




Please cite the Published Version

Zhang, X, Han, L , Davies, S , Sobeih, T, Han, L and Dancey, D  (2025) A novel energy-efficient spike transformer network for depth estimation from event cameras via cross-modality knowledge distillation. *Neurocomputing*, 658. 131745 ISSN 0925-2312

DOI: <https://doi.org/10.1016/j.neucom.2025.131745>

Publisher: Elsevier

Version: Published Version

Downloaded from: <https://e-space.mmu.ac.uk/642860/>

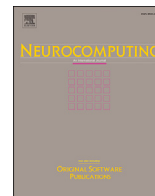
Usage rights:  [Creative Commons: Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

Additional Information: This is an open access article published in *Neurocomputing*, by Elsevier.

Data Access Statement: Data will be made available on request.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)



A novel energy-efficient spike transformer network for depth estimation from event cameras via cross-modality knowledge distillation

Xin Zhang^a, Liangxiu Han^{a,*} , Sergio Davies^a , Tam Sobehi^a, Lianghao Han^b, Darren Dancey^a

^a Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M15 6BH, UK

^b Department of Computer Science, Brunel University, Uxbridge UB8 3PH, UK

HIGHLIGHTS

- Novel energy-efficient Spike Transformer for depth estimation using event cameras.
- Purely spike driven transformer with spike-based attention and residual mechanisms.
- Fusion depth head combines multi-stage features for fine-grained predictions.
- Cross-modality knowledge distillation from DINOv2 enhances SNN training.

ARTICLE INFO

Communicated by Y. Wu

Keywords:

Spiking neural networks (SNNs)
Event camera
Transformer
Depth estimation
Knowledge distillation
Neuromorphic computing
Energy-efficient computing

ABSTRACT

Depth estimation is a critical task in computer vision, with applications in autonomous navigation, robotics, and augmented reality. Event cameras, which encode temporal changes in light intensity as asynchronous binary spikes, offer unique advantages such as low latency, high dynamic range, and energy efficiency. However, their unconventional spiking output and the scarcity of labeled datasets pose significant challenges to traditional image-based depth estimation methods. To address these challenges, we propose a novel energy-efficient Spike-Driven Transformer Network (SDT) for depth estimation, leveraging the unique properties of spiking data. The proposed SDT introduces three key innovations: (1) a purely spike-driven transformer architecture that incorporates spike-based attention and residual mechanisms, enabling precise depth estimation with minimal energy consumption; (2) a fusion depth estimation head that combines multi-stage features for fine-grained depth prediction while ensuring computational efficiency; and (3) a cross-modality knowledge distillation framework that utilises a pre-trained vision foundation model (DINOv2) to enhance the training of the spiking network despite limited data availability. Experimental evaluations on synthetic and real-world event datasets demonstrate the superiority of our approach, with substantial improvements in Absolute Relative Error (49 % reduction) and Square Relative Error (39.77 % reduction) compared to existing models. The SDT also achieves a 70.2 % reduction in energy consumption (12.43 mJ vs. 41.77 mJ per inference) and reduces model parameters by 42.4 % (20.55 M vs. 35.68 M), making it highly suitable for resource-constrained environments. This work represents the first exploration of transformer-based spiking neural networks for depth estimation, providing a significant step forward in energy-efficient neuromorphic computing for real-world vision applications.

1. Introduction

Depth estimation is a fundamental task in computer vision, underpinning applications such as autonomous driving, robotics, agricultural monitoring, and environmental analysis [1]. Traditionally, state-of-the-art depth prediction has relied on standard frame-based cameras combined with artificial neural networks (ANNs) [2–4]. However, these

approaches are often limited by latency, power consumption, and dynamic range constraints inherent to conventional imaging sensors.

Event-based cameras have emerged as a promising alternative, inspired by biological vision systems. These sensors asynchronously capture changes in brightness at each pixel, resulting in high temporal resolution, low latency, low power consumption, and a wide dynamic range [5–10]. Their unique capabilities have enabled new possibilities

* Corresponding author.

Email address: l.han@mmu.ac.uk (L. Han).

in fields such as 3D scanning, robotic vision, and automotive applications [11–13]. Despite these advantages, event-based cameras produce spiking data that is inherently noisy and lacks the mature processing algorithms available for conventional images.

Spiking Neural Networks (SNNs), also known as the third generation of neural networks [14], are well-suited for processing the discrete spike streams generated by event cameras. SNNs mimic biological neurons by transmitting information via discrete spikes, rather than continuous values as in traditional ANNs [15]. This makes SNNs a natural fit for event-based data, and recent research has begun to explore their potential for vision tasks [16–18].

The choice of SNNs for event-based depth estimation is motivated by several unique advantages beyond general energy efficiency. First, there is a fundamental synergy in data representation: event cameras produce asynchronous, sparse, binary data [5,8,9], and SNNs process information in precisely the same manner [19]. This allows SNNs to process event streams more directly, avoiding the information loss and computational overhead associated with converting events into dense, frame-like representations for ANNs [5]. Second, SNNs are inherently temporal processors whose membrane dynamics integrate signals over time [20], making them well-suited to capturing the rich temporal dynamics of event data required for motion-based depth cues. This event-driven computation—processing only upon spike arrival—enables sparse activation, reducing unnecessary operations and supporting low-latency inference on neuromorphic hardware [21]. These properties are critical for real-time applications such as autonomous navigation and robotics operating under energy and latency constraints [12,13].

Nevertheless, the application of SNNs and event-based cameras to depth estimation remains in its early stages [10]. Two major challenges persist: (1) the lack of robust SNN backbones specifically designed for extracting features from spike data for depth estimation, and (2) the generally lower performance of SNNs compared to their ANN counterparts in complex vision tasks.

The lack of SNN backbone designed for spike data depth estimation. The event-based camera generates continuous spike streams in a binary irregular data structure that possesses ultrahigh temporal features. SNNs are applicable to event camera datasets and are able to improve the depth estimation performance by exploiting advanced architectures of ANN, such as ResNet like SNNs and Spiking Recurrent Neural Networks [20,22–24]. Vision transformer [25,26] (ViT), is currently the most popular ANN structure and is based on a self-attention mechanism to capture long-distance dependencies, especially spatio-temporal features in images/videos. It improves the performance of AI in many computer vision tasks such as image classification/ segmentation [27–29], object detection [30] and depth estimation [31,32]. Transformer-based SNNs are a new form of SNN combining transformer architectures with spiking neurons, offering great potential to break the performance bottleneck on spike stream data. In Zhang et al. [33], the authors used the original ViT structure as a backbone to extract features from both spatial and temporal domains in spike data. The result demonstrated the suitability of the transformer for extracting spatio-temporal features. However, the original transformer structure has a large number of multiplication operations and excessive computational energy consumption compared to SNNs. In Zhou et al. [34,35], the authors proposed a pure spike driven self-attention and residual connection to avoid non-spike computations. This was a major step forward in the potential use of transformers for depth estimation from spike data.

SNN model performance. One of the biggest challenges with SNNs currently is their inability to achieve equivalent training performance on spiking data compared to ANNs on non-spiking data. Gradient-based backpropagation is a powerful algorithm for training ANNs, but since spiking data is non-differentiable it cannot be used directly with SNNs [36]. Converting ANN to SNN is a solution but it may introduce errors of uncertainty or lose the temporal information of spikes [20]. Meanwhile, the number of event-based datasets is small compared to the static images used in traditional ANN training, making SNNs prone to overfitting

and limiting their generalisation ability [36]. Knowledge distillation is a technique in deep learning to transfer knowledge from the teacher model to the student model. It allows training of a lightweight model (student model) to be as accurate as a larger model (teacher model). Currently, there are already some ANN models trained with massive data that can achieve zero-shot for depth estimation [37–39]. Logically, the accuracy of these models has the potential to be transferred to the SNN model during training.

In this work, we propose a novel energy-efficient spiking transformer network for depth estimation, leveraging cross-modality knowledge distillation to combine the biological efficiency of SNNs with the advanced feature extraction capabilities of a visual foundation model (DINOv2). To the best of our knowledge, this is the first exploration of a transformer-based SNN for depth estimation, marking a significant advancement in the field. The proposed framework comprises three key components, each contributing uniquely to its overall effectiveness:

- (1) We introduce a novel energy-efficient spike-driven transformer that eliminates conventional floating-point operations through carefully designed spike-based attention and residual mechanisms. This network incorporates two essential components: a spiking patch embedding module that converts raw event data into spike-based tokens while preserving temporal-spatial information, and spiking transformer blocks that integrate Spiking Self-Attention (SSA) and Spiking MLP for efficient feature processing. This design significantly reduces energy consumption while ensuring robust performance.
- (2) We develop a fusion depth estimation head that combines features from multiple transformer stages for fine-grained depth prediction. This head is intentionally hybrid: it uses conventional ConvBN and upsampling (MAC-based) operations to preserve the numerical precision required for the dense regression task of depth estimation. This design choice allows us to separate the energy accounting between the purely spike-driven backbone and the hybrid head.
- (3) We propose a single-stage cross-modality knowledge distillation framework that leverages a large vision foundation model (DINOv2) to enhance SNN training with limited data. By utilising domain loss and semantic loss, our framework effectively transfers knowledge from both final and intermediate layers of DINOv2 to the spike-driven transformer.

2. Related works

This section presents a literature review of existing research in monocular depth estimation, SNNs, and knowledge distillation, highlighting the key challenges that motivate our work.

2.1. Image-based and event-based monocular depth estimation

Depth estimation from images aims to measure the distance of each pixel relative to the camera. Monocular depth estimation is a challenging but promising technology. It has the advantage of only requiring one image unlike traditional depth estimation, which makes it more practical for applications where it is not possible to take a pair of images, such as on mobile devices. Depending on the type of data used, we can divide monocular depth estimation into Image-based and Event-based methods [3]. Image-based monocular depth estimation is more common as it estimates depth using the information in RGB images, which are easy to collect and process. This makes them well-suited for depth estimation in challenging conditions, such as low light and fast motion [18], although event-based data is harder to collect and process.

The latest developments in deep learning have made it possible to develop monocular depth estimation models that can achieve satisfactory accuracy and robustness [3,40,41]. Similar to other deep learning models, these models typically consist of a generalised encoder that extracts abstract features from context information and a decoder that recovers

depth information from the features. For RGB images, in Laina et al. [42], the authors used ResNet-50 as an encoder and novel up-sampling blocks as a decoder to estimate depth from a single RGB image. In Laina et al. [42], the authors utilised ViT instead of convolutional networks as the backbone for a depth estimation task. Experiments have found that transformers are able to provide finer and more globally consistent predictions than traditional convolutional networks. For event data, the research is still in its infancy. The authors [18] presented a new deep learning model called E2Depth that can estimate depth from event cameras with high accuracy. A fully convolutional neural network based on the U-Net architecture [43] was used in this work. In Nam et al. [44], a multiscale encoder was used to extract features from mixed-density event stacking and an upscaling decoder was used to predict the depth. The transformer structure has also been used in event-based monocular depth estimation. In Liu et al. [45], EReFormer was proposed to estimate depth from event cameras with superior accuracy based on transformers.

However, these models predominantly utilise traditional deep learning frameworks, overlooking the unique potential of event-based data. Existing research identifies two key challenges that remain unaddressed. The first challenge lies in the unique characteristics of event camera data. This requires algorithms that can process data in real-time and maintain temporal accuracy [5,6]. The second challenge is the scarcity of spiking training data. High-quality, labeled datasets tailored for SNNs, particularly for tasks like depth estimation, remain limited. The acquisition and labeling of event-based data are both complex and resource-intensive, further constraining the availability of training resources. To address this limitation, knowledge distillation offers a promising solution. This involves transferring knowledge from a well-trained artificial neural network (ANN). The ANN acts as a “teacher”, guiding the SNN, or “student”, to learn effectively with limited event-based data.

2.2. Spiking neural networks (SNNs)

Unlike traditional deep learning models that convey information using continuous decimal values, SNNs use discrete spike sequences to calculate and transmit information. Spiking neurons receive continuous values and convert them into spike sequences. A number of different spiking neuron models have been proposed. The Hodgkin-Huxley model is one of the first models that describes the behaviour of biological neurons [46], and is fundamental to explaining how spikes flow in neurons, but the model is too complex to implement in silicon. The Izhikevich model [47], which simplifies the Hodgkin-Huxley model, is a two-dimensional model that describes the dynamics of the membrane potential of a neuron. The leaky integrate-and-fire (LIF) neuron is another simple neuron model that is widely used in neuroscience and SNNs. It is simpler than Izhikevich model but captures the essential features of how neurons work. It can be used to build SNNs and implemented in very-large-scale integrations (VLSI) [48]. The membrane potential of the LIF neuron is governed by the following equation:

$$dv/dt = I - v/\tau, \text{ for } v < v_{\text{threshold}} \quad (1)$$

where v is the membrane potential, t is time, τ is a time constant, and I is the input current. The input current I can be either excitatory, which makes the membrane potential more positive, or inhibitory, which makes it more negative. If $v \geq v_{\text{threshold}}$, the neuron fires a spike and then resets its membrane potential to a predefined reset value. The LIF neuron model is simple and computationally efficient, making it suitable for hardware implementations. In this work, LIF is used to build the proposed model.

Similar to ANNs, as the depth of SNNs increases, their performance significantly improves [22,23,49]. Currently, most SNNs have borrowed structures from ANNs, which can be categorised into two main groups: CNN-based and Vision Transformer (ViT) -based SNNs. ResNet, as the most successful CNN model has been extensively studied to extend the depth of SNNs [22,23]. SEW ResNet [22] overcomes the vanishing/exploding gradient problem in SNNs by using a technique called

spike-timing dependent plasticity (STDP). It has been shown to be effective in a variety of tasks, including image classification and object detection. However, convolutional networks possess translation invariance and local dependency, but their calculations have a fixed receptive field, limiting their ability to capture global dependencies. In contrast, ViTs [25] are based on self-attention mechanisms that can capture long-distance dependencies. They are based on the Transformer architecture, which was originally developed for natural language processing tasks.

ViT-based SNNs represent a novel form of SNNs that combine the transformer architecture with SNNs, providing great potential to break through the performance bottleneck of SNNs. Yao et al. [50], and Zhou et al. [34], proposed two different Spike-Driven Self-Attention models. To avoid multiplication, they utilised only mask and addition operations, which are efficient and have low computational energy consumption. Zhou et al. [35], proposed Spikingformer, modifying the residual connection to be purely event-driven, making it energy efficient while improving performance.

However, the application of ViT architectures to Spiking Neural Networks (SNNs) for depth estimation is an emerging field [33] facing significant challenges. Key among these are the difficulties in training pure transformer-based SNN models and the limited availability of paired event-based depth data essential for robust training. Knowledge distillation presents a promising approach to mitigate such challenges, particularly data scarcity, by transferring knowledge from well-pretrained models. Accordingly, this work proposes a knowledge distillation method to leverage ANN model knowledge for SNN-based depth estimation.

2.3. Knowledge distillation for SNN

Knowledge distillation is a model compression technique that transfers knowledge from a large teacher model to a smaller student model, enabling efficient training with limited resources [51]. It has been shown to be effective for improving SNN performance: Kushawaha et al. [52] transferred knowledge from a large to a small SNN for image classification; He et al. [36] further boosted student SNN accuracy; Qiu et al. [53] reduced the ANN-SNN performance gap; and [54] first explored cross-modality distillation for SNN depth estimation using RGB data. While these studies [36,53,54] demonstrate clear benefits, existing approaches still face limitations: (1) most focus on classification rather than dense prediction tasks like depth; (2) cross-modality transfer between conventional images and event data remains underexplored; and (3) many require training a separate teacher, adding computational overhead. Currently, large foundation models have become the new deep learning hotspot [55]. A large foundation model is trained on a vast quantity of data at scale (often by self-supervised learning or semi-supervised learning) so that the learned features can be used directly for various downstream tasks or knowledge distillation. For example, Dense Prediction Transformers (DPT) [39] are a type of ViT designed for depth prediction tasks, trained on 1.4 million images for monocular depth estimation. DINOv2 [37] used ViT-Giant, a larger version of ViT with 1 billion parameters. It is more powerful than previous ViT models and outperforms previous self-supervised learning methods in a variety of computer vision tasks, especially for depth estimation. In this work, for the first time, we will explore transferring knowledge from a large foundation model (DINOv2) to SNNs for depth estimation.

3. The proposed method

We propose a novel energy-efficient spike transformer network for depth estimation via cross-modality knowledge distillation. The flowchart of the method is illustrated in Fig. 1, encompassing three primary components: (1) Spike-Driven Transformer, (2) Fusion Depth Estimation Head, and (3) Knowledge Distillation.

The rationale for our proposed method centres on three key innovations:

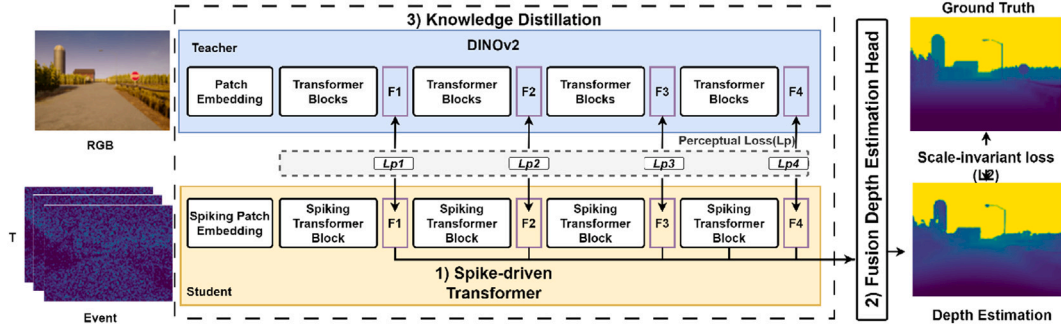


Fig. 1. Overview of the proposed method, illustrating the three main components: (1) spike-driven transformer, (2) fusion depth estimation head, and (3) knowledge distillation.

- (1) We introduce a Spike-Driven Transformer architecture. This architecture replaces conventional computationally intensive floating-point matrix multiplications with binary spike-based computations. This is achieved through spike-based attention and residual mechanisms, which not only reduce energy consumption but also maintain high performance in capturing long-range dependencies.
- (2) We propose a novel fusion depth estimation head designed to integrate features from multiple transformer stages for precise and robust depth estimation. Compared to existing methods, our approach overcomes the limitations of CNN based architectures, which often lose critical spatial information due to downsampling. By leveraging transformers' ability to retain dimensional consistency and integrating features at multiple levels, the fusion head achieves superior depth estimation accuracy. Additionally, it is fully compatible with spike-based computation models, making it both efficient and biologically plausible, providing a significant advantage for real-world applications requiring precision and robustness in challenging environments.
- (3) To address the limited training data available for SNNs, we leverage knowledge from DINOv2, a large vision foundation model, through a novel single-stage cross-modality distillation framework. Rather than requiring separate training phases or an additional teacher model, our approach directly transfers relevant features from RGB to event data domains, enabling efficient training while preserving the spike-based computation paradigm.

3.1. Spike-driven transformer

The proposed spike transformer aligns with the foundational structure of the original ViT, encompassing a Spiking Patch Embedding and Spiking Transformer Block. Given an event sequence, $I \in \mathbb{R}^{T \times C \times H \times W}$, the spike patch embedding is used to convert the input into a sequence of tokens that can be processed by the transformer architecture, where the event input is projected as spike-form patches $X \in \mathbb{R}^{T \times N \times D}$, $N = \frac{H}{8} \times \frac{W}{8}$. Then, the spiking patches X are passed to the multi spiking transformer blocks (L). Considering that we have used knowledge distillation from the large model, this method uses only a minimum number of blocks as $L = 4$. Inspired by Zhou et al. [34,35], in order to avoid non-spike computations in traditional deep learning architectures, a Spiking Self Attention (SSA) and a Spiking MLP block are used in spiking transformer blocks.

3.1.1. Spiking patch embedding

In the original ViT [25], the patch embedding is used to represent an image as a sequence of tokens. This is done by dividing the image into a grid of patches and flattening each patch into a vector. In this work, we implement this operation through a convolution batch norm (ConvBN), Max pooling (MP) and multistep LIF (MLIF) combination. The structure is shown in Fig. 2. Given an input sequence as $I \in \mathbb{R}^{T \times H \times W}$, after the processing of picking patch embedding, I is split into an image patches

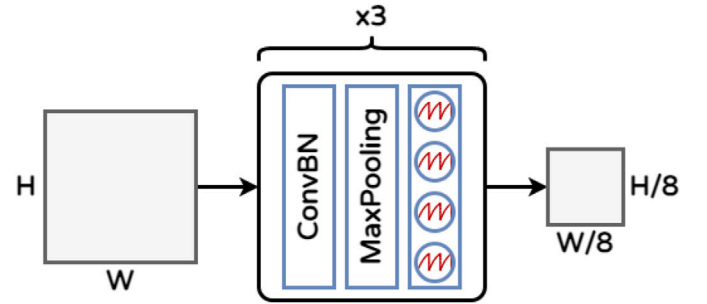


Fig. 2. The architecture of the Spiking Patch Embedding module. This module converts the input event sequence into spike-form patches using a combination of Convolution-Batch Normalization (ConvBN), Max Pooling (MP), and Multi-step Leaky Integrate-and-Fire (MLIF) operations.

$I_{patches} \in \mathbb{R}^{T \times N \times D}$. This process can be formulated as:

$$I_{patches} = \text{MLIF}(\text{MP}(\text{ConvBN}(I))) \quad (2)$$

where the ConvBN applies a 2D convolution with a 3×3 kernel (stride = 1) followed by batch normalization to patch the input. MP (Max Pooling) is used to down-sample the feature size to patch size. The MLIF module is designed to simulate the multi-step dynamics of an LIF neuron to convert continuous feature maps into spike-based representations. In practice, this means that the MLIF operation iterates the LIF process over several discrete time steps to produce a spike train, rather than a single output. The corresponding pseudocode for the MLIF operation is shown in Fig. 3.

where I is assumed to be a sequence of input values over T time steps. The term $(1 - (1/\tau))$ approximates the decay of the potential. The neuron emits a spike when v reaches or exceeds V_{th} , and then v is reset to v_{reset} . The number of operations can be greater than 1. When multiple blocks are used, the number of output channels gradually increases and the size of the feature is halved, eventually matching the embedding dimension of the patch in ViT.

3.1.2. Spiking transformer block

The Spiking Transformer Block is structured to incorporate both a Spiking Self Attention (SSA) mechanism and a Spiking MLP block, as illustrated in Fig. 4.

Guided by the findings in [35], we position an MLIF before the ConvBN within the residual mechanism to omit floating-point multiplication and mixed-precision calculations during the ConvBN operation. This adjustment also enables ConvBN to replace conventional linear layers and batch normalization seamlessly. The SSA operation can be

Algorithm 1 MLIF Algorithm

```

1: function MLIF( $I, T, \tau, V_{th}, v_{reset}$ )
2:    $v \leftarrow 0$ 
3:    $spikes \leftarrow []$ 
4:   for  $t = 1$  to  $T$  do
5:      $v \leftarrow (1 - \frac{1}{\tau}) \cdot v + I[t]$     ▷ assuming  $\Delta t = 1$ 
6:     if  $v \geq V_{th}$  then
7:        $s \leftarrow 1$ 
8:        $v \leftarrow v_{reset}$     ▷ reset potential after spike
9:     else
10:       $s \leftarrow 0$ 
11:    end if
12:    Append  $s$  to  $spikes$ 
13:  end for
14:  return  $spikes$ 
15: end function

```

Fig. 3. Pseudocode for the MLIF operation.

mathematically described as:

$$\begin{aligned}
Q &= \text{MLIF}_Q(\text{ConvBN}_Q(X')) \\
K &= \text{MLIF}_K(\text{ConvBN}_K(X')) \\
V &= \text{MLIF}_V(\text{ConvBN}_V(X')) \\
\text{SSA}(Q, K, V) &= \text{ConvBN}(\text{MLIF}(QK^T V * s))
\end{aligned} \tag{3}$$

The Query (Q), Key (K), and Value (V) matrices are generated by processing input features through learnable transformations (e.g., ConvBN layers) followed by distinct spiking neuron layers (e.g., MLIF layers), as detailed in Eq. (3). This process yields $Q, K, V \in \mathbb{R}^{T \times N \times D}$ as pure spike data, containing only binary values (0 or 1). The Spiking Self-Attention (SSA) mechanism leverages the inherently non-negative nature of these spike-form Q and K matrices to produce a non-negative attention map. This characteristic allows SSA to directly aggregate relevant features while disregarding irrelevant ones, thereby making the conventional softmax function redundant. A scaling factor, s , is employed to adjust the magnitude of the matrix multiplication results within the SSA operation, without altering the fundamental properties of the attention mechanism itself. The Spiking MLP block, also a component of the transformer architecture, consists of a residual connection and a combination of MLIF and ConvBN operations.

3.2. Fusion depth estimation head

The task of depth estimation requires generating pixel-wise depth predictions from encoded features. A common approach is a simple Fully Convolutional Network (FCN) head that processes only the final encoder features. While computationally efficient, this often fails to preserve fine

spatial details necessary for accurate depth maps. A key challenge in adapting transformers for dense prediction is that, unlike CNNs which naturally produce a multi-scale feature hierarchy through progressive downsampling, standard Vision Transformers (ViTs) maintain a constant spatial resolution of tokens throughout their layers. To address this, we propose a fusion depth estimation head that explicitly combines features from multiple stages of the spike-driven transformer backbone.

While the spatial resolution of tokens remains fixed, the effective receptive field and semantic level of the features evolve through the transformer blocks. Early layers capture local, fine-grained details, whereas deeper layers, through successive self-attention operations, integrate information across the entire token set to learn more global, abstract, and semantic representations. Therefore, fusing features from different stages allows the decoder to leverage both high-resolution structural details (from early layers) and robust semantic context (from later layers), which is critical for high-quality depth estimation. This multi-stage fusion strategy is not ad hoc but follows established best practices in state-of-the-art transformer architectures for dense prediction. Models like DPT (Dense Prediction Transformers) [39] and SegFormer [56] have successfully demonstrated that combining features from multiple transformer blocks significantly improves performance in tasks like depth estimation and semantic segmentation. Our fusion head adapts this proven concept to our spike-driven backbone. The structure of the fusion head for depth estimation is shown in Fig. 5.

The first step of the fusion head is to assemble the internal features in transformer blocks into image-like feature representations. The feature representations are then fused into the final dense prediction with skip connections. A generic upsampling structure is used to restore the feature representations to original data size. Given an input feature as $F_i \in \mathbb{R}^{T \times H/8 \times W/8}$, $i=1, 2, 3, 4$. The depth estimation head can be formulated as follows:

$$\begin{aligned}
Y_2 &= (\text{ConvBN}(\text{Up}(F_1)) + \text{Up}(F_2)) \\
Y_3 &= (\text{ConvBN}(\text{Up}(Y_2)) + \text{Up}(F_3)) \\
Y_4 &= (\text{ConvBN}(\text{Up}(Y_3)) + \text{Up}(F_4)) \\
Y &= \text{Sigmoid}(Y_4)
\end{aligned} \tag{4}$$

This design enables the network to combine high-level semantic information from deeper layers with fine-grained spatial details from earlier layers, leading to more accurate depth predictions while maintaining compatibility with knowledge distillation from vision foundation models.

3.3. Knowledge distillation

Knowledge distillation is particularly challenging for SNNs due to two main factors: (1) the binary nature of spike data differs fundamentally from the continuous values used in traditional ANNs, and (2) the limited availability of labeled event camera data makes training challenging. To address these challenges, we propose a single-stage cross-modality knowledge distillation framework that leverages DINOv2 [37], a large-scale vision foundation model, to guide our SNN training.

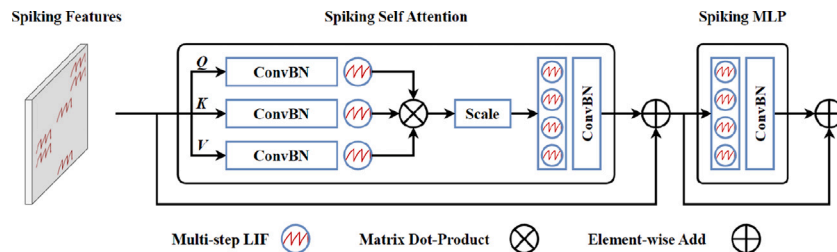


Fig. 4. The architecture of the Spiking Transformer Block, detailing the integration of the Spiking Self Attention (SSA) mechanism and the Spiking MLP block.

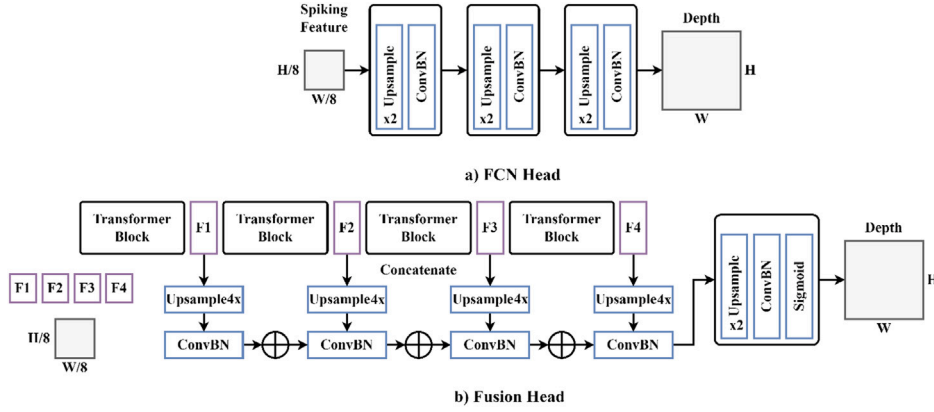


Fig. 5. Architecture of the proposed fusion depth estimation head. It integrates multi-stage features from the spike-driven transformer blocks (F_1, F_2, F_3, F_4) through progressive upsampling and skip connections to generate the final depth map.

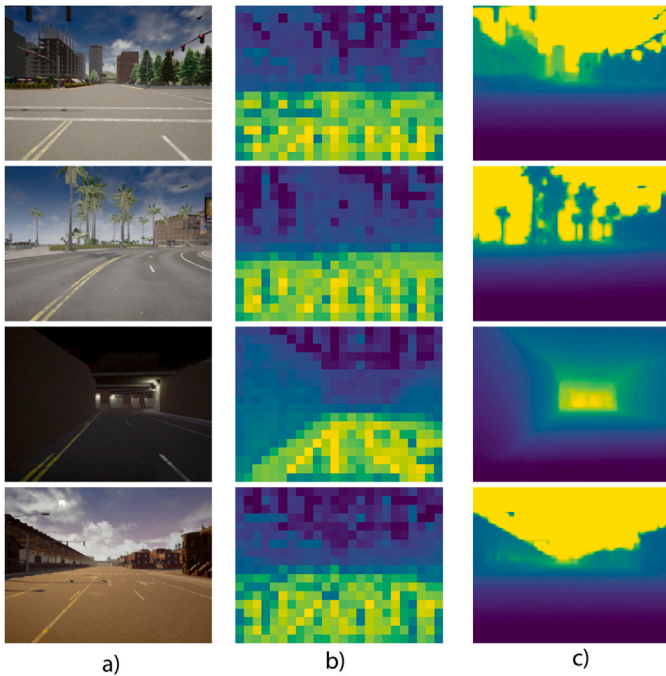


Fig. 6. (a) Sample RGB image. (b) Visualization of self-attention map from DINOv2 features. (c) Depth estimation results using a linear probe applied to frozen DINOv2 features.

Our choice of DINOv2 as the teacher model is motivated by several key advantages:

- (1) **Architectural Compatibility:** DINOv2’s Vision Transformer (ViT) architecture closely aligns with our model’s structure, facilitating effective knowledge transfer due to its similar feature representations and computational patterns.
- (2) **Rich Feature Representations:** Pre-trained on 142 million diverse images, DINOv2 has demonstrated state-of-the-art performance in depth estimation tasks on benchmark datasets such as NYU [57] and SUN RGB-D [58]. As shown in Fig. 6, DINOv2’s self-attention patterns and depth estimation capabilities on our dataset suggest that it can provide valuable guidance during the knowledge distillation process.
- (3) **Zero-shot Generalisation:** DINOv2’s strong zero-shot learning capabilities enable effective knowledge transfer even when dealing with limited event camera data.

Algorithm 2 Spike knowledge distillation algorithm

Input:

- x_batch : One batch Spiking images;
 - x_{rgb_batch} : One batch Mathced RGB images;
 - T : Teacher Network;
 - S : Student Network;
 - H : Depth estimation head Network;
- 1: set $T.params = S.params$;
 - 2: set $T.Frozen()$ # Frozen Teacher’s params;
 - 3: **for** x, x_{rgb} **in** x_batch, x_{rgb_batch} **do** # One batch training
 - 4: $x'_{rgb} = T(x_{rgb})$
 - 5: $x' = S(x)$
 - 6: $D = H(x')$
 - 7: $loss = \sum_{i=1}^4 L_p(x'_i, x'_{i,rgb}) + L_2(D, Target)$
 - 8: $loss.backward()$ # Back-propagate
 - 9: Update($S.params$) # Student params update by knowledge distillation
 - 10: **end for**

Fig. 7. The knowledge distillation algorithm, illustrating the process of transferring knowledge from the DINOv2 teacher model to the student SNN using a combined loss function.

The knowledge distillation process can be shown in Fig. 1. We freeze the DINOv2 (Lightblue) as a teacher model. The output features from DINOv2 are considered as targets in our training. To ensure compatible feature dimensions, we upsample the RGB input images by a factor of 1.75, resulting in teacher model features of size $x'_{rgb} \in \mathbb{R}^{d \times H/8 \times W/8}$.

Fig. 7 shows the knowledge distillation algorithm. Our distillation framework employs a fusion loss function that combines two complementary components:

- **Feature Perceptual Loss (Lp):** Measures the distance between student and teacher feature representations, ensuring the SNN learns similar feature patterns. The Perceptual Loss [59] is used here to help capture high-level semantic differences between teacher and student representations, going beyond pixel-level comparisons.
- **L2 loss function:** A scale-invariant metric [60], specifically designed for monocular depth estimation to address the inherent scale ambiguity problem. This scale-invariant loss is particularly important as it focuses on relative depth relationships rather than absolute values,

aligning with the fundamental nature of monocular depth estimation where absolute scale cannot be determined from a single view.

The fusion loss function is defined by the following equations:

$$\mathcal{L}_{p_i} = \frac{1}{C \times H \times W} \|x_i - x'_i\|_2^2$$

$$\mathcal{L}_2 = \frac{1}{n} \sum_i (D_i^t - D_i^p)^2 - \frac{1}{n^2} \left(\sum_p D_i^t - D_i^p \right)^2 \quad (5)$$

Where \mathcal{L}_{p_i} is the feature perceptual loss between features for pixel i . $D_i^t - D_i^p$ is the difference between predicted and ground truth depth for pixel i , and n is total number of pixels with a dimension of $H \times W$. D_i^t is the ground-truth depth, and D_i^p is the predicted depth. It makes the loss invariant to uniform scaling of the depth predictions, allowing the network to learn consistent relative depth relationships even when absolute scale cannot be determined. This aligns with human depth perception, which relies heavily on relative rather than absolute depths.

4. Experiments

To test our model, we conducted two experiments to demonstrate the effectiveness of the proposed SNN. We first introduce the details of datasets used in this experiment. Then, we evaluate our method's performance, including accuracy and energy consumption, on both real and synthetic event data to demonstrate its robustness and generalisability. Finally, comprehensive ablation studies are conducted to investigate the impact of each component.

4.1. Datasets

For model evaluation, we utilise two datasets comprising both real and synthetic data.

DENSE Datasets: The first dataset is a synthetic dataset from Zhang et al. [33], which is generated from the DENSE dataset [61], including clear depth maps and intensity frames at 30 FPS under a variety of weather and illumination conditions. To obtain spike streams with high temporal resolution, the video is interpolated to generate intermediate RGB frames between adjacent 30-FPS frames. With absolute intensity information among RGB frames, each sensor pixel can continuously accumulate the light intensity with the spike generation mechanism, producing spike streams with a high temporal resolution (128×30 FPS) that is 128 times the video frame rate. The 'spike' version of the DENSE dataset (namely DENSE spike) contains eight sequences, five for training, and three for evaluation. Each sequence consists of 999 samples, and each sample is a tuple of one RGB image, one depth map, and one

spike stream. Each spike stream is simulated between two consecutive images, generating a binary sequence of 128 spike frames (with a size of 346×260 each) to depict the continuous process of dynamic scenes.

DSEC Datasets: The second dataset, DSEC [62], is a real event dataset that provides stereo dataset in driving scenarios. It contains data from two monochrome event cameras and two global shutter colour cameras in favourable and challenging illumination conditions. Hardware synchronised LiDAR data is also provided for depth prediction. The dataset contains 41 sequences collected by driving in a variety of illumination conditions and provides ground truth disparity for the depth estimation evaluation. In this work, 29 sequences (70 %) are used for model training and 12 are used for evaluation. Each sequence consists of 200–900 samples, and each sample is a tuple of one RGB image, one depth map (dense disparity), and one spike stream with 16 spike frames and size of 480×640 . Fig. 8 presents the two data samples used in this work.

4.2. Experiment design

4.2.1. Model performance

In this section, we evaluate the depth estimation performance and energy consumption of our SNN on the synthetic (DENSE) and real datasets (DSEC) and compare it with three competing dense prediction networks, namely U-Net [43], E2Depth [61] and Spike-T [33]. U-Net employs 2D convolutional layers as its encoder and focuses on spatial feature extraction, while E2Depth applies ConvLSTM layers that combine CNN and LSTM to capture the spatial and temporal features. The Spike-T employs transformer-based blocks to learn the spatio-temporal features simultaneously. These models therefore constitute our immediate and direct competitors. To ensure a fair and direct comparison, all baseline models were re-implemented and trained from scratch on our specific datasets and data processing pipeline using their publicly available source code and recommended hyperparameters.

The network's total energy consumption is the sum of energy from its spike-based Accumulate (AC) operations and any conventional Multiply-Accumulate (MAC) operations.

These calculations assume 45 nm hardware [63], with an energy cost of $E_{MAC} = 4.6$ pJ per MAC operation and $E_{AC} = 0.9$ pJ per AC operation. The total energy is calculated as:

$$E_{\text{model}} = \sum_{l \in \text{MAC_layers}} E_{MAC} \times \text{FLOP}_l + \sum_{l \in \text{AC_layers}} E_{AC} \times \text{SOP}_l \quad (6)$$

Here, FLOP_l is the number of floating-point MAC operations in a conventional layer l . For spike-based layers, the number of Synaptic

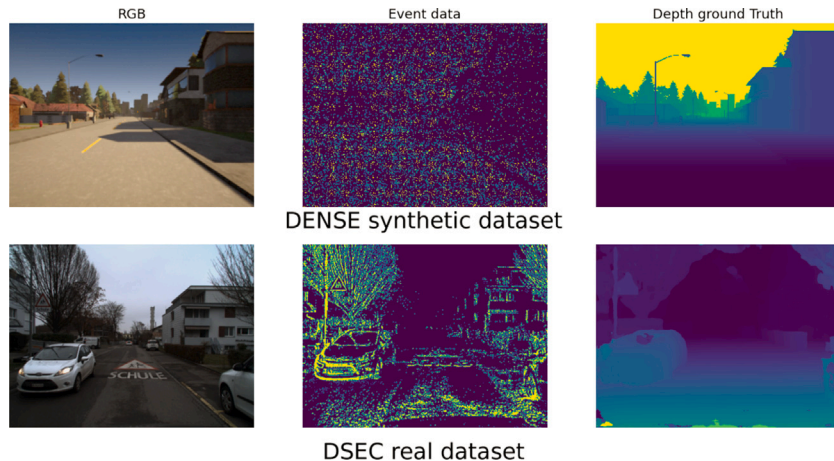


Fig. 8. Examples from the DENSE (synthetic) and DSEC (real-world) datasets used for evaluation. Each sample typically includes an RGB image, the event stream and a corresponding depth map.

Table 1

Quantitative performance comparison on synthetic (DENSE) and real (DSEC) datasets. Symbols ↓ / ↑ indicate that lower / higher values are better. Param (M): parameters (millions). Mean firing rate (f_r) and Power: theoretical energy per inference (45 nm estimates). †Models evaluated on DSEC with replicated temporal frames (16→128).

Model	Dataset	Abs Rel ↓	Sq Rel ↓	RMS log ↓	SI log ↓	$\delta < 1.25$ †	$\delta < 1.25^2$ †	$\delta < 1.25^3$ †	Spike-driven	Param (M)	Mean FR	Power (mJ)
U-Net	DENSE	2.89	72.25	0.19	1.73	0.39	0.50	0.58	No	31.20	0.88	72.93
	DSEC	1.203	3.281	0.181	1.210	0.142	0.309	0.502				
E2Depth	DENSE	9.91	96.01	0.30	1.70	0.21	0.31	0.45	No	10.71	0.85	59.25
	DSEC†	9.909	96.011	0.299	1.697	0.209	0.309	0.448				
Spike-T	DENSE	1.57	39.77	0.17	0.91	0.50	0.65	0.74	No	35.68	0.65	41.77
	DSEC†	2.853	51.757	0.321	0.751	0.164	0.341	0.483				
Proposed	DENSE	0.80	8.32	0.17	0.46	0.53	0.68	0.76	Yes	20.55	0.35	12.43
	DSEC	1.000	0.999	0.105	0.212	0.387	0.500	0.583				

Operations (SOP) is estimated based on the firing rate f_r (the proportion of non-zero elements in the spike matrix), the number of time steps T , and the equivalent FLOPs of the layer:

$$SOP_l = f_r \times T \times FLOP_l \quad (7)$$

4.2.2. Ablation study

An ablation study is detailed in this subsection to investigate the contributions of two novel components within our model. They are the fusion depth estimation head and the knowledge distillation technique. We compare: (i) Linear FCN Head – a single-scale decoder analogous to standard lightweight heads attached to frozen encoders (e.g., typical DINOv2 usage), (ii) full architecture without knowledge distillation (W/O KD), and (iii) proposed multi-stage fusion + KD. Their respective impacts on model performance are dissected and discussed.

4.2.3. Implementation details

All models were trained and evaluated on a single NVIDIA RTX A6000 GPU. Our proposed model and baseline model U-Net were trained for 200 epochs with a batch size of 4 using the AdamW optimizer with a weight decay of 0.05. The learning rate was initialised to 1e-5 and followed a cosine annealing schedule. For the Spike-T and E2Depth, we utilised the officially provided pre-trained weights from their respective authors, as the full original training parameters were not available. The models and weights were obtained from their official codebases at <https://github.com/Leozhangjiyuan/MDE-SpikingCamera> and https://github.com/uzh-rpg/rpg_e2depth respectively. The architecture of our proposed model is detailed in Section 3, with key components illustrated in Figs. 1–4. The spike-driven backbone consists of 4 spiking transformer blocks with an embedding dimension of 384. To facilitate reproducibility, our source code and pre-trained weights will be made publicly available at <https://gitlab.com/han-research/spike-transformer-for-depth>.

4.3. Metrics

Several metrics are selected to evaluate the performance of the proposed method, including Absolute Relative Error (Abs Rel.), Squared Relative Error (Sq Rel.), Mean Absolute Error (MAE), Root Mean Square Logarithmic Error (RMSE log) and the Accuracy metric (Acc. δ). The formulations are as follows:

Absolute Relative Error (Abs Rel.) ↓ computes average errors on the normalized depth map for every pixel, formulated as:

$$Abs\ Rel. = \frac{1}{N} \sum_p \frac{|D_p - \hat{D}_p|}{|D_p|} \quad (8)$$

It normalizes the value of depth to the range [0,1].

Square Relative Error (Sq Rel.) ↓, formulated as

$$Sq\ Rel. = \frac{1}{N} \sum_p \frac{|D_p - \hat{D}_p|^2}{|D_p|} \quad (9)$$

which focuses on large depth errors due to its squared numerator.

Mean Absolute Error (MAE) ↓ can be formulated as:

$$MAE = \frac{1}{N} \sum_p |D_p - \hat{D}_p| \quad (10)$$

Root Mean Square Error (RMSE log) ↓ is a classic metric for pixel prediction error and the logarithmic version can be denoted as

$$RMSE = \sqrt{\frac{1}{N} \sum_p |\log D_p - \log \hat{D}_p|^2} \quad (11)$$

The Accuracy (Acc δ) ↑ as δ denotes the percentage of all pixels D_p that satisfy max:

$$Acc = \left(\frac{\hat{D}_p}{D_p}, \frac{D_p}{\hat{D}_p} \right) < thr \quad (12)$$

where $thr = 1.25, 1.25^2, 1.25^3$ [7].

4.4. Experiment results

4.4.1. Model performance

Our experimental investigation encompasses both quantitative performance and energy consumption analyses, utilising synthetic (DENSE) and real-world (DSEC) datasets. Several metrics were employed to evaluate the outcomes comprehensively Table 1.

On DENSE dataset, the results consistently indicate that the proposed method outperforms the alternative approaches across nearly all evaluated metrics. Notably, substantial reductions are observed in the absolute relative error (Abs.Rel) and squared relative error (Sq.Rel)—critical metrics in depth estimation tasks. The Abs.Rel for the proposed method, recorded at 0.80, represents reductions of 72 %, 91.9 %, and 49 % compared to U-Net (2.89), E2Depth (9.91), and Spike-T (1.57), respectively. Similarly, the Sq.Rel of our method, at 8.32, demonstrates reductions of 88.5 %, 91.3 %, and 79 % relative to U-Net (72.25), E2Depth (96.01), and Spike-T (39.77).

In addition to error reduction, the proposed method achieves modest increases in accuracy metrics ($\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$), attaining values of 0.53, 0.68, and 0.76, respectively. These values are slightly higher than those achieved by the competing methods, reinforcing the method's superior capability in depth estimation tasks.

In terms of power consumption, the proposed method demonstrates a marked advantage in computational efficiency. A key contributor to this is the sparse activation inherent to SNNs, reflected in the low mean firing rate (FR) of our model. This metric quantifies the event-driven nature of the computation, where energy is consumed only for active neurons. This sparsity allows our model to reduce theoretical power consumption by up to 82.9 % compared to the dense operations of U-Net (12.43 mJ vs. 72.93 mJ). Furthermore, our knowledge distillation framework enables a more compact architecture, reducing the parameter count by 42.4 % compared to the Spike-T method (20.55 M vs. 35.68 M).

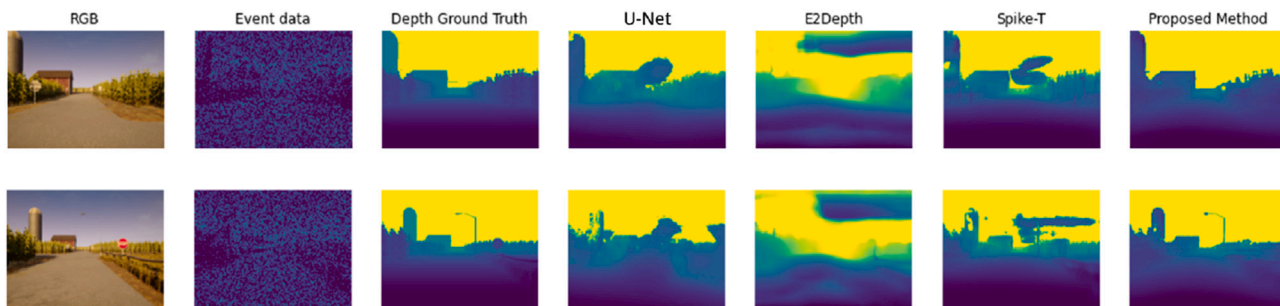


Fig. 9. Visual comparison of predicted depth maps on the DENSE validation set. The proposed model recovers sharper object boundaries, thin structures, and more globally consistent depth (far-range stability and reduced bleeding) than U-Net, E2Depth, and Spike-T, which exhibit smoothing, edge erosion, or depth discontinuities.

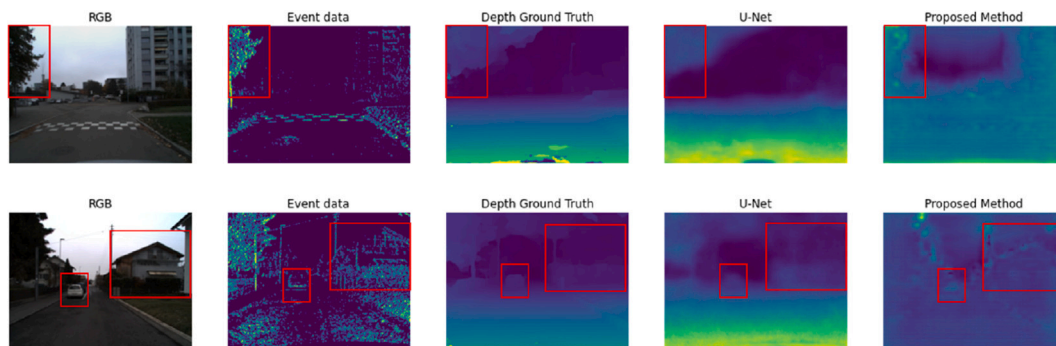


Fig. 10. Depth estimation results under low-light conditions on the DSEC dataset. The proposed model preserves thin roadside structures (e.g., trees, poles), sharp object boundaries (buildings, vehicles), and delivers more stable far-range depth than baseline methods, despite the sparsity and noise of the event stream.

Table 2

Ablation study results: Quantitative performance comparison on the synthetic (DENSE) dataset. Symbols \downarrow and \uparrow indicate that a lower value and higher value are preferable, respectively.

	Abs Rel \downarrow	Sq Rel \downarrow	RMS log \downarrow	SI log \downarrow	$\delta < 1.25$ \uparrow	$\delta < 1.25^2$ \uparrow	$\delta < 1.25^3$ \uparrow
Linear FCN Head	2.85	51.76	0.32	0.75	0.16	0.34	0.48
W/O KD	3.52	85.10	0.25	2.15	0.31	0.42	0.51
Proposed	0.80	8.32	0.17	0.46	0.53	0.68	0.76

These experimental results show that our proposed method can more effectively capture the spatiotemporal characteristics of irregular continuous spike data streams, delivering satisfactory accuracy. This is further illustrated in Fig. 9, which depicts the visualization results from multiple comparison models on a validation synthetic dataset. The visualization demonstrates that, unlike the U-Net and Spike-T methods which can predict details yet misestimate depth, or the E2Depth method that produces blurry outcomes losing fine details, our method effectively manages to capture more intricate details, including minute structures, sharp edges, and contours.

In addition, in order to validate the generalisability of the model, we evaluate the proposed model on DSEC real event dataset. It is noteworthy that while the DENSE synthetic dataset encompasses 128 spike frames, the real-world DSEC dataset contains only 16 frames. Our model and the SpikeU-Net model require retraining on this reduced dataset. However, the SNN-based methods (E2Depth and Spike-T) are unable to be retrained due to insufficient training parameters, necessitating the replication of DSEC data to 128 frames to accommodate their setups. Consequently, the performance of E2Depth and Spike-T is expectedly low.

The results, as shown in Table 1, demonstrate that the proposed model excels across all metrics in comparison to the E2Depth, Spike-T, and U-Net models. This superior performance is evident particularly in terms of metrics such as Abs Rel, Sq Rel, RMS log, and SI log, as well

as in the accuracy metrics ($\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$), where higher scores are indicative of better performance. These findings underscore the effectiveness of the proposed model in handling real-event data from spiking cameras.

Fig. 10 shows the visualization result in a low light environment. Our method effectively identifies features such as trees and houses along the roadside, as well as vehicles located in the center of the road.

4.4.2. Ablation study

This subsection presents an ablation study conducted to evaluate the effectiveness of the proposed Fusion Depth Estimation Head and Knowledge Distillation (KD) modules.

Table 2 reports the quantitative performance comparison on the synthetic datasets (DENSE) in the ablation study. As demonstrated by the results, all accuracy metrics exhibit a decline when employing the linear FCN (Fully Convolutional Network) head for depth estimation. Specifically, Absolute Relative (Abs Rel) error increased from 0.80 to 2.85, while the Squared Relative (Sq Rel) error escalated from 8.32 to 51.76. Fig. 11 illustrates the visualization results of using two different heads. The image becomes notably blurrier and loses details when employing the linear FCN head, which relies solely on the final features generated by the transformer. Conversely, our fusion head integrates multi-scale features, thereby facilitating superior recovery of details compared to the linear FCN head.

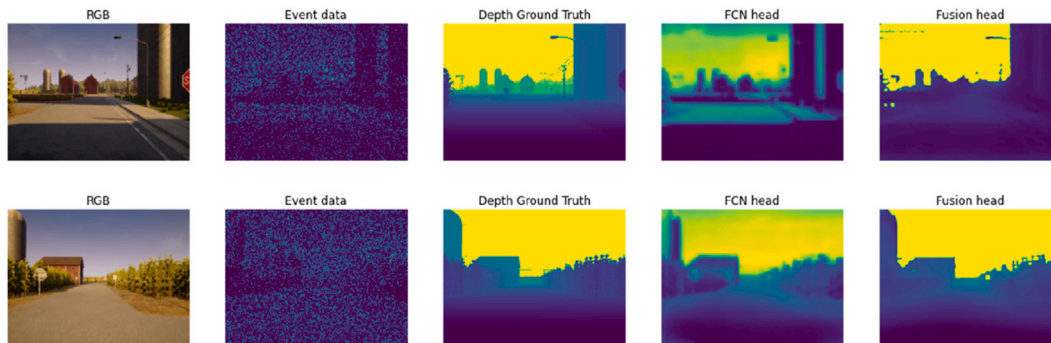


Fig. 11. Visualization results on validation synthetic (DENSE) dataset by using FCN head and proposed fusion depth estimation head.

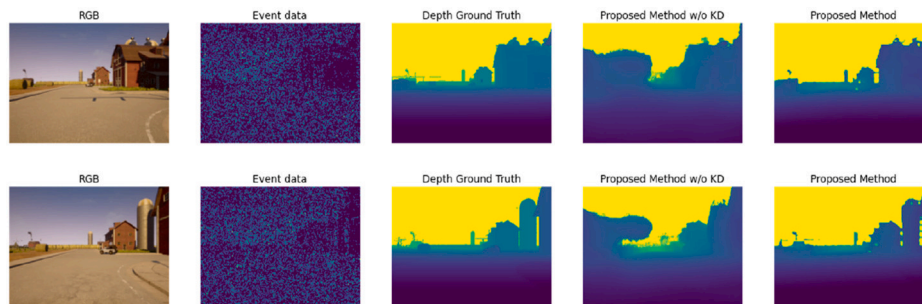


Fig. 12. Visualization results on validation synthetic (DENSE) dataset with and without knowledge distillation.

The visualization of results with and without knowledge distillation is depicted in Fig. 12. The results without knowledge distillation approximate those of the baseline models from Experiment 1. The presence of noise in the spike data leads to less accurate depth estimations for certain segments of the point cloud. However, employing knowledge distillation enables the model to predict the depths of distant clouds more accurately, a benefit attributed to the enhanced inductive capabilities derived from the substantial foundational model.

5. Discussion

The development of efficient and accurate depth estimation methods for event cameras remains a significant challenge in computer vision [5,6]. Our work addresses this challenge through three key innovations: (1) a purely spike-driven transformer architecture, (2) fusion depth estimation head and (3) a novel knowledge distillation framework. The experimental results demonstrate the improvements in both accuracy and energy efficiency, we discuss the implications, limitations, and potential impact of our approach in this section.

The satisfactory performance of the spike transformer architecture can be attributed to several factors. Beyond mere energy savings, the architecture’s success stems from the fundamental synergy between SNNs and event data. By implementing spike-driven residual learning and self-attention mechanisms [34,35], our model effectively captures the temporal dynamics inherent in event camera data while maintaining the computational efficiency characteristic of SNNs [15]. This is evidenced by the significant reductions in absolute relative error (49 % reduction) and squared relative error (39.77 % reduction) compared to the state-of-the-art Spike-T model [33]. The elimination of floating-point operations in the transformer portion substantially reduces power consumption [64,65], making our approach more practical for real-world applications where energy efficiency is crucial.

Our innovative fusion depth estimation head integrates multi-scale transformer features, preserving fine details and global structure, addressing limitations of traditional methods and ensuring robust

performance. Traditional approaches often struggle to maintain fine-grained spatial information while processing temporal event data [3]. Our fusion head addresses this by effectively combining features from multiple transformer stages, as demonstrated by the improved preservation of detail in our depth predictions compared to standard linear FCN heads [43]. This multi-scale feature integration is particularly beneficial for capturing both fine details and global scene structure [66], contributing to the overall robustness of our depth estimates. As noted in the introduction, the fusion head is intentionally not purely spike-based. This decision was driven by the requirement for high accuracy in depth estimation tasks, as pure spike-based operations currently face limitations for precise value prediction.

Our hybrid approach allows us to leverage the energy efficiency of spike-based computing in the feature extraction stages while maintaining the high accuracy requirements of depth estimation through conventional computation in the fusion head.

Our knowledge distillation framework represents a significant advancement in addressing the longstanding challenge of training SNNs with limited data [36]. While the ablation study in Table 2 shows that our model without knowledge distillation (“W/O KD”) performs comparably to the U-Net baseline, this result precisely underscores the necessity and impact of our distillation approach. The key contribution is not that a spike-driven transformer alone is superior, but that our specific knowledge distillation framework elevates its performance to a state-of-the-art level. The technical innovation lies in our single-stage, cross-modality distillation strategy, which directly transfers knowledge from a large, pre-trained vision foundation model (DINOv2) to the SNN student without requiring a custom-trained teacher model. This is enabled by a carefully designed fusion loss function that combines a Feature Perceptual Loss (L_p) with a scale-invariant L2 loss. This combination is novel and critical: the perceptual loss aligns high-level semantic features across the disparate RGB and event-data domains, while the scale-invariant loss specifically addresses the inherent scale ambiguity of monocular depth estimation. As shown in Fig. 12, this framework allows the SNN to learn robust representations that generalise even to

challenging scenarios, such as estimating the depth of distant objects, which would be difficult to learn from the limited event data alone.

Despite the promising results, our approach has several limitations that merit explicit discussion. (1) Dependence on knowledge distillation: the model's state-of-the-art performance is heavily dependent on knowledge distillation from a large foundation model; this reliance introduces complexity into the training pipeline and may limit applicability where a suitable teacher is unavailable. (2) Scalability constraints: training larger backbones or scaling to substantially larger datasets would sharply increase computational cost, memory footprint (activation, membrane and optimizer states), and training instability. Although spike operations are energy-efficient at inference, current software stacks (GPU/TPU kernels, autograd, scheduler support) are not yet optimised to exploit event sparsity during large-scale training, yielding sub-linear hardware utilisation, higher gradient variance, and more frequent issues such as silent or exploding neurons. (3) Dataset availability: publicly available event-based depth datasets remain limited in both scale and scene diversity compared to large RGB depth corpora; this scarcity constrains architecture scaling, hyperparameter exploration, and statistically robust generalisation assessment, reinforcing the need for cross-modality distillation to import broader semantic priors. (4) Real-world deployment: reported energy gains are analytic (45 nm MAC/AC cost models) and realised efficiency will depend on hardware factors (memory hierarchy, spike packet congestion, IO bandwidth, leakage, clock gating efficacy, quantisation resilience, temporal jitter, sensor noise, device variability); additionally the MAC-based fusion/upsampling depth head (retained for continuous numeric precision) breaks end-to-end spike purity and may become a bottleneck on strictly event-driven accelerators.

Several avenues for future research emerge from our findings. First, investigating methods to achieve comparable accuracy with pure spike-based fusion mechanisms remains an important challenge, though this may require fundamental advances in spike-based computing precision. Second, evaluation on neuromorphic hardware platforms (such as SpiNNaker 1/2 [67,68], BrainScales [69], or TrueNorth [70]) would provide valuable insights into real-world performance and energy efficiency. Especially SpiNNaker 2 [68], unlike traditional methods, is specifically designed to handle operations such as multiplications, which are generally inefficient for spiking computations. The introduction of such platforms highlights the necessity for hybrid approaches, where specific operations may leverage conventional hardware optimisations while retaining the efficiency of spike-driven designs.

Additionally, our architecture's ability to effectively process temporal event data suggests potential applications beyond depth estimation, such as object tracking [7] or motion estimation [11]. The success of our knowledge distillation approach also raises interesting questions about the broader applicability of foundation models in training efficient SNNs for various computer vision tasks [55].

The integration of event cameras in autonomous systems and robotics applications continues to grow [12], driven by their advantages in terms of latency, dynamic range, and power efficiency [9]. Our work demonstrates that by combining the biological inspiration of SNNs with modern deep learning architectures and knowledge distillation techniques, we can develop more efficient and accurate methods for processing event camera data. While our hybrid approach represents a careful balance between computational efficiency and accuracy requirements, it provides valuable insights for the broader development of neuromorphic computing systems. The success of this approach suggests that future developments in spike-based computing may benefit from similar pragmatic trade-offs between pure neuromorphic computation and task-specific performance requirements.

6. Conclusion

In this paper, we have introduced a novel energy-efficient Spike Transformer network for depth estimation, leveraging spiking camera data. The proposed architecture integrates spike-driven residual learning

and spiking self-attention mechanisms, creating a transformer framework that operates entirely within the spike domain. This innovative design achieves significant computational efficiency, with an 82.9 % reduction in power consumption compared to conventional methods (from 72.93 mJ to 12.43 mJ per inference). Additionally, our single-stage knowledge distillation framework, leveraging large foundational ANN models such as DINOv2, enables robust training of SNNs even in the presence of limited data. Extensive evaluations on synthetic and real-world datasets demonstrate the efficacy of our approach, with significant improvements in key performance metrics, including a 49 % reduction in Absolute Relative Error and a 39.77 % reduction in Square Relative Error compared to the state-of-the-art SpikeT model. The architecture further enhances efficiency with a 42.4 % reduction in parameters (20.55 M versus 35.68 M), making it particularly well-suited for resource-constrained environments. By combining high accuracy with remarkably low power requirements, our spike-based design is well-suited for practical applications. Future work will focus on extending this research to broader real-world scenarios, including deployment on dedicated SNN processors and further validation with diverse datasets. These efforts aim to unlock the full potential of Spike Transformers in applications such as autonomous navigation, robotics, and energy-efficient vision systems, paving the way for advanced neuromorphic computing in practical settings.

CRedit authorship contribution statement

Xin Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Liangxiu Han:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Sergio Davies:** Writing – review & editing, Supervision, Methodology. **Tam Sobeih:** Writing – review & editing, Supervision, Methodology, Data curation. **Lianghao Han:** Writing – review & editing. **Darren Dancey:** Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered potential competing interests:

Liangxiu Han reports that financial support was provided by Engineering and Physical Sciences Research Council. Liangxiu Han reports that financial support was provided by Biotechnology and Biological Sciences Research Council. Liangxiu Han reports that financial support was provided by Innovate UK. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Funding acknowledgement: EPSRC (EP/X013707/1), BBSRC (BB/R019983/1, BB/S020969/1), Innovate UK (Grant No.10091423).

Data availability

Data will be made available on request.

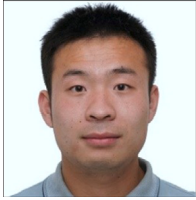
References

- [1] U. Rajapaksha, F. Sohel, H. Laga, D. Diepeveen, M. Bennamoun, Deep learning-based depth estimation methods from monocular image and videos: a comprehensive survey, *ACM Comput. Surv.* 56 (2024) 1–51.
- [2] C. Zhao, Q. Sun, C. Zhang, Y. Tang, F. Qian, Monocular depth estimation based on deep learning: an overview, *Sci. China Technol. Sci.* 63 (9) (2020) 1612–1627, <https://doi.org/10.1007/s11431-020-1582-8>
- [3] Y. Ming, X. Meng, C. Fan, H. Yu, Deep learning for monocular depth estimation: a review, (2021) <https://doi.org/10.1016/j.neucom.2020.12.089>
- [4] H. Laga, L.V. Jospin, F. Boussaid, M. Bennamoun, A survey on deep learning techniques for stereo-based depth estimation, *IEEE Trans. Pattern Anal. Mach. Intell.* (2022) <https://doi.org/10.1109/TPAMI.2020.3032602>.

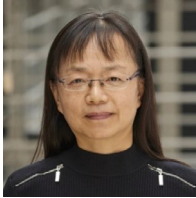
- [5] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A.J. Davison, J. Conradt, K. Daniilidis, D. Scaramuzza, Event-based vision: a survey, *IEEE Trans. Pattern Anal. Mach. Intell.* (2022) <https://doi.org/10.1109/TPAMI.2020.3008413>.
- [6] M.H. Tayarani-Najaran, M. Schmuken, Event-based sensing and signal processing in the visual, auditory, and olfactory domain: a review, *Front. Neural Circuits* 15 (2021) 610446.
- [7] S. Chen, M. Guo, Live demonstration: CeleX-V: a 1m pixel multi-mode event-based sensor, (2019) 1682–1683. ISSN: 2160-7516, <https://doi.org/10.1109/CVPRW.2019.00214>
- [8] P. Lichtsteiner, C. Posch, T. Delbruck, A 128×128 120 dB $15\mu\text{s}$ latency asynchronous temporal contrast vision sensor, *IEEE J. Solid-State Circuits* 43 (2008) 566–576.
- [9] C. Posch, D. Matolin, R. Wohlgenannt, A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS, *IEEE Journal of Solid-State Circuits* (2011) <https://doi.org/10.1109/JSSC.2010.2085952>.
- [10] J. Furmonas, J. Liobe, V. Barzdenas, Analytical Review of Event-Based Camera Depth Estimation Methods and Systems, MDPI, 2022.
- [11] X. Huang, M. Halwani, R. Muthusamy, A. Ayyad, D. Swart, L. Seneviratne, D. Gan, Y. Zweiri, Real-Time Grasping Strategies Using Event Camera, Springer, 2022.
- [12] H. Cao, G. Chen, J. Xia, G. Zhuang, A. Knoll, Fusion-Based Feature Attention Gate Component for Vehicle Detection Based on Event Camera, *IEEE*, 2021.
- [13] G. Chen, H. Cao, J. Conradt, H. Tang, F. Rohrborn, A. Knoll, Event-based neuromorphic vision for autonomous driving: a paradigm shift for bio-inspired visual sensing and perception, *IEEE Signal Process. Mag.* 37 (4) (2020) 34–49.
- [14] W. Maass, Networks of spiking neurons: the third generation of neural network models, *Neural Networks* 10 (1997) 1659–1671.
- [15] T.H. Rafi, A Brief Review on Spiking Neural Network—a Biological Inspiration, 2021 Preprints.
- [16] C. Lee, A.K. Kosta, A.Z. Zhu, K. Chaney, K. Daniilidis, K. Roy, Spike-flownet: event-based optical flow estimation with energy-efficient hybrid neural networks, in: *European Conference on Computer Vision*, 2020, pp. 366–382, https://doi.org/10.1007/978-3-030-58526-6_22
- [17] D. Auge, J. Hille, E. Mueller, A. Knoll, A survey of encoding techniques for signal processing in spiking neural networks, *Neural Process. Lett.* 53 (6) (2021) 4693–4710, <https://doi.org/10.1007/s11063-021-10562-2>
- [18] L. Cordone, B. Miramond, S. Ferrante, Learning from event cameras with sparse spiking convolutional neural networks 2021. ISSN: 2161-4407, <https://doi.org/10.1109/IJCNN52387.2021.9533514>
- [19] C.D.E.A. Schuman, Opportunities for neuromorphic computing algorithms and applications, *Nat. Comput. Sci.* 2 (2022) 10–19.
- [20] P.U. Diehl, G. Zarella, A. Cassidy, B.U. Pedroni, E. Neftci, Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware, in: *2016 IEEE International Conference on Rebooting Computing*, 2016, pp. 1–8, <https://doi.org/10.1109/ICRC.2016.7738691>
- [21] M.E.A. Davies, Advancing neuromorphic computing with Loihi 2, *IEEE Micro* 41 (2021) 42–53.
- [22] W. Fang, Z. Yu, Y. Chen, T. Huang, T. Masquelier, Y. Tian, Deep residual learning in spiking neural networks, (2021).
- [23] Y. Hu, H. Tang, G. Pan, Spiking deep residual networks, *IEEE Trans. Neural Netw. Learn. Syst.* 34 (8) (2021) 5200–5205, <https://doi.org/10.1109/TNNLS.2021.3119238>
- [24] B. Yin, F. Corradi, S.M. Bohté, Effective and Efficient Computation with Multiple-Timescale Spiking Recurrent Neural Networks, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–8. <https://doi.org/10.1145/3407197.3407225>
- [25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale 2020. *ArXiv:2010.11929*.
- [26] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, D. Tao, A survey on vision transformer, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (1) (2022) 87–110, <https://doi.org/10.1109/TPAMI.2022.3152247>
- [27] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: transformer for semantic segmentation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.
- [28] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: a video vision transformer, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.
- [29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, *ArXiv:2103.14030*, 2021.
- [30] Z. Sun, S. Cao, Y. Yang, K.M. Kitani, Rethinking transformer-based set prediction for object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3611–3620.
- [31] J. Yang, L. An, A. Dixit, J. Koo, S.I. Park, Depth estimation with simplified transformer, *ArXiv:2204.13791*, 2022.
- [32] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, S. Mattoccia, Monovit: Self-Supervised Monocular Depth Estimation with a Vision Transformer, 2022, pp. 668–678. <https://doi.org/10.1109/3DV57658.2022.00077> ISSN: 2475-7888.
- [33] J. Zhang, L. Tang, Z. Yu, J. Lu, T. Huang, Spike transformer: monocular depth estimation for spiking camera, in: *European Conference on Computer Vision*, Springer, 2022, pp. 34–52.
- [34] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. Yan, Y. Tian, L. Yuan, Spikformer: When spiking neural network meets transformer, *ArXiv:2209.15425*, 2022.
- [35] C. Zhou, L. Yu, Z. Zhou, Z. Ma, H. Zhang, H. Zhou, Y. Tian, Spikingformer: spike-driven residual learning for transformer-based spiking neural network, *ArXiv:2304.11954*, 2023.
- [36] X. He, D. Zhao, Y. Li, G. Shen, Q. Kong, Y. Zeng, Improving the performance of spiking neural networks on event-based datasets with knowledge transfer 2023. *ArXiv:2303.13077*.
- [37] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.Y. Huang, S.W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synaev, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, Dinov2: learning robust visual features without supervision, *ArXiv:2304.07193*, 2023.
- [38] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer, *ArXiv:1907.01341*, 2020.
- [39] R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, *ArXiv:2103.13073*, 2021.
- [40] F. Khan, S. Salahuddin, H. Javidnia, Deep Learning-Based Monocular Depth Estimation Methods—A State-of-the-Art Review, *Multidisciplinary Digital Publishing Institute*, 2020, <https://doi.org/10.3390/s20082272>
- [41] Q. Li, J. Zhu, J. Liu, R. Cao, Q. Li, S. Jia, G. Qiu, Deep learning based monocular depth prediction: datasets, methods and applications, *ArXiv:2011.04123*, 2020.
- [42] I. Laina, C. Ruppel, V. Belagiannis, F. Tombari, N. Navab, Deeper depth prediction with fully convolutional residual networks, *ArXiv:1606.00373*, 2016.
- [43] O. Ronneberger, P. Fischer, T. Brox, U-NET: convolutional networks for biomedical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.
- [44] Y. Nam, M. Mostafavi, K.J. Yoon, J. Choi, Stereo Depth from Events Cameras: Concentrate and Focus on the Future, *IEEE, New Orleans, LA, USA*, 2022, pp. 6104–6113. <https://doi.org/10.1109/CVPR52688.2022.00602>
- [45] X. Liu, J. Li, X. Fan, Y. Tian, Event-based monocular dense depth estimation with recurrent transformers, *ArXiv:2212.02791*, 2022b.
- [46] A.L. Hodgkin, A.F. Huxley, A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve, *Wiley-Blackwell*, 1952.
- [47] E.M. Izhikevich, Simple Model of Spiking Neurons, *IEEE*, 2003.
- [48] H.Y. Hsieh, K.T. Tang, VLSI implementation of a bio-inspired olfactory spiking neural network, *IEEE Trans. Neural Netw. Learn. Syst.* 23 (2012) 1065–1073.
- [49] H. Zheng, Y. Wu, L. Deng, Y. Hu, G. Li, Going deeper with directly-trained larger spiking neural networks, *ArXiv:2011.05280*, 2020.
- [50] M. Yao, J. Hu, Z. Zhou, L. Yuan, Y. Tian, B. Xu, G. Li, Spike-driven transformer, *ArXiv:2307.01694*, 2023.
- [51] J. Gou, B. Yu, S.J. Maybank, D. Tao, Knowledge distillation: a survey 2021. *ArXiv:2006.05525*.
- [52] R.K. Kushawaha, S. Kumar, B. Banerjee, R. Velmurugan, Distilling spikes: knowledge distillation in spiking neural networks, *ArXiv:2005.00288*, 2020.
- [53] H. Qiu, M. Ning, Z. Song, W. Fang, Y. Chen, T. Sun, Z. Ma, L. Yuan, Y. Tian, Self-architectural knowledge distillation for spiking neural networks, *Neural Netw.* 178 (2024) 106475.
- [54] J. Liu, Q. Zhang, J. Li, M. Lu, T. Huang, S. Zhang, Unsupervised spike depth estimation via cross-modality cross-domain knowledge transfer, *ArXiv:2208.12527*, 2022a.
- [55] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M.S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, 2022, *ArXiv:2108.07258*.
- [56] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, Segformer: simple and efficient design for semantic segmentation with transformers, *Adv. Neural Inf. Process. Syst.* 34 (2021) 12077–12090.
- [57] D.H.P.K. Nathan Silberman, R. Fergus, Indoor segmentation and support inference from RGB-D images, in: *ECCV*, 2012.
- [58] S. Song, S.P. Lichtenberg, J. Xiao, Sun rgb-d: a rgb-d scene understanding benchmark suite, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567–576.
- [59] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part II 14*, Springer, 2016, pp. 694–711.
- [60] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, *ArXiv:1411.4734*, 2015.
- [61] J. Hidalgo-Carri6, D. Gehrig, D. Scaramuzza, Learning monocular dense depth from events 2020. *ArXiv:2010.08350*.
- [62] M. Gehrig, W. Aarents, D. Gehrig, D. Scaramuzza, Dsec: a stereo event camera dataset for driving scenarios, *IEEE Robot. Autom. Lett.* 6 (3) (2021) 4947–4954, <https://doi.org/10.1109/LRA.2021.3068942>
- [63] M. Horowitz, Computing 2019s energy problem (and what we can do about it), in: *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, IEEE, 2014, pp. 10–14.
- [64] B. Han, K. Roy, Deep spiking neural network: energy efficiency through time based coding, in: *European Conference on Computer Vision*, 2020, pp. 388–404.
- [65] E. Lemaire, L. Cordone, A. Castagnetti, P.E. Novac, J. Courtois, B. Miramond, An analytical estimation of spiking neural networks energy efficiency, (2022).
- [66] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, C. Feichtenhofer, Multiscale vision transformers, *ArXiv:2104.11227*, 2021.
- [67] S.B. Furber, F. Galluppi, S. Temple, L.A. Plana, The Spinnaker project, *Proc. IEEE* 102 (5) (2014) 652–665.
- [68] C. Mayr, S. Hoepfner, S. Furber, Spinnaker 2: a 10 million core processor system for brain simulation and machine learning—keynote presentation, in: *Communicating Process Architectures 2017 & 2018*, IOS Press, 2019, pp. 277–280.

- [69] C. Pehle, S. Billaudelle, B. Cramer, J. Kaiser, K. Schreiber, Y. Stradmann, J. Weis, A. Leibfried, E. Müller, J. Schemmel, The BrainScaleS-2 accelerated neuromorphic system with hybrid plasticity, *Front. Neurosci.* 16 (2022) 795876.
- [70] M.P. Lohr, C. Jarvers, H. Neumann, Complex neuron dynamics on the IBM TrueNorth neurosynaptic system, (2020).

Author biography



Xin Zhang is an Associate Researcher at Manchester Metropolitan University (MMU). He received the B.S. degree from the PLA Academy of Communication and Commanding, China, in 2009, and the Ph.D. degree in Cartography and Geographic Information System from Beijing Normal University (BNU), China, in 2014. His current research interests include remote sensing image processing and deep learning.



Liangxiu Han received the Ph.D. degree in computer science from Fudan University, Shanghai, China, in 2002. She is currently a Professor of Computer Science with the School of Computing, Mathematics, and Digital Technology at Manchester Metropolitan University. Her research areas lie in the development of novel big data analytics and intelligent architectures (e.g., parallel and distributed computing, cloud/service-oriented computing) and their application to diverse domains using large datasets from biomedical images, environmental sensors, and network traffic. She is a Principal Investigator or Co-Investigator on numerous research projects in these fields.



Sergio Davies is a Senior Lecturer in the Department of Computing & Mathematics at Manchester Metropolitan University. He received his Ph.D. in computer science from the University of Manchester in 2012, working on the SpiNNaker project. Dr. Davies continued his research on neuromorphic computing as a postdoctoral researcher within the Human Brain Project (HBP) until 2016. After leading research projects in industry, he returned to academia in 2019. His research is centered on spiking neural networks, with a current focus on their application to computer network security. He is a member of the IET and IEEE, and a Fellow of the HEA.



Tam Sobeih is a Research Associate with the School of Computing, Mathematics, and Digital Technology at Manchester Metropolitan University. Leveraging his prior industry experience, his interests lie in providing real-world solutions to complex problems through the development of novel intelligent architectures, big data analytics, and artificial intelligence applications.



Lianghao Han received his Ph.D. in Engineering from the University of Cambridge, UK, in 2005. He was a Professor of Biomedical Engineering in the Medical School at Tongji University, P.R. China, and is currently a Senior Research Fellow at the Department of Computer Science, Brunel University London. His research focuses on Medical Image Analysis, Machine Learning, and Deep Learning for Healthcare.



Darren Dancey is the Head of the Department for Computing and Mathematics at Manchester Metropolitan University. He holds a Ph.D. in Artificial Neural Networks. His recent work has concentrated on creating collaborations and knowledge exchange between universities and industry, particularly with the SME sector. He has led several large projects funded by Innovate UK, the Digital R&D Fund for the Arts, and the European Research Council. He is also on the organising committee for the Manchester Raspberry Pi Jam and sits on the BCS Manchester branch committee.