NeuroGrasp: Multimodal Neural Network With Euler Region Regression for Neuromorphic Vision-Based Grasp Pose Estimation

Hu Cao, Guang Chen, Zhijun Li, *Fellow, IEEE*, Yingbai Hu, and Alois Knoll, *Senior Member, IEEE*

***Abstract*— Grasp pose estimation is a crucial procedure in robotic manipulation. Most of the current robot grasp manipulation systems are built on frame-based cameras like RGB-D cameras. However, the traditional frame-based grasp pose estimation methods have encountered challenges in scenarios such as low dynamic range and low power consumption. In this work, a neuromorphic vision sensor—dynamic and active-pixel vision sensor (DAVIS)—is introduced to the field of robotic grasp. DAVIS is an event-based bio-inspired vision sensor that records asynchronous streams of local pixel-level light intensity changes, called events. The strengths of DAVIS are it can provide high temporal resolution, high dynamic range, low power consumption, and no motion blur. We construct a neuromorphic vision-based robotic grasp dataset with 154 moving objects, named NeuroGrasp, which is the first RGB-Event multimodality grasp dataset (to the best of our knowledge). This dataset records both RGB frames and the corresponding event streams, providing frame data with rich color and texture information and event streams with high temporal resolution and high dynamic range. Based on the NeuroGrasp dataset, we further develop a multimodal neural network with a specific Euler region regression sub-network (ERRN) to perform grasp pose estimation. Combined with frame-based and event-based vision, the proposed method achieves better performance than the method that only takes RGB frames or event streams as input on the NeuroGrasp dataset.**

***Index Terms*— Euler region regression sub-network (ERRN), grasp pose estimation, multimodal fusion, vision-based robotic manipulation.**

I. Introduction

GRASP pose estimation plays an important role in robotic manipulation. The emergence of advanced sensors, such as Microsoft Kinect, has enriched robot perception systems. In recent years, deep-learning-based methods have been widely applied in robotic manipulation [1]–[5]. The success of deep learning has driven approaches that leverage large volumes of training data to perform complex tasks [2], [6]. However, grasp datasets collected in the physical environment are relatively scarce. Dexnet [7] has explored the use of simulated data in grasp pose estimation to alleviate this problem. Another challenge is maintaining a balance between computational cost and the power available within embedded robot systems. Current state-of-the-art robotic grasp manipulation systems [8]–[10] usually leverage frame-driven RGB-D cameras as the perception sensors. The traditional frame-driven cameras capture the environmental information by generating a series of discrete frames at a fixed frequency, providing rich color and texture information. However, frame-based cameras suffer the challenges of high computing time and storage consumption [11]. In this article, we build a dynamic sensing pipeline using a neuromorphic vision sensor: dynamic and active-pixel vision sensor (DAVIS-346). DAVIS is a camera model which consists of a dynamic vision sensor (DVS) (event-based sensor) synchronized with an RGB frame-based sensor. DAVIS can synchronously record RGB data and the corresponding event streams. Specifically, it only transmits the local pixel-level changes caused by the change in lighting intensity within a scene at the time they occur, like a bio-inspired retina [12]. Concretely, the change in light intensity is very effective for detecting moving objects. Fig. 1 presents a comparison between conventional images and the corresponding event frames. Events are timestamped with the precision of around a microsecond. A single event is defined as the tuple $\{t, x, y, p\}$, where t is the timestamp of the event, x and y are the pixel coordinates of the event in 2-D space, and $p = \pm 1$ is the polarity of the event which is the sign of the brightness change. Compared with frame-based cameras, the neuromorphic vision sensors have properties that are complementary to RGB sensors, including very high temporal resolution, high dynamic range (120 dB), and low power consumption [13]. In our previous work [14], we introduced an event-based grasping dataset (E-Grasping) for robotic grasp pose estimation. However, the E-Grasping dataset only records event streams and contains fewer grasp objects. In this work, we use DAVIS as a perception sensor to construct a more challenging dataset in the practical environment. This dataset includes both RGB data and the corresponding event streams.

Manuscript received December 10, 2021; revised April 30, 2022; accepted May 7, 2022. Date of publication June 2, 2022; date of current version June 14, 2022. This work was supported in part by the Shanghai Municipal Science and Technology Major Project through the Zhejiang Lab and the Shanghai Center for Brain Science and Brain-Inspired Technology under Grant 2018SHZDZX01; in part by the National Natural Science Foundation of China under Grant 82072021 and Grant 61906138; in part by the Shanghai Rising Star Program under Grant 21QC1400900; and in part by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement 945539 (Human Brain Project SGA3). The Associate Editor coordinating the review process was Dr. Qing Wang. *(Corresponding author: Guang Chen.)*

Hu Cao, Yingbai Hu, and Alois Knoll are with the Chair of Robotics, Artificial Intelligence and Real-Time Systems, Technische Universität München, 80333 Munich, Germany.

Guang Chen is with the Department of Computer Science and the School of Automotive Studies, Tongji University, Shanghai 201804, China (e-mail: guangchen@tongji.edu.cn).

Zhijun Li is with the Wearable Robotics and Autonomous Systems Laboratory, University of Science and Technology of China, Hefei 230022, China.

Digital Object Identifier 10.1109/TIM.2022.3179469

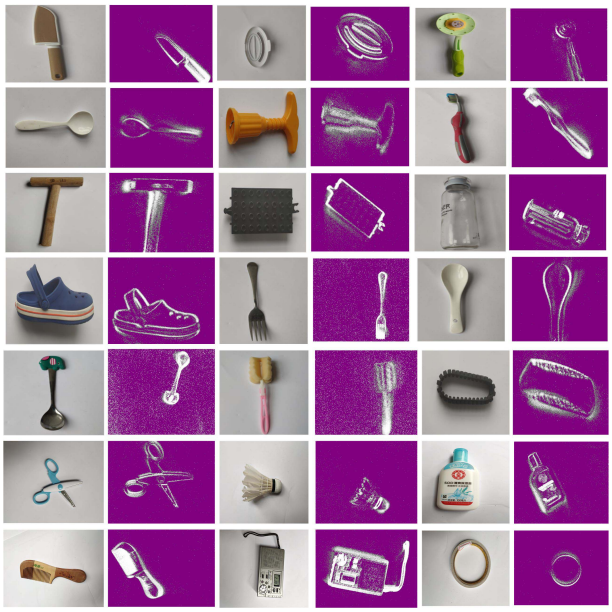


Fig. 1. Samples from NeuroGrasp dataset: a list of the selected RGB images and the corresponding event frames.

Early works for robotic grasp mainly rely on template matching to perform grasp pose estimation. In unstructured environments where objects vary in shape and appearance, template-matching algorithms cannot work effectively. Taking 2-D images instead of the 3-D model as input is more convenient to predict grasp pose [1], [15]. Based on 2-D images, many researchers have applied deep convolutional neural networks in robotic grasp pose estimation and achieved great success. In [1], a sliding window detection framework is used for 2-D robotic grasp pose estimation. Specifically, image sequences are fed into convolutional neural network to extract features, and the highest output confidence score of all grasp candidates is chosen as final prediction results. The drawback of this method is the high computation cost. To speed up these algorithms, end-to-end methods are developed [9], [16]–[18]. Concurrently, the authors take RGB or RGB-D images as input to perform regression or classification on grasp rectangles and achieve significant improvements on Cornell Grasping Dataset [15]. Compared with the conventional frame-based grasping, neuromorphic vision-based grasping is still in its infancy.

For event-based robotic grasp pose estimation, it faces two main problems: lack of data and effective algorithms. To cope with these challenges, we collected a manually labeled multimodality (RGB-Event) robotic grasping dataset, NeuroGrasp dataset, and developed a multimodal neural network to explore how to fuse the valuable feature context of RGB frames and events to improve the performance. Specifically, with the use of frequency-based encoding method [19], events generated from DAVIS can be fed into convolutional neural networks for subsequent grasp pose prediction. To take advantage of DAVIS, we use convolutional filters to fuse the valuable feature context of events and RGB images to improve the prediction performance. Furthermore, an Euler region regression sub-network (ERRN) is introduced to predict the orientation of grasp objects by adding an imaginary and a real fraction to the regression network. This strategy builds a closed mathematical space to avoid singularities that may occur in single-angle estimation [20]. The experimental results show that the proposed method achieves better performance than the method that only takes a single-mode signal as input.

Our main contributions can be summarized as follows.

1) We collect an RGB-Event multimodality grasp dataset named NeuroGrasp from a real-world experiment environment, which will promote the research on neuromorphic vision sensors for robotic grasp pose estimation.
2) We develop a novel multimodal neural network to fuse the valuable feature context of RGB images and events to improve the performance. An ERRN is also introduced for more accurate pose estimation.
3) Extensive experiments on the E-Grasping and NeuroGrasp datasets demonstrate that the proposed method outperforms the method that only takes RGB frames or event streams as input.

The main content of this article will cover seven parts. Section II briefly reviews related works about grasping datasets, frame-based, and event-based grasp pose estimation methods. Section III illustrates the setup of neuromorphic robot manipulation system. Section IV describes the details of our grasp pose estimation architecture. Section V presents the dynamic grasping dataset, and Section VI gives the experimental results and analysis on two dynamic grasping datasets, E-Grasping and NeuroGrasp. Finally, we conclude our work in Section VII.

II. Related Work

A. Datasets

At present, the Cornell grasp dataset [15] and Dex-Net dataset [7] are collected for analyzing grasp quality with parallel plate gripper (PPG). The Cornell Grasp dataset recorded with an RGB-D camera consists of 885 images of 280 different objects. It is widely used by researchers and greatly contributes to the robotic grasp research field. The grasp dataset demonstrates 8019 labeled grasp rectangles, including several good grasp positions (5110) and bad grasp positions (2909) for each view of an object. The point cloud data and background image of each image are also provided. The Dex-Net dataset is collected by UC Berkeley Automation Lab [7]. Dex-Net provides synthetic point clouds and grasp annotations based on 3-D objects and has been extended to three versions of Dex-Net 1.0, Dex-Net 2.0, and Dex-Net 3.0. Dex-Net 1.0 includes over 10 000 different 3-D object models and contains about 2.5 million grasp labels, and Dex-Net 2.0 is a dataset of more than 6.7 million synthetic point clouds and corresponding labels. Since Dex-Net 3.0 is built for studying suction grasp, we do not describe it in detail in this article. Moreover, a simulated dataset, named the Jacquard Dataset [21], is created from CAD models through simulation. In this dataset, more than 50k images of 11k objects are collected and 1 million unique grasp rectangles are labeled. However, these datasets are all focused on RGB-D data. In our early work [14], we constructed an event stream-based

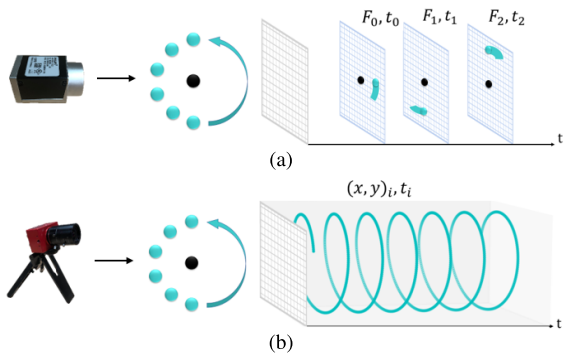


Fig. 2. Comparison of output between standard frame-based camera and neuromorphic vision sensor. (a) Frame-based camera captures images at a fixed frame rate. (b) Neuromorphic vision sensor captures emitted events caused by moving objects asynchronously.

grasping dataset (E-Grasping) using an event-based dynamic active vision sensor (DAVIS). Using an SMP filter to track led markers, all objects are labeled automatically. The disadvantage of this dataset is that it only records the event streams for grasp objects. In this work, we build a more challenging multimodality grasp dataset with more grasp objects.

B. Frame-Based Grasp Pose Estimation

Research on robotic grasp pose estimation has made important advances over the past 20 years. Early works [1], [22] trained grasp detector based on the sliding window, which is very time-consuming. In [23] and [24], the authors reduced inference time by learning their methods on a discrete set of grasp candidates. However, these approaches ignore some potential grasps. Other methods like [8], [25] used end-to-end CNN-based algorithms to regress a single grasp for an input image, but these approaches tend to estimate the average grasp pose of objects. In [9], a grasp region proposal network is incorporated for grasp pose estimation based on Faster RCNN [26]. Furthermore, the authors of [16] proposed a single-stage real-time grasp network with the orientation anchor box mechanism, which achieves the outstanding performance of both speed and accuracy. For object overlapping scenes, an ROI-based method is developed in [27]. The experimental results showed that their algorithm can effectively deal with object overlapping scenes. Since the ground truths in the grasp pose are not exhaustive, Chen *et al.* [28] introduced a grasp path to generate mapped grasp for convolutional multigrasp prediction which improved grasp accuracy in real-world scenarios. In [29], the authors presented a highly accurate and real-time grasp detection system with a rotation ensemble module (REM). Some ideas of this network design are inspired by YOLO9000 [30]. Another works [10], [18], [31] deployed the neural network to generate grasps with high-resolution images. Their model solves the problem of pixel-wise robotic grasp pose estimation. Moreover, [17] and [32] used the fusion method to perform grasp prediction and achieved better performance. However, the above methods usually take RGB or RGB-D images as input to perform regression or classification on grasp rectangles, and we will explore the potential application of robotics by focusing on neuromorphic vision sensors (DAVIS346).

C. Event-Based Grasp Pose Estimation

Recently, the development of event-based neuromorphic vision technology provides an alternative sensing scheme for many vision fields. Some attempts have been made in the field of object detection [19], [33]. For robotic grasp, a method including perception, reasoning, and control is proposed to solve the problem of picking and placing in mobile robots [34]. Based on embedded DVS, this method can pick up the object and move it to its correct position. In [35], the authors proposed a dynamic vision-based finger system for slip detection and suppression. This fingering system can detect object slips better under illumination and vibration with a threshold algorithm. For vision-based measurement applications, a dynamic-vision-based approach for tactile sensing is introduced in [4]. Furthermore, the authors of [14] constructed an Event-based Dataset and developed an event-based deep neural network to predict grasp pose. However, compared with the conventional frame-based vision, neuromorphic vision is still in its infancy and generally offers a lower spatial resolution.

In this work, we introduce a multimodal neural network to perform robotic grasp pose estimation, which combines frame-based vision and event-based vision. We evaluate our model on two dynamic robotic grasping datasets, E-Grasping and NeuroGrasp. The experimental results demonstrate that our model is capable of predicting exactly grasping rectangular shapes.

III. Neuromorphic Grasping System

A. Neuromorphic Vision Sensor

A neuromorphic vision sensor is a bio-inspired sensor, which mimics the working principle of biological neurons found in the visual cortex of mammalians [12]. The traditional frame-based vision cameras sense the environment by producing a series of frames that sample the light intensity at discrete time intervals. Neuromorphic vision sensors record asynchronous event streams of the change in light intensity of a given pixel. It allows the sensor to measure the per-pixel changes caused by motion in a scene at the time of occurrence. The difference between the two hardware systems is presented in Fig. 2. A stream of sparse spatial–temporal events can be represented by $e_i(x_i, y_i, t_i, p_i)_{i\in[1,N]}$, which means that an event is triggered at pixel location $l_i = (x_i, y_i)$ when the intensity change at a pixel occurs, i.e.,

$$\Delta L(l_i, t_i) = L(l_i, t_i) - L(l_i, t_i - \Delta t_i) \tag{1}$$

where $L(\cdot)$ is the brightness log function, and Δt_i is the time interval between the current event and the last event at the same pixel. Specifically, the temporal contrast threshold $\pm T$ ($T > 0$) is set for the intensity change to be reached

$$\Delta L(l_i, t_i) = p_i T \tag{2}$$

where $p \in \{+1, -1\}$ is the polarity of event, which represents the brightness change. $p = +1$ denotes the increase in brightness intensity and -1 denotes the decrease. As an emerging bio-inspired vision sensor, event-based neuromorphic vision sensors have several promising properties—low energy consumption, low latency, high dynamic range, and

high temporal resolution. In this work, we will explore the potential of neuromorphic vision sensors in the field of robotic grasp pose estimation.

B. System Setting

A neuromorphic vision sensor (DAVIS 346) collects the event data through lighting intensity changing, so the object needs to maintain movement within the field of view. The DAVIS 346 sensor is attached to the gripper of a robot arm (hand–eye system) to simulate the real trajectory during grasping. The PPG is widely mounted on the end of the robot arm, and our grasping dataset is built following the Cornell Grasping dataset [15]. At first, we only consider flat objects as grasping objects. Moreover, most grasping objects can be considered as flat objects when the objects are placed on the table with a proper direction. Compared with building a 3-D grasping point cloud, this approach can reduce the cost of storage and calculation. The grasping information of flat objects with a PPG can be demonstrated as a rectangle. The width of rectangles presents the distance between gripper plates, the height represents the range of compatible grasping, and the center is placed at a particular point on the table, which presents the grasping point. In addition, the rectangle must be rotated to a particular angle to increase the capability of grasp. This rectangle only provides the pose of PPG when it contacts the table and tries to grasp the object.

C. Problem Definition

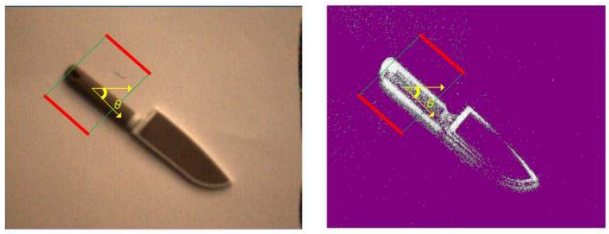
Given RGB images and event streams of different objects, the grasp pose estimation algorithm needs to learn how to find a successfully grasp configuration G for each object. As described in [1], a 5-D grasp representation can be mapped into the 7-D configuration for robotic grasp execution on a real scene. In this work, we take RGB images and events generated by a neuromorphic vision sensor (DAVIS346) as input to predict the 5-D grasp configuration of a robot with a parallel-plate gripper. As shown in Fig. 3, the grasp pose can be formulated as follows:

$$G = \{x, y, w, h, \theta\}^T \tag{3}$$

where (x, y) is the central coordination of grasp rectangle, w corresponds to the maximum distance between parallel plates, h represents the height of parallel plates of the robot, and θ is the angle of grasp rectangle with respect to the horizontal axis.

IV. Grasp Pose Estimation

In this section, we presented a multimodal neural network architecture for grasp pose estimation. The overall framework of this method is shown in Fig. 4. The method consists of two branches, one branch extracts feature representations from RGB images and another focuses on extracting feature representations from event streams. The extracted features are fused and fed into the subsequent network. Furthermore, three task-specific subnetworks are added to perform grasp angle estimation, object classification, and bounding box regression on the feature outputs, respectively. We will describe the details of each component of grasping network.



(a) (b)

Fig. 3. 5-D grasp configuration. (a) Grasp configuration is presented in RGB images. (b) Grasp configuration is presented in the corresponding event frames.

A. Event Representation

Event streams generated by the neuromorphic vision sensor are sparse and asynchronous, which cannot be processed by the traditional computer vision method, such as CNN-based algorithms [12]. Therefore, we use frequency-based encoding method to pre-process event sequences to output matrix for the CNN to extract deep feature.

Given that many more events would occur near an object's edges because edges of the moving object tend to be the edges of illumination in the image, we use the event frequency as the spike coding to strengthen the profile of the object. At the same time, noise caused by the sensor can be significantly filtered out due to its low occurrence frequency at a particular pixel within a given time interval [19]. Concretely, we count the spike occurrence at each pixel (x, y), and based on this, we calculate the spike coding value using the following activation function:

$$\sigma(n) = 255 \cdot 2 \cdot \left(\frac{1}{1 + e^{-n}} - 0.5\right) \tag{4}$$

where n is the total number of occurred spikes (positive or negative) at pixel (x, y) within given interval, and $\sigma(n)$ is the spike coding value of this pixel in the event sequences.

B. Multimodal Fusion

After the event streams are processed by the frequency-based encoding method, we use a 7×7 convolution layer and a 3×3 max-pool layer to transfer the matrix of event sequences into unified scale feature maps. As shown in Fig. 4, the features from event-based and RGB-based networks are fused to generate new feature maps. Since neuromorphic vision sensors can capture dynamic features of the object with high time resolution, combining event streams and RGB data can enhance the spatial–temporal context around grasp objects for improving the detection performance. After that, we use ResNet as a feature extractor to learn deep feature representation for grasp pose estimation. The basic building block of ResNet is the residual block, which is designed as the incorporation of a skip connection with conventional CNN. Referring to [36], feature pyramid network (FPN) is also used to get multiscale features with a top-down pathway and lateral connections. We build a pyramid from P_3 to P_7, where each pyramid level has $C = 256$ channels. To obtain proper grasp pose, a two-vector of grasp angle regression targets, a length K one-hot vector

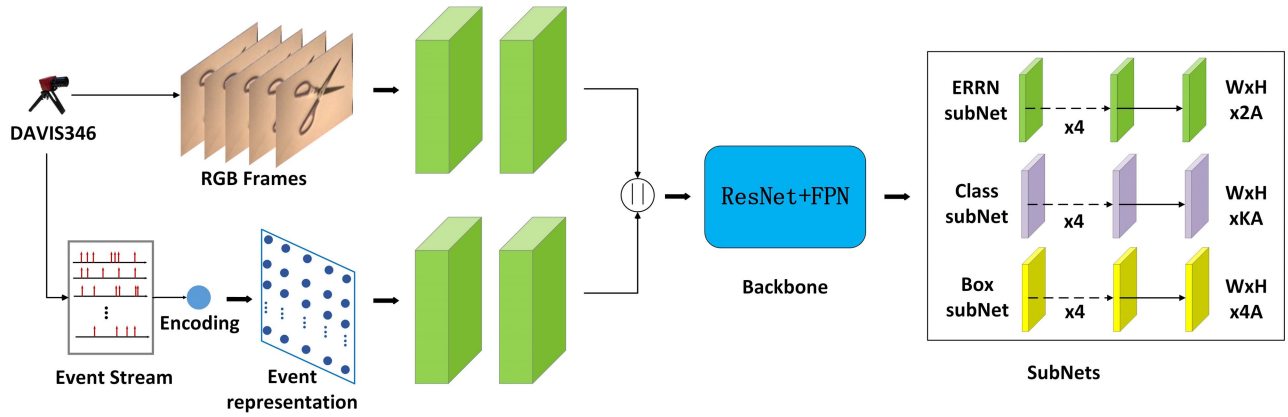


Fig. 4. Multimodal neural network architecture. The network includes two branches, one branch extracts feature representations from RGB images and another focuses on extracting feature representations from event streams. The extracted features are fused and fed into the architecture formed by feature pyramid network (FPN) on top of a feedforward ResNet to generate multiscale features. The outputs of the network are composed of the orientation angle, object classification, and the corresponding grasp poses.

of classification targets, where K denotes the number of grasp object classes, and a four-vector of box regression targets are assigned to each anchor. An imaginary and a real fraction is directly embedded into the network to estimate the grasp angle. The grasping rectangle with adding a complex angle $\arg(|r|e^{i\theta})$ can be defined as

$$\begin{aligned} t_x &= \frac{(x_g - x_a)}{w_a} \\ t_y &= \frac{(y_g - y_a)}{h_a} \\ t_w &= \log\left(\frac{w_g}{w_a}\right) \\ t_h &= \log\left(\frac{h_g}{h_a}\right) \\ t_\theta &= \arg\left(|r|e^{i\theta}\right) = \arctan_2(t_{\mathrm{Im}}, t_{\mathrm{Re}}) \end{aligned} \tag{5}$$

where x and y are the center coordinates of the grasping rectangle. w and h denote the width and height, respectively. θ is the orientation angle, which is represented by the form of an imaginary parameter t_{Im} and real fraction parameter t_{Re}. Variables x_g, x_a, and t_x are for ground-truth box, anchor box, and regression offsets between the anchor box and the ground-truth box, respectively.

C. Euler Region Regression SubNet

The Euler region regression subnet is responsible for predicting the orientation angle of each grasping object. A fully convolutional network (FCN) is applied to each feature pyramid level. Specifically, FCN consists of four 3×3 convolution layers with 256 filters, followed by a 3×3 convolution layer with $2A$ filters, where $A = 9$. The orientation angle can be computed from regression parameters t_{Im} and t_{Re} using $\arctan_2(t_{\mathrm{Im}}, t_{\mathrm{Re}})$. As shown in Fig. 5, instead of directly predicting the angle θ, we estimate the grasp pose of the object by adding an imaginary and a real fraction to the Euler region regression subnet. This strategy builds a closed mathematical space resulting in better generalization ability of the model.

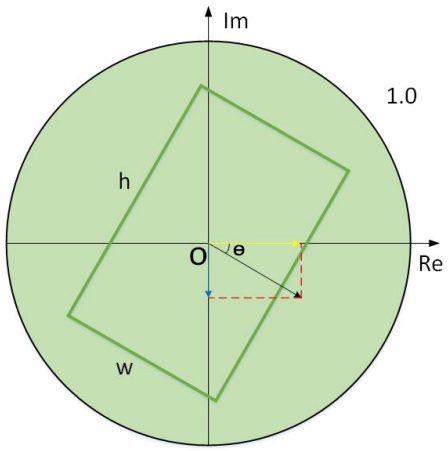


Fig. 5. Grasp orientation angle regression. The oriented grasp pose is predicted based on the complex angle represented by an imaginary and a real fraction.

D. Class and Box SubNet

In parallel with the Euler region regression subnet, two small FCNs are attached to each pyramid level for classification and bounding box regression, respectively. The structure of the two subnets is identical to the Euler region regression subnet except that the classification subnet outputs KA predictions and the box regression subnet produces $4A$ predictions. In the classification subnet, the probability of grasping objects for each of the A anchors and K object classes is inferred by finally passing sigmoid activations. Furthermore, the box regression subnet produces four outputs to regress the offsets between the anchor and the ground-truth box.

E. Loss Function

The multitask loss function of our grasp pose estimation network is defined as follows:

$$L = L_{\mathrm{cls}} + L_{\mathrm{reg}} + L_{\mathrm{euler}}. \tag{6}$$

The loss function L consists of three parts, in which L_{cls} represents the classification loss, L_{reg} denotes the box

regression loss, and L_{euler} is the Euler region regression loss. To improve the robustness of the network, we refer to the design of optimization loss function L_{cls} and L_{reg} in [36]. Moreover, we extend the concepts of L_{reg} by an Euler region regression part L_{euler} to get the use of closed complex number space. The specific formulations are as follows:

$$\begin{aligned} L_{\text{cls}} &= \frac{\lambda_1}{N}\sum_{i=1}^{N} l_{\text{cls}}(p_i, t_i) \\ L_{\text{reg}} &= \frac{\lambda_2}{N}\sum_{i=1}^{N} t_i' \sum_{j\in\{x,y,w,h\}} l_{\text{reg}}\left(v_{ij}', v_{ij}\right) \\ L_{\text{euler}} &= \frac{\lambda_3}{N}\sum_{i=1}^{N} t_i' \sum_{k\in\{Im,Re\}} l_{\text{reg}}\left(\theta_{ik}', \theta_{ik}\right) \end{aligned} \tag{7}$$

where l_{cls} is the focal loss, and l_{reg} represents the smooth L_1 loss. In addition, N is the number of anchors, p_i is computed by the sigmoid function to represent the probability distribution of various classes, and t_i is the corresponding label of the category. v_{ij}' and v_{ij} denote the predicted offset vector and the corresponding vector of ground-truth, respectively. For the Euler region regression loss, we assume that the difference between the predicted complex number and the ground truth is always located on the unit circle with $|r| = 1$. Specifically, the orientation angle θ is regressed by the form of an imaginary Im and real fraction Re. λ_1, λ_2, and λ_3 are hyper-parameters for controlling the trade-off of different losses.

V. Dynamic Robotic Grasping Dataset

For robotic grasping pose estimation, the number of available grasping datasets is limited. The most famous common RGB-D grasping datasets are Cornell, Dexnet, and Jacquard, which are used to compare the state-of-the-art algorithms. To facilitate the application of event-based neuromorphic vision sensors in robotics, an automatically annotated event-based grasping dataset (E-Grasping) is proposed in our previous work [14]. However, compared with traditional vision, event-based research is still in its infancy. In this work, we present a manually labeled dynamic robotic grasping dataset named NeuroGrasp. Compared with the E-Grasping dataset proposed in [14], NeuroGrasp is the first event-based multimodality dataset for grasp pose estimation. The dataset can be found in https://github.com/HuCaoFighting/DVS-GraspingDataSet.

A. Dataset Recording

The dataset is collected using a neuromorphic vision sensor (DAVIS346) with a 346 × 260-pixel resolution. DAVIS346, also known as DVS or event-based camera, is a camera model consisting of a DVS synchronized with an RGB frame-based sensor. We use DAVIS346 to capture 154 grasp objects by recording event-based and RGB frame-based streams separately. The entire dataset is about 4620.42 s in length and contains 14 141.7 M events, making the dataset more diverse and challenging.

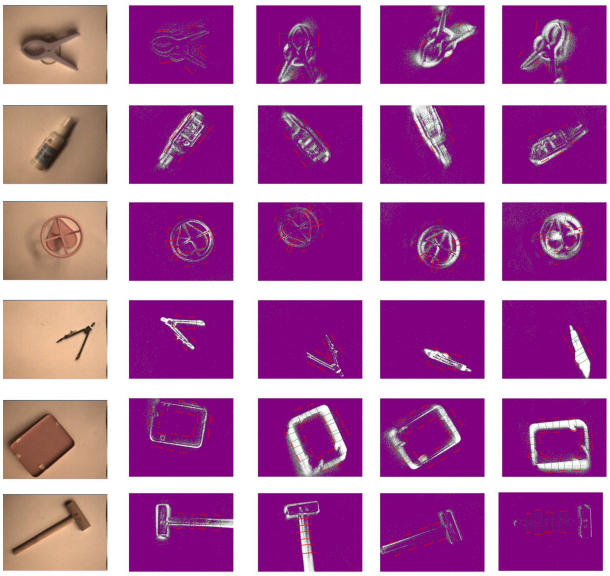


Fig. 6. Grasp annotations: six grasping objects with different poses and views are selected for display. The first column is RGB images, and the remaining columns are the labeled grasping objects' event data with different poses.

B. Dataset Annotation

After manual filtering of unusable data, the NeuroGrasp dataset contains 8753 RGB images and corresponding event streams of 154 different objects with various scales, orientations, and locations. Each image is manually labeled with multiple ground-truth grasp rectangles corresponding to possible grasp configurations, as shown as Fig. 6. However, the annotations are comprehensive and representative examples of good grasp candidates and do not cover all potential grasps. The rating score is affected by the denseness of each object's label. The standard file format in our benchmark is presented in Table I. The dataset contains original binary data, raw event data, RGB images, timestamp files for each frame of RGB images, and labels. We also build a multiobject grasping dataset for testing the generalization ability of our algorithm on a more realistic and cluttered scene. In a multiobject grasping dataset, a single image has three to five different objects with various orientations or poses.

C. Dataset Analysis

In Table II, we summarize the public datasets and our NeuroGrasp dataset. The most common grasping dataset is Cornell, which is collected in a real-world environment. The DexNet and Jacquard datasets are larger than Cornell's. However, both DexNet and Jacquard datasets are generated by simulation, so that large amounts of synthetic data and labels can be produced. The E-Grasping dataset is our previous work [14], which is labeled by tracking led markers. Since the size of the E-Grasping dataset is small, we extend the version of the event-based grasping dataset named NeuroGrasp, which comprises 8753 images with the resolution of 346 × 260 pixels of 154 different novel real objects. In the NeuroGrasp dataset, both RGB images and corresponding event streams

TABLE I
INTRODUCTION OF STANDARD FILE FORMAT IN OUR NEUROGRASP DATASET

File name	Description	Format
original data (.aedat)	original data	raw binary data
events (.txt)	One event per line	(timestamp, x, y, p)
RGB images (.png)	RGB frame-based data	PNG images
timestamp file for RGB frames (.txt)	One timecode per line	(frameNumber, timestamp)
labels (.txt)	One ground-truth measurement per line	(x1, y1, x2, y2, x3, y3, x4, y4)

TABLE II
SUMMARY OF THE PUBLIC GRASPING DATASETS

Dataset	**Modality**	**Objects**	**Images**
Cornell	RGB-D	240	885
Dexnet	Depth	1500	6.7M
Jacquard	RGB-D	11K	54K
E-Grasping	Event Stream	91	18.2k
NeuroGrasp	RGB+Event Stream	154	8753

are recorded, and it is conducive to facilitating event-based robotic grasping research.

VI. EXPERIMENTS AND ANALYSIS

We present the experimental results of the proposed multimodal neural network on the E-Grasping dataset [14] and NeuroGrasp dataset.

A. Implementation Details

In our experiment setup, DAVIS 346 is attached to the end of the robot arm to ensure relative motion between the grasping object and the sensing sensor. The motion speed is controlled under 10–50 mm/s. The experimental dataset is randomly divided into training data and test data in a ratio of 8:2. In the training period, we train the grasping network end to end for 30 epochs on two Nvidia GTX2080Ti GPUs with 22 GB memory. We define the initial learning rate as 0.0005. The weight decay and momentum are set to 0.0001 and 0.9, respectively. The network is implemented using TensorFlow with cudnn-7.5 and Cuda-10.0 packages.

B. Evaluation Metrics

In this work, the widely used rectangle metric is selected to evaluate grasping pose estimation methods. In particular, a prediction of grasp is regarded as valid when it satisfies the following two conditions:

1) **Angle difference:** the difference in the grasp orientation angle between the predicted grasp and the ground truth is within 30°.
2) **Jaccard index:** the intersection over union (IOU) of ground truth and the predicted grasp is more than 25%, as shown in the following equation:

$$J\left(g_p, g_t\right) = \frac{|g_p \cap g_t|}{g_p \cup g_t} \tag{8}$$

TABLE III
ACCURACY (%) OF DIFFERENT METHODS ON THE E-GRASPING DATASET PROPOSED IN [14]

Method	**Light Condition**	**Input**	**Accuracy**(%)
[14]	Light Dark	Event Streams	97.8 96.2
Ours	Light Dark	Event Streams	**98.9** **96.7**

where g_p is the area of the predicted grasp rectangle, and g_t denotes the area of the ground truth. The intersection of predicted grasp rectangle and ground-truth rectangle and the union of predicted grasp rectangle and ground truth rectangle are calculated by $g_p \cap g_t$ and $g_p \cup g_t$, respectively.

C. Results

We explore the performance of the proposed multimodal neural network in different input data and analyze the experimental results of different grasping pose estimation algorithms. The grasping performance are summarized in Tables III and IV.

1) Experimental Results on E-Grasping Dataset: To facilitate comparison with [14], we train our model with the event streams as an input on the E-Grasping dataset. Compared with [14], the proposed grasp pose estimation method achieves better performance with an accuracy of 98.9%. For different lighting conditions, the proposed model can adapt well to the changes in brightness. Furthermore, both the proposed model and [14] have better performance in brighter conditions.

2) Experimental Results on NeuroGrasp Dataset: We compare our model with the event-based method [14] and frame-based method [9] on the NeuroGrasp dataset. Since DAVIS346 can simultaneously output two separate event streams and RGB images, we develop a multimodal neural network to fuse the valuable feature context of event streams and RGB images. The experimental results demonstrate that the proposed multimodal method has a better generalization ability and achieves the best performance with an accuracy of 80.6%.

In Fig. 7, the grasping poses' prediction results are presented. The ground-truth grasping rectangles are in the first row, the top-1 prediction results are visualized in the second row, and the multigrasp results are depicted in the third row. In the multigrasp case, our grasping pose estimation model can predict grasping poses from the features of different objects. The predicted results of these objects demonstrate

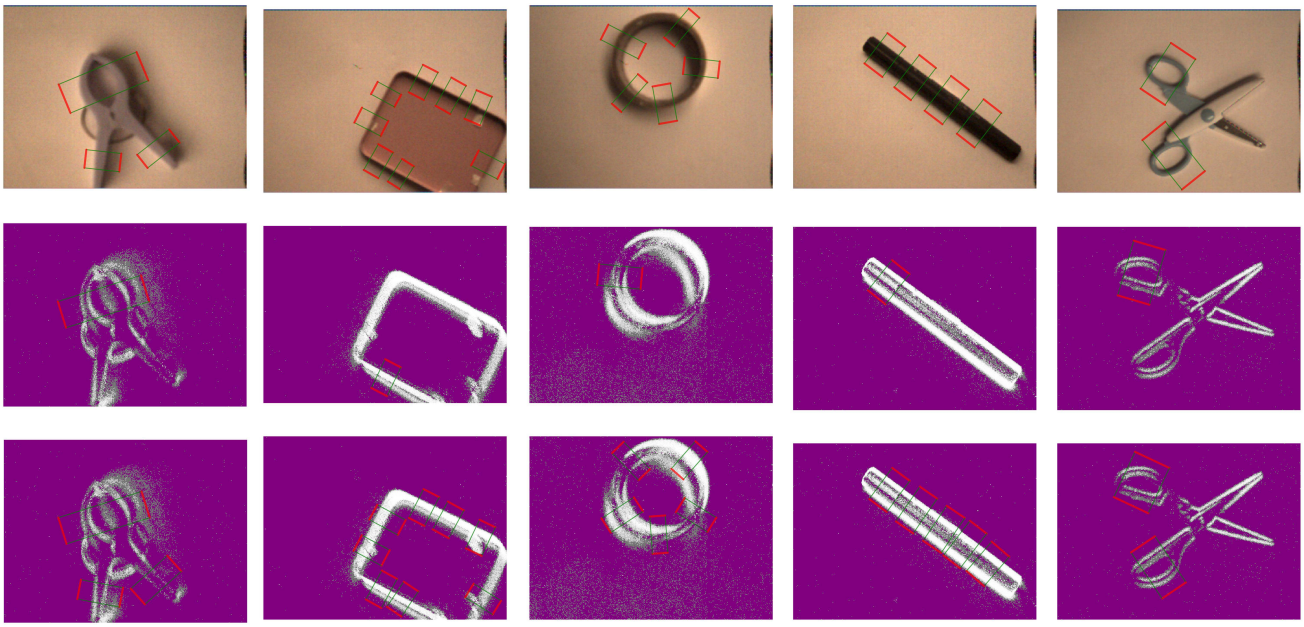


Fig. 7. Prediction results of the proposed grasping network. The first row is the ground truth. The second row is the top-1 grasp outputs for several objects. The third row is the multigrasp results (best viewed in color).

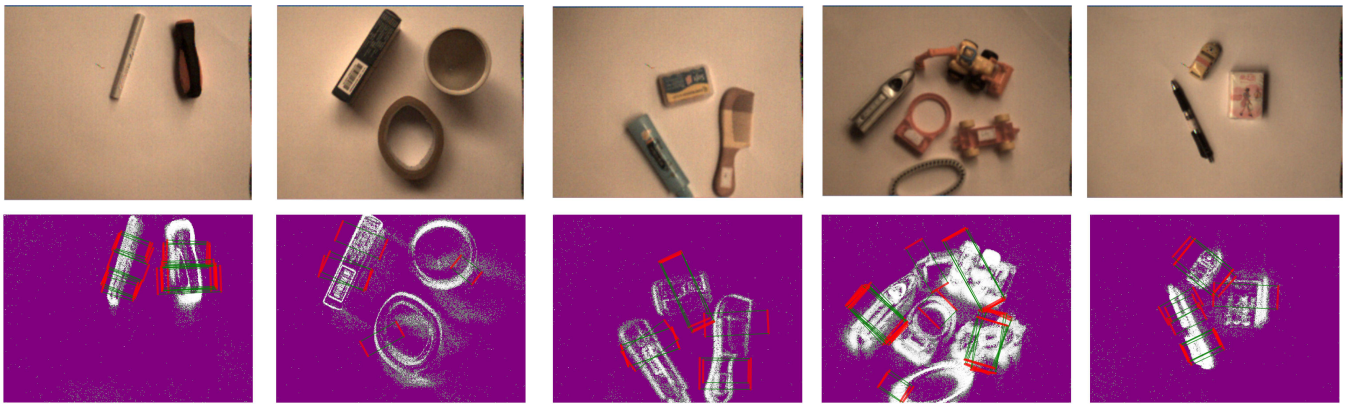


Fig. 8. Prediction results of multiple grasping objects. The first row is the RGB images. The second row is the grasp outputs of the corresponding event streams for several objects (best viewed in color).

TABLE IV
ACCURACY (%) OF DIFFERENT METHODS ON THE NEUROGRASP DATASET

Method	Input	Backbone	Accuracy(%)
[14]	Event Streams	Vgg-16	41.2
[9]	RGB Frames	ResNet-50	52.7
Ours	Event Streams	ResNet-50	53.2
	RGB Frames		76.5
	Event + RGB		**80.6**

that our grasping pose estimation method can predict grasp configuration effectively.

3) Single-Modal Versus Multimodal: In Table IV, grasp prediction results with different input data are presented. For each input data, we use ResNet-50 as backbone to explore the impact of input modality on algorithm performance. Due to the lack of rich appearance features such as color and texture, the grasping pose estimation accuracy based on event streams is lower than RGB frames. However, event streams can provide valuable information with high temporal resolution and high dynamic range, which are complementary to RGB signals. In this work, we use convolutional neural network to learn to fuse information from RGB frames and event streams. By combining RGB frames and event streams, the prediction accuracy is improved by about 4%. The proposed fusion method outperforms the method that only takes RGB frames or event streams as input. To validate the generalization ability of our method, the model trained on the NeuroGrasp dataset is used to test in multigrasp and multiobject environments. The prediction results are presented in Figs. 7 and 8. The model is trained on a single object dataset, but can still predict the grasp pose of multiple objects and multigrasp with various orientations. The results demonstrate the excellent generalization ability and robustness of our method.

4) Effect of Dataset: We train our grasping pose estimation algorithm on both the E-Grasping dataset and NeuroGrasp dataset. Because the annotation method and quantity of label data of two datasets are different, this will affect the prediction accuracy. For the E-Grasping dataset, the same method achieves a higher precision on the E-Grasping dataset than on

TABLE V
ACCURACY (%) OF DIFFERENT BACKBONES ON THE NEUROGRASP DATASET

Method	Input	Backbone	Accuracy(%)
Ours	Event Streams	ResNet-50	53.2
		ResNet-101	54.8
	RGB Frames	ResNet-50	76.5
		ResNet-101	83.0
	Event + RGB	ResNet-50	80.6
		ResNet-101	**83.8**

TABLE VI
NETWORK PARAMETERS' COMPARISON OF DIFFERENT METHODS

Model	Parameter size (Approx.)	Accuracy(%)	Speed(fps)
Single Input	113.95 million	76.5	15
Fusion Input	113.98 million	**80.6**	13

the NeuroGrasp dataset as the size of the labeled ground-truth box is larger and the number of grasping objects is fewer. The NeuroGrasp dataset is more challenging.

5) Effect of Model Scale: In Table V, we discuss the effect of network deepening on model performance. It can be seen from Table V that the performance of the model combined with ResNet-101 is better than that combined with ResNet-50. Furthermore, the proposed fusion method improves the prediction accuracy by about 4% on ResNet-50, but less on ResNet-101. The reasons for this issue can be summarized as follows.

1) Since the method used in this article is early fusion (feature-level fusion at the early layers of the network), the early fused features become more abstracted as the network deepens, thus leading to less effective results.
2) The high detection accuracy achieved by the grasping model on ResNet-101 makes it difficult to further improve the performance. However, while the performance of large model (ResNet-101) is higher, the model complexity is also bigger. Therefore, the performance improvement of multimodal fusion based on ResNet-50 is more promising for application.

6) Complexity Analysis: The comparison of the network parameters between the method with single-modal input and the proposed algorithm is listed in Table VI. With the addition of 0.03 M parameters, the proposed fusion method improves the prediction accuracy by about 4% and achieves the running speed of 13 fps. Our fusion method has a good balance between accuracy and speed.

D. Ablation Study

We provide an ablation study to discuss the impact of the Euler region regression subnet (ERRN), objects in clutter, and failure cases analysis. All the results are based on the ResNet-50 backbone and trained on the NeuroGrasp dataset.

1) Effect of Euler Region Regression SubNet: To explore the effect of the Euler region regression subnet (ERRN) for grasp pose learning, we use ResNet-50 as the backbone to train our model with and without ERRN on the NeuroGrasp dataset. The performances are presented in Table VII. The experimental results illustrate that the prediction accuracy can be improved by about 3% in the best case (RGB input), which demonstrates the effectiveness of the proposed ERRN subnetwork.

TABLE VII
IMPACT OF ERRN SUBNET ON THE PERFORMANCE (%) OF THE NEUROGRASP DATASET

Input	Model	Accuracy(%)
Event Streams	Without ERRN	52.9
	With ERRN	53.2
RGB Frames	Without ERRN	73.3
	With ERRN	76.5
Event Streams + RGB Frames	Without ERRN	78.2
	With ERRN	**80.6**

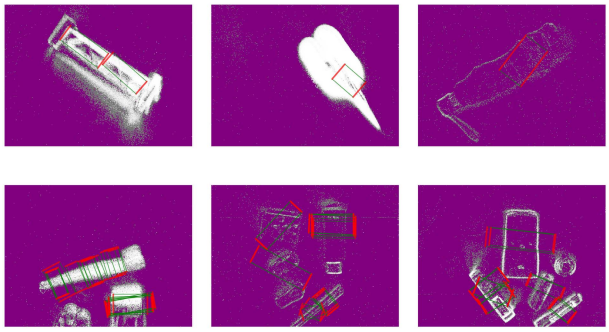


Fig. 9. Failed detection cases. The first row is detection failure cases of single grasping object, and the second row is failure cases of objects in clutter (best viewed in color).

2) Objects in Clutter: For validating the generalization ability of our method, we use the ResNet-50-based model to test on a more realistic and cluttered scene, where a single view has two to five different objects with various orientations or poses. The test results are shown in Fig. 8. In complex scenarios, the proposed method can predict the grasp pose of multiple objects simultaneously and have a good generalization ability.

3) Failure Cases Analysis: Some failed prediction cases are selected to be shown in Fig. 9. It can be seen that the shadow of the grasping object also produces events and affects the prediction results. Some grasping objects with dense events may cause the model to fail to recognize their contour shapes, which leads to the failure of grasping prediction. At the same time, objects that fail to generate enough events also cannot be predicted very well.

E. Discussion

Compared with the traditional frame-based cameras, event-based neuromorphic vision sensors have several advantages.

1) Energy-Friendly and Low Latency: Since event-based neuromorphic vision sensors only process the triggered events and do not need global exposure of the frame, they consume less energy and have a lower latency. Such properties make it more suitable for real-time applications.

2) High Temporal Resolution: For event-based neuromorphic vision sensors, changes can be captured and timestamped to microsecond. This property meets the fast response requirements of the controller in robotics.

3) High Dynamic Range (HDR): The event-based neuromorphic vision sensors have an HDR (120 dB), which outperforms the frame-based cameras (60 dB). Under a light-changing scene, event-based sensors would perform better.

4) Capturing Grasping Object's Edges: The event-based neuromorphic vision sensor can filter out redundant information and capture the grasping object's shapes and edges. The object's shapes and edges are beneficial for grasping and are complementary to frame-based sensors.

VII. Conclusion

In this article, we construct a dynamic robotic grasping dataset named NeuroGrasp. To the best of our knowledge, it is the first event-based multimodality robotic grasping dataset. Based on this dataset, we introduce a multimodal deep neural network for grasping pose estimation with combining frame-based vision and event-based vision. Furthermore, an Euler region regression sub-network (ERRN) is proposed to obtain more accurate orientation angle estimation. The proposed multimodal method is evaluated on the E-Grasping and NeuroGrasp datasets. The experimental results indicate that the proposed method has a better performance and generalization ability. We demonstrate that a neuromorphic sensor will improve both the versatility and the precision of robotic grasp pose estimation.

References

[1] 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, 2015.

[2] 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.

[3] 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, Feb. 2020.

[4] 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.

[5] P. Payeur, C. Pasca, A.-M. Cretu, and E. M. Petriu, "Intelligent haptic sensor system for robotic manipulation," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 4, pp. 1583–1592, Aug. 2005.

[6] 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.

[7] J. Mahler *et al.*, "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *Proc. Robot., Sci. Syst.*, Boston, MA, USA, 2017.

[8] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 769–776.

[9] 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.

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

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

[12] G. Chen, H. Cao, J. Conradt, H. Tang, F. Rohrbein, and A. Knoll, "Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 34–49, Jul. 2020.

[13] G. Chen *et al.*, "Neuromorphic vision-based fall localization in event streams with temporal-spatial attention weighted network," *IEEE Trans. Cybern.*, early access, May 9, 2022, doi: 10.1109/TCYB.2022.3164882.

[14] 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 Neurorobot.*, vol. 14, p. 51, Oct. 2020.

[15] 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.

[16] 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.

[17] U. Asif, J. Tang, and S. Harrer, "Densely supervised grasp detector (DSGD)," in *Proc. AAAI*, 2018, pp. 8085–8093.

[18] 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*.

[19] G. Chen *et al.*, "Multi-cue event information fusion for pedestrian detection with neuromorphic vision sensors," *Frontiers Neurorobot.*, vol. 13, p. 10, Apr. 2019.

[20] M. Simon, K. Amende, A. Kraus, J. Honer, and H. M. Gross, "Complexer-YOLO: Real-time 3D object detection and tracking on semantic point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2018, pp. 1190–1199.

[21] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," 2018, *arXiv:1803.11469*.

[22] 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.

[23] 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)*, 2016, pp. 4461–4468.

[24] 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.

[25] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, "A hybrid deep architecture for robotic grasp detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1609–1614.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, Cambridge, MA, USA, vol. 1, 2015, pp. 91–99.

[27] 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.

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

[29] D. Park, Y. Seo and S. Y. Chun, "Rotation ensemble module for detecting rotation-invariant features," 2018, *arXiv:1812.07762*.

[30] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[31] S. Wang, X. Jiang, J. Zhao, X. Wang, W. Zhou, and Y. Liu, "Efficient fully convolution neural network for generating pixel wise robotic grasps with high resolution images," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2019, pp. 474–480.

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

[33] H. Cao, G. Chen, J. Xia, G. Zhuang, and A. Knoll, "Fusion-based feature attention gate component for vehicle detection based on event camera," *IEEE Sensors J.*, vol. 21, no. 21, pp. 24540–24548, Nov. 2021.

[34] 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.

[35] 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.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

Hu Cao received the M.Eng. degree in vehicle engineering from Hunan University, Changsha, China, in 2019. He is currently pursuing the Ph.D. degree in computer science with the Informatics-6, with the Chair of Robotics, Artificial Intelligence and Real-Time Systems, Technische Universität München, Munich, Germany.

He is also a member of the Informatics 6—Chair of Robotics, Artificial Intelligence and Real-Time Systems, Technische Universität München, Munich, Germany. His research interests include computer vision, neuromorphic engineering, robotics, and deep learning.

Guang Chen received the B.S. and M.Eng. degrees in mechanical engineering from Hunan University, Changsha, China, in 2008 and 2011, respectively, and the Ph.D. degree from the Faculty of Informatics, Technical University of Munich, Germany in 2016.

He is currently a Research Professor with Tongji University, Shanghai, China, and a Senior Research Associate (Guest) with the Technical University of Munich, Munich, Germany. His research interests include computer vision, image processing and machine learning, and bio-inspired vision with applications in robotics and autonomous vehicle. He was a Research Scientist at fortiss GmbH, a research institute of the Technical University of Munich, from 2012 to 2016, and a Senior Researcher with the Chair of Robotics, Artificial Intelligence and Real-time Systems, Technical University of Munich, from 2016 to 2017.

Dr. Chen was a recipient of the Program of Tongji Hundred Talent Research Professor 2018.

Zhijun Li (Fellow, IEEE) received the Ph.D. degree in mechatronics from Shanghai Jiao Tong University, Shanghai, China, in 2002.

From 2003 to 2005, he was a Post-Doctoral Fellow with the Department of Mechanical Engineering and Intelligent systems, The University of Electro-Communications, Tokyo, Japan. From 2005 to 2006, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and with Nanyang Technological University, Singapore. Since 2017, he has been a Professor with the Department of Automation, University of Science and Technology of China, Hefei, China, where he has been the Vice Dean of the School of Information Science and Technology since 2019. His current research interests include wearable robotics, tele-operation systems, nonlinear control, and neural network optimization.

Dr. Li has been the Co-Chairs of IEEE SMC Technical Committee on Bio-mechatronics and Bio-robotics Systems (B^2S) and IEEE RAS Technical Committee on Neuro-Robotics Systems since 2016. He is serving as the Editor-at-Large for *Journal of Intelligent & Robotic Systems* and an Associate Editor for several IEEE TRANSACTIONS.

Yingbai Hu received the M.Sc. degree in control engineering from the South China University of Technology, Guangzhou, China, in 2017. He is currently pursuing the Ph.D. degree in computer science with the Informatics 6-Chair of Robotics, Artificial Intelligence and Real-time Systems Technical University of Munich, Munich, Germany.

He is also a member of the Informatics 6-Chair of Robotics, Artificial Intelligence and Real-time Systems Technical University of Munich. His research interests include neural network optimization and control in robotics, robot learning, surgical robotics, and reinforcement learning.

Alois Knoll (Senior Member, IEEE) received the Diploma (M.Sc.) degree in electrical/communications engineering from the University of Stuttgart, Stuttgart, Germany, in 1985, and the Ph.D. degree *(summa cum laude)* in computer science from the Technical University of Berlin (TU Berlin), Berlin, Germany, in 1988.

He served as the Faculty for the Department of Computer Science, TU Berlin, until 1993. He joined the University of Bielefeld, Bielefeld, Germany, as a Full Professor, and the Director of the research group Technical Informatics until 2001. Since 2001, he has been a Professor with the Department of Informatics, Technische Universität München (TUM), Munich, Germany. He was also on the Board of Directors of the Central Institute of Medical Technology, TUM (IMETUM). From 2004 to 2006, he was the Executive Director of the Institute of Computer Science, TUM. His research interests include cognitive, medical, and sensor-based robotics, multiagent systems, data fusion, adaptive systems, multimedia information retrieval, model-driven development of embedded systems with applications to automotive software and electric transportation, and simulation systems for robotics and traffic.