(e_{ij} \in [0, 1]^2\). The function \(W(e_{ij})\) is modeled by a \(d\)-order B-Spline in \(m = 2\) dimensions with \(k \times k\) learnable weight matrices equally spaced in \([0, 1]^2\). During the evaluation, the function interpolates between these learnable weights according to the value of \(e_{ij}\). In this work, we choose \(d = 1\) and \(k = 5\).  

Max Pooling Max pooling splits the input space into \(g_x \times g_y \times g_t\) voxels, \(V\), and clusters nodes in the same voxel. At the output, each non-empty voxel has a node, located at the rounded mean of the input node positions, and its feature equal to the maximum of the input nodes features.

\[
n^i_j = \max_{n \in V_i} n^i_j \quad \text{and} \quad n^i_p = \frac{1}{\alpha} \left[ \frac{\alpha}{|V_i|} \sum_{n \in V_i} n^i_p \right]
\]

(4)

Here multiplying by \(\alpha = [H, W, \frac{1}{r}]^T\) scales the mean to the original resolution. To compute the new edges, it forms a union of all edges connecting the cluster centers and removes duplicates. This operation can result in bidirectional edges between output nodes if at least one node from voxel A connected to one of voxel B and vice versa. The combination of max-pooling and position rounding has two main benefits: First, it allows the implementation of highly efficient LUT-SC convolutions, and second, it enables update pruning, which further reduces computation, discussed in Sec. 4.1. For our pooling layers we select \((g_x, g_y, g_t) = (56/2^i, 40/2^i, 1)\), where \(i\) is the index of the pooling layer. As seen in Sec. 4.1, selecting \(g_t = 1\) is crucial to obtain high performance, since it accelerates the information mixing in the network.  

Detection head Inspired by the YOLOX detection head, we design a series of (LUT-SC, BN, ReLU) blocks which progressively compute a bounding box regression \(f_{reg} \in \mathbb{R}^4\), class score \(f_{cls} \in \mathbb{R}^{n_{cls}}\) and object score \(f_{obj} \in \mathbb{R}\) for each output node. We then decode the bounding box location as in [14], but relative to the voxel location in which the node resides. This results in a sparse set of output detections.  

3.3. Asynchronous Operation

As in [27, 32, 45], after training, we deploy our network in an asynchronous mode. The conversion to asynchronous mode happens in three steps: (i) look-up-table spline convolution caching and batch norm fusing, (ii) network activation initialization, and (iii) update propagation and pruning.  

Look-up-Table Spline Convolution Caching Spline Convolutions generate the highest computational burden in our method since they involve evaluating a multi-variate, matrix-valued function, and performing a matrix-vector multiplication. Following the implementation in [12], com-
putting a single message between neighbors requires
\[ C_{\text{msg}} = (2[d + 1]^m - 1)c_{\text{in}}c_{\text{out}} + (2c_{\text{in}} - 1)c_{\text{out}}, \]  
(5)
floating point operations (FLOPS), where the first term computes the interpolation of the weight matrix, and the second computes the matrix-vector product. Here the first term dominates due to the highly superlinear dependence on \( d \) and \( m \). Our LUT-SC eliminates this term. We recognize that the edge attributes \( e_{ij} \) only depend on the relative spatial node positions. Since events are triggered on a grid, and the distance between neighbors is bounded, these edge attributes can only take on a finite number of possible values. Therefore, instead of recomputing the interpolated weight at each step, we can precompute all weight matrices once and store them in a look-up table. This table stores the relative offsets of nodes together with their weight matrix. We thus replace the message propagation equation with
\[
\mathbf{n}_{ij}^t = \mathbf{W}
\]
\[ + \sum_{(j,i) \in E} W_{ij} \mathbf{n}_{j}^t \]  
(6)
\[ W_{ij} = LUT(dx, dy), \]  
(7)
where \( dx \) and \( dy \) are the relative 2D positions of nodes \( i \) and \( j \). Note that this transformation reduces the complexity of our convolution operation to \( C_{\text{msg}} = (2c_{\text{in}} - 1)c_{\text{out}} \) which is on the level of the classical graph convolution (GC) used in [27]. However, crucially, LUT-SC still retains the relative spatial awareness of Spline Convolutions, since \( W_{ij} \) changes with the relative position and is thus more expressive than GCs. After caching, we fuse the weights computed above with the batch norm layer immediately following each convolution, thereby eliminating its computation from the tally. After pooling, ordinarily, node positions would not have the property that they lie on a grid anymore, as their coordinates get set to the centroid location. However, since we apply position rounding, we can apply LUT-SC caching in all layers of the network.

**Network Activation Initialization** Before asynchronous processing, we pass a dense graph through our network and cache the intermediate activations at each layer. While in convolution layers we cache the activation, i.e., the results of sums computed from Eq. (6), in max pooling layers we cache (i) the indices of input nodes used to compute the output feature for each voxel, (ii) a list of currently occupied output voxels, and (iii) a partial sum of node positions and node counts per voxel, to efficiently update output node positions after pooling.

**Update Propagation and Pruning** When a new event is inserted, we recompute messages in all layers of the network. We do this to achieve an output identical to the output the network would have computed if the complete graph with one event added was processed from scratch. The propagation rules are outlined in Fig. 3.

**Input layer:** In the early layers of the network before pooling (C, top row), we only need to compute messages (green arrow) which are required to compute the feature of the new node in the network. Moreover, as the network depth increases, the number of messages stays constant, allowing us to stack multiple layers at the input with minimal computation increase. After pooling (C bottom row), bidirectional edges may be encountered, and thus update propagation follows the rules outlined in A and B. This is because, it may happen that for voxels \( V_i \) and \( V_j \) one edge goes from \( V_i \) to \( V_j \) and vice versa, with both edges being temporally ordered. The union operation of max pooling would then form a bidirectional edge between the output clusters. At each layer, we maintain a running list of unchanged nodes (gray) and changed nodes (cyan), and whether their position has changed, the feature has changed, or both.

**Convolution Layers:** In a convolution layer (A), if the node has a different position (top), we recompute that node’s feature and resend a message from that node to all
Table 1. Features affecting computational complexity.

| Ablation                        | mAP↑ | MFLOPS/ev↓ |
|---------------------------------|------|------------|
| w/o early aggregation           | 21.2 | 6.27       |
| w/o multi-layer input           | 30.0 | 4.56       |
| baseline                        | 31.8 | 4.58       |

Table 2. Features affecting accuracy of the method.

4. Experiments

Training Details In all our experiments, we use the AdamW optimizer [31] with a learning rate of 0.01 and weight decay of $10^{-5}$. We train each model for 150'000 iterations with a batch size of 64. We randomly crop the events to 75% of the full resolution, and randomly translate them by up to 10% of the full resolution. We use the YOLOX loss [14], which includes an IOU loss, class loss, and a regression loss, discussed in [14]. To stabilize training, we also use exponential model averaging (EMA) [23].

Datasets We evaluate our method on the N-Caltech101 detection [36], and the Gen1 Detection Dataset [8]. N-Caltech101 consists of recording by a DAVIS240 [3] undergoing a saccadic motion in front of a projector, projecting samples of Caltech101 [11] on a wall. In post-processing, bounding boxes around the visible boxes were hand placed. The Gen1 Driving Dataset is a more challenging, large-scale detection dataset targeting an automotive setting. It was recorded with an ATIS sensor [39] with a resolution of $304 \times 240$, two classes, 228,123 annotated cars, and 27,658 annotated pedestrians. As [38] we remove bounding boxes diagonals below 30 and sides below 20 pixels from Gen1.

4.1. Ablation Studies

Here we motivate the use of the features discussed previously. We split our ablation studies into two parts, i.e. those targeting the efficiency (Tab. 1), or the accuracy (Tab. 2) of the method. For all experiments, we use the model illustrated in Fig. 2 as a baseline, and report means average precision (mAP) scores [29] on the validation set of Gen1 [8].

Ablations on Efficiency A key enabling factor for using 2D LUT-SCs lies in transitioning from 2D to 3D convolutions, which we investigate by training a model with 3D Spline Convolutions (row 1 in Tab. 1). This result does not yet take into account update pruning, which is discussed later. With an mAP of 31.84, it achieves a 0.05 higher mAP than our baseline (bottom row). Using 3D convolutions yields a slight improvement in accuracy since it uses more information, but does not allow us to perform an efficient lookup, yielding 150.87 MFLOPS per new event. Using 2D convolutions (row 2) reduces the computation to 79.6 MFLOPS/ev due to the dependence on the dimension $d$ in Eq. (5), which is further reduced to 17.3 MFLOPS/ev after implementing LUT-SCs (row 3). Despite the small decrease in performance due to 2D convolutions, we gain a factor of 8.7 in terms of FLOPS per event.

Next, we investigate pruning. We recompute the FLOPS of the previous model by terminating update propagation after max-pooling layers, illustrated in Fig. 3 (c), and reported in row 4 of Tab. 1. We find that this reduces the computational complexity from 17.3 to 16.3 MFLOPS/ev. This reduction comes from removing the orange messages in Fig. 3 A (bottom). Implementing node position in Eq. 4 (row 5), allows us to fully prune updates. The final method only requires 4.56 MFLOPS/ev. Node position rounding reduces mAP only by 0.01, justifying its use.

Ablations on Accuracy We found that two features of our network had a major impact on performance: First, we applied early temporal aggregation, i.e., using $g_t = 1$, which sped up training and led to higher accuracy. We train another model which pools the temporal dimension more gradually by setting $g_t = 8/2^i$, where $i$ is the index of the pooling layer. This model only reached an mAP of 21.2 (Tab. 2, first row), after reducing the learning rate to 0.002 to enable stable training. This highlights that early pooling plays an important role since it improves our result by 10.6
mAP. We believe that it is important for mixing features quickly so that they can be used in lower layers.

Next, we investigate using multiple layers before the max-pooling layer. We train another model which only has a single input layer, replacing the Layer in Fig. 2 with a (LUT-SC, BN, ReLU) block. This yielded a performance of 30.0 mAP which is 1.8 mAP lower than the baseline. The computational complexity is only marginally lower, which is explained by Fig. 3 C (top). We see that adding layers at the input only generates few additional messages. This highlights the benefits of using a directed event graph.

4.2. Max Pooling

In this section, we take a closer look at the pruning mechanism. We find that almost all pruning happens in the very first max pooling layer. This motivates the placement of the pooling layer at the early stages of the network, which allows us to skip most computations when pruning happens. Also, since the subgraph is still small in the early layers, it is easy to prune the entire update tree. We interpret this case as ”event filtering” and investigate this filter in Fig. 4.

When applied to raw events (Fig. (a)) we obtain filtered events (Fig. (b)), i.e., events that passed through the first max pooling layer. We observe that max-pooling makes the events more uniformly distributed over the image plane. This is also supported by the density plot in Fig. 4 (c), which shows that the distribution of the number of events per pixel shifts to the left after filtering, removing events in regions where there are too many. This behavior can be explained by the pigeon-hole principle when applied to max-pooling layers. Max-pooling usually only uses a fraction of its input nodes to compute the output feature. The number of input nodes used by the max-pooling layer is upper bounded by its output channel dimension, \( c_{\text{out}} \), since it could at maximum only use one feature from each input node. As a result, max-pooling selects at most \( c_{\text{out}} \) nodes for each voxel, resulting in more uniformly sampled events.

To study the effect of the output channel dimension on filtering, we train four models with \( c_{\text{out}} \in \{8, 16, 24, 32\} \), where our baseline model had \( c_{\text{out}} = 16 \). We report the mAP, MFLOPS/ev, and fraction of events after filtering, \( \phi \) averaged over Gen1, in Fig. 4 (d). As predicted, we find that increasing \( c_{\text{out}} \) increases mAP, MFLOPS, and \( \phi \). However, increase happens at different rates. While MFLOPS and \( \phi \) grow roughly linearly, mAP growth slows down significantly after \( c = 24 \). Interestingly, by selecting \( c_{\text{out}} = 8 \) we still achieve an mAP of 30.6, while only using 21% of events. This type of filtering has interesting implications for future work. An interesting question would be whether events that are not pruned carry salient and interpretable information.

4.3. Comparison with State of the Art

Finally, we compare our method against state-of-the-art dense and asynchronous methods and report results in Table 3 on the N-Caltech101 and Gen1 test sets. We evaluate four versions of our model: nano (-N), small (-S), medium (-M), and large (-L). These differ in the number of features in the layer blocks 3, 4 and 5 and in the detection heads, and have 32, 64, 92, and 128 channels in these layers respectively. Our baseline from before has 64. We compare against the following state-of-the-art methods:

**Dense Recurrent Methods** In this category, RED [38] and ASTM [26] are the state-of-the-art, and feature recurrent architectures. We also include MatrixLSTM+YOLOv3 [5] which features a recurrent, learnable representation and a YOLOv3 detection head.

**Dense Feed-forward Methods** The work in [26] provides results on Gen1 for the dense feed-forward methods which we term Events+RRC [6], Inception+SDD [21] and Events+YOLOv3 [24]. These use dense event representations with the RRC, SSD, or YOLOv3 detection head.

**Spiking Methods** We compare against the spiking network Spiking DenseNet [7], which uses an SSD detection head.

**Asynchronous Methods** Here we compare against state-of-the-art methods AEGNN [45] and NVS-S [27], both graph-based, AsyNet [32] which uses submanifold sparse convolutions [20] and YOLE [4], which uses an asynchronous CNN. All of these methods deploy their networks in an
while still outperforming AEGNN by 10 mAP. Our smallest network, nano, is even 3.8 times more efficient achieving a 14.1 mAP higher performance than AEGNN. This opens the door to efficient, and accurate object detection in edge-case scenarios.

From Tab. 3 we can make multiple observations. First, we find that recurrent dense methods RED and ASTM net outperform our L model by 7.9 mAP and 14.6 mAP respectively. However, since these methods are dense, we observe significantly higher computation when compared to our method (4712 vs. 1.36 for our nano model). We believe that deeper networks and recurrence are two major factors that help performance in their methods. By contrast, our large model with 32.1 mAP outperforms the recurrent method MatrixLSTM [5] by 1.1 mAP and also has 120 times fewer FLOPS. We also find that our large model with an mAP of 32.1 outperforms dense feed-forward methods Events+RRC [6] (30.7), Inception+SSD [21] (30.1) and Events+YOLOv3 [24] (31.2). When compared to the spiking network [7] we find that our method has a 13.1 mAP higher score. The low performance of the SNN is expected to increase as better learning strategies become available to the community. Finally, we compare against sparse methods. We find that our small model outperforms all methods in terms of computation, with around 13% times fewer MFLOPS/ev than the runner-up AEGNN [45]. It also achieves a 14.1 mAP higher performance than AEGNN. Our smallest network, nano, is even 3.8 times more efficient while still outperforming AEGNN by 10 mAP.

On the N-Caltech101 dataset, our small model outperforms state-of-the-art dense and sparse methods, achieving 70.2 mAP, which is 5.9 mAP higher than the runner-up AsyNet [32]. Moreover, this model has lower computational complexity than state-of-the-art AEGNN [45]. Our large model achieves the highest score with 73.2 mAP. Our nano model achieves the lowest computation of 2.28 MFLOPS/ev, which is 3.25 times lower than that of AEGNN and has a 3.4 higher mAP. It is higher than that for Gen1 since N-Caltech101 has more classes.

### 5. Limitations and Future Work

While making an important step toward highly efficient and robust event processing, we believe that further improvements can be made. In particular, we found that the methods currently outperforming our event-based detector are all recurrent, a feature that was not studied in this work. We thus believe that our method could also benefit from recurrence since it can overcome issues especially when few events are present. Combining this approach with additional sensors, such as LiDARs or RGB cameras, is also a promising research direction. They can provide strong priors, which may increase its performance and reduce its complexity if shallower networks can be utilized.

### 6. Conclusion

While traditionally, data-driven event-based methods have relied on converting events into dense image-like representations, these still do not effectively model the asynchronous and sparse nature of event data, which increases computation. Despite recent progress on reducing this complexity, asynchronous methods have not yet delivered on high accuracy due to their shallow networks. In this work, we introduced a series of network design choices that allow us to design deeper neural networks without sacrificing complexity. We achieve this by introducing novel look-up-based convolutions, performing update pruning, and using directed event graphs in early layers which reduce the complexity of our approach by a factor of 33.1 compared to the baseline. Thus, on the Gen1 and N-Caltech101 datasets, our method achieves a 11.5 point higher mAP than state-of-the-art asynchronous methods, with higher efficiency, and it even outperforms several dense, feed-forward methods. This opens the door to efficient, and accurate object detection in edge-case scenarios.
7. Acknowledgement

This work was supported by Huawei Zurich, the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) Robotics (grant number 51NF40_185543), and the European Research Council (ERC) under grant agreement No. 864042 (AGILE-FLIGHT).

8. Appendix

8.1. Overview

Here we will discuss in more detail the equivalence of the update rules discussed in Sec. M-3.3 (Sec. A-8.2). Then we will analyze per layer statistics in more detail in Sec. A-8.3, such as per layer computation, node updates and filtering. In what follows, we will mark sections, and equations from the main manuscript with a "M-", and sections and equations here with "A-".

8.2. Asynchronous Operation

While the work in [45] flagged changed nodes in each layer, and recomputed each one of them by collecting all messages from the surrounding nodes of the updated nodes, in this work we take a more efficient route. We start off by considering the Eq. M-6 from the main text, which we repeat here:

\[
\begin{align*}
\mathbf{n}_j^t &= \mathbf{W} \mathbf{n}_j^t + \sum_{(j,i) \in E} W_{ij} \mathbf{n}_j^t \\
W_{ij} &= LUT(dx, dy)
\end{align*}
\]  

(8)

(9)

Here, \(dx\) and \(dy\) are

\[
\begin{align*}
\begin{bmatrix} dx \\ dy \end{bmatrix} &= \begin{bmatrix} n_{j, x} - n_{i, x} \\ 2r - n_{i, y} - n_{j, y} \end{bmatrix} + \frac{1}{2}
\end{align*}
\]

(10)

and only take on a finite number of values. In the following we discuss what messages need to be recomputed if the position and the feature of the node change.

8.2.1 Position Change

Assume the position of a node \(\mathbf{n}^k\) changes, then we essentially update the weight matrices \(W_{ij}\) since the \(dx\) and \(dy\) change in Eq. A-(10). We thus need to modify weight matrices of the form \(W_{kj}\) and \(W_{ik}\). We will indicate the update weights with \(W^*\). Since edges \((j, i)\) are used in Eq. A-(8), we conclude that only nodes with edges \((j, k)\) and \((k, i)\) are affected. While the edges \((j, k)\) correspond to messages sent from source neighbors \(\mathbf{n}^j\) to node \(\mathbf{n}^k\), the edges \((k, i)\) correspond to messages sent from \(\mathbf{n}^k\) to its destination neighbors \(\mathbf{n}^i\). For this reason, only node \(\mathbf{n}^k\) and its destination neighbors \(\mathbf{n}^i\) need to be recomputed. We discuss the update to each of them separately:

**Changed node \(\mathbf{n}^k\):** Eq. A-(8) for \(\mathbf{n}^k\) becomes

\[
\mathbf{n}_j^k = \mathbf{W} \mathbf{n}_j^k + \sum_{(j,k) \in E} W_{kj} \mathbf{n}_j^k
\]

(12)

where we need to recompute all messages in the sum. The number of summations is the number of the source neighbors of \(\mathbf{n}^k\)

\[
\mathcal{N}_{src}(k) = \{\mathbf{n}^j | (j, i) \in E\}
\]

(13)

which we define as \(N_{k,src} = |\mathcal{N}_{src}(k)|\) so the amount of computation to perform the above step is

\[
C_{recomp,k} = N_{k,src}(2c_{in} - 1)c_{out} + (N_{k,src} - 1)c_{out}
\]

(14)

Here the first term corresponds to performing the matrix-vector multiplications, and the second to summing the outputs. These exactly correspond to the messages illustrated as green arrows in Fig. A-5

**Update messages to destination nodes:** Eq. A-(8) for nodes \(\mathbf{n}^i \in \mathcal{N}_{dst}(k)\) becomes

\[
\mathbf{n}_j^i = \mathbf{W} \mathbf{n}_j^i + \sum_{(j,i) \in E} W_{ij} \mathbf{n}_j^i + W_{ik} \mathbf{n}_j^k - W_{ik} \mathbf{n}_j^k
\]

(15)

(16)

Figure 5. Details about update mechanism. We analyze the update computation necessary when a node \(\mathbf{n}_c\) is updated, and differentiate between position and feature updates. Green arrows denote updated messages from the green source neighbors of \(\mathbf{n}_c\), denoted as \(\mathcal{N}_{src}(c)\). Orange arrows denote update messages from node \(\mathbf{n}_c\) to its yellow destination arrows, \(\mathcal{N}_{dst}(c)\). Nodes can belong to both sets, and these are illustrated with a yellow-green gradient.
Here we replace the old message by subtracting off the old message and adding on the new one. In total this results in

\[ C_{\text{update pos. } k} = N_{k, \text{dst}}(2c_{\text{in}} - 1)c_{\text{out}} + 2N_{k, \text{dst}}c_{\text{out}} \]

Here the first term computes the new message, a total of \( N_{k, \text{dst}} \div |N_{\text{dst}}(k)| \) times, and the second term computes the subtraction and addition for each output message. These exactly correspond to the messages illustrated as orange arrows in Fig. A-5.

**Total:** In total, the computation necessary for a single node position change is

\[ C_{\text{pos. change } k} = C_{\text{recomp. } k} + C_{\text{update pos. } k} \] (17)

8.2.2 Feature Change

Assume the node feature of node \( n_f^k \) changes, indicating the change of \( n_f^k \). We will denote the changed feature by \( n_f^{k,*} \). Inspecting Eq. A-(8) we find that now only nodes \( n_f^j \) with \( (k, i) \in E \) are affected, which is what we previously called the destination neighbors \( N_{\text{dst}}(k) \). Additionally, we need to update the root linear term \( W n_f^k \), which simply amounts to a computation

\[ C_{\text{root}} = (2c_{\text{in}} - 1)c_{\text{out}} + 2c_{\text{out}} \] (18)

The update message terms instead follow the updated Eq. A-(8)

\[ n_f^j = W n_f^j + \sum_{(j,i) \in E} W_{ij} n_f^i + W_{ik} n_f^{k,*} - W_{ik} n_f^k \] (19)

as in Eq. A-(15), this amounts to a computation

\[ C_{\text{update feat. } k} = N_{k, \text{dst}}(2c_{\text{in}} - 1)c_{\text{out}} + 2N_{k, \text{dst}}c_{\text{out}} \] (20)

**Total:** In total, the computation for updating the node feature is

\[ C_{\text{feat. change } k} = C_{\text{root}} + C_{\text{update feat. } k} \] (21)

8.3. Per Layer Analysis

In this section, we will have a deeper look at the per-layer metrics, of our network which will give us more insights into our method and also inspire future research in building better and more efficient neural network architectures. In what follows, we will analyze three quantities: (i) per-layer FLOPS (ii) per-layer changed nodes, and (iii) per-layer probability of filtering. We will discuss each in turn.

8.3.1 Per Layer FLOPS

We analyze the number of FLOPS performed per layer, by tallying the number of FLOPS at each layer, averaged over the Gen1 dataset [8], as was done in Sec. M-4.1 and M-4.2 in the main manuscript. We visualize the distribution of FLOPS per layer in Fig. A-6. We can make a couple of observations: First, pooling layers contribute only a minimal amount of computation. That is because they are only concerned with updating node positions, something that only happens for few nodes, as will be seen later. Secondly, computation in the first three layers of the neural network is very small. This is because the number of nodes to be updated remains small, due to the directedness of the event graph. Thirdly, computation is highest in lower layers, such as the detection heads and Layer blocks 4 and 5. This is because in these layers, we need to recompute new messages for a larger number of nodes, as will be seen later. The layer with the highest computation are layers Head1.ClsConv and Head1.PredConv. These two layers have (i) the highest number of node changes (as seen later), and they also have the highest number of input and output features. Finally, output layers such as the object, regression and class predictors all have low computational complexity. This is because, despite having a high number of changed nodes, they only have few (maximally 4) output channels. Next, we will discuss the origin of these numbers by studying the number...
of node changes per layer.

### 8.3.2 Per Layer Changed Nodes

The above FLOPS are caused by changes in nodes features and node positions. However, it is not clear which source is dominant. To further reduce complexity, such information is vital. For this reason, we show the break-down of the number of nodes with changed features and changed positions in Fig. A-7 right and left respectively. At each layer we count the number of changed nodes which enter that layer. We see that most nodes changes are due to feature changes. For position changes, (left) we see that at most one node has a position change, and this corresponds to the nodes which have the new node in their voxel. We see that only 6.3% of node insertions lead to position changes after the first pooling layer. This value diminishes to 2.6%, 1.1% and 0.3% after each pooling layer. We thus conclude that most of the network computation is due to message updates of the form seen in Fig. A-5 in the bottom row. By contrast, node feature changes happen more often. After each convolution layer, the number of changes nodes increases, and after each pooling layer it decreases. We also find that most node changes happen in the high resolution detection head (Head1). This is because this head has more nodes at its disposal (a total of $g_x \times g_y$, i.e. 140 for Head1, and 35 for Head2). However, we also see that only a fraction of these nodes are updated (max 11.8).

### 8.3.3 Per Layer Probability of Filtering

In a final step, we analyze the filtering mechanism described in Sec. M-4.2 in more detail, by investigating the probability of a node feature and node position change being present at each layer. As previously discussed. To measure this, we count for how many samples there is at least one position or feature change at each layer, and we illustrate this in Fig. 8. While we see the same trend for the node position as before, for the node feature it is different. Here only 27.8% of node features are unchanged after the first pooling layer, and this number is minimally impacted in lower layers. A slight increase in the second Layer block is observed, and this has to do with the concatenation of the node position at each layer. By concatenating the node position, previously unchanged node features can experience a small update. However, as mentioned in the main text, when taking this into account, pruned updates make up 73% of the total.

### 8.4. Licenses

We use the Prophesee Gen1 dataset [8] under the “PROPHESEE GEN1 AUTOMOTIVE DETECTION DATASET LICENSE TERMS AND CONDITIONS” found at the URL https://www.prophesee.ai/2020/01/24/prophesee-gen1-automotive-detection-dataset/. N-Caltech101 [36] is used under the “Creative Commons Attribution 4.0 license” and downloaded at https://www.garrickorchard.com/datasets/n-caltech101.
At each layer we count the number of changed nodes which enter that layer. Layers on the same network branch are have the same color, and skip branch connections are indicated with arrows of the corresponding color.

References

[1] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. In Conference of Computer Vision and Pattern Recognition (CVPR), pages 7388–7397, 2017. 3

[2] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. IEEE Int. Conf. Comput. Vis. (ICCV), 2019. 2, 3

[3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240x180 130dB 3µs latency global shutter spatiotemporal vision sensor. IEEE JSSC, 49(10):2333–2341, 2014. 2, 6

[4] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Asynchronous convolutional networks for object detection in neuromorphic cameras. In CVPRW, 2019. 7, 8

[5] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. 2020. 2, 7, 8

[6] Nicholas F. Y. Chen. Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion. In CVPRW, 2018. 2, 7, 8

[7] Loic Cordone, Benoit Miramond, and Philippe Thierion. Object detection with spiking neural networks on automotive event data. International Joint Conference on Neural Networks (IJCNN), 2022. 2, 3, 7, 8

[8] Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. arXiv e-prints, abs/2001.08499, 2020. 6, 8, 10, 11

[9] Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. 2022. 2, 3

[10] Davide Falanga, Kevin Kleber, and Davide Scaramuzza. Dynamic obstacle avoidance for quadrotors with event cameras. Science Robotics, 5(40):eaaz9712, 2020. 3

[11] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In CVPR, 2004. 6

[12] Matthias Fey, Jan Eric Lenssen, Frank Weichert, and Heinrich Müller. Splinecnn: Fast geometric deep learning with continuous b-spline kernels. Conference of Computer Vision and Pattern Recognition (CVPR), 2018. 2, 4

[13] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. IEEE TPAMI, 2020. 2

[14] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. 2021. 1, 2, 4, 6

[15] Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In IEEE Int. Conf. Comput. Vis. (ICCV), 2019. 2

[16] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. 1

[17] Mathias Gehrig, Sumit Bam Shrestha, Daniel Mouritzen, and Davide Scaramuzza. Event-based angular velocity regression with spiking networks. IEEE International Conference on Robotics and Automation (ICRA), 2020. 3
[47] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: Histograms of averaged time surfaces for robust event-based object classification. In Conference of Computer Vision and Pattern Recognition (CVPR), pages 1731–1740, 2018.

[48] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In IEEE Int. Conf. Comput. Vis. (ICCV), October 2019.

[49] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In Conference of Computer Vision and Pattern Recognition (CVPR), pages 16155–16164, 2021.

[50] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In Conference of Computer Vision and Pattern Recognition (CVPR), 2019.