IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.
Cite as T-ASE paper.

Neuromorphic Event Camera-Based Object Recognition and Grasping Position Detection Using a Transfer Learning-Enhanced Multi-Task Model

Muhammad Hamza Zafar, *Graduate Student Member, IEEE*, Syed Kumayl Raza Moosavi, and Filippo Sanfilippo, *Senior Member, IEEE*

***Abstract*—Object recognition and grasping position detection are critical tasks in robotic manipulation, particularly when operating in dynamic and unstructured environments. This paper presents the Channel Sharpening Attention-based Adaptive Inception Network (CSA-AInceptNet), a novel multi-task learning model designed for these tasks using event camera data. The proposed architecture integrates channel sharpening attention with adaptive inception networks to enhance feature extraction and improve robustness. The model's performance is evaluated on two state-of-the-art event camera datasets, E-Grasp and Neuro-Grasp. On the E-Grasp dataset, CSA-AIncepNet achieves a remarkable accuracy of 99.47% and a mean Intersection over Union (IoU) of 0.9370, significantly surpassing existing methods. On the Neuro-Grasp dataset, leveraging transfer learning, the model attains 98.58% accuracy and a mean IoU of 0.4897, demonstrating strong generalization capabilities across datasets. Comparative analyses and ablation studies further validate the effectiveness of the proposed architecture, highlighting its superiority over conventional models like ConvNeXt, DarkNet, DenseNet, and VGG16. The results establish CSA-AIncepNet as a robust solution for event-based object recognition and grasping detection, paving the way for advancements in human-robot collaboration and dynamic robotic manipulation.**

***Note to Practitioners*—This work provides a practical solution for improving object recognition and grasping position detection in robotic systems, particularly in unpredictable and fast-changing real-world environments. By leveraging event camera data, the proposed approach enables robots to efficiently identify objects and determine optimal grasping positions, even under challenging conditions. The results highlight the model's ability to outperform existing methods, making it highly suitable for applications such as human-robot collaboration and precise object handling. This advancement has significant implications for industries like manufacturing, logistics, and healthcare, where robots must interact with objects quickly and accurately. Practitioners can adopt this method to enhance robotic performance, reduce errors, and improve operational efficiency. Future work could focus on testing the model in more complex environments and adapting it for real-time deployment in dynamic settings.**

Received 17 December 2024; revised 5 May 2025 and 2 July 2025; accepted 7 August 2025. Date of publication 13 August 2025; date of current version 27 August 2025. This article was recommended for publication by Associate Editor Y. Wu and Editor X. Liu upon evaluation of the reviewers' comments. *(Corresponding author: Filippo Sanfilippo.)*

Muhammad Hamza Zafar and Syed Kumayl Raza Moosavi are with the Department of Engineering Sciences, University of Agder, 4879 Grimstad, Norway (e-mail: muhammad.h.zafar@uia.no; syed.k.moosavi@uia.no).

Filippo Sanfilippo is with the Department of Engineering Sciences, University of Agder, 4879 Grimstad, Norway, and also with the Department of Software Engineering, Kaunas University of Technology, 51368 Kaunas, Lithuania (e-mail: filippo.sanfilippo@uia.no).

Digital Object Identifier 10.1109/TASE.2025.3598695

***Index Terms*—Event camera, multi-task model, object recognition, grasping position detection, transfer learning.**

I. Introduction

GRASP pose estimation is an essential aspect of robotic manipulation, enabling robots to interact effectively with objects in their environment. This capability is particularly crucial for applications ranging from industrial automation to assistive robotics. A fundamental challenge in robotic manipulation is the traditionally separate training and execution of object recognition and grasp pose estimation–two intricately linked tasks that determine a robot's ability to interact effectively with its environment. While these tasks are inherently connected in practice (a robot must first recognize what an object is before determining how to grasp it [1], [2], [3]), most existing approaches treat them as isolated problems requiring separate models, pipelines, and computational resources. This separation creates inefficiencies, inconsistencies between recognition and grasping strategies, and integration challenges that limit performance in dynamic, real-world environments.

The inspiration for multi-task learning in robotic manipulation stems from the biological principle that the human visual system seamlessly integrates object recognition with motor planning. When humans reach for an object, they simultaneously process "what" the object is and "how" to grasp it through shared neural pathways, suggesting that these tasks can mutually enhance each other when learned jointly [1]. In robotic systems, object recognition provides semantic understanding that can inform optimal grasping strategies–for instance, recognizing a fragile glass versus a robust tool fundamentally changes the required grasping approach. Conversely, spatial features learned for grasp pose estimation, such as surface orientations and geometric properties, can provide valuable geometric cues that enhance object classification accuracy. This symbiotic relationship motivates the development of unified multi-task architectures that can leverage shared representations to improve both tasks while reducing computational overhead–a critical consideration for resource-constrained robotic platforms.

Neuromorphic computing has emerged as a promising paradigm for addressing the computational limitations inherent in traditional robotic perception systems. Inspired by the structure and function of biological neural networks, neuromorphic approaches offer advantages in terms of energy

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

efficiency, real-time processing capabilities, and adaptive learning mechanisms [4]. Event-based cameras, a key component of neuromorphic systems, capture visual information as discrete events triggered by pixel-level intensity changes, rather than capturing full frames at fixed intervals. This approach significantly reduces data redundancy and computational requirements while providing high temporal resolution and dynamic range. The integration of neuromorphic principles with multi-task learning architectures presents an opportunity to develop more efficient and capable robotic manipulation systems that can operate effectively in dynamic, unstructured environments.

In recent years, deep learning has gained widespread application in robotic manipulation tasks [5], [6], [7], driven by its ability to utilize large datasets for solving complex problems [6], [8]. Deep convolutional neural networks (CNNs) have been successfully applied to grasp pose estimation, beginning with approaches like the sliding window detection framework proposed by Lenz et al. [5]. This method extracts features from image sequences and selects the grasp candidate with the highest confidence score. While effective, this approach has high computational costs, presenting a significant limitation for real-time robotic applications. To address computational inefficiency and streamline grasp prediction, end-to-end approaches were subsequently developed [9], [10]. These methods often use RGB or RGB-D images for regression or classification of grasp rectangles and have shown significant performance improvements on benchmark datasets like the Cornell Grasping Dataset [11]. Language driven grasp detection also developed but still uses RGB data for this purpose [12].

However, one major challenge in robotic manipulation remains is balancing computational efficiency with the limited power of embedded robotic systems. Current state-of-the-art systems [13], [14], [15], [16] predominantly use RGB-D cameras as perception sensors. These cameras capture environmental information as a series of discrete frames at a fixed frequency, providing rich color and texture details. However, this frame-based approach faces issues like high computational time and storage demands, as noted by [4] and [17]. Multi-task learning approaches offer a potential solution to these computational challenges by sharing feature representations across related tasks, thereby reducing overall model complexity and inference time while maintaining or even improving individual task performance. Another persistent challenge in the field is the limited availability of grasp datasets collected from real-world environments. To address this limitation, approaches like DexNet [18] have explored using simulated data for grasp pose estimation, with varying degrees of success in transferring to real-world scenarios.

A. Literature Review

Over the past two decades, significant progress has been made in robotic grasp pose estimation, evolving from computationally expensive methods to more efficient approaches that maintain or improve accuracy.

Early methods, such as those proposed by Lenz et al. [5] and Saxena et al. [19], used sliding window techniques to train grasp detectors. While innovative at the time, these approaches suffered from high computational costs that limited their practical application in real-time robotic systems. To improve efficiency, Johns et al. [20] and Morrison et al. [21] developed methods that reduced inference time by focusing on a discrete set of grasp candidates. However, this approach often overlooked potential grasps by limiting the search space. Other studies [8], [25] introduced end-to-end CNN-based algorithms to predict a single grasp for an input image, but these often generalized to the average grasp pose for objects, lacking the specificity needed for diverse object manipulation. Building on object detection frameworks, Chu et al. [14] proposed using a grasp region proposal network inspired by Faster RCNN [22] for grasp pose estimation. This approach allowed for the detection of multiple grasp candidates within an image. Further advancing both speed and accuracy, Zhang et al. [9] introduced a real-time grasp network with an orientation anchor box mechanism.

For handling overlapping objects—a common challenge in real-world scenarios–an ROI-based method was developed by Zhang et al. [23], demonstrating effectiveness in such complex scenes. Chen et al. [24] proposed a grasp path approach for convolutional multigrasp prediction, improving accuracy in real-world scenarios to address the lack of comprehensive ground truths in grasp pose datasets. More recent innovations include a real-time grasp detection system with a rotation ensemble module (REM) inspired by YOLOv9 [25], delivering high accuracy while maintaining computational efficiency. Studies like those by Cao et al. [10], [15] used neural networks for grasp prediction with high-resolution images, solving pixel-wise estimation challenges. Fusion-based methods, such as those by Asif et al. [26] and Wu et al. [27], also enhanced performance in grasp prediction by combining multiple data sources.

While the approaches discussed thus far largely rely on RGB or RGB-D images to predict grasp rectangles [28], there is growing interest in exploring the potential of neuromorphic vision sensors like the DAVIS346. The emergence of event-based neuromorphic vision technology offers a new approach for various vision tasks, with several methods [29], [30], [31] developed specifically for object detection using this technology. For robotic grasping, Mirus et al. [32] proposed a framework combining perception, reasoning, and control for mobile robots to pick and place objects. Using embedded dynamic vision sensors (DVS), this method allows robots to grasp and relocate objects accurately even under challenging conditions like rapid motion or varying illumination. Additional applications of event-based vision in robotics include dynamic vision-based finger systems for slip detection and suppression [33], which perform well under challenging conditions like illumination changes and vibrations. For tactile sensing, a dynamic vision-based measurement approach was presented by Naeini et al. [34]. Furthermore, Li et al. [35] developed an event-based dataset and a deep neural network model specifically for grasp pose prediction, showcasing the potential of this emerging technology.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

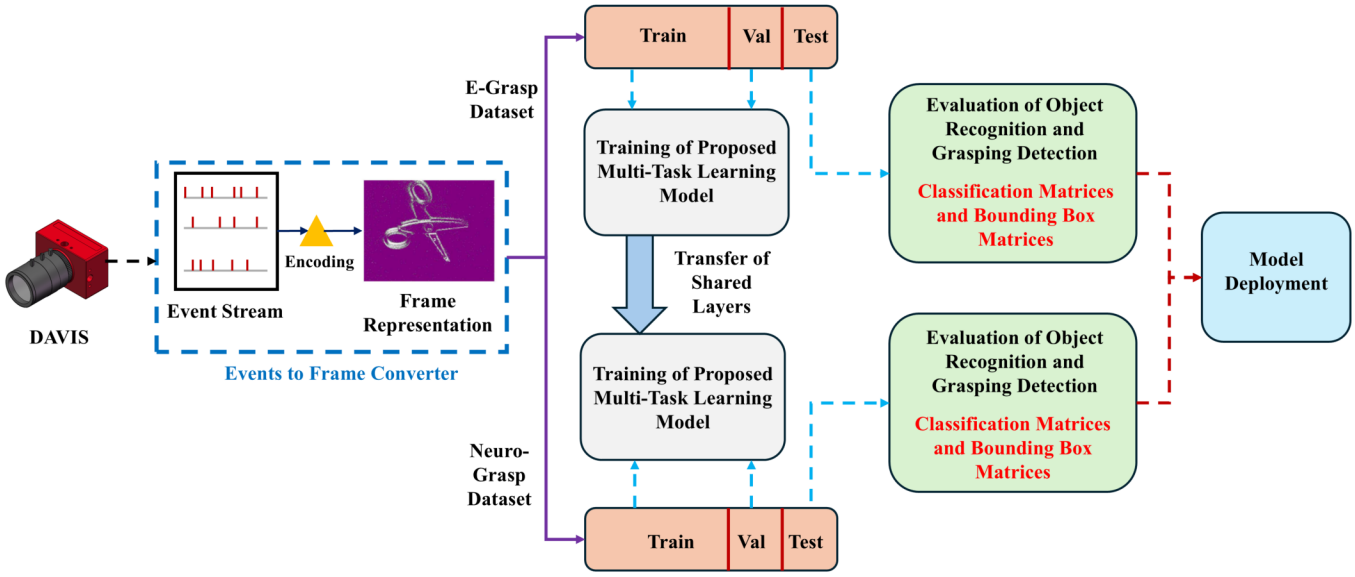


Fig. 1. Detailed Flow Description of Proposed Transfer Learning Assisted Multi-Task Learning Model using Channel Sharpening Attention based Adaptive Inception Network.

However, compared to conventional frame-based methods, event-based vision still offers lower spatial resolution and remains in its early stages of development. This presents both challenges and opportunities for advancing the field of robotic grasp pose estimation using neuromorphic vision sensors.

B. Contributions and Paper Organisation

Despite these advances, significant research gaps remain in the field of robotic grasp pose estimation. First, most existing approaches treat object recognition and grasp detection as separate tasks with independent models, resulting in the inefficiencies and inconsistencies outlined above. This separation is particularly problematic in resource-constrained robotic systems where computational efficiency is paramount. Second, while multi-task learning has shown promise in other computer vision domains, its application to event-based vision for simultaneous object recognition and grasp detection remains largely unexplored. Finally, existing models fail to fully leverage the temporal dynamics and sparse representation advantages that event cameras offer, particularly for operation in challenging conditions like rapid movement or varying illumination.

Our work addresses these challenges by proposing the Channel Sharpening Attention-based Adaptive Inception Network (CSA-AInceptNet), a novel multi-task learning model that unifies object recognition and grasp position detection in a single architecture. By leveraging the unique properties of event camera data and incorporating channel sharpening attention mechanisms with adaptive inception networks, our approach enhances feature extraction while maintaining computational efficiency. This integration enables robots to simultaneously recognize objects and determine optimal grasp positions in dynamic, unstructured environments–a capability critical for advancing human-robot collaboration and autonomous manipulation tasks. The detailed abstract description of the proposed work is shown in Fig. 1.

The contributions of this work are:

- A novel multi-task learning framework, the Channel Sharpening Attention-based Adaptive Inception Network (CSA-AInceptNet), designed to simultaneously perform object recognition and grasp position estimation. This unified approach enhances the efficiency and accuracy of robotic grasping tasks by leveraging shared representations for both tasks.
- Development of a transfer learning methodology that effectively bridges the domain gap between different event-based datasets, enabling knowledge transfer from a source dataset E-Grasp to a more complex target dataset Neuro-Grasp. Our approach demonstrates how architectural innovations in the CSA-AInceptNet model enable efficient feature extraction that generalizes across datasets with minimal fine-tuning, achieving 98.58% accuracy on the challenging Neuro-Grasp dataset despite its increased complexity and different event data characteristics.
- An extensive evaluation of the proposed model against existing state-of-the-art methods. Our analysis highlights the superior performance of CSA-AInceptNet in terms of accuracy and computational efficiency, validating its effectiveness in addressing the challenges of robotic grasp pose estimation and object recognition.

II. Proposed Methodology

A. Dataset Description

1) E-Grasp Dataset: The dataset used for training the model consists of two main parts: the base dataset and the annotation dataset. The base dataset is created by applying three event-to-frame encoding techniques (Frequency, SAE, and LIF) to process event streams from a DAVIS camera into event frames. A sliding window with a 20ms interval is employed to accumulate event information and generate frames, which allows for better integration with deep learning

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

algorithms. The dataset includes approximately 91 objects and was recorded in both high and low light conditions. The annotation dataset is generated by tracking LED markers using an SMP filter, which creates bounding boxes for good or bad grasps. These annotations are extended from single-grasp to multi-grasp scenarios, allowing for automatic pose annotation with high time resolution (milliseconds) [35]. The complete Event-Grasping dataset encompasses 91 objects with approximately 18,200 annotated frames, each containing high-precision grasping rectangles that can be used to train robust robotic grasping algorithms specialized for the high-speed, low-latency capabilities of neuromorphic vision.

2) Neuro-Grasp Dataset: The dataset is collected using a DAVIS346 neuromorphic vision sensor, which combines an event-based Dynamic Vision Sensor (DVS) with an RGB frame-based sensor, offering a resolution of 346×260 pixels. This sensor captures both event-based and RGB streams separately to record a diverse range of 154 grasping objects, leading to a dataset that spans 4620.42 seconds and contains 14,141.7 million events. The dataset is particularly challenging due to its diversity in object scales, orientations, and locations. After manually filtering out unusable data, the resulting NeuroGrasp dataset includes 8753 RGB images and corresponding event streams. Each image is annotated with ground-truth grasp rectangles, representing potential good grasp configurations, though not covering all possible grasps. The dataset includes binary data, raw event data, RGB images, timestamp files, and labels, all in a standardized format. Additionally, a multi-object grasping dataset is created to test the algorithm's ability to generalize in more complex and cluttered environments, where images contain three to five objects with varying poses and orientations [36].

B. Problem Formulation

In industrial and robotic applications, fast and efficient object recognition and grasp detection are crucial for tasks such as manipulation and assembly. Event cameras, which record changes in pixel intensities asynchronously, provide a promising solution for recognizing objects in dynamic environments due to their high temporal resolution and reduced data redundancy. However, processing this spatio-temporal data for multitask learning–object classification and bounding box detection–presents significant challenges. This project aims to develop a multitask learning model that can simultaneously recognize objects and detect graspable regions using event camera-based frames.

Let $\mathbf{E} \in \mathbb{R}^{H\times W\times T\times 3}$ represent the event camera frame data, where H, W, and T denote the height, width, and temporal dimensions, respectively. The three channels correspond to the pixel coordinates x, y, and polarity p, where $p = \{-1, +1\}$ indicates the direction of brightness change. For simplicity in the frame representation, we consider the frame $\mathbf{E} \in \mathbb{R}^{H\times W\times 3}$, where H and W are the spatial dimensions and the third dimension represents the aggregated event data across time.

1) Object Recognition Task: Let $\mathcal{X} \in \mathbb{R}^{H\times W\times 3}$ denote the input image (event camera frame), and $\mathbf{y}_{\text{class}} \in \{0, 1\}^C$ be the one-hot encoded class label vector for object classification, where C is the total number of object classes. The objective of the classification task is to find a function $f_{\text{class}} : \mathcal{X} \to \mathbf{y}_{\text{class}}$, which minimizes the classification loss:

$$\mathcal{L}_{\text{class}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} \tag{1}$$

where $y_{i,c}$ is the ground truth label and $\hat{y}_{i,c}$ is the predicted probability for class c for the i-th input.

2) Bounding Box Detection Task (Grasp Detection): For each frame, we aim to predict the bounding box $\mathbf{b} = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$, which represents the grasping region of the object. The model learns a function $f_{\text{bbox}} : \mathcal{X} \to \mathbf{b}$, and the objective is to minimize the mean squared error (MSE) between the predicted and true bounding boxes:

$$\mathcal{L}_{\text{bbox}} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\mathbf{b}_i^{\text{true}} - \mathbf{b}_i^{\text{pred}}\right)^2\right) \tag{2}$$

3) Multitask Learning Objective: The final objective is to optimize both tasks simultaneously using a weighted sum of the classification and bounding box losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{class}}\mathcal{L}_{\text{class}} + \lambda_{\text{bbox}}\mathcal{L}_{\text{bbox}} \tag{3}$$

where λ_{class} and λ_{bbox} are the weights assigned to the classification and bounding box regression tasks, respectively. These weights balance the importance of each task during training.

C. Initial Feature Extraction

To efficiently handle high-dimensional data and prepare it for the adaptive inception module, the initial layers employ convolutional and pooling operations. These operations reduce the dimensionality of the input space while preserving critical spatial features [37]. By transforming the data into a more compact representation, these layers enhance computational efficiency and improve parallelism, setting the stage for the multiscale feature extraction capabilities of the adaptive inception module [38].

The input feature map, $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, represents data for each frame, where $H = 224$, $W = 224$, and $C = 3$, corresponding to the height, width, and number of channels (RGB), respectively. The spatial features are extracted using convolutional and pooling layers, enabling hierarchical information capture.

$$F_{\text{conv1}} = \sigma\left(W_{\text{conv1}} * X + b_{\text{conv1}}\right) \tag{4}$$

where W_{conv1} and b_{conv1} are the weights and biases, $*$ denotes convolution, and σ is the activation function (e.g., ReLU).

This is followed by max-pooling:

$$F_{\text{pool1}} = \text{MaxPool}(F_{\text{conv1}}) \tag{5}$$

For subsequent layers:

$$F_{\text{conv2}} = \sigma\left(W_{\text{conv2}} * F_{\text{pool1}} + b_{\text{conv2}}\right) \tag{6}$$

$$F_{\text{pool2}} = \text{MaxPool}(F_{\text{conv2}}) \tag{7}$$

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

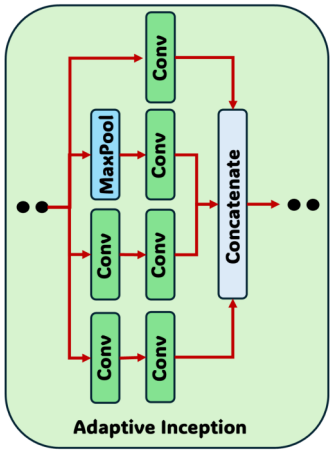


Fig. 2. Architecture of Adaptive Inception Network Module.

D. Adaptive Inception Module

Adaptive attention mechanisms dynamically emphasize significant event streams while suppressing redundant pixel data, enabling efficient processing. This strategy aligns with the Channel-Sharpening Attention (CSA) mechanism, which refines channel-wise features and effectively handles high-dimensional data by focusing on critical elements [38]. We explore the adaptive inception module, which processes the output of the initially extracted feature map $\mathbf{F}_{\text{pool2}}$ through four parallel branches, achieving greater computational efficiency compared to the adaptive attention mechanism. The output from this module becomes the input to the channel-sharpening block. The outputs of these branches are concatenated to form the final feature map [38]. The structure of adaptive inception module is shown in Fig. 2. The details of this adaptive inception module is detailed below:

1×1 *Convolution Branch (Conv1)*

This branch applies a 1×1 convolution to capture local features and reduce dimensionality:

$$\mathbf{F}_{\text{conv1}} = \sigma\left(W_{\text{conv1}} * \mathbf{F}_{\text{input}} + b_{\text{conv1}}\right) \tag{8}$$

3×3 *Convolution Branch (Conv2)*

This branch first applies a 1×1 convolution for dimensionality reduction, followed by a 3×3 convolution to extract spatial features:

$$\mathbf{F}_{\text{conv2}} = \sigma\left(W_{\text{conv2}} * \left(W_{\text{reduce2}} * \mathbf{F}_{\text{input}} + b_{\text{reduce2}}\right) + b_{\text{conv2}}\right) \tag{9}$$

5×5 *Convolution Branch (Conv3)*

This branch first applies a 1×1 convolution for dimensionality reduction, followed by a 5×5 convolution to capture larger-scale features:

$$\mathbf{F}_{\text{conv3}} = \sigma\left(W_{\text{conv3}} * \left(W_{\text{reduce3}} * \mathbf{F}_{\text{input}} + b_{\text{reduce3}}\right) + b_{\text{conv3}}\right) \tag{10}$$

Pooling Branch (Conv4)

This branch applies max pooling, followed by a 1×1 convolution to project the pooled features:

$$\mathbf{F}_{\text{conv4}} = \sigma\left(W_{\text{conv4}} * \text{MaxPool}(\mathbf{F}_{\text{input}}) + b_{\text{conv4}}\right) \tag{11}$$

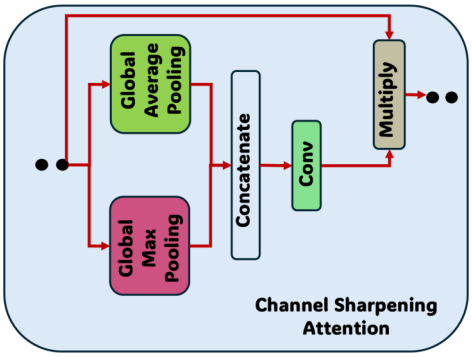


Fig. 3. Architecture of Channel Sharpening Attention Module.

The outputs of the four branches are concatenated to form the final output feature map:

$$\mathbf{F}_{\text{adaptive}} = \text{Concat}\left(\mathbf{F}_{\text{conv1}}, \mathbf{F}_{\text{conv2}}, \mathbf{F}_{\text{conv3}}, \mathbf{F}_{\text{conv4}}\right) \tag{12}$$

E. Channel Sharpening Attention (CSA)

CSA mechanism enhances the most important channels within the feature map, allowing the network to selectively focus on significant features while suppressing irrelevant ones. This is achieved by leveraging global spatial information across the feature map through a combination of global average pooling and global max pooling across the spatial dimensions H and W, followed by a sigmoid gating mechanism to generate attention weights. These attention weights are then used to sharpen the channel-wise feature representation by performing an element-wise multiplication with the original feature map [39], [40]. The architecture of channel sharpening attention module is shown in Fig. 3.

The global average pooling operation is defined as:

$$\mathbf{F}_{\text{avg}}(c) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{F}_{\text{adaptive}}(h, w, c) \tag{13}$$

Here, H, W, and c represent the height, width, and channel index, respectively, while $\mathbf{F}_{\text{adaptive}}(h, w, c)$ is output of the adaptive inception module, denotes the feature map value at spatial position (h, w) for channel c. This operation computes the average of all spatial values for each channel.

Similarly, the global max pooling operation is defined as:

$$\mathbf{F}_{\text{max}}(c) = \max_{h,w} \mathbf{F}_{\text{adaptive}}(h, w, c) \tag{14}$$

This operation selects the maximum value from all spatial positions for each channel. Both pooling methods, $\mathbf{F}_{\text{avg}}$ and $\mathbf{F}_{\text{max}}$, capture distinct spatial information for each channel. The feature maps F_{avg} and F_{max}, representing the channel-wise average pooling and max pooling, respectively, are concatenated to form F_{concat}:

$$F_{\text{concat}} = \text{Concat}(F_{\text{avg}}, F_{\text{max}}) \tag{15}$$

This concatenated feature is passed through a convolutional layer with a sigmoid activation function to generate the attention weights for each channel:

$$F_{\text{attention}} = \sigma\left(W_{\text{conv2d}} F_{\text{concat}} + b_{\text{conv2d}}\right) \tag{16}$$

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

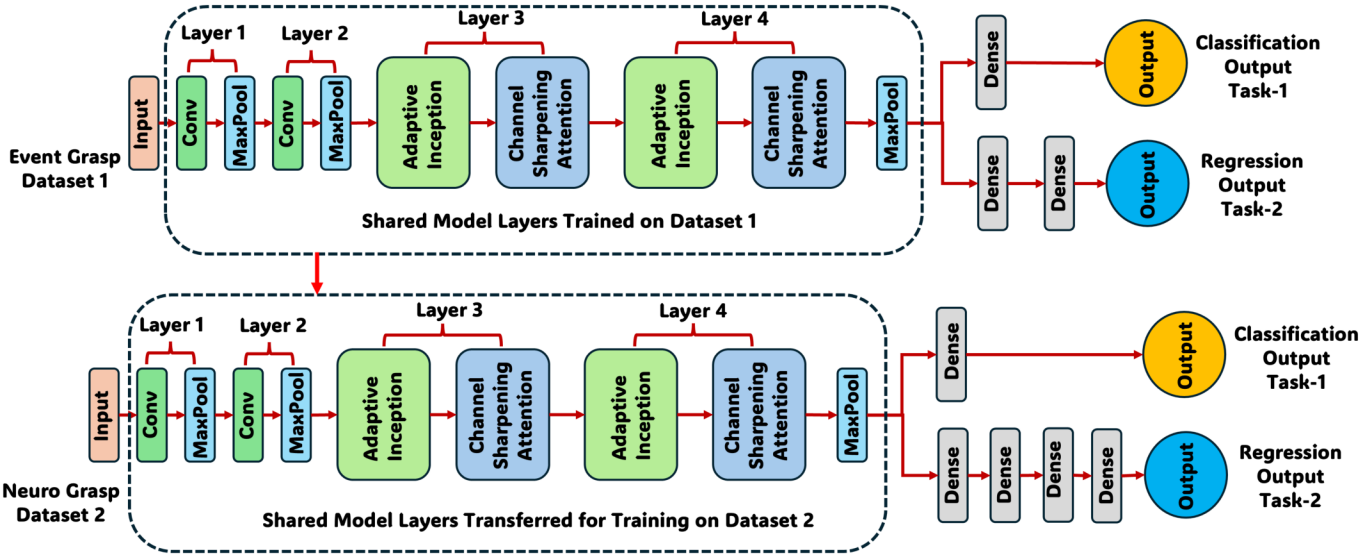


Fig. 4. Transfer Learning for Training on Dataset 2.

The final step of CSA is to apply the computed attention map to the original feature map $\mathbf{F}_{\text{adaptive}}$ via element-wise multiplication, which sharpens the most informative channels and suppresses the irrelevant ones:

$$F_{\text{sharpened}} = F_{\text{adaptive}} \odot F_{\text{attention}} \tag{17}$$

F. Fine-Tuning Using Transfer Learning

In this study, the model is initially trained on E-Grasp dataset, which consists of 91 classes, and each containing bounding box per image for grasping detection. The training process spans 100 epochs, during which the model demonstrates effective generalization. Upon achieving satisfactory performance on E-Grasp, transfer learning (TL) is applied to fine-tune the model for the Neuro-Grasp dataset, which includes 151 classes, each with multiple bounding boxes.

The fine-tuning process leveraged the shared layers from the E-Grasp-trained model, as shown in Fig. 4. By reusing these shared layers, the model effectively transferred the learned features from the E-Grasp dataset to the Neuro-Grasp, ensuring a more efficient adaptation process. This approach also helped to mitigate the computational complexity typically associated with training models on high-dimensional datasets with multiple bounding boxes.

To empirically validate the benefits of TL with shared layers, we conducted a comparative analysis between training the model with and without shared layers. The results demonstrated that using shared layers in the TL process significantly improved computational efficiency without compromising performance. The model trained with TL not only required fewer computational resources but also converged faster, highlighting the effectiveness of this approach for complex, multi-class, and multi-bounding box datasets.

G. Adaptive Feature Learning for Multitask Objectives

By combining multiple scales of feature representation and refining them through channel-wise attention, the *Adaptive Inception Module with Channel Sharpening Attention* provides a powerful backbone for multitask learning. In the context of object recognition and grasping detection, this module is used to extract features that are beneficial for both object recognition and grasp detection. The multi-scale processing captures different levels of abstraction, while the attention mechanism ensures that only the most important features are passed to the classification and bounding box regression heads, improving overall performance.

H. MultiTask Learning Model

Multitask Learning (MTL) is an approach where a model is trained to perform multiple related tasks simultaneously. The main motivation behind MTL is that by sharing knowledge across tasks, the model can generalize better and learn more robust representations. In this work, multitask learning is applied to simultaneously perform **object recognition** and **bounding box regression** for grasp detection from event camera frames.

Multitask learning in this framework involves optimizing two tasks with a shared feature extraction backbone:

- **Object Recognition**: Predicts the class label of the object from a predefined set of classes.
- **Bounding Box Regression**: Predicts the bounding box coordinates for the graspable region of the object.

The model, therefore, has two output heads: one for classification and one for bounding box regression. Both tasks share the same underlying feature extraction network, which is the *Adaptive Inception Module with Channel Sharpening Attention*. This shared network enables the model to leverage common features that are useful for both tasks, such as spatial structure, texture, and object boundaries.

I. Hyperparameters

The hyperparameters used in this multitask learning model are carefully selected to balance the demands of both object

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

TABLE I
HYPERPARAMETERS OF THE MULTITASK CSA-AINCEPTNET MODEL

Layer #	Layer Name	Type	Filters/Units	Kernel Size	Stride	Activation	Notes
Input and Initial Convolutional Layers							
0	Input	Input	–	–	–	–	Shape=(224, 224, 3)
1	Conv1 MaxPool1	Conv2D MaxPooling2D	64 –	7×7 3×3	2×2 2×2	ReLU –	padding='same' padding='same'
2	Conv2 MaxPool2	Conv2D MaxPooling2D	192 –	3×3 3×3	1×1 2×2	ReLU –	padding='same' padding='same'
Adaptive Inception Module 1 with Channel Sharpening Attention							
3.1	1×1 Branch	Conv2D	64	1×1	1×1	ReLU	padding='same'
3.2	3×3 Branch	Conv2D (reduce) Conv2D	96 128	1×1 3×3	1×1 1×1	ReLU ReLU	padding='same' padding='same'
3.3	5×5 Branch	Conv2D (reduce) Conv2D	16 32	1×1 5×5	1×1 1×1	ReLU ReLU	padding='same' padding='same'
3.4	MaxPool Branch	MaxPooling2D Conv2D	– 32	3×3 1×1	1×1 1×1	– ReLU	padding='same' padding='same'
3.5	CSA Mechanism	AvgPool + MaxPool Conv2D	– Same as input	– 1×1	– 1×1	– Sigmoid	Global pooling Feature reweighting
Adaptive Inception Module 2 with Channel Sharpening Attention							
4.1	1×1 Branch	Conv2D	128	1×1	1×1	ReLU	padding='same'
4.2	3×3 Branch	Conv2D (reduce) Conv2D	128 192	1×1 3×3	1×1 1×1	ReLU ReLU	padding='same' padding='same'
4.3	5×5 Branch	Conv2D (reduce) Conv2D	32 96	1×1 5×5	1×1 1×1	ReLU ReLU	padding='same' padding='same'
4.4	MaxPool Branch	MaxPooling2D Conv2D	– 64	3×3 1×1	1×1 1×1	– ReLU	padding='same' padding='same'
4.5	CSA Mechanism	AvgPool + MaxPool Conv2D	– Same as input	– 1×1	– 1×1	– Sigmoid	Global pooling Feature reweighting
Final Feature Extraction and Output Heads							
5	MaxPool3 Flatten	MaxPooling2D Flatten	– –	3×3 –	2×2 –	– –	padding='same' –
6	Classification Head	Dense Dense	128 num of classes	– –	– –	ReLU Softmax	– name='classification'
7	Bounding Box Head	Dense Dense Dense	32 32 4	– – –	– – –	ReLU ReLU Linear	– – name='bounding_box'
Training Parameters							
Optimizer			Adam (learning_rate=0.001)				
Classification Loss			Categorical Crossentropy (weight=2.0)				
Bounding Box Loss			Mean Squared Error (weight=0.5)				
Batch Size			16				
Epochs			100				
Model Checkpoint			Yes (monitor='val_loss', save_best_only=True)				
Total Parameters			15.7M				

recognition and bounding box detection from event camera frames. The input image size is standardized to $224 \times 224 \times 3$, ensuring compatibility with the convolutional layers, while the batch size of 16 optimizes computational efficiency during training. The model classifies objects across 91 categories and uses the Adam optimizer with a learning rate of 0.001, ensuring stable and effective convergence over 100 epochs. The Adaptive Inception Module utilizes various convolutional filters (1×1, 3×3, 5×5) and pooling layers, with the number of filters ranging from 32 to 192 to capture multi-scale features. Fully connected layers for both the classification and bounding box heads use 128 and 32 units, respectively and the bounding box head also includes a secondary layer with 32 units. A dropout rate of 0.5 is applied to reduce overfitting in fully connected layers. Loss weighting is used to emphasize the classification task (weight 2.0) over bounding box regression (weight 0.5), ensuring that both tasks are optimized concurrently in the multitask learning framework. These hyperparameters allow the model to learn both tasks while balancing complexity and performance effectively. The hyperparameters of the proposed model are shown in Table. I.

III. RESULTS AND DISCUSSION

A. Performance Matrices

1) Classification Accuracy: The accuracy of object classification is measured by the fraction of correctly predicted object

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

classes.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\arg\max(\hat{y}_i) = \arg\max(y_i)\right) \tag{18}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

2) Intersection Over Union (IoU) for Bounding Boxes: IoU serves as a fundamental metric for quantifying the spatial accuracy of predicted bounding boxes in object detection tasks. Given two bounding boxes—the ground truth box $\mathbf{b}_{\text{true}}$ and the predicted box $\mathbf{b}_{\text{pred}}$—the IoU metric is mathematically defined as the ratio of their intersection area to their union area:

$$\text{IoU} = \frac{|\mathbf{b}_{\text{true}} \cap \mathbf{b}_{\text{pred}}|}{|\mathbf{b}_{\text{true}} \cup \mathbf{b}_{\text{pred}}|} \tag{19}$$

For rectangular bounding boxes in 2D space, the intersection area can be computed as:

$$\begin{aligned}|\mathbf{b}_{\text{true}} \cap \mathbf{b}_{\text{pred}}| = \max(0, x_{\min}^{\text{overlap}} - x_{\max}^{\text{overlap}})\cdot \\ \times \max(0, y_{\min}^{\text{overlap}} - y_{\max}^{\text{overlap}})\end{aligned} \tag{20}$$

where $(x_{\min}^{\text{overlap}}, y_{\min}^{\text{overlap}})$ and $(x_{\max}^{\text{overlap}}, y_{\max}^{\text{overlap}})$ represent the top-left and bottom-right corners of the intersection region, respectively.

The IoU value ranges from 0 (no overlap) to 1 (perfect overlap). As the IoU approaches 1, the predicted bounding box more precisely aligns with the ground truth, indicating superior detection accuracy. In the context of robotic grasping, high IoU values ensure that the system correctly identifies the exact graspable regions of target objects.

The optimization objective encompasses maximizing object classification accuracy while concurrently enhancing bounding box localization quality (IoU) for graspable object regions. This dual optimization is crucial for enabling real-time operation with event-based camera data, which is essential for time-sensitive robotic manipulation tasks requiring both speed and precision.

3) Jaccard Threshold Accuracies: The Jaccard threshold, equivalent to an IoU threshold, provides a systematic framework for evaluating detection performance at varying levels of spatial precision. This approach establishes discrete benchmarks for quantifying the model's ability to correctly localize objects.

We define a set of threshold values $\Theta = \{0.25, 0.30, 0.35, 0.40\}$, representing increasingly stringent requirements for spatial overlap. For each threshold $\theta \in \Theta$, we calculate the accuracy as:

$$\text{Accuracy}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\text{IoU}(\mathbf{b}_{\text{true}}^{i}, \mathbf{b}_{\text{pred}}^{i}) \geq \theta\right) \tag{21}$$

where N represents the total number of test samples, $\mathbb{1}(\cdot)$ is the indicator function that outputs 1 when the condition is satisfied and 0 otherwise, $\mathbf{b}_{\text{true}}^{i}$ denotes the ground truth bounding box for the i-th sample, and $\mathbf{b}_{\text{pred}}^{i}$ denotes the predicted bounding box for the i-th sample

This multi-threshold evaluation provides a comprehensive performance profile:

- At $\theta = 0.25$ (25% overlap): This permissive threshold captures the model's ability to roughly locate objects, yielding higher accuracy values but potentially including predictions with significant spatial imprecision.
- At $\theta = 0.30$ (30% overlap): This threshold represents a moderate requirement for spatial accuracy, balancing between permissiveness and strictness.
- At $\theta = 0.35$ (35% overlap): This threshold demands greater precision, typically resulting in decreased accuracy but improved reliability of positive detections.
- At $\theta = 0.40$ (40% overlap): This stringent threshold requires substantial spatial alignment, typically producing lower accuracy metrics but with significantly higher confidence in the precision of successful detections.

This stratified evaluation framework enables precise characterization of the precision-recall trade-off inherent in the model's performance. Lower thresholds prioritize recall (detecting most graspable regions), while higher thresholds emphasize precision (ensuring detected regions are accurately localized). This approach allows for adaptive parameter selection based on specific robotic manipulation requirements, such as prioritizing grasp reliability versus operation speed in different task contexts.

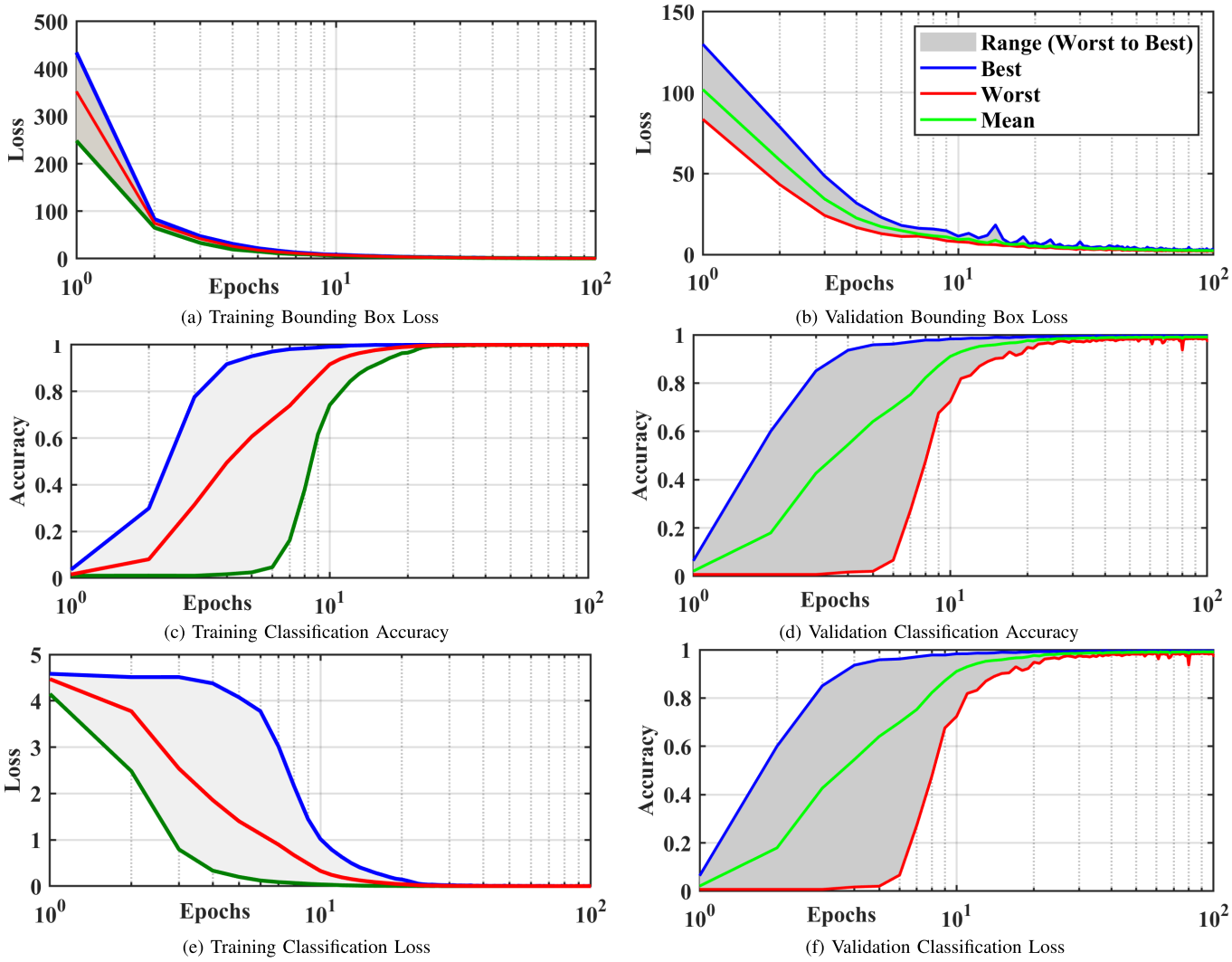
TABLE II
MODEL CONFIGURATIONS USED FOR ABLATION STUDY

Model	Layer 1	Layer 2	Layer 3	Layer 4	Layer 3 and 4 with SA
M1	✓	✓	✓	✓	×
M2	✓	✓	✓	×	×
M3	✓	✓	×	×	✓
M4	×	×	✓	×	×
M5	✓	✓	×	×	×

B. Ablation Study

The ablation study conducted on the E-Grasp and Neuro-Grasp datasets highlights the performance of different model configurations in object recognition and grasping position detection tasks, as shown in Table. III. For ablation study 5 different variations of the model are used, as shown in Table. II. Among the models evaluated, the proposed CSA-AIncepNet (M-1) consistently outperformed the others, achieving the highest accuracy, precision, recall, F1-score, and mean Intersection over Union (IoU) across all datasets. Specifically, in the E-Grasp dataset, CSA-AIncepNet demonstrated exceptional performance with an accuracy of 99.47% and a mean IoU of 0.9370, surpassing the other models by a notable margin. In the Neuro-Grasp dataset, the model's performance was similarly impressive, achieving 98.58% accuracy and a mean IoU of 0.4897 when transfer learning was applied. In contrast, the models that relied on simpler architectures like One-CSA and One-Inception (M-2) and Inception Network (M-4) showed lower performance, particularly on the Neuro-Grasp dataset without transfer learning, where accuracies dropped to around 70%. Additionally, models such as Spatial Attention-Inception (M-3) and Multi-CNN Network (M-5) exhibited reduced performance across both datasets, with M-5 particularly struggling, as indicated by its lower mean IoU values. The study demonstrates the superiority of CSA-AIncepNet in handling complex object recognition and grasping tasks,

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.



(a) Training Bounding Box Loss
(b) Validation Bounding Box Loss
(c) Training Classification Accuracy
(d) Validation Classification Accuracy
(e) Training Classification Loss
(f) Validation Classification Loss

Fig. 5. Best, Worst and Mean Curve of Training and Validation Losses and Accuracies after running the proposed model for 30 times.

TABLE III
COMBINED ABLATION STUDY ON E-GRASP AND NEURO-GRASP DATASETS

Dataset	Model	Accuracy	Precision	Recall	F1-Score	Mean IoU
E-Grasp	M-1	0.9947	0.9950	0.9948	0.9947	0.9370
	M-2	0.9942	0.9944	0.9942	0.9942	0.9079
	M-3	0.9750	0.9758	0.9750	0.9749	0.8967
	M-4	0.9835	0.9843	0.9835	0.9836	0.8976
	M-5	0.9750	0.9757	0.9750	0.9750	0.8498
Neuro-Grasp (No Transfer Learning)	M-1	0.7016	0.7022	0.7015	0.7014	0.3970
	M-2	0.6934	0.6928	0.6930	0.6931	0.3722
	M-3	0.6788	0.6779	0.6787	0.6784	0.3566
	M-4	0.6843	0.6839	0.6843	0.6842	0.3685
	M-5	0.6322	0.6326	0.6321	0.6322	0.3488
Neuro-Grasp (Transfer Learning)	M-1	0.9858	0.9855	0.9857	0.9855	0.4897
	M-2	0.9672	0.9670	0.9671	0.9670	0.4607
	M-3	0.9585	0.9483	0.9484	0.9483	0.4520
	M-4	0.9614	0.9612	0.9613	0.9612	0.4490
	M-5	0.9434	0.9435	0.9434	0.9435	0.4391

especially when leveraging transfer learning for the challenging Neuro-Grasp dataset. These results underscore the effectiveness of combining channel sharpening attention and adaptive inception networks for improving model robustness and generalization in diverse grasping scenarios. The proposed model is run 30 times to check the performance of the model for more generalization. These runs' training, validation losses and accuracies are shown in Fig. 5. The visualisation of Predicted vs True Classes and Grasping bounding boxes for the proposed model is shown in Fig. 6.

C. Comparison With Different Loss Function Weights

In the analysis of the performance across different loss weights for multi-task learning, the results demonstrate the

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

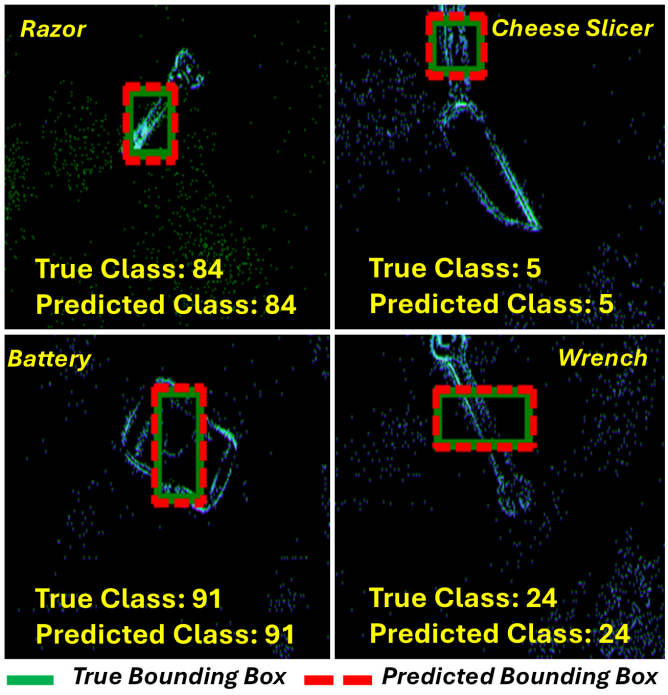


Fig. 6. Visualisation of Predicted vs True Object Class and the predicted vs True Bounding boxes of grasping position for the proposed model.

TABLE IV
CONFIGURATIONS OF DIFFERENT CLASSIFICATION AND REGRESSION LOSS WEIGHTS FOR MULTI-TASK LEARNING (CL IS FOR CLASSIFICATION LOSS WEIGHT AND RL IS FOR REGRESSION LOSS WEIGHT)

C-1	CL=1; RL=1
C-2	CL=1; RL=0.5
C-3	CL=2; RL=0.5
C-4	CL=0.5; RL=1
C-5	CL=0.5; RL=2
C-6	CL=0.5; RL=0.5

impact of varying the classification (CL) and regression (RL) loss weights on model performance, as shown in Table. IV. The analysis of different models on different loss function weights for E-Grasp dataset is shown in Table. V and for Neuro-Grasp dataset is shown in Table. VI. For the E-Grasp dataset, model M-1 (CSA-AIncepNet) consistently achieved high accuracy and mean IoU across all loss weight configurations, with the best performance observed in C-1 (CL=1, RL=1), where the model reached 99.42% accuracy and a mean IoU of 0.9151. This indicates a balanced approach to both classification and regression tasks. Interestingly, the performance of model M-2 (One-CSA and One-Inception) was also robust, achieving the highest accuracy (99.45%) in C-2 (CL=1, RL=0.5) with a mean IoU of 0.8957, suggesting that slightly lowering the regression loss weight can yield better performance for some models. Models M-3 (Spatial Attention-Inception), M-4 (Inception Network), and M-5 (Multi-CNN Network) showed varied performance across configurations, with M-3 generally performing lower on accuracy (83.49%) in C-1 and improving as the loss weights were adjusted. For the Neuro-Grasp dataset after transfer learning, M-1 achieved the best accuracy (98.58%) and mean IoU (0.4897) in C-1, further confirming that a balanced loss weight approach is effective. However, as the regression loss weight increased relative to classification (e.g., in C-5 and C-6), accuracy and mean IoU dropped, particularly for models such as M-5, which struggled to adapt to these changes. These results suggest that the optimal configuration for multi-task learning may depend on the model and dataset, with more balanced loss weights generally leading to superior performance.

D. Comparison of Multi-Task Versus Separate Tasks

We have done a detailed comparative analysis of the proposed Multi-Task Model with the Separate Training of Each Task. The evaluation Matrices and the Size of the models is presented in Table. VII. The results demonstrate that dedicated single-task models consistently outperform the multi-task approach on individual tasks, highlighting the inherent trade-offs in multi-task learning architectures. On the E-Grasp dataset, the single-task object recognition model achieves 99.68% accuracy compared to 99.47% for the multi-task version, while the dedicated grasping model reaches 94.93% IoU versus 93.70% for the multi-task approach. This performance gap becomes more pronounced on the challenging Neuro-Grasp dataset, where single-task models show substantial improvements (74.23% vs 70.16% accuracy, and 44.21% vs 39.70% IoU without transfer learning), indicating that task interference is more severe when dealing with complex, cross-domain data. However, the multi-task model provides significant computational efficiency gains, requiring only 14.2M parameters compared to 16.3M parameters if both single-task models (12.7M + 3.6M) were deployed separately, representing a 13% parameter reduction. The transfer learning results on Neuro-Grasp reveal dramatic improvements for both approaches, with the multi-task model achieving 98.58% accuracy and single-task models reaching 99.04% accuracy, demonstrating that while task interference persists even with pre-training, the performance gap narrows considerably, making the multi-task approach more attractive for practical deployment scenarios where computational efficiency and unified inference pipelines are prioritized over marginal performance gains on individual tasks.

E. Comparison With Different Models

The experimental results on the E-Grasp dataset, as shown in Table. VIII, demonstrate that the proposed CSA-AIncepNet model outperforms state-of-the-art methods across all evaluation metrics. In the classification task, CSA-AIncepNet achieved the highest accuracy (99.47%), precision (99.50%), recall (99.48%), and F1-score (99.47%), showcasing its superior capability to accurately recognize objects. Additionally, the model significantly outperformed competitors in bounding box detection, achieving the highest Mean IoU (93.7%) and exhibiting minimal localization error (0.629). These results highlight the effectiveness of the CSA-AIncepNet architecture in simultaneously addressing object recognition and grasping position detection tasks, proving its robustness and efficiency

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

TABLE V

PERFORMANCE METRICS (ACCURACY AND MEAN IOU) OF MODELS ACROSS DIFFERENT LOSS WEIGHTS ON E-GRASP DATASET

Model	**C-1**		**C-2**		**C-3**		**C-4**		**C-5**		**C-6**	
	Acc.	**Mean IoU**	**Acc.**	**Mean IoU**	**Acc.**	**Mean IoU**	**Acc.**	**Mean IoU**	**Acc.**	**Mean IoU**	**Acc.**	**Mean IoU**
M-1	0.9942	0.9151	0.9956	0.9118	0.9937	0.9111	0.9901	0.9088	0.9948	0.9371	0.9945	0.9125
M-2	0.9945	0.8878	0.9931	0.8957	0.9937	0.9103	0.9934	0.8966	0.9942	0.9080	0.9942	0.9044
M-3	0.8349	0.8610	0.9931	0.8874	0.9766	0.8871	0.9527	0.8980	0.9750	0.8967	0.9714	0.8919
M-4	0.9874	0.8914	0.9687	0.8859	0.9871	0.8859	0.9794	0.8855	0.9835	0.8977	0.9854	0.8930
M-5	0.9904	0.8462	0.9857	0.8449	0.9901	0.8463	0.9918	0.8395	0.9750	0.8498	0.9791	0.8409

TABLE VI

PERFORMANCE METRICS (ACCURACY AND MEAN IOU) OF MODELS ACROSS DIFFERENT LOSS WEIGHTS ON NEURO-GRASP DATASET AFTER TRANSFER LEARNING

Model	**C-1**		**C-2**		**C-3**		**C-4**		**C-5**		**C-6**	
	Acc.	**Mean IoU**	**Acc.**	**Mean IoU**	**Acc.**	**Mean IoU**	**Acc.**	**Mean IoU**	**Acc.**	**Mean IoU**	**Acc.**	**Mean IoU**
M-1	0.9562	0.4750	0.9611	0.4775	0.9657	0.4797	0.9462	0.4605	0.9858	0.4897	0.9739	0.4841
M-2	0.9382	0.4469	0.9435	0.4492	0.9480	0.4530	0.9091	0.4349	0.9672	0.4607	0.9181	0.4401
M-3	0.9297	0.4184	0.9346	0.4410	0.9393	0.4451	0.9018	0.4262	0.9585	0.4520	0.9126	0.4313
M-4	0.9326	0.4255	0.9375	0.4380	0.9421	0.4429	0.9041	0.4241	0.9614	0.4490	0.9150	0.4292
M-5	0.9151	0.4059	0.9197	0.4282	0.9246	0.4323	0.8876	0.4147	0.9434	0.4391	0.8982	0.4197

TABLE VII

EVALUATION OF MULTI-TASK MODEL WITH SEPARATE TASK MODELS

Dataset	Model	Accuracy	Mean IoU	Parameters
E-Grasp	Proposed	0.9947	0.9370	14.2 M
	Object Recognition	0.9968	×	12.7 M
	Grasping Position	×	0.9493	3.6 M
Neuro-Grasp (Without TL)	Proposed	0.7016	0.3970	14.2 M
	Object Recognition	0.7423	×	12.7 M
	Grasping Position	×	0.4421	3.6 M
Neuro-Grasp (With TL)	Proposed	0.9858	0.4897	14.2 M
	Object Recognition	0.9904	×	12.7 M
	Grasping Position	×	0.5399	3.6 M

TABLE VIII

COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART ON E-GRASP DATASET

	Classification Matrices				Bounding Box Matrices		Jaccard Accuracy for Bounding Boxes			
Model	Accuracy	Precision	Recall	F1-Score	Mean IoU	LE	25%	30%	35%	40%
ConvNeXt [41]	0.8824	0.9093	0.8824	0.8882	0.724	0.276	0.9761	0.9676	0.9552	0.9396
DarkNet [42]	0.9813	0.983	0.9813	0.9816	0.8273	0.1727	0.9951	0.9937	0.9909	0.9868
DenseNet [43]	0.9376	0.9406	0.9376	0.9384	0.635	0.365	0.9552	0.9313	0.9058	0.8673
VGG16 [44]	0.9813	0.9822	0.9813	0.9814	0.8263	0.1737	0.9973	0.9956	0.9929	0.9893
CSA-AIncepNet(Proposed)	0.9947	0.995	0.9948	0.9947	0.937	0.0629	0.9996	0.9991	0.9988	0.9978

TABLE IX

COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART ON NEURO-GRASP DATASET WITH TRANSFER LEARNING

	Classification Matrices				Bounding Box Matrices		Jaccard Accuracy for Bounding Boxes				
Model	Accuracy	Precision	Recall	F1-Score	Mean IoU	LE	25%	30%	35%	40%	Parameters
ConvNeXt [41]	0.8911	0.8912	0.8910	0.8912	0.3688	0.6312	0.3959	0.3688	0.3138	0.2891	14.2M
DarkNet [42]	0.9559	0.956	0.9598	0.9559	0.0.3279	0.6721	0.0.3871	0.3408	0.3144	0.2867	16.4M
DenseNet [43]	0.9142	0.9141	0.9138	0.9141	0.2978	0.7022	0.3401	0.2933	0.2584	0.2304	15.9M
VGG16 [44]	0.9736	0.9735	0.9732	0.9736	0.4298	0.5702	0.412	0.3588	0.3105	0.2944	13.5M
CSA-AIncepNet(Proposed)	0.9858	0.9855	0.9857	0.9855	0.4897	0.5103	0.5644	0.4956	0.4123	0.3705	15.7M

compared to existing methods such as DarkNet, DenseNet, ConvNeXt, and VGG16.

On the Neuro-Grasp dataset, as shown in Table. IX, where transfer learning was employed, CSA-AIncepNet maintained its superiority, achieving the best performance across classification and bounding box metrics. The model recorded an accuracy of 98.58%, precision of 98.55%, recall of 98.57%, and F1-score of 98.55%, which are significantly higher than other techniques. For bounding box detection, CSA-AIncepNet achieved a Mean IoU of 48.97% and demonstrated improved localization performance with a lower error rate compared to competing models. These results illustrate the model's ability to generalize effectively to new datasets through transfer learning, outperforming established methods such as VGG16, DarkNet, and ConvNeXt. The findings validate the architectural design choices of CSA-AIncepNet, making it a robust solution for multi-task learning in object recognition and grasping detection.

F. Discussion

The results from the ablation study and comparative analysis clearly demonstrate the superior performance of CSA-AIncepNet in object recognition and grasping position detection tasks. On the E-Grasp dataset, the proposed model outperformed all baseline and state-of-the-art models in both

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

classification and bounding box metrics. The incorporation of channel sharpening attention and adaptive inception layers proved instrumental in enhancing model accuracy and robustness. Notably, CSA-AIncepNet achieved an accuracy of 99.47% and a mean IoU of 0.9370, reflecting its ability to precisely localize and classify objects even in complex scenarios. In contrast, models such as Spatial Attention-Inception and Multi-CNN Network struggled to achieve comparable performance, highlighting the limitations of simpler architectures in capturing intricate object features.

On the Neuro-Grasp dataset, CSA-AIncepNet maintained its edge, particularly when transfer learning was applied. The model achieved an impressive accuracy of 98.58% and a mean IoU of 0.4897, significantly outperforming competing models. These results underscore the model's strong generalization capability and its ability to adapt learned features to a new dataset. Other models, such as DenseNet and VGG16, showed a marked drop in performance on Neuro-Grasp, particularly in bounding box metrics, where CSA-AIncepNet exhibited superior localization accuracy. The findings validate the architectural design of CSA-AIncepNet, which integrates attention mechanisms and adaptive feature extraction to handle the challenges of multi-task learning effectively.

Our experiments demonstrated that event-based cameras provide substantial benefits for neural network processing beyond the commonly cited advantages of reduced power consumption and storage requirements. The sparse, binary nature of event data significantly streamlined our preprocessing pipeline by eliminating complex operations such as illumination normalization and color space transformations that are typically required for conventional frame-based imagery, thus reducing computational bottlenecks during both training and inference. Event streams inherently highlight object boundaries and moving features, providing our network with higher information density per input sample, which enabled our model to achieve convergence with fewer training iterations compared to frame-based alternatives. Furthermore, the continuous temporal nature of event data facilitated more stable feature tracking across sequential inputs, improving the network's ability to maintain consistent grasp point predictions despite object motion or viewpoint changes. The binary, sparse representation of events substantially reduced GPU memory requirements during the training phase, allowing us to implement deeper network architectures and larger batch sizes than would be feasible with conventional RGB input data. Perhaps most significantly, when transferring our pretrained model from the E-Grasp to the Neuro-Grasp dataset, we observed that the domain gap was markedly smaller than typically encountered in RGB-based systems, requiring fewer fine-tuning epochs to achieve comparable performance. This suggests that event-based representations may offer more generalizable features for robotic manipulation tasks, potentially reducing the data collection burden for deployment in novel environments.

IV. Conclusion

In this paper, we proposed the Channel Sharpening Attention-based Adaptive Inception Network (CSA-AInceptNet), a novel multi-task learning model for object recognition and grasping position detection using event camera data. The integration of channel sharpening attention and adaptive inception networks demonstrated significant improvements in feature extraction and task performance. The model was rigorously evaluated on two state-of-the-art event camera datasets, E-Grasp and Neuro-Grasp, achieving exceptional results. On the E-Grasp dataset, CSA-AIncepNet achieved a classification accuracy of 99.47% and a mean IoU of 0.9370, outperforming existing methods by a substantial margin. On the Neuro-Grasp dataset, with transfer learning, the model demonstrated strong generalization capabilities, achieving 98.58% accuracy and a mean IoU of 0.4897, further validating its robustness.

Comparative analyses and ablation studies highlighted the superiority of CSA-AIncepNet over established architectures such as ConvNeXt, DarkNet, DenseNet, and VGG16, as well as simpler baseline configurations. The results underscore the importance of incorporating advanced attention mechanisms and adaptive feature extraction techniques in multi-task learning models for event camera data. This work establishes CSA-AIncepNet as a robust and effective solution for dynamic object recognition and grasping tasks, providing a solid foundation for future advancements in robotic manipulation, human-robot collaboration, and autonomous systems.

References

[1] F. Sanfilippo, H. Zhang, K. Y. Pettersen, G. Salvietti, and D. Prattichizzo, "ModGrasp: An open-source rapid-prototyping framework for designing low-cost sensorised modular hands," in *Proc. 5th IEEE RAS/EMBS Int. Conf. Biomed. Robot. Biomechatronics*, Aug. 2014, pp. 951–957.

[2] F. Sanfilippo, H. Zhang, and K. Y. Pettersen, "The new architecture of ModGrasp for mind-controlled low-cost sensorised modular hands," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2015, pp. 524–529.

[3] S. K. R. Moosavi, M. H. Zafar, and F. Sanfilippo, "A review of the state-of-the-art of sensing and actuation technology for robotic grasping and haptic rendering," in *Proc. 5th Int. Conf. Inf. Comput. Technol. (ICICT)*, Mar. 2022, pp. 182–190.

[4] B. Chakravarthi, A. A. Verma, K. Daniilidis, C. Fermüller, and Y. Yang, "Recent event camera innovations: A survey," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 342–376.

[5] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, nos. 4–5, pp. 705–724, Apr. 2015.

[6] D. Liu, X. Tao, L. Yuan, Y. Du, and M. Cong, "Robotic objects detection and grasping in clutter based on cascaded deep convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, 2022.

[7] B. Cheng, W. Wu, D. Tao, S. Mei, T. Mao, and J. Cheng, "Random cropping ensemble neural network for image classification in a robotic arm grasping system," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6795–6806, Sep. 2020.

[8] G. Chen, K. Chen, L. Zhang, L. Zhang, and A. Knoll, "VCANet: Vanishing-Point-Guided context-aware network for small road object detection," *Automot. Innov.*, vol. 4, no. 4, pp. 400–412, Nov. 2021.

[9] H. Zhang, X. Zhou, X. Lan, J. Li, Z. Tian, and N. Zheng, "A real-time robotic grasping approach with oriented anchor box," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 51, no. 5, pp. 3014–3025, May 2021.

[10] H. Cao, G. Chen, Z. Li, J. Lin, and A. Knoll, "Lightweight convolutional neural network with Gaussian-based grasping representation for robotic grasping detection," 2021, *arXiv:2101.10226*.

[11] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3304–3311.

[12] A. D. Vuong et al., "Language-driven grasp detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17902–17912.

[13] Y. Zhang et al., "DSNet: Double strand robotic grasp detection network based on cross attention," *IEEE Robot. Autom. Lett.*, vol. 9, no. 5, pp. 4702–4709, May 2024.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

[14] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3355–3362, Oct. 2018.

[15] H. Cao, G. Chen, Z. Li, J. Lin, and A. Knoll, "Residual squeeze-and-excitation network with multi-scale spatial pyramid module for fast robotic grasping detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13445–13451.

[16] L. Tong, K. Song, H. Tian, Y. Man, Y. Yan, and Q. Meng, "A novel RGB-D cross-background robot grasp detection dataset and background-adaptive grasping network," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–15, 2024.

[17] G. Gallego et al., "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.

[18] J. Mahler et al., "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017, *arXiv:1703.09312*.

[19] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, Feb. 2008.

[20] E. Johns, S. Leutenegger, and A. J. Davison, "Deep learning a grasp function for grasping under gripper pose uncertainty," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 4461–4468.

[21] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int. J. Robot. Res.*, vol. 39, nos. 2–3, pp. 183–201, Mar. 2020.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[23] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "ROI-based robotic grasp detection for object overlapping scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4768–4775.

[24] L. Chen, P. Huang, and Z. Meng, "Convolutional multi-grasp detection using grasp path for RGBD images," *Robot. Auto. Syst.*, vol. 113, pp. 94–103, Mar. 2019.

[25] D. Park, Y. Seo, and S. Y. Chun, "Real-time, highly accurate robotic grasp detection using fully convolutional neural network with rotation ensemble module," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 9397–9403.

[26] U. Asif, J. Tang, and S. Harrer, "Densely supervised grasp detector (DSGD)," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8085–8093.

[27] G. Wu, W. Chen, H. Cheng, W. Zuo, D. Zhang, and J. You, "Multi-object grasping detection with hierarchical feature fusion," *IEEE Access*, vol. 7, pp. 43884–43894, 2019.

[28] D.-C. Hoang et al., "Collision-free grasp detection from color and depth images," *IEEE Trans. Artif. Intell.*, vol. 5, no. 11, pp. 5689–5698, Nov. 2024.

[29] X. Wang et al., "Object detection using event camera: A MoE heat conduction based detector and a new benchmark dataset," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, Jun. 2025, pp. 29321–29330.

[30] J. Cao, X. Zheng, Y. Lyu, J. Wang, R. Xu, and L. Wang, "Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 9026–9032.

[31] Y. Peng, H. Li, Y. Zhang, X. Sun, and F. Wu, "Scene adaptive sparse transformer for event-based object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16794–16804.

[32] F. Mirus, C. Axenie, T. C. Stewart, and J. Conradt, "Neuromorphic sensorimotor adaptation for robotic mobile manipulation: From sensing to behaviour," *Cognit. Syst. Res.*, vol. 50, pp. 52–66, Aug. 2018.

[33] R. Muthusamy, X. Huang, Y. Zweiri, L. Seneviratne, and D. Gan, "Neuromorphic event-based slip detection and suppression in robotic grasping and manipulation," *IEEE Access*, vol. 8, pp. 153364–153384, 2020.

[34] F. B. Naeini et al., "A novel dynamic-vision-based approach for tactile sensing applications," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 1881–1893, May 2020.

[35] B. Li, H. Cao, Z. Qu, Y. Hu, Z. Wang, and Z. Liang, "Event-based robotic grasping detection with neuromorphic vision sensor and event-grasping dataset," *Frontiers Neurorobotics*, vol. 14, p. 51, Oct. 2020.

[36] H. Cao, G. Chen, Z. Li, Y. Hu, and A. Knoll, "NeuroGrasp: Multimodal neural network with Euler region regression for neuromorphic vision-based grasp pose estimation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[37] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.

[38] A. Bitton, H. C. Duwek, and E. E. Tsur, "Adaptive attention with a neuromorphic hybrid frame and event-based camera," in *Proc. IEEE 21st Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI*CC)*, Dec. 2022, pp. 242–247.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[40] T.-K. Yen et al., "Tracking-assisted object detection with event cameras," 2024, *arXiv:2403.18330*.

[41] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986.

[42] R. Ningthoujam, K. Pritamdas, and L. S. Singh, "Edge detective weights initialization on darknet-19 model for YOLOv2-based facemask detection," *Neural Comput. Appl.*, vol. 36, no. 35, pp. 22365–22378, Dec. 2024.

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[44] D. Theckedath and R. R. Sedamkar, "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks," *Social Netw. Comput. Sci.*, vol. 1, no. 2, p. 79, Mar. 2020.

Muhammad Hamza Zafar (Graduate Student Member, IEEE) received the bachelor's degree in electronics engineering and the master's degree in electrical engineering from the Capital University of Science and Technology, Islamabad, Pakistan, in 2015 and 2022, respectively. He is currently pursuing the Ph.D. degree in robotics and AI with the Department of Engineering Sciences, University of Agder, Grimstad, Norway. His research interests include robotics, machine vision, human–robot teaming, and swarm optimization. He is an IEEE Student Section Representative of Norway Section. He is a reviewer of several international conferences and journals.

Syed Kumayl Raza Moosavi received the bachelor's degree in mechatronics engineering and the master's degree in artificial intelligence from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2016 and 2023, respectively. He is currently pursuing the Ph.D. degree in robotics and AI with the Department of Engineering Sciences, University of Agder, Grimstaad, Norway. His research interests include robotics, artificial intelligence, and disaster management. He is a reviewer of several international conferences and journals.

Filippo Sanfilippo (Senior Member, IEEE) received the Ph.D. degree in engineering cybernetics from Norwegian University of Science and Technology (NTNU), Norway, with a focus on intelligent control approaches for robotic manipulators. He is currently appointed as a Professor at the Faculty of Engineering and Science, University of Agder (UiA), Grimstad, Norway. He has a vast experience in participating in European research programs and various national projects from the Research Council of Norway (RCN). He has authored and co-authored several technical papers in various journals and conferences. His research interests include robotics, wearables, human–robot teaming, artificial intelligence, and control theory. He is a member of the IEEE Region 8 Chapter Coordination Committee, the Conference Coordination Committee, the IEEE Public Visibility Committee, the IEEE R8 Awards and Recognitions Committee, and the Professional and Educational Activities Committee. He is the former Chair of the IEEE Norway Section. He is the Chair of the IEEE Robotics and Automation, Control Systems, and Intelligent Transportation Systems Joint Chapter. He is the Chair of Norway Section Life Members Affinity Group. He is a reviewer of several international conferences and journals.