



A lightweight 3D-CNN for event-based human action recognition with privacy-preserving potential

Title	A lightweight 3D-CNN for event-based human action recognition with privacy-preserving potential
Author(s)	Sefidgar Dilmaghani, Mehdi;Fowley, Frank;Corcoran, Peter
Publication Date	2026-02-02
Publisher	Institute of Electrical and Electronics Engineers
Repository DOI	https://doi.org/10.1109/ACCESS.2026.3660117

RESEARCH ARTICLE

A Lightweight 3D-CNN for Event-Based Human Action Recognition With Privacy-Preserving Potential

MEHDI SEFIDGAR DILMAGHANI¹, FRANK FOWLEY, AND PETER CORCORAN¹, (Fellow, IEEE)

Department of Electronic Engineering, College of Science and Engineering, University of Galway, Galway, H91 TK33 Ireland

Corresponding author: Mehdi Sefidgar Dilmaghani (mehdi.sefidgardilmaghani@universityofgalway.ie)

This work was supported in part by Research Ireland under the PRIVISENSE project (Grant No. IRCLA/2023/1992).

ABSTRACT This paper presents a lightweight three-dimensional convolutional neural network (3D-CNN) for human activity recognition (HAR) using event-based vision data. Privacy preservation is a key challenge in human monitoring systems, as conventional frame-based cameras capture identifiable personal information. In contrast, event cameras record only changes in pixel intensity, providing an inherently privacy-preserving sensing modality. The proposed network effectively models both spatial and temporal dynamics while maintaining a compact and computationally efficient design. To address class imbalance and enhance generalization, focal loss with class reweighting and targeted data augmentation strategies are employed. The model is trained and evaluated on a composite dataset derived from the Toyota Smart Home and ETRI datasets. Experimental results demonstrate an F1-score of 0.9415 and an overall accuracy of 94.17%, outperforming benchmark 3D-CNN architectures such as C3D, ResNet3D, and MC3_18 by up to 3%. It also has approximately 10 times fewer parameters than the second best benchmark network. These results highlight the potential of event-based deep learning for developing accurate, efficient, and privacy-aware human action recognition systems suitable for real-world applications.

INDEX TERMS Human action recognition (HAR), event cameras, 3D convolutional neural network (3D-CNN), privacy preservation, spatiotemporal feature extraction, lightweight deep learning.

I. INTRODUCTION

Human Action Recognition (HAR), using cameras and other sensors such as radar, plays a critical role in the care of elderly, sick, and disabled individuals by reducing the need for constant monitoring by human caregivers. HAR also has a wide range of other applications, including intelligent surveillance, human-computer interaction, and smart environments [1]. Recognizing an action requires not only understanding what is happening in each frame at a given moment but also how it changes over time to make an accurate prediction. As an example, pouring hot water into a cup could be part of making tea, instant coffee, or another drink, so the action needs to be observed over time to determine the exact outcome. Other challenges also exist in making

HAR practical for real-world applications. These include limitations in the sensors, and in the algorithms [2].

Regarding sensors, most current HAR algorithms rely on frame-based cameras, which must continuously capture frames to monitor human activities. These systems not only need to store the frames but also process them using preprocessing and deep learning algorithms, requiring substantial memory and computational resources. A more critical limitation is that such cameras record personal information in private spaces, raising privacy concerns and often conflicting with regulations and user consent [3].

On the algorithm side, significant challenges also remain. Networks that achieve high accuracy have complex architectures and require considerable computational and memory resources. Deploying such algorithms is even more challenging, and their design must be highly efficient and optimized. Consequently, there is a growing need for lightweight, high

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar¹.

accuracy, and privacy-preserving HAR solutions that can operate efficiently without compromising recognition performance [4]. This need is particularly critical for indoor elderly-care applications, where HAR systems must run on resource-constrained devices deployed in home environments, while ensuring privacy and continuous operation.

Recent advances in deep learning have shown great potential in addressing these challenges. Deep neural networks have been developed with a strong ability to learn both spatial and temporal patterns from large datasets [5]. New algorithms have also improved the accuracy of HAR systems, making it possible to build models that can understand, predict, and respond to human behavior in real time. In addition, significant research has focused on designing compact neural networks suitable for deployment on edge devices [6].

To address privacy concerns, event cameras—also known as neuromorphic or dynamic vision sensors—offer a promising alternative to conventional frame-based cameras. In contrast to standard cameras, which capture full intensity frames at fixed intervals, event cameras record only changes in pixel intensity, producing a stream of events that represent motion. Because they do not capture detailed texture or color information, event cameras inherently preserve privacy, making them particularly suitable for sensitive HAR applications [5].

Despite the potential of novel AI algorithms and event cameras, research on combining these two for HAR applications remains relatively limited. Most existing HAR methods have been developed for frame-based video data using convolutional neural networks (CNNs). While these models have demonstrated strong performance, they face limitations beyond those mentioned above; in particular, CNNs are not well suited to capturing temporal information [7].

The few existing approaches that leverage event data often rely on complex architectures, large model sizes, or extensive preprocessing pipelines to convert asynchronous events into other representations. These requirements reduce the advantages of event cameras and limit their applicability in real-time or embedded systems. Consequently, there is a growing need for simpler, yet effective, deep learning models specifically designed for event-based data [8].

To address these challenges, we have designed and trained a three-dimensional convolutional neural network (3D-CNN) to analyze the output of event cameras. 3D-CNNs have the same capability as conventional CNNs in capturing spatial information, while also extracting temporal features from consecutive frames. Using event cameras helps reduce privacy risks in HAR. It does not guarantee complete privacy protection; however, it adds complexity to accessing private data. Aggregating event streams into 2D matrices for use in the 3D-CNN, once embedded in the main network, further reduces privacy risks compared to conventional cameras that provide direct access to frames containing personal data, thereby offering an additional layer of protection.

In this work, we specifically target indoor human activity recognition for elderly people, with an emphasis on

lightweight models suitable for deployment in home environments. The goal of this paper is to propose an algorithm that analyzes event camera data which is inherently privacy-preserving with a model that is accurate, and compact. To meet these requirements, our contributions are [9]:

- Propose a lightweight algorithm for indoor HAR tailored to elderly people,
- Proposing a 3D-CNN specifically designed for HAR that captures both spatial and temporal information,
- Training the algorithm using event data which supports privacy preservation inherently,
- Achieving efficient and robust recognition with a lightweight network without requiring large computational resources.

The rest of the paper is organized as follows. Section II reviews related work on HAR. Section III presents the proposed methodology in detail. Section IV describes the experimental results and discusses the findings. Finally, Section V concludes the paper and outlines directions for future research.

II. LITERATURE REVIEW

Human action recognition has been an important problem in computer vision, driven by applications in surveillance, healthcare, and human-computer interaction. Numerous approaches have been proposed over the past decades to address this challenge. However, several issues remain unresolved and continue to be active areas of research, particularly privacy concerns [10] and the difficulty of accurately recognizing visually similar activities [11]. The following section provides a review of related work in this domain.

A. TRADITIONAL IMAGE AND VIDEO PROCESSING METHODS FOR HAR

Prior to the vast developments in deep learning based methods, HAR was based on hand-crafted spatio-temporal features, bag-of-features, and trajectory representations. Space-time interest points (STIPs) detect localized spatio-temporal saliency and were foundational for early video descriptors [12]. Dollár et al. proposed sparse spatio-temporal cuboid features and demonstrated robust behavior/gesture recognition with these local descriptors [13]. Dense trajectory methods and their improvement via camera-motion compensation became a dominant representation for action recognition because they encode long-term motion patterns robustly [14]. These methods report high performance on curated benchmarks of their era (e.g., improved dense trajectories reach competitive accuracies on UCF-101 / HMDB benchmarks) but remain sensitive to viewpoint, illumination and require manual feature design. The strengths and limitations of these approaches are summarized in the comprehensive survey proposed in [15].

B. DEEP LEARNING WITH FRAME-BASED CAMERAS FOR HAR

Deep learning moved HAR away from engineered features to learned spatio-temporal representations. It introduced architectures such as Two-Stream CNNs [16], large-scale video CNNs [17], Temporal Segment Networks (TSN) [18], and Inflated 3D ConvNets (I3D) [19], which demonstrated strong performance but at significant computational cost.

The two-stream fusion of an RGB (spatial) net and an optical-flow (temporal) net achieved strong gains on standard benchmarks. The fused two-stream model reported 88.0% mean accuracy on UCF-101 and 59.4% on HMDB-51 when using SVM fusion; temporal stream alone was 83.7% on UCF-101 in the original experiments. These results established that explicit modeling of motion via optical flow substantially boosts accuracy [16]. TSN introduced sparse but global temporal sampling and aggregation. The paper reports 94.9% accuracy on UCF-101 and 71.0% on HMDB-51. TSN also offers efficient variants that trade off accuracy and speed [18]. Work on large-scale datasets and architectures established design patterns and the need for scale [17]. Inflated 3D ConvNets (I3D) extended 2D filters to 3D and, trained on the Kinetics dataset, set new state-of-the-art baselines for learned spatio-temporal modeling [19].

C3D, a 3D ConvNet trained end-to-end on video volumes, reported 85.2% on UCF-101 which was one of the best prior 3D convolutional results at the time and strong gains on Sports-1M. These results demonstrated the value of spatio-temporal convolutions [20]. R(2+1)D is an architecture that factorizes 3D convs into separate spatial and temporal convolutions. Its variants reached state-of-the-art levels on large video datasets. Reported video level accuracies include 73.3% on Sports-1M and strong Kinetics numbers in the original study [21]. SlowFast introduces dual pathways (Slow and Fast). Depending on instantiation, SlowFast variants achieve 75–80% accuracy on Kinetics-400. The work also reports FLOPs, inference-view tradeoffs, and shows consistent gains over single-path baselines [22]. Temporal Shift Module (TSM) is a lightweight temporal module enabling 2D backbones to capture temporal information with low overhead. Reported Kinetics performance for TSM-ResNet50 is 74.1% accuracy. A mobile TSM variant with MobileNet-V2 achieved 69.5% accuracy on Kinetics, demonstrating feasibility for edge devices [23]. Direct numeric comparisons must be read carefully. That is because evaluation protocols including number of clips, multi-crop testing, pretraining, and dataset splits differ for different projects. Transformer token-based and well-tuned 3D convs often lead on Kinetics scale benchmarks, while two-stream and TSN remain strong and more efficient on some tasks [16], [18].

Although the mentioned architectures achieve high accuracy in certain cases, they do not address privacy concerns. Since all of them require frames containing personal information for training and inference, they are not suitable for

real-world applications, especially under regulations such as the GDPR.

C. PRIVACY-PRESERVING HAR

The aforementioned HAR methods require frames containing personal information for correct functioning, which raises privacy concerns. Alternative solutions include RF-based systems such as WiSee [24], TARF [25], WiFall [26], and the radar-based TinyRadarNN [27], which achieve competitive accuracy without visual recording.

WiSee [24] was an early system that used WiFi Doppler signals for whole-home gesture recognition, highlighting the through-wall capability and privacy benefits of RF sensing. The paper reports an average accuracy of 94% for nine gestures in both in-room and through-wall tests, with reduced performance when multiple people are present. Commercial mmWave (60 GHz / 77 GHz) radars and recent research prototypes demonstrate real-time action and gesture recognition using micro-Doppler features and lightweight classifiers, making radar a practical and privacy-preserving choice for in-home HAR [27], [28], [29].

TinyRadarNN [27] is a compact, low-power, radar-based gesture classifier with only 46k parameters and a memory footprint of 92 KB. On its 11-gesture dataset, it achieves 86.6% accuracy in multi-user cross-subject tests and 92.4% in single-user experiments, while consuming just 21 mW during inference. This strong performance underscores radar's potential for embedded HAR [27], [30], [31].

Comprehensive reviews of radar, WiFi, and RFID-based HAR approaches, their machine learning pipelines, and challenges such as domain adaptation and multimodal fusion are provided in [25] and [26]. Although RF and radar methods demonstrate strong accuracy on constrained gesture sets and operate with low power consumption, they are often evaluated on different gesture vocabularies and under controlled conditions, complicating cross-method comparisons [32].

D. EVENT-BASED VISION FOR HAR

Event cameras, or neuromorphic sensors, generate a stream of events rather than standard image frames. They record asynchronous brightness changes at the pixel level with microsecond latency and high dynamic range. These properties make them attractive for low-power and privacy-sensitive applications, including HAR [5]. Several systems [33], benchmarks [34], and datasets [35] have been developed for this purpose, demonstrating the potential of event-based sensing for efficient HAR.

Early event-based gesture and activity recognition systems converted event streams into 2D matrix representations for conventional processing [36]. More recent benchmarks and dedicated action-recognition datasets demonstrate growing interest in event-based HAR, introducing evaluation protocols that emphasize temporal ordering rather than static event accumulation [35].

E. 3D CNNs WITH EVENT CAMERAS FOR HAR

Algorithms that use both spatial and temporal information are effective for tasks such as HAR. In this area, spatio-temporal deep learning with event-based data has become an important approach. These methods often use 3D-CNNs [20], voxel-based models [37], asynchronous networks [38], and learned event representations [39]. By combining spatial features with changes over time, these techniques can better understand motion and activities in complex environments.

A common approach for applying deep video models to event-based data is to convert the asynchronous event stream into dense spatio-temporal representations, such as voxel grids, event spike tensors, or time surfaces. These representations are then processed using 2D/3D or 3D CNN backbones, such as C3D and I3D. While C3D and related 3D CNN architectures effectively capture spatio-temporal features, they are computationally expensive and memory-intensive [19], [20].

Gehrig et al. formalized a set of grid-based event representations, including event frames, voxel grids, and event spike tensors, which serve as inputs to conventional CNNs. Their work highlighted the trade-off between preserving temporal resolution and managing computational cost [39]. Recent event-specific architectures explore voxel, graph, and transformer-based representations, e.g., voxel-graph CNNs, and volumetric transformers, to exploit event sparsity. However, models employing full 3D convolutional stacks still have high parameter counts and computational demands, motivating the development of sparsity-aware, quantized, or lightweight alternatives [15], [37], [39].

F. LIGHTWEIGHT DEEP LEARNING MODELS FOR VISION TASKS

Efforts to reduce computational cost have focused on developing compact, high accuracy models for resource constrained platforms, leading to lightweight HAR architectures suitable for edge deployment. Representative model families include MobileNet, which employs depthwise separable convolutions [40]; MobileNetV2, which introduces inverted residuals and linear bottlenecks [41]; ShuffleNet, which utilizes channel shuffling and group convolutions [42]; and EfficientNet, which applies compound scaling based on neural architecture search [43]. These architectures achieve favorable accuracy–latency trade offs, making them well suited for efficient on-device inference.

These models have been adapted for tasks such as detection and segmentation, and extended to video through 2D backbones with temporal fusion or efficient 3D temporal modules. When integrated with sparse event encodings or lightweight temporal components, such architectures form strong foundations for efficient HAR systems, including event based pipelines. Mobile oriented backbones when combined with temporal modules such as TSM, provide favorable accuracy efficiency trade offs. For example, an online TSM variant using a MobileNetV2 backbone achieved approximately 69.5% accuracy on the Kinetics dataset while remaining

practical for edge deployment. These results highlight the potential of mobile backbones for event or radar based low power HAR applications [44].

G. SUMMARY OF HAR MODELS

Table 1 summarizes the key HAR models reviewed. It lists the model type, the data modality used, and the reported performance metrics. The table illustrates the progression from traditional handcrafted features to deep learning approaches, including 3D CNNs, two-stream networks, transformer-based models, and lightweight architectures suitable for resource-constrained environments. It also highlights computational aspects, such as model size and efficiency, for each approach.

TABLE 1. Reported performance of representative HAR models.

Model	Dataset	Accuracy / Top-1	Citation
Two-Stream	UCF-101	88.0%	[16]
	HMDB-51	59.4%	
C3D	UCF-101	85.2%	[20]
R(2+1)D	Sports-1M	73.3%	[21]
TSN	UCF-101	94.9%	[18]
	HMDB-51	71.0%	
I3D	Kinetics-400	72%	[19]
SlowFast	Kinetics-400	79.8%	[22]
TSM-Res50	Kinetics-400	74.1%	[23]
TSM-MobileNet	Kinetics-400	69.5%	[23]
WiSee	9 Gestures	94%	[24]
TinyRadarNN	11 Gestures	86–92%	[31]
Swin-T	DailyDVS-200	48.1%	[30]
Spikformer	DailyDVS-200	36.9%	[30]
EVMamba	DailyDVS-200	+8.8% vs. TSM	[32]

H. SUMMARY OF GAPS IN LITERATURE

Despite the considerable achievements in the field of HAR, especially following recent developments in AI algorithms, the review presented in this section indicates that several areas still require further attention and remain far from the ideal state. The following fields, in particular, need more focus:

Privacy–accuracy trade-offs: Radar and WiFi based methods protect privacy better than camera systems but still are less accurate in recognizing detailed actions and functioning across different scenes. Combining multiple sensing methods is still at an early stage [25], [28]. These approaches can reach high accuracy for small, fixed gesture sets (e.g., WiSee: 94%; TinyRadarNN: 86–92%), but fair comparison with camera based methods is difficult because they use different datasets and problem setups. The main challenge is to match the detail of camera based recognition while keeping user privacy [2].

Efficient spatio-temporal modeling for events: Voxelization and 3D CNNs preserve temporal details but result in large models with high computational cost. Sparsity-aware, asynchronous, and spiking approaches show promise but still require thorough comparison with compact ANN-based alternatives [37], [39].

Compute vs. temporal fidelity: 3D CNNs reach high accuracy on large video datasets such as Kinetics, however require heavy computational and memory resources. Lighter models such as TSM or MobileNet cut these costs with reasonable drops in accuracy. This trade-off is still not well studied for event based and RF data [45]. While efficient image networks are widely used, their utilization for temporal modeling in HAR is limited. Adapting architectures such as MobileNet for event or RF inputs is challenging and remains an open area for research [40].

Event-based dataset scale and model gap: Most event HAR datasets are small or can be analyzed using simple, non-temporal methods. There is a clear shortage of large, realistic datasets that require temporal modeling and generalization across subjects and scenes [46], [47]. Recent results on the DailyDVS-200 dataset show that transformer and token-based models currently perform best on large scale event HAR, but top-1 accuracy remains below 50% for cross-subject evaluation. While models such as EVMamba show promising improvements, they still lag behind dense frame based methods on simpler tasks. These results highlight the requirement for better event specific architectures and larger cross domain benchmarks [48].

III. METHODOLOGY

This section explains the details of the algorithm that's proposed to recognize the human activities. The algorithm is a 3D-CNN that extracts the information from the raw stream of events generated by a neuromorphic sensor. The reason behind using the 3D-CNN rather than other algorithms is its strength in analyzing both spatial and temporal information, simultaneously.

The designed algorithm is a lightweight network which leads to faster training process and easier implementation. Beyond the advantages of the proposed network, output data of the event cameras offer inherent privacy-preserving potential. Although in the proposed method, the raw events still need to be collected into 2D matrices, which resemble frames, accessing personal information remains more complex than with frame-based cameras. The proposed pipeline consists of preprocessing input data, feeding them into a lightweight yet expressive 3D-CNN model, and training the network. A self-attention is incorporated besides the 3D convolutional layers to enhance feature representation of the network if required. Other strategies in the algorithm design include utilization of focal loss and class reweighting. The details are explained in this section.

A. EVENT DATA REPRESENTATION

The main dataset used to train the network consists of RGB videos of varying lengths. These videos are first converted into raw event data, which are then accumulated into 2D matrices at a rate of 30 fps. Consequently, the total number of 2D matrices—hereafter referred to as event frames—varies across videos. However, in 3D-CNNs for proper processing,

the number of frames must be consistent across batches of videos. To solve this problem, a consistent number of event frames per video are downsampled uniformly, and the intermediate frames are discarded. This consistent rate is set fixed at 10 frames per video as this rate achieved a good balance between accuracy and GPU load. Moreover, as will be shown later, lower rates resulted in a considerable drop in accuracy, whereas higher rates achieved no improvement. Detailed information about the dataset preparation is as follows.

1) TRAINING DATASET

The reason for using an RGB dataset rather than an event-based one is the limited availability of suitable event-based datasets for human activity recognition (HAR) applications. Available datasets such as DailyDVS-200 [49] and HARDVS [50] are important contributions; however, neither the subjects in these datasets are elderly people nor are the environments and activities limited to home settings. Even among non-event datasets, most are constrained by data privacy concerns, minimal scene diversity, the requirement for expensive sensor equipment, and an insufficient number of participants [51].

Natural realism is a major challenge in building effective HAR datasets. Scripted scenes performed by actors often lack spontaneity and natural occlusions. As noted by Htun et al. [51], datasets designed mainly for object detection but without sufficient human-object interactions are “context-biased,” which reduces the performance of HAR models. In addition, background environments in home-like settings are difficult to modify, leading to limited variation in object placement, such as furniture or appliances. Virtual-reality simulations can help address this issue, but realism remains limited. Finally, most reviewed datasets show little correspondence between activity classes and the instruments used to perform them.

The Toyota Smart Home (TSH) trimmed dataset and the ETRI dataset were selected as the most suitable sources for compiling the training data for the 3D-CNN model. Both datasets contain a large number of video samples, cover a wide range of annotated activities, and include challenges such as occlusions and lighting variations. Common subclasses were identified and extracted from both datasets, and subsequently merged into six classes for training, as summarized in Table 2. To create a balanced training dataset, ETRI videos were segmented into shorter clips: videos of 5–10 seconds were split in half, those of 10–15 seconds into thirds, and those longer than 20 seconds into four equal segments. All videos were resized to a frame resolution of 640×480 pixels. To address differences in dataset size, 720 video samples were selected from each ETRI class, and 280 samples were randomly drawn from each TSH class, resulting in a balanced dataset of 1,000 samples per HAR class.

TABLE 2. Balanced training dataset compiled from ETRI and TSH sources.

Class	Samples from TSH	Samples from ETRI	Total
Cooking	280	720	1,000
Drinking	280	720	1,000
Eating	280	720	1,000
Getting Up	280	720	1,000
Sitting Down	280	720	1,000
Washing Up	280	720	1,000

2) DATASET SPLITS

The dataset was partitioned into training, validation, and test sets using a 70:20:10 ratio. Splits were created using PyTorch's `random_split` function with a fixed random seed to ensure reproducibility and roughly balanced class distributions across the sets. The training DataLoader shuffled the data to improve model generalization, whereas validation and test DataLoaders did not shuffle to ensure consistent evaluation.

3) DATA AUGMENTATION

The data augmentation strategy is targeted and designed to improve generalization while addressing class imbalance. While standard resizing, padding, and normalization is used for all of the classes, for underrepresented classes which include "Eating" and "Washing Up", random horizontal flip, rotation, affine transform, and Gaussian blur is used.

4) DOMAIN GAP BETWEEN SIMULATED AND REAL EVENT DATA

While converting RGB videos into event representation, enables the use of larger datasets, a domain gap exists compared to real neuromorphic cameras, which capture events asynchronously with microsecond latency and high dynamic range. Consequently, the network might require fine-tuning when deployed on actual event cameras to account for differences in temporal precision, event sparsity, and sensor noise characteristics.

B. NETWORK ARCHITECTURE

The proposed lightweight 3D-CNN uses a series of 3D convolutional layers each being followed by a batch normalization and pooling layer. An average pooling and a fully connected layer for classification are the final layers.

1) 3D CONVOLUTION LAYERS

The network includes five sequential 3D convolutional blocks with increasing channel sizes: 1, 16, 32, 64, 128, and 256. Each convolution is followed by BatchNorm3d, ReLU activation, and MaxPool3d to downsample spatial dimensions while preserving the temporal dimension (`kernel_size=(1,2,2)`). This architecture allows the model to learn both spatial and temporal features concurrently.

2) CLASSIFICATION HEAD

The final classifier consists of a dropout layer for regularization, a global average pooling for dimensionality reduction, and a fully connected layer that maps the pooled features to class probabilities. This minimal design ensures a balance between efficiency and performance.

3) FEATURE EXTRACTION

The 3D convolution layers operate over the frame sequence to extract spatial-temporal features. The final average pooling layer reduces the feature map size to provide a compact representation of the input video for the classification head.

4) LOSS FUNCTION AND OPTIMIZATION

The model is trained using Focal Loss, which addresses class imbalance and emphasizes on the examples which are more difficult to classify. The focal loss is defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (1)$$

where p_t is the predicted probability of the true class, α_t is a weighting factor inversely proportional to class frequencies to handle class imbalance, and $\gamma = 2.0$ is a focusing parameter that down weights easy examples. This combination allows the network to focus on videos which are harder to classify.

The AdamW optimizer is used for training, chosen for its adaptive learning rate and decoupled weight decay, which improves generalization. Early stopping is applied to prevent overfitting during training.

5) MODEL EFFICIENCY AND DESIGN CHOICES

The main concern during the development process was to keep the model lightweight with limited parameters to ensure faster training and inference, and easier implementation. The final average pooling removes the requirement for large fully connected layers. Self-attention module is optional and can be toggled to balance complexity with performance. The operation can be conducted efficiently on limited GPU resources (e.g., batch size 32 with pinned memory and `num_workers=8`).

C. TRAINING STRATEGY

During the training procedure the AdamW optimizer learning rate is fixed to 0.0009 and `weight_decay` is $1e-4$. Early stopping is designed with a patience of 20 epochs based on F1 score improvements. Best-performing model is saved and later evaluated on a held-out test set. Training and validation losses are logged and saved as plots and will be presented in the results and discussion section.

D. PRIVACY-PRESERVING CHARACTERISTICS

By collecting the event-based data into low-resolution grayscale 2D matrices, the model inherently provides a degree of privacy as the event cameras do not capture identifiable details. This makes the approach more suitable for privacy-sensitive applications such as healthcare or surveillance.

IV. RESULTS AND DISCUSSION

In this section, we first present and analyze the results of training, validation, and testing of the proposed structure on the prepared dataset. Next, we train three widely used 3D-CNN architectures to enable a fair comparison between the performance of our proposed structure and state-of-the-art models.

A. DESCRIPTION OF EXPERIMENT

The model is trained on grayscale sequences of 2D matrices extracted from stream of event data. Each sequence (video) comprises 10 temporally spaced frames, representing a short-duration action segment. A lightweight 3D CNN architecture is employed, incorporating optional self-attention, with focal loss to counteract class imbalance. Early stopping is applied to avoid overfitting. Random shuffling ensures fair train-validation-test splits.

B. DESCRIPTION OF DATASET

All training, validation, and testing are conducted on a custom event dataset comprising short videos of human activities. The dataset contains six classes: cooking, washing dishes, sitting down, getting up, eating, and drinking. These classes are selected in pairs with similar characteristics to ensure a more challenging classification task. Each class includes 1,000 sequences, with each sequence represented as a folder containing image frames. Data augmentation techniques, as described in the methodology section, are applied to increase variability. Each input sequence consists of 10 grayscale frames, resized and padded to a resolution of 128×128 .

C. EVALUATION METRICS

Standard multi-class classification metrics are employed for evaluation. The following metrics are used:

- Focal Loss – Applied during both training and validation. The details of this loss are explained in the methodology section.
- Accuracy – Measures the overall proportion of correctly predicted samples.
- F1-Score – The harmonic mean of precision and recall, computed with weighted averaging to account for potential class imbalance.
- Confusion Matrix – Provides a visualization of misclassifications across classes.
- Training Time – Reported to enable fair comparison between the proposed method and state-of-the-art networks.

D. HARDWARE AND TRAINING ENVIRONMENT

All experiments are conducted on a workstation equipped with an NVIDIA GeForce RTX 3090 Ti GPU with 24 GB of VRAM, an Intel Core i7-7700K CPU @ 4.20 GHz (4 cores, 8 threads), and 64 GB of system memory. The system runs Ubuntu 22.04.5 LTS (kernel 6.8.0-79-generic) with NVIDIA

driver version 570.169. The software environment is based on Python 3.10.12 and PyTorch 2.5.1+cu121, which is compiled with CUDA 12.1 and linked against cuDNN 9.0.1. The batch Size is 32 for all data loaders, epochs are set to 1000 with early stopping with patience of 100 epochs. Learning Rate is 0.0009 and AdamW is used as optimizer.

E. QUANTITATIVE RESULTS

The proposed model achieved a test accuracy of 94.17% and a test F1-score of 0.9415 on unseen data, demonstrating very good generalization. During training and validation, the best performing model—saved at epoch 648—achieved a validation F1-score of 0.9409, indicating that the model performs even better on the test set, which is a strong sign of robustness and generalization rather than overfitting. The total training time to reach the optimal model was 322 minutes. The short training time besides the smallest number of parameters compared to three benchmarked methods, demonstrates its suitability for real-time and resource-constrained applications.

Figure 1 illustrates the training and validation loss curves. The overall trend of the validation loss closely follows that of the training loss and does not exhibit a consistent upward drift, indicating that overfitting did not occur. The intermittent fluctuations and spikes observed in the validation loss are attributed to batch-level statistical variation, which is expected when using small batch sizes in conjunction with batch normalization. Despite these fluctuations, the validation F1-score progresses smoothly and consistently upward, confirming that the training process remained stable.

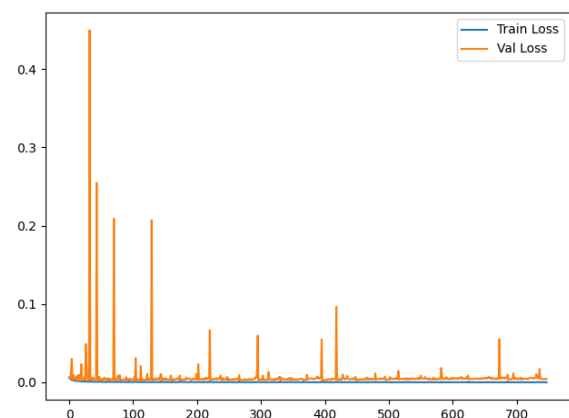


FIGURE 1. Training and validation loss graphs the proposed method.

As shown in Figure 2, the confusion matrix further supports the model's strong performance across all activity classes. Most predictions are concentrated along the diagonal, indicating high true positive rates with only minor misclassifications.

The classes shown in the matrix are indexed as Cooking: 0, Drinking: 1, Eating: 2, Getting up: 3, Sitting down: 4, Washing up the dishes: 5. As shown in the following sections,

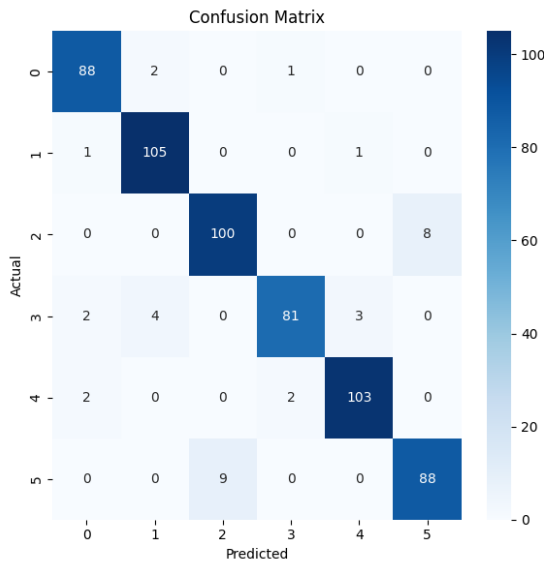


FIGURE 2. Confusion matrix showing classification performance of the proposed method.

our approach achieves higher per-class accuracy, particularly for visually or temporally similar classes, highlighting the model's ability to capture key spatiotemporal features.

Figures 1 and 2 show high overall performance and good balance across class distributions, indicating effective generalization to unseen samples. The lightweight design and reduced parameter count contribute significantly to training efficiency.

F. COMPARISON WITH BASELINE MODELS

In addition to evaluating the proposed model independently, its performance is also compared against three high-performing networks within the same class of 3D-CNNs. These baseline models were originally designed for three-channel RGB video input. Since the event-based input data in this research is structured similarly to single-channel grayscale frames, the input layers of these networks are modified to be compatible with one channel instead of three. Accordingly, the original pretrained weights—tuned for RGB input—could not be reused, due to their reliance on inter-channel correlations. Therefore, all benchmark models were retrained from scratch on the same dataset with identical splits. The training pipeline, augmentation techniques, and employed Focal Loss were the same as those used in the proposed network to ensure a fair comparison.

1) BASELINE MODELS

Several alternative models and architectural paradigms could be considered for comparison with the proposed method. Approaches based on different input representations, such as Spiking Neural Networks (SNNs) [52] designed to directly process raw asynchronous event streams, and represent a

promising direction for event-based HAR. However, such models require fundamentally different training strategies and face distinct optimization challenges. Moreover, the goal of the comparative evaluation in this work is to assess architectural performance under a consistent input representation. Comparing methods that differ both in network architecture and input modality would introduce multiple confounding factors, making it difficult to draw clear conclusions.

Among models that employ input representations similar to ours, other than the 3D-CNNs, strong architectures such as EVMamba [53], a state-space approach for HAR, have been proposed. However, these models typically operate in a significantly larger parameter regime (e.g., approximately 75 million parameters for EVMamba), which contradicts one of the main objectives of this work: designing a compact and computationally efficient architecture. Consequently, we select comparison baselines that share the same input representation and 3D-CNN architectural paradigm. Although some of them operate in a large parameter regime, their inclusion provides a meaningful comparison within the same modeling paradigm, isolating architectural design choices from representation differences. The three 3D-CNN models selected for benchmarking are: C3D, MC3_18, and ResNet3D. The models are briefly introduced below, followed by a quantitative comparison of their performance.

a: C3D

Convolutional 3D or C3D is developed by the researchers at Meta [20]. It is a widely used architecture for spatiotemporal feature learning in videos. It utilizes 3D convolutional and pooling operations to jointly capture spatial and temporal information across consecutive frames of videos. This enables more effective modeling of motion dynamics compared to traditional 2D ConvNets, which process frames independently and lack temporal awareness. C3D employs a homogeneous architecture with $3 \times 3 \times 3$ convolution kernels throughout, which has been empirically shown to be effective for modeling dynamic visual patterns. Trained on large-scale video datasets such as Sports-1M, C3D provides compact, efficient, and transferable features that perform well across various video analysis tasks. Due to its simplicity and strong benchmark performance, it serves as a standard reference model for evaluating new spatiotemporal architectures, including the one proposed in this work.

b: R3D

3D ResNet (R3D) [54], extends the 2D residual learning framework to spatiotemporal data by incorporating 3D convolutions across all layers, enabling direct modeling of motion dynamics in video. In contrast to 2D CNNs that process individual frames, R3D learns spatiotemporal features jointly by convolving over both space and time. This archi-

texture preserves the benefits of residual connections—such as improved optimization and gradient flow—while adapting them for volumetric video input. Although full 3D convolutions increase computational overhead, empirical studies show that R3D consistently outperforms 2D counterparts on large scale video benchmarks, validating the importance of explicitly modeling temporal structure for effective video understanding.

c: MC3_18

MC3-18 [21] is a hybrid architecture that combines elements of both 2D and 3D convolutional networks for video analysis. Based on the ResNet-18 [21] backbone, it integrates 3D convolutions in the early layers to capture short-term temporal dynamics, while employing 2D convolutions in the later layers to reduce computational cost. This selective use of spatiotemporal modeling results in a balanced trade-off between representational power and efficiency. Empirical studies demonstrate that MC3-18 achieves performance close to fully 3D architectures such as R3D, while using significantly fewer parameters and computations. On standard action recognition benchmarks, it outperforms 2D models and provides a competitive yet efficient baseline for spatiotemporal feature learning.

2) QUANTITATIVE COMPARISON

Table 3 presents the evaluation metrics for the baseline models and the proposed structure, as can be seen there our method performs better than all models.

TABLE 3. Performance comparison of different networks.

Network	F1-Score	Accuracy	Val Loss	Parameters (M)	Time (min)
C3D	0.6408	69.17%	0.0104	78.0	74
ResNet3D	0.9131	91.33%	0.4619	33.0	344
MC3_18	0.8622	86.67%	0.3388	11.13	948
Proposed Method	0.9415	94.17%	0.0049	1.5	323

As shown in the table, the proposed method outperforms all other architectures across most metrics, with the exception of training time. In this case, C3D achieves the shortest training duration, followed by the proposed method. However, despite its fast training, C3D has more parameters and exhibits the weakest performance across all other evaluation metrics, indicating that low training time alone does not make it a viable option for this application.

Figure 3 illustrates the confusion matrices for all methods. As shown, the proposed method exhibits the highest values along the main diagonal, indicating superior performance in correctly classifying the target classes compared to the other approaches.

Beyond retraining representative baselines on our dataset, we examined recently proposed state of the art HAR architectures. While these methods report strong performance on large-scale benchmarks, they typically operate with substantially higher parameter counts. In contrast, this work focuses

on compact architectures suitable for resource-constrained deployment, and therefore limits comparisons to models with comparable computational complexity.

Table 4 reports the number of parameters of the proposed method and representative state of the art HAR networks.

TABLE 4. Comparison of parameter counts of the proposed method and the state of the art HAR models.

Method	Network	Parameters (M)
BIKE [55]	VLM	230
Hiera-H [56]	Visual Transformer	672
L-SPM8/14/18 [57]	Visual Transformer	304
Text4Vis [58]	VLM	230
TubeViT-H [59]	Visual Transformer	635
UMT-L [60]	Video Foundation Model	304
VideoMAE [61]	Masked Autoencoder	633
VideoMAE V2 [62]	Masked Autoencoder	999
Proposed Method	ANN	1.5

G. QUALITATIVE RESULTS

Figure 4 shows three misclassified examples. For each video sample, five out of the ten frames are presented. As can be seen, even for a human observer it is not possible to identify the correct activity.

H. ABLATION STUDIES

1) EFFECT OF NETWORK SIZE

A smaller version of the proposed network with half the channel size, leads to a 4% drop in both accuracy and F1-score. Conversely, enlarging the network by doubling the channels size does not improve performance; instead, both accuracy and F1-score decrease by about 1%, while training cost increases. These results indicate that the chosen network size achieves a good balance between capacity and efficiency and is well-suited to the dataset.

2) EFFECT OF FRAME RATE

The event stream was converted into 2D matrices and then sampled at a rate of 10 frames per each video. Reducing the rate to 5 frames per video decreased the accuracy to 89.33% and the F1-score to 0.8939. Although the training time at this frame rate (78 minutes) was considerably shorter than the 323 minutes required at 10 frames, the nearly 5% drop in accuracy made this trade-off undesirable. Increasing the rate to 20 resulted in a 2% reduction in both accuracy and F1-score, while the training time increased to 332 minutes. These results suggest that 10 frames per video is sufficient to capture the necessary temporal context for most actions. The observed drop in performance at higher frame rates per video may appear unexpected; however, it can be attributed to the fixed temporal kernel size of the 3D convolutional layers. While these kernels effectively capture short-term tempo-

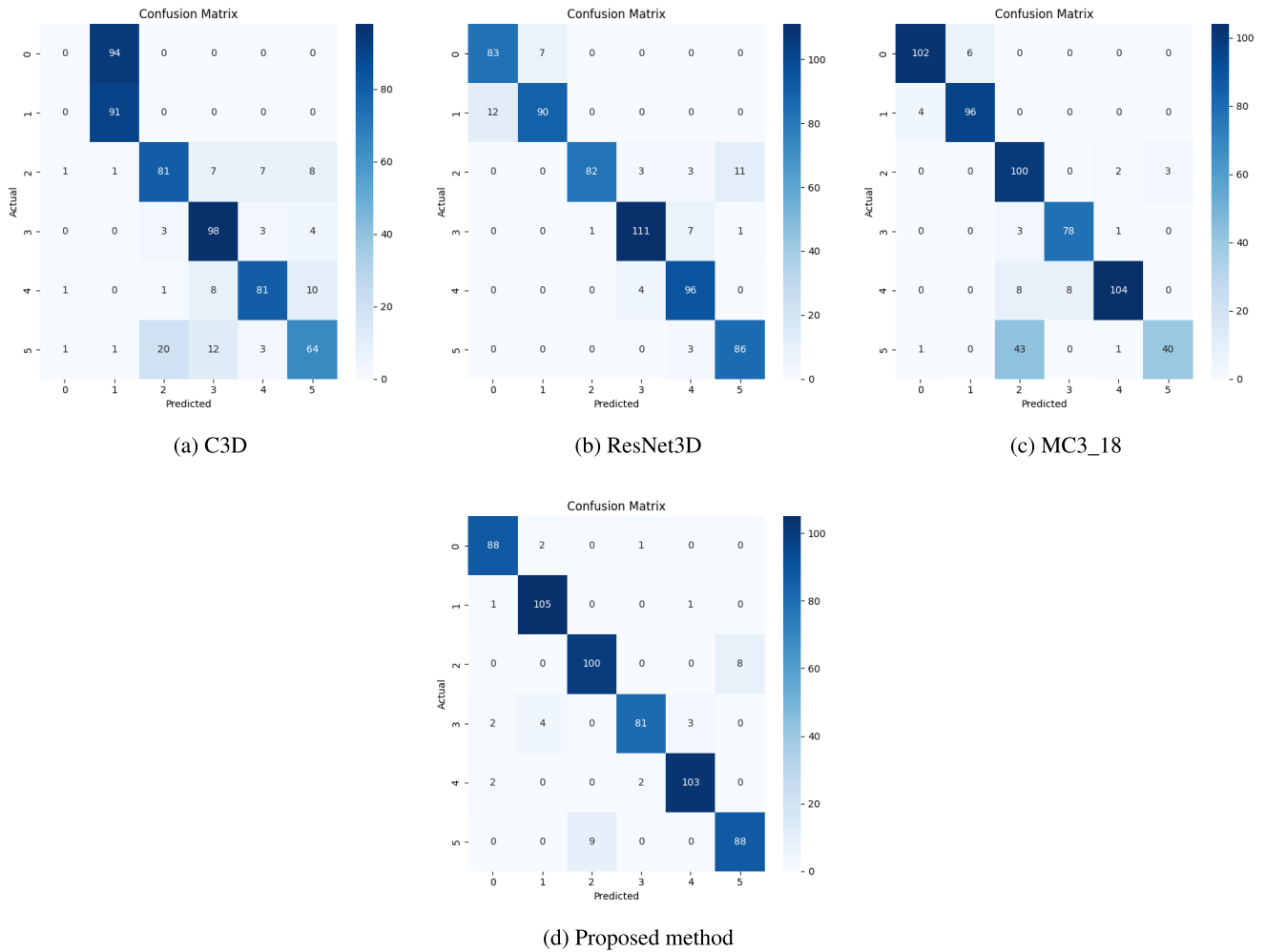


FIGURE 3. Confusion matrices of networks: (a) C3D, (b) ResNet3D, (c) MC3_18, (d) proposed method.

ral dynamics, spreading the same action over more frames reduces the density of informative events per frame relative to the receptive field of the network. Consequently, the network may struggle to extract spatiotemporal features from each frame, leading to slight accuracy degradation. In addition, denser sequences introduce redundant information that can amplify minor inconsistencies between consecutive frames, slightly increasing noise in the learned features. Table 5 summarizes the effect of varying the frame rate and network size on performance and training time. As shown in the table, the best accuracy is achieved by the proposed original architecture.

TABLE 5. Performance comparison of different network configurations.

Network Configuration	F1-Score	Accuracy	Best Val Loss	Training Time
Half Channels Size	0.9032	90.33%	0.0027	118 min
Double Channels Size	0.9332	93.33%	0.0024	231 min
Half Frames per Video	0.8939	89.33%	0.0031	79 min
Double Frames per Video	0.9283	92.83%	0.0046	331 min
Original Network Architecture	0.9415	94.17%	0.0049	323 min

3) DISCUSSION OF STRENGTHS AND LIMITATIONS

a: STRENGTHS

- The proposed compact 3D CNN architecture achieves fast inference while maintaining strong generalization across different datasets.
- Class imbalance is effectively mitigated through the use of focal loss and data augmentation strategies.
- Using event-based input data rather than conventional RGB frames offers potential privacy-preserving advantages.

b: LIMITATIONS

- Performance may degrade under severe class imbalance or when trained on datasets with a significantly larger number of classes.
- Subtle or overlapping motion patterns remain challenging to classify.
- Self-attention yields only marginal gains, and full transformer models were limited by computational resources.

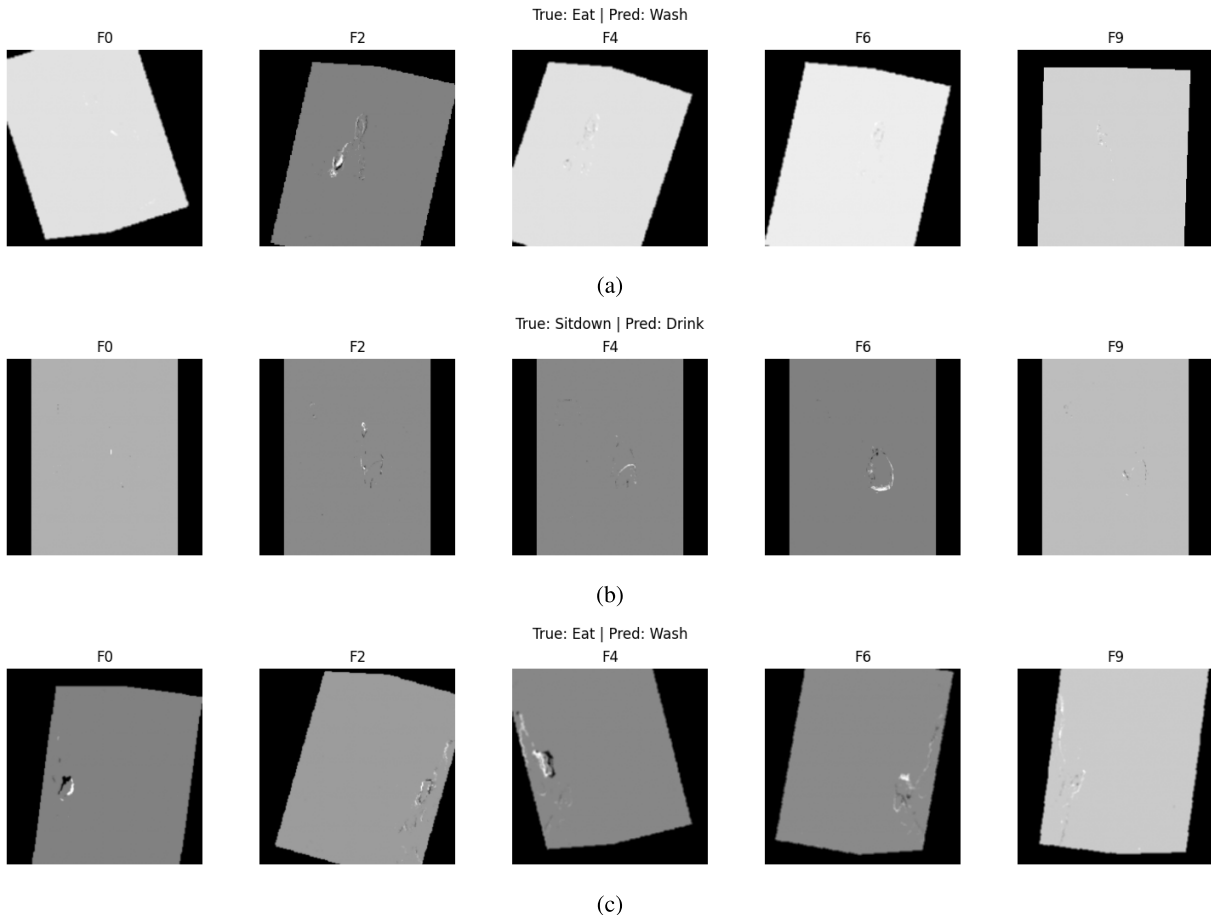


FIGURE 4. Example frames from misclassified video samples classified by the proposed network.

V. CONCLUSION

This work presented a lightweight 3D CNN architecture for human activity recognition using event-based vision data. By leveraging the inherent advantages of event cameras, the proposed method addresses key limitations of traditional frame-based HAR approaches, including privacy concerns. The network's compact design, coupled with focal loss and targeted augmentation strategies, enables effective handling of class imbalance while maintaining high training efficiency which is an essential requirement for deployment.

Experimental evaluations show that the proposed model outperforms remains significantly lighter than widely adopted 3D-CNN baselines, including C3D, ResNet3D, and MC3_18. With an F1-score of 0.9415 and an accuracy of 94.17%, the model demonstrates strong generalization capability, particularly in distinguishing visually and temporally similar activities. These results confirm its effectiveness in extracting discriminative spatiotemporal features from low-resolution event-based inputs. In addition to its high accuracy, the model has a short training time.

Future work will investigate more advanced attention mechanisms, end-to-end event-stream processing without intermediate frame conversion—particularly through SNN

approaches and adaptive temporal resolution strategies to further enhance recognition accuracy while preserving computational efficiency.

In summary, the proposed lightweight 3D-CNN provides an effective, privacy-preserving, and resource-efficient solution for real-time HAR, making it well suited for applications in healthcare, surveillance, and smart environments where both performance and privacy are paramount.

REFERENCES

- [1] L. Bibbò and M. Vellasco, "Human activity recognition (HAR) in healthcare," *Appl. Sci.*, vol. 13, no. 24, p. 13009, 2023. [Online]. Available: <https://www.mdpi.com/2076-3417/13/24/13009>
- [2] G. Bholia and D. K. Vishwakarma, "A review of vision-based indoor HAR: State-of-the-art, challenges, and future prospects," *Multimedia Tools Appl.*, vol. 83, no. 1, pp. 1965–2005, Jan. 2024.
- [3] T. F. N. Bukht, H. Rahman, M. Shaheen, A. Algarni, N. A. Almujaali, and A. Jalal, "A review of video-based human activity recognition: Theory, methods and applications," *Multimedia Tools Appl.*, vol. 84, no. 17, pp. 18499–18545, Jul. 2024.
- [4] T. Wang, J. Guo, B. Zhang, G. Yang, and D. Li, "Deploying AI on edge: Advancement and challenges in edge intelligence," *Mathematics*, vol. 13, no. 11, p. 1878, Jun. 2025. [Online]. Available: <https://www.mdpi.com/2227-7390/13/11/1878>
- [5] W. Shariff, M. S. Dilmaghani, P. Kieley, M. Moustafa, J. Lemley, and P. Corcoran, "Event cameras in automotive sensing: A review," *IEEE Access*, vol. 12, pp. 51275–51306, 2024.

- [6] B. Nguyen, Y. Coelho, T. Bastos, and S. Krishnan, "Trends in human activity recognition with focus on machine learning and power requirements," *Mach. Learn. Appl.*, vol. 5, Sep. 2021, Art. no. 100072. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666827021000360>
- [7] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects," *Comput. Biol. Med.*, vol. 149, Oct. 2022, Art. no. 106060. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0010482522007739>
- [8] A. Kar, V. Raj, and G.-M. Su, "Event camera guided visual media restoration & 3D reconstruction: A survey," 2025, *arXiv:2509.09971*.
- [9] M. S. Dilmaghani, F. Fowley, and P. Corcoran, "A lightweight 3D-CNN for event-based human action recognition with privacy-preserving potential," 2025, *arXiv:2511.03665*.
- [10] L. Bigelli, C. Contoli, V. Freschi, and E. Lattanzi, "Privacy preservation in sensor-based human activity recognition through autoencoders for low-power IoT devices," *Internet Things*, vol. 26, Jul. 2024, Art. no. 101189. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2542660524001306>
- [11] D. Garcia-Gonzalez, D. Rivero, E. Fernandez-Blanco, and M. R. Luaces, "Deep learning models for real-life human activity recognition from smartphone sensor data," *Internet Things*, vol. 24, Dec. 2023, Art. no. 100925. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2542660523002482>
- [12] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, Sep. 2005.
- [13] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Aug. 2005, pp. 65–72.
- [14] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.
- [15] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, Jun. 2010.
- [16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 568–576.
- [17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
- [18] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 20–36.
- [19] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, May 2017, pp. 4724–4733.
- [20] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
- [21] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 6450–6459.
- [22] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Aug. 2019, pp. 6201–6210.
- [23] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jan. 2019, pp. 7082–7092.
- [24] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw. - MobiCom*, 2013, pp. 27–38.
- [25] Z. Yang, C. Wu, W. Xi, L. Wang, and Y. Liu, "Tarf: Enabling fine-grained RF-based activity recognition," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 97–103, 2018.
- [26] Y. Wang, K. Wu, and L. M. Ni, "WiFall: Device-free fall detection by wireless networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 581–594, Feb. 2017.
- [27] M. Scherer, M. Magno, J. Erb, P. Mayer, M. Eggimann, and L. Benini, "TinyRadarNN: Combining spatial and temporal convolutional neural networks for embedded gesture recognition with short range radars," 2020, *arXiv:2006.16281*.
- [28] I. Ullmann, R. G. Guendel, N. C. Kruse, F. Fioranelli, and A. Yarovoy, "A survey on radar-based continuous human activity recognition," *IEEE J. Microw.*, vol. 3, no. 3, pp. 938–950, Jul. 2023.
- [29] Z. Gu, X. He, G. Fang, C. Xu, F. Xia, and W. Jia, "Millimeter wave radar-based human activity recognition for healthcare monitoring robot," 2024, *arXiv:2405.01882*.
- [30] Q. Wang, Z. Xu, Y. Lin, J. Ye, H. Li, G. Zhu, S. A. A. Shah, M. Bennamoun, and L. Zhang, "Dailydvs-200: A large-scale event-based dataset for daily human actions," 2024, *arXiv:2407.05106*.
- [31] M. Scherer, S. Galler, and A. Koelpin, "Tinyradarm: A tiny neural network for gesture recognition with fmcw radar," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
- [32] Y. Yang, X. Li, and L. Wang, "EVMamba: Efficient event-based human action recognition with mamba networks," 2024, *arXiv:2408.12345*.
- [33] A. Amir, B. Taba, D. V. D. Berg, T. Melano, J. L. McKinsty, C. Di Nolfo, T. K. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. DeBole, S. K. Esser, T. Delbrück, M. Flickner, and D. S. Modha, "A low power, fully event-based gesture recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2017, pp. 7388–7397.
- [34] G. Gao, Y. Chen, J. He, and J. Cheng, "An event-based benchmark for human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2019, pp. 201–210.
- [35] J. Zhang, H. Yu, and L. Song, "Dailydvs-200: A neuromorphic dataset for human action recognition," *Sensors*, vol. 21, no. 15, p. 5020, 2021.
- [36] K. Paul, M. S. Dilmaghani, C. Ryan, J. Lemley, and P. Corcoran, "Neuromorphic sensing for yawn detection in driver drowsiness," *Proc. SPIE*, vol. 12701, pp. 287–294, May 2023, doi: [10.1117/12.2680327](https://doi.org/10.1117/12.2680327).
- [37] Y. Deng, H. Chen, H. Liu, and Y. Li, "A voxel graph CNN for object classification with event cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1162–1171.
- [38] S. Schaefer, D. Gehrig, and D. Scaramuzza, "AEGNN: Asynchronous event-based graph neural networks," 2022, *arXiv:2203.17149*.
- [39] D. Gehrig, A. Loquercio, K. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5632–5642.
- [40] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
- [41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
- [42] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6848–6856.
- [43] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.
- [44] N. C. Tay, T. Connie, T. S. Ong, A. B. J. Teoh, and P. S. Teh, "A review of abnormal behavior detection in activities of daily living," *IEEE Access*, vol. 11, pp. 5069–5088, 2023.
- [45] K. Jekel et al., "Mild cognitive impairment and deficits in instrumental activities of daily living: A systematic review," *Alzheimer's Res. Therapy*, vol. 7, no. 1, p. 17, 2015.
- [46] A. Vicente-Sola, D. L. Manna, P. Kirkland, G. Di Caterina, and T. Bihl, "Spiking neural networks for event-based action recognition: A new task to understand their advantage," 2022, *arXiv:2209.14915*.
- [47] Q. Wang, Z. Xu, Y.-M. Lin, J. Ye, H. Li, G. Zhu, S. A. A. Shah, M. Bennamoun, and L. Zhang, "DailyDVS-200: A comprehensive benchmark dataset for event-based action recognition," in *Proc. ECCV*, Jun. 2024, pp. 55–72.
- [48] F. X. Gaya-Morey, C. Manresa-Yee, and J. M. Buades-Rubio, "Deep learning for computer vision based activity recognition and fall detection of the elderly: A systematic review," *Appl. Intell.*, vol. 54, no. 19, pp. 8982–9007, Oct. 2024.
- [49] Q. Wang, Z. Xu, Y. Lin, J. Ye, H. Li, G. Zhu, S. A. A. Shah, M. Bennamoun, and L. Zhang, "DailyDVS-200: A comprehensive benchmark dataset for event-based action recognition," 2024, *arXiv:2407.05106*.
- [50] X. Wang, Z. Wu, B. Jiang, Z. Bao, L. Zhu, G. Li, Y. Wang, and Y. Tian, "HARDVS: Revisiting human activity recognition with dynamic vision sensors," 2022, *arXiv:2211.09648*.

- [51] S. N. N. Htun, S. Egami, and K. Fukuda, "A survey and comparison of activities of daily living datasets in real-life and virtual spaces," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Atlanta, GA, USA, Jan. 2023, pp. 1–7.
- [52] A. Vicente-Sola, D. L. Manna, P. Kirkland, G. D. Caterina, and T. J. Bihl, "Spiking neural networks for event-based action recognition: A new task to understand their advantage," *Neurocomputing*, vol. 611, p. 128657, Aug. 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231224014280>
- [53] X. Wang, S. Wang, P. Shao, B. Jiang, L. Zhu, and Y. Tian, "Event stream based human action recognition: A high-definition benchmark dataset and algorithms," 2024, *arXiv:2408.09764*.
- [54] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3154–3160.
- [55] W. Wu, X. Wang, H. Luo, J. Wang, Y. Yang, and W. Ouyang, "Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models," 2022, *arXiv:2301.00182*.
- [56] C. Ryali, Y.-T. Hu, D. Bolya, C. Wei, H. Fan, P.-Y. Huang, V. Aggarwal, A. Chowdhury, O. Poursaeed, J. Hoffman, J. Malik, Y. Li, and C. Feichtenhofer, "Hiera: A hierarchical vision transformer without the bells-and-whistles," 2023, *arXiv:2306.00989*.
- [57] C. Pan, R. Hou, H. Yu, Q. Wang, S. Velipasalar, and M. Khabza, "SVT: Supertoken video transformer for efficient video understanding," 2023, *arXiv:2304.00325*.
- [58] W. Wu, Z. Sun, and W. Ouyang, "Revisiting classifier: Transferring vision-language models for video recognition," in *Proc. 37th AAAI Conf. Artif. Intell. 55th Conf. Innov. Appl. Artif. Intell. 13th Symp. Educ. Adv. Artif. Intell.*, vol. 37, 2023, pp. 2847–2855, doi: [10.1609/aaai.v37i3.25386](https://doi.org/10.1609/aaai.v37i3.25386).
- [59] A. Piergiovanni, W. Kuo, and A. Angelova, "Rethinking video ViTs: Sparse video tubes for joint image and video learning," 2022, *arXiv:2212.03229*.
- [60] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, and Y. Qiao, "Unmasked teacher: Towards training-efficient video foundation models," 2023, *arXiv:2303.16058*.
- [61] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," 2022, *arXiv:2203.12602*.
- [62] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "VideoMAE v2: Scaling video masked autoencoders with dual masking," 2023, *arXiv:2303.16727*.



MEHDI SEFIDGAR DILMAGHANI received the B.Sc. degree in electronics engineering from the University of Tabriz, in 2012, the M.Sc. degree from the K. N. Toosi University of Technology, in 2016, and the Ph.D. degree from the University of Galway, in 2025. Prior to his Ph.D. research, he focused on electronic implementation of signal processing algorithms and wavelets. During and after his Ph.D., his main research interests have been in computer vision, deep learning, and neuro-morphic sensors (event cameras) for high-accuracy and privacy-preserving applications.



FRANK FOWLEY received the M.Sc. degree from Dublin City University, in 2012, and the Ph.D. degree from University College Dublin, in 2024. His early academic research concentrated on developing secure and efficient architectural designs for cloud computing. More recently, his work has focused on applying deep learning computer vision models to human-centered applications.



PETER CORCORAN (Fellow, IEEE) holds the personal chair position in electronic engineering with the College of Science and Engineering, University of Galway. He was the Co-Founder of several start-up companies, notably FotoNation. He has over 600 technical publications and patents, over 100 peer-reviewed journal articles, and 120 international conference papers. He is the co-inventor of more than 300 granted U.S. patents. He is an IEEE Fellow, recognized for his contributions to digital camera technologies, notably in-camera red-eye correction, and facial detection. He is a member of the IEEE Consumer Electronics Society, for over 25 years. He is the former Editors-in-Chief of *IEEE Consumer Electronics Magazine*.

• • •