Fusion-Based Feature Attention Gate Component for Vehicle Detection Based on Event Camera

Article *in* IEEE Sensors Journal · September 2021

DOI: 10.1109/JSEN.2021.3115016

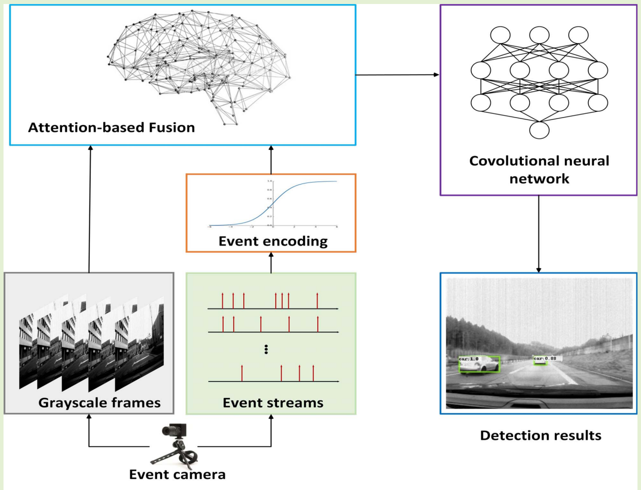
Hu Cao
Technische Universität München
11 PUBLICATIONS **97** CITATIONS
SEE PROFILE

Jiahao Xia
University of Technology Sydney
11 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Fusion-Based Feature Attention Gate Component for Vehicle Detection Based on Event Camera

Hu Cao, Guang Chen, *Member, IEEE*, Jiahao Xia, Genghang Zhuang, and Alois Knoll, *Senior Member, IEEE*

***Abstract*—In the field of autonomous vehicles, various heterogeneous sensors, such as LiDAR, Radar, camera, etc, are combined to improve the vehicle ability of sensing accuracy and robustness. Multi-modal perception and learning has been proved to be an effective method to help vehicle understand the nature of complex environments. Event camera is a bio-inspired vision sensor that captures dynamic changes in the scene and filters out redundant information with high temporal resolution and high dynamic range. These characteristics of the event camera make it have a certain application potential in the field of autonomous vehicles. In this paper, we introduce a fully convolutional neural network with feature attention gate component (FAGC) for vehicle detection by combining frame-based and event-based vision. Both grayscale features and event features are fed into the feature attention gate component (FAGC) to generate the pixel-level attention feature coefficients to improve the feature discrimination ability of the network. Moreover, we explore the influence of different fusion strategies on the detection capability of the network. Experimental results demonstrate that our fusion method achieves the best detection accuracy and exceeds the accuracy of the method that only takes single-mode signal as input.**



***Index Terms*—Vehicle detection, multi-modal fusion, feature attention gate component (FAGC), event camera.**

I. Introduction

RELIABLE perception system can provide the state and pose of the objects for autonomous vehicles. Vehicle detection plays an important role in the field of autonomous driving. For autonomous vehicle, it is equipped with various sensors, such as camera, lidar and radar, to sense the environment. By combining a variety of heterogeneous sensors, autonomous vehicles can sense obstacles and avoid accidents [1]–[4]. The frame-based camera acquires the visual data as a sequence of frames at a fixed frequency. Currently, thanks to the breakthrough of deep learning technology, the frame-based object detection methods, such as [5]–[7], have achieved excellent performance. However, frame-based cameras still suffer from the challenges of overexposure and motion blur in the high light and fast motion [8]. In this work, we try to introduce event camera for vehicle detection task. Event camera captures the pixel-level changes caused by motion and brightness changing. Different from frame-based camera, the event camera outputs high temporal resolution and high dynamic range (120dB) event streams [9], [10]. The comparison of the output between frame-based camera and event camera is presented in Fig 1.

Some works have been proposed to investigate the application potential of event cameras. In [11], the authors use event camera to predict optical flow by using the data from the MVSEC [12] dataset collected by themselves. The first

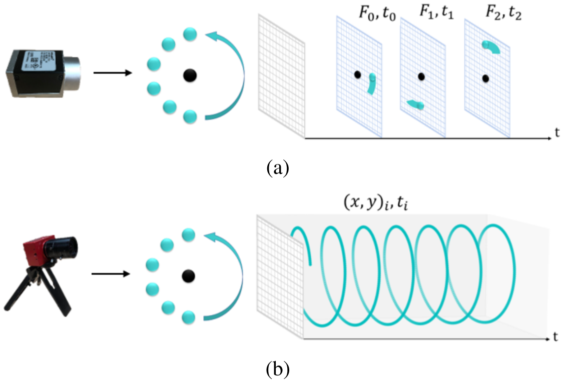
Manuscript received July 21, 2021; revised August 20, 2021; accepted August 22, 2021. Date of publication September 22, 2021; date of current version October 29, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61906138, in part by Shanghai Automotive Industry Sci-Tech Development Program under Grant 1838, 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), and in part by the scholarship from China Scholarship Council (CSC) under Grant 201906130181. The associate editor coordinating the review of this article and approving it for publication was Prof. Chao Tan. *(Corresponding author: Guang Chen.)*

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

Guang Chen is with the School of Automotive Studies, Tongji University, Shanghai 200092, China (e-mail: tj_autodrive@hotmail.com).

Jiahao Xia is with the School of Electrical and Data Engineering, FEIT, University of Technology Sydney, Ultimo, NSW 2007, Australia.

Digital Object Identifier 10.1109/JSEN.2021.3115016



(a)

(b)

Fig. 1. Comparison of the output between standard frame-based camera and event camera [15]. (a) The frame-based camera captures images at a fixed frame rate. (b) The event camera emits events caused by the moving objects asynchronously.

Event-based semantic segmentation is introduced in [13]. Xception-based convolutional neural network (CNN) is trained on Ev-Seg dataset to learn segmentation from events. The researchers also apply the event camera to perform the end-to-end steering angle prediction [14]. Recently, neuromorphic vision based safe driving system is built in [15] and [16]. In [15], the driver drowsiness detection is completed through facial motion analysis by using event camera. And, a new database and baseline evaluation method is proposed in [16]. For event-based object detection, several works, such as [17]–[19], have been done for vehicle or pedestrian detection. However, these methods focus on how to improve the detection accuracy of event-based detector. Since the event streams lack appearance features such as texture and color information, it is difficult to achieve high object detection accuracy only by using event streams as input. Now, there is still a lack of research on how to fuse frame-based and event-based multi-modal features. Hence, it is necessary to study the fusion of event information and other input signal.

In this work, we introduce a fusion-based feature attention gate component (FAGC) for vehicle detection based on event camera. To take advantages of grayscale frames with texture features and events with high dynamic range, both grayscale frames and event streams are fed into the network to fuse together and complement each other. Based on this mechanism, the experimental results on the labeled DDD17 dataset [20], [21] indicate that the detection accuracy of vehicle detection network with FGAC is significantly improved, which is better than the method that only takes grayscale frames as input or only takes event streams as input. Our detailed contributions are as follows:

- A vehicle detection method based on event camera is introduced in this work for autonomous vehicle perception.
- We develop a feature attention gate component (FAGC) to fuse grayscale-based features and event-based features to improve the performance of the vehicle detector. The impact of different fusion strategies and event representations are discussed.
- The experimental results on the labeled DDD17 dataset show that the detection accuracy of vehicle detector can be significantly improved by combining the frame-based vision and event-based vision.

II. Related Work

A. Methods of Frame-Based Object Detection

Early frame-based object detection methods are based on handcrafted features, such as Histogram of oriented gradients (HOG) [22] and aggregate channel features (ACF) [23]. However, with the rise of deep learning, a large number of object detection algorithms based on deep learning have emerged. Current deep learning-based object detection algorithms are mainly divided into one-stage [5], [6], [24]–[26] and two-stage detectors [7], [27], which have been applied in many fields [28]–[30]. For two-stage detectors, Faster-RCNN [7] and Mask-RCNN [27] achieve high detection accuracy based on region proposal network (RPN). Compare with two-stage methods, the one-stage detectors achieve a better balance between accuracy and speed, such as YOLO [6], SSD [5] and Retinanet [24]. However, YOLO [6], SSD [5] and RetinaNet [24] are all Anchor-based object detectors. Recently, anchor-free methods have been developed rapidly and achieved excellent performance, such as Centernet [25] and FCOS [26]. Compared with the frame-based object detection method, the event-based method is still in its preliminary stage.

B. Methods of Event-Based Object Detection

For event-based vision, several works attempt to apply the event camera in various fields, such as intelligent transportation system [31] and robotic grasping [30]. Compared with frame-based object detection, a small amount of research works has been done on event-based object detection [17]–[19], [21], [32]–[34]. More event-based related works can be found in [2] and [8]. In [17], the authors use grayscale frames to pass through the state-of-the-art object detector to generate the pseudo-labels which are used for training the detector model taking the events as input. And, a joint detection framework is introduced in [21] to combine the frame-based and event-based vision for autonomous driving. Different from focusing the vehicle detection under ego-motion in the work [17], [21], the study in [19] concentrates on the pedestrian detection in the field of intelligent transportation system. The fusion method based on confidence map is proposed in [19] to improve the pedestrian detection accuracy. Moreover, in order to take full advantage of the event information, multi-cue event information fusion are developed in [18] for pedestrian detection. Recently, [32] and [33] attempted to use rgb-based detector to improve the performance of the event-based detector. And, the event-based detection method and a high-resolution large-scale dataset are introduced in [34]. The experiment results demonstrate the effectiveness of their method.

III. Method

A. Event Camera

Event camera is a bio-inspired vision sensor, also known as neuromorphic vision sensor and dynamic vision sensor, that works in ways that mimic the perception paradigm of

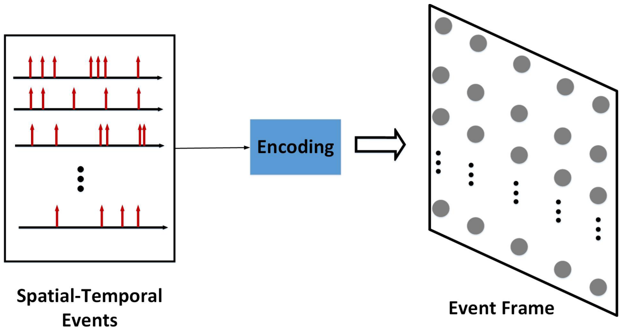


Fig. 2. Event representation: the spatial-temporal events are processed by encoding method to generate the event frame.

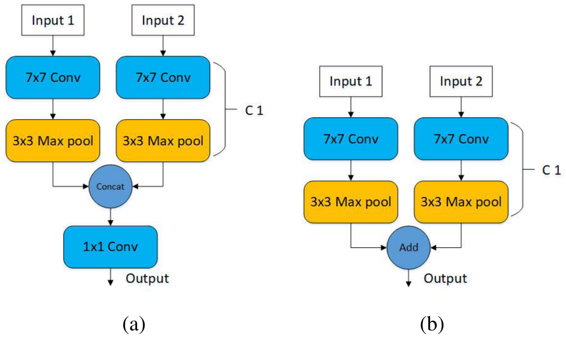


Fig. 3. Comparison of fusion strategy between soft fusion and hard fusion. (a) Soft fusion, (b) Hard fusion.

the biological retina [2]. Conventional frame-based cameras output a series of frames at a fixed frequency. Unlike frame-based cameras, event cameras output data in microseconds and asynchronously, as shown in Fig. 1. An event is triggered only if the brightness change at the same pixel position exceeds a certain threshold. A sparse spatial-temporal event stream can be mathematically represented as:

$$E = \{e_i\}_{i\in[1,N]}, \quad e_i = [x_i, y_i, t_i, p_i]^T \tag{1}$$

where, N represents the number of e_i contained in the event stream E. (x, y) is the coordinates of the triggered pixel position. t and p denote the corresponding triggering timestamp and event polarity, respectively. And, $p \in \{+1, -1\}$ represents the brightness change, $+1$ denotes increase and -1 denotes decrease.

In this work, Dynamic and Active Pixel Vision Sensor (DAVIS) is used for sensing object. DAVIS consists of a grayscale frame-based camera and an event camera, so that it can simutaneously output grayscale images and event streams. To take full advantage of the grayscale frames with texture features and event data with high dynamic range, we combine the two data to improve the accuracy of vehicle detection. Since the asynchronous event stream cannot be directly processed by convolutional neural network, we use the frequency-based [17], [18] encoding method to preprocess it into event frames before feeding it into the network. Frequency-based encoding method can be formulated as:

$$P(n) = 255 \cdot 2 \cdot (\frac{1}{1+e^{-n}} - 0.5) \tag{2}$$

where, n represents the total number of the triggered events (*positive or negative*) at location (x, y), and $P(n)$ denotes the corresponding transformed pixel value. As presented in Fig. 2, the triggered spatial-temporal events are processed by frequency-based encoding method to generate the event frame. Specifically, each pixel value in the event frame is obtained by using Eq. 2 to calculate the events generated within $20ms$.

B. Fusion Strategy

In this work, we explore the impact of different fusion strategies on the detection accuracy of the network. In Fig. 3, two fusion strategies are presented: soft fusion (Fig. 3(a)) and hard fusion (Fig. 3(a)). Instead of merging grayscale frames and events directly, we let the network learn which features need to be fused. Therefore, both the grayscale frames and the event frames are fed into the C_1 block to automatically learning to extract features, F_{gray} and F_{event}. $C1$ block consists of a convolutional filter with kernel size of 7 × 7 and a max pooling layer with kernel size of 3 × 3.

1) Hard Fusion: for hard fusion, it denotes the element-wise sum of extracted feature maps, which can be defined as follows:

$$F_{hard} = f_{add}(F_{gray}, F_{event}) \tag{3}$$

where, f_{add} represents the function of element-wise addition.

2) Soft Fusion: For soft fusion, it represents feature fusion by using learned parameters. In particular, both F_{gray} and F_{event} are concatenated together, then, the convolutional filter with kernel size of 1 × 1 is applied to learn the weight parameters for feature fusion and unify dimension. The process can be expressed as follows:

$$F_{soft} = f_{concat}(F_{gray}, F_{event}) \otimes conv_{1\times1} \tag{4}$$

where, f_{concat} and $\otimes$ denote concatenate and convolution operation, respectively. Different from hard fusion and soft fusion, feature attention gate component (FAGC) combines hard fusion and attention mechanism to fuse grayscale-based features and event-based features, so as to significantly improve the vehicle detection accuracy.

3) Feature Attention Gate Component (FAGC): Attention mechanism has been applied in computer vision and worked very well, such as [35]–[39]. In this work, the extracted grayscale-based features F_{gray} and event-based features F_{event} are fed into feature attention gate component (FAGC) to extract valuable features. The block diagram of the feature attention gate component (FAGC) is presented in Fig. 4. Both the grayscale-based features and the event-based features pass through a convolutional filter with kernel size of 3 × 3 and a ReLU activation function to get transformed contextual information. Then, the transformed features are fused by element-wise addition:

$$F_{fuse} = f_{add}((F_{gray}, F_{event}) \otimes conv_{3\times3}) \tag{5}$$

Furthermore, in order to identify salient feature regions and suppress unrelated background regions, event-based features are used as gate signals, and 5 × 5 convolution followed by

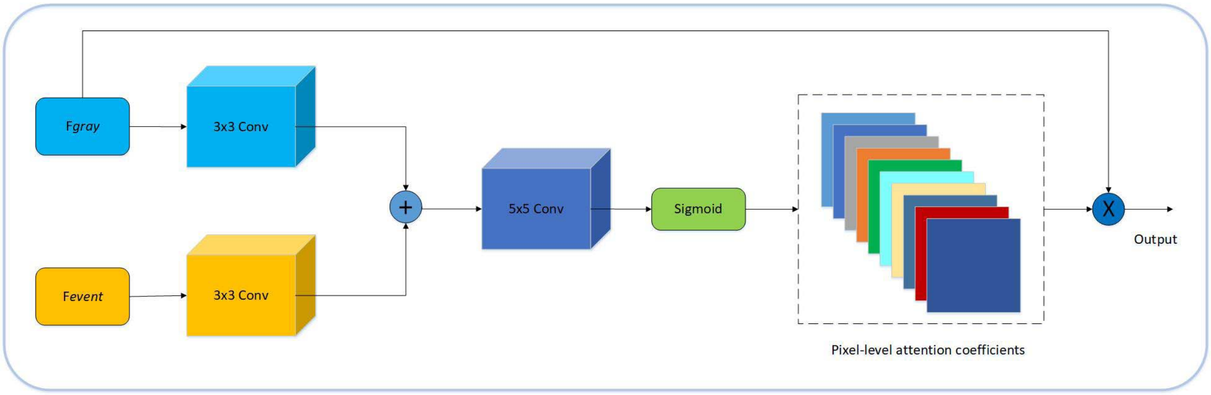


Fig. 4. Feature attention gate component (FAGC): both grayscale-based features and event-based features are fed into FAGC to generate the pixel-level attention coefficients.

a Sigmoid activation function is used to generate pixel-level attention coefficients from the fused features. The output of the feature attention gate component (FAGC) is the element-wise multiplication of the input grayscale-based features and the pixel-level attention coefficients:

$$F_{output} = \sigma(F_{fuse} \otimes conv_{5\times 5}) \cdot F_{gray} \quad (6)$$

Based on this mechanism, the object features will be enhanced to further improve the detection accuracy of the network. The impact of different fusion methods will be discussed in detail in section IV-D.

C. Network Architecture

The vehicle detection framework used in this work is built on the basis of [24], as shown in Fig. 5. The image size of 532 × 400 grayscale frames and event frames are fed into Resnet [40] to extract meaningful features. Both event-based features and grayscale-based features are fused by feature attention gate component (FAGC). The fused features are collected to pass through the feature pyramid network (FPN) [29] to obtain deep features for detecting vehicles of different scales. The vehicle detection network is composed of resnet, feature attention gate component (FAGC), feature pyramid network (FPN) and detection head subnets. It can be formulated as follows:

$$\begin{aligned} L_c^k &= f_{Resnet}(x_{gray}, x_{event}) \\ F^k &= f_{FAGC}(L_c^k) \\ P^n &= f_{FPN}(F^k) \\ Y^n &= f_{Head}(P^n) \end{aligned} \quad (7)$$

where, x_{gray} and x_{event} represent the grayscale frame input and event frame input, respectively. $\left\{L_c^k\right\}_{k=1,c\in[gray,event]}^4$ denotes the extracted grayscale-based features and event-based features. $\{F_k\}_{k=1}^4$ is the fused features generated by feature attention gate component (FAGC). $\{P_n\}_{n=1}^5$, and $\{Y_n\}_{n=1}^5$ denote the fused multi-resolution features and prediction outputs, respectively. And, the functions of feature attention gate component (FAGC), Resnet, feature pyramid network (FPN) and detection head subnets are represented by f_{FAGC}, f_{Resnet}, f_{FPN}, and f_{Head}.

1) Resnet: In this work, we use Resnet-50 [40] as the backbone network. Resnet-50 is composed of four layers, represented as $\{L_1, L_2, L_3, L_4\}$, where, the feature map resolution is continuously down-sampled from L_1 to L_4 layer, and the feature resolution remains the same in each layer. Feature attention gate component (FAGC) mentioned above is inserted between two layers. By combining the residual learning and feature attention gate component (FAGC), the more strong semantic and valuable features can be extracted.

2) Feature Pyramid Network (FPN): Similar to the previous works [24], [29], feature pyramid network (FPN) is used to fuse the features generated from $\{C_k\}_{k=1}^4$ to improve the detection robustness of vehicles of different sizes. The outputs $\{P_n\}_{n=1}^4$ are produced by top-down pathway and lateral connections. And, the last level feature map P_5 is produced by applying a 3×3 convolutional layer with stride 2 on the P_4. Multi-level feature maps $\{P_n\}_{n=1}^5$ will be fed into the detection head subnets for prediction.

3) Detection Head Subnets: After processing by feature pyramid network (FPN), two separate subnets are applied for classification and box regression. Refer to [24], each subnet consists of four 3×3 convolutional layers with 256 filters. For classification subnet, followed by a 3×3 convolutional layers with KA filters, followed by sigmoid activations, it outputs KA binary predictions. For box regression subnet, followed by a 3×3 convolutional layers with $4A$ filters, it outputs $4A$ offset predictions. A is set as 9 in this work. Specific offset parameters of the bounding box can be represented as follows:

$$\begin{aligned} t_x' &= \frac{(x' - x_a)}{w_a}, \\ t_y' &= \frac{(y' - y_a)}{h_a}, \\ t_w' &= log(\frac{w'}{w_a}), \\ t_h' &= log(\frac{h'}{h_a}), \end{aligned} \quad (8)$$

where, x, y, w, h represent the center coordinates, width and height of the bounding box. Variables t', x', x_a denote the prediction regression offsets, predicted bounding box and anchor box, respectively.

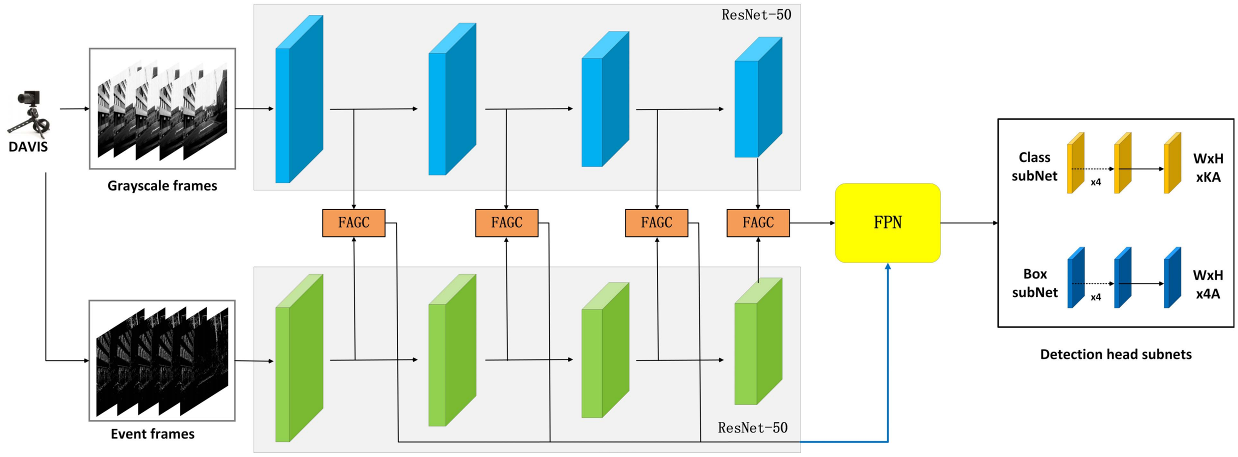


Fig. 5. The architecture of our vehicle detection network. The network consists of resnet, feature attention gate component (FAGC), feature pyramid network (FPN) and detection head subnets.

D. Loss Function

The loss function of our vehicle detection network consists of classification and regression loss function. The total loss function L can be represented as follows:

$$L=\frac{\lambda_1}{N}\sum_{i=1}^{N} l_{cls}(p_i,t_i)+\frac{\lambda_2}{N}\sum_{i=1}^{N} t_i' \sum_{j\in\{x,y,w,h\}} l_{reg}(v_{ij}',v_{ij}) \quad (9)$$

where N denotes the number of anchors. Specifically, focal loss l_{cls} and giou loss l_{reg} are used in this work. The hyper-parameter λ_1 and λ_2 control the trade-off of classification and regression losses. $\lambda_1=\lambda_2=1$ are used in our experiments.

IV. Experiment

The experiments of our vehicle detection network are performed on the labeled DDD-17 dataset [20], [21]. The results indicate that the fusion-based feature attention gate component (FAGC) can improve the detection accuracy of vehicle detector. And, we also discuss the influence of different fusion strategies and event representations on the detection performance of the network.

A. Dataset

In order to verify the effectiveness of our fusion method, the experiments are conducted on the DDD17 dataset. DDD17 [20] uses DAVIS to record both grayscale frames and event streams. The comparison of grayscale frame and the corresponding event frame is presented in Fig. 6. The dataset is collected on highway and city scenes from Switzerland to Germany. Since DDD17 is established for end-to-end learning, it does not contain the labels of object detection, while the authors of [21] manually labeled the vehicles of the dataset based on the original raw data. The detailed description are summarized in Tab. I. On account of our model requires both event-based and frame-based data, and DDD17 is a challenging data set, we use the labeled DDD17 as the benchmark to compare the performance of the different fusion strategies on vehicle detection. The labeled DDD17 dataset contains 3154 frames. We used 2241 frames as the training set and 913 frames as the test set. In order to train more robust models, data augmentation methods such as flipping and color enhancement are used to increase the diversity of data samples.

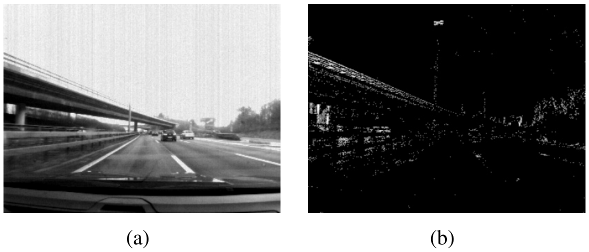


Fig. 6. Comparison of grayscale frame and event frame. (a) grayscale frame, (b) event frame.

TABLE I
DETAILED DESCRIPTION OF THE RECORDED DATA IN THE LABELED DDD17 DATASET

Recorded data	Condition	Length (s)	Type
1487339175	day	347	test
1487417411	day	2096	test
1487419513	day	1976	train
1487424147	day	3040	train
1487430438	day	3135	train
1487433587	night-fall	2335	train
1487593224	day	524	test
1487594667	day	2985	train
1487597945	night-fall	50	test
1487598202	day	1882	train
1487600962	day	2143	test
1487608147	night-fall	1208	train
1487609463	night-fall	101	test
1487781509	night-fall	127	test

B. Evaluation Metrics

We use the common metric, AP, to evaluate the performance of the different vehicle detectors. The value of AP denotes the area under the Precision-Recall curve. Recall, Precision and

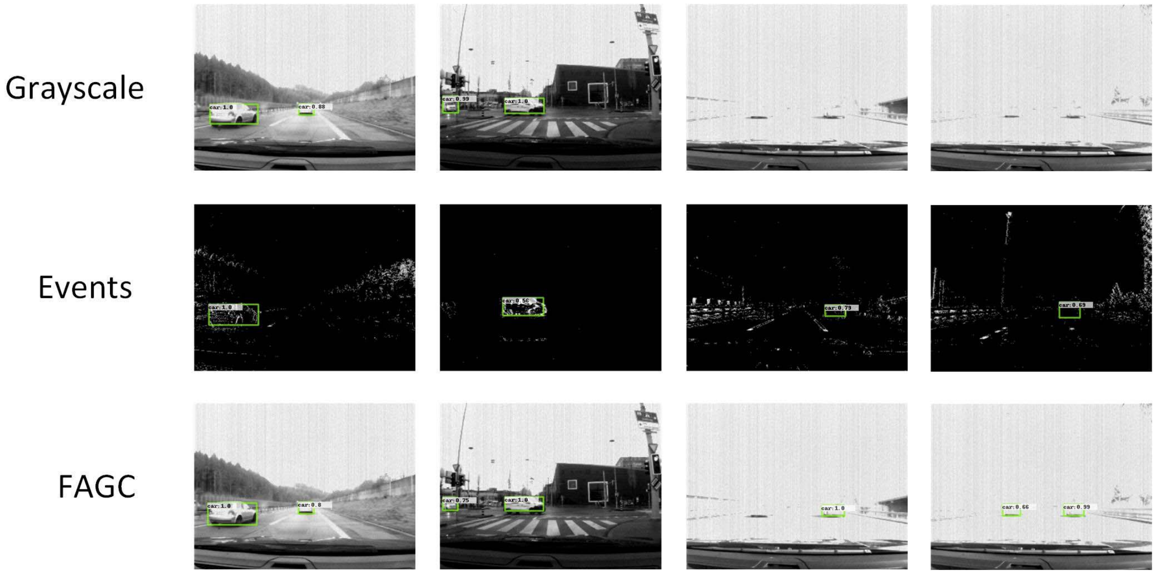


Fig. 7. The selected detection results of vehicle detection model on the labeled DDD17 dataset. The first, second and third rows display the detection results of grayscale, events and FAGC, respectively.

TABLE II
THE DETECTION RESULTS OF DIFFERENT EVENT REPRESENTATIONS ON THE LABELED DDD17 DATASET

Methods	Input Modality	AP(%)	FPS
LIF	Events	13.9	14
SAE	Events	51.1	14
Frequency	Events	**52.3**	14

TABLE III
THE DETECTION RESULTS OF DIFFERENT FUSION STRATEGIES ON THE LABELED DDD17 DATASET

Methods	Input Modality	AP(%)	FPS
Baseline	Events	52.3	14
Baseline	Grayscale	79.6	14
MTC	Events	47.8	14
Hard fusion	Events & Grayscale	77.2	12
Soft fusion	Events & Grayscale	79.4	12
FAGC	Events & Grayscale	**81.6**	8

IOU (Intersection over Union) are expressed as the following:

$$\begin{aligned} Recall &= \frac{tp}{tp+fn}, \\ Precision &= \frac{tp}{tp+fp} \\ IOU &= \frac{detections \cap groundtruth}{detections \cup groundtruth} \\ &= \frac{tp}{tp+fp+fn} \end{aligned} \tag{10}$$

where, tp represents the true positive samples, meaning the correctly predicted vehicles. Similarly, fp and fn denote the false positive samples and false negative samples, respectively. After Recall, Precision, and IOU are calculated, we can use the area under the Precision-Recall cure (AP) to summarize the performance of the detector. In particular, the AP at $IOU = 0.5$ [41] is used as our evaluation metric.

C. Implement Details

In training period, we train the vehicle detection network end to end for 30 epochs on a Nvidia Tesla V100 GPU with 32GB memory. We define the initial learning rate as 0.01. Weight decay and momentum are set to 0.0001 and 0.9, respectively. The network is implemented using pytorch-1.7.0 with cudnn-7.5 and cuda-10.0 pacakges.

D. Quantitative Analysis

1) Effect of Event Representation: Our vehicle detection network can take different image-like event representations as input. We compare the performance of three more representative event encoding methods, Frequency [17], LIF ((Leaky Integrate-and-Fire) [42] and SAE (Surface of Active Events) [43]. The results on the labeled DDD17 dataset are presented in Tab. II. Compared with other two encoding methods, Frequency-based event representation achieve the best performance with the accuracy of 52.3%. Therefore, we use Frequency as our event preprocess method in the subsequent experiments.

2) Impact of Different Fusion Strategies: We explore the impact of MTC (Merged-Three-Channel) [18], hard fusion, soft fusion and FAGC. MTC is a channel-level fusion strategy. In this work, the three channels of MTC frames are consists of $[Frequency, SAE, LIF]$. The Retinanet based on resnet-50 with the grayscale frames and frequency-based event representation as input are the baselines. Specifically, we use the pre-trained weight of Resenet-50 on ImageNet to initialize the model parameters and train the vehicle detector with MTC, hard fusion, soft fusion and FAGC respectively on the labeled DDD17 dataset. The experiment results are given in Tab. III. As can be seen from the Tab III, the network gets 79.6% vehicle detection accuracy by taking grayscale frames as input. In order to enable the event streams to be processed by CNN, the frequency-based [17] encoding method is used to regularize the events into event frames to pass through the network. Using only event data as input, the network achieves

TABLE IV
The Detection Results of Different Methods on the Labeled DDD17 Dataset

Methods	Input Modality	AP(%)	FPS
FasterRCNN [7]	Grayscale	80.2	3
SSD [5]	Grayscale	73.1	12
Yolo [6]	Grayscale	70.2	15
Retinanet [24]	Grayscale	79.6	14
FAGC	Events & Grayscale	**81.6**	8

a detection accuracy of 52.3%. Compared with grayscale-based vehicle detector, the accuracy of event-based vehicle detector is significantly lower than that of grayscale-based because of the lack of appearance information such as texture in the event data. However, due to the advantages of high dynamic range and high temporal resolution of event data, it can alleviate the motion blur of grayscale frames under high illumination and high speed motion. Therefore, the feature attention gate component (FAGC) is developed to fuse event data with grayscale frames. The test results indicate that the performance of the vehicle detection network based on MTC, hard fusion and soft fusion is basically not improved. However, based on FAGC, the network achieves the best detection accuracy of 81.6%, which outperforms the method that only takes grayscale frames or events as input.

3) Comparision With Grayscale-Based Detectors: We compare our model with several selected frame-based object detectors [5]–[7], [24]. All experiments are conducted on the labeled DDD17 dataset and the tested results are summarized in Tab. IV. Specifically, the results of [5]–[7] are referred from [21]. Compared with one-stage object detectors [5], [6], [24], the vehicle detection network with FAGC achieves a significant improvement in detection accuracy, while the running speed is reduced. Moreover, FAGC also has better performance over the two-stage object detector [7], which demonstrates the effectiveness of our method.

E. Qualitative Analysis

The selected detection results are visualized in Fig. 7. The detection results of grayscale, events and FAGC are presented in the first, second and third rows respectively.

1) Normal Detections: It can be seen from the Fig. 7, the vehicle detection results of the first two columns demonstrate that the detection performance of the detector based on grayscale is stronger than that based on event under the normal light condition. And the detection accuracy is similar between the grayscale-based vehicle detector and the fusion-based (FAGC) vehicle detector.

2) Overexposure Detections: In Fig. 7, the vehicle detection results in the third column show that the grayscale-based detector is weaker than the event-based detector under high illumination conditions. Moreover, the vehicle detection results of the last column indicate that the FAGC-based fusion method can achieve better performance when the detection performance of the detector is not good either grayscale-based or event-based.

3) Evaluation on Day and Night-Fall Condition: In order to further explore the effectiveness of vehicle detectors, Frequency-based, grayscale-based (Retinanet-Gray) and FAGC-based vehicle detectors are tested respectively on day, night-fall, and all (day and night-fall) conditions. The test results are summarised in the Table V. Both event-based and grayscale-based detectors achieve stable detection performance under day and night-fall conditions. And, the proposed FAGC-based detector can achieve more robust generalized ability through fusion of events and grayscale frames. The main reason for this result is that the high temporal resolution and high dynamic range of events can alleviate the challenge of grayscale frames due to overexposure, low light and high speed motion.

TABLE V
Evaluation on Day and Night-Fall Condition

Methods	Input Modality	day	night-fall	All
Frequency	Events	49.4	67.1	52.3
Retinanet-Gray	Grayscale	77.9	**87**	79.6
FAGC	Events & Grayscale	**80.5**	86.2	**81.6**

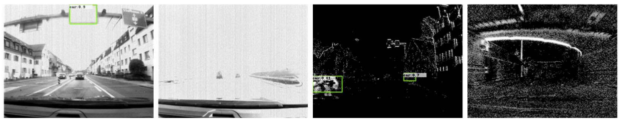


Fig. 8. Failed detection cases: the first two cases are false detection results based on grayscale frames and the last two cases are false detection resultss based on event frames.

4) Failure Cases Analysis: Some failed detection cases are displayed in the Fig. 8. For graycale-based vehicle detector, the model incorrectly detects traffic sign as vehicle and performs poorly under high light conditions. Although the event data filtered most of the background, a large number of events were generated by some roadside obstacles in the process of perceiving the environment, leading to incorrect detection results output by the model. In addition, when passing a scene such as a bridge, a large number of events will be generated by the outline of the bridge, resulting in little information of the vehicle. Compared with traditional vision, event-based vision research is still in the preliminary stage, so further development of this technology is needed to make it mature gradually.

V. Conclusion

In this work, we introduce a fully convolutional neural network with feature attention gate component (FAGC) to perform vehicle detection. Both grayscale frames and event streams are fused together to improve the detection accuracy of the network. To better fuse the frame-based and event-based vision, hard fusion and soft fusion are discussed. Based on hard fusion and attention mechanism, FAGC is developed to combine the grayscale frames with texture and events with high dynamic range to improve the discrimination ability of the model. By integrating the FAGC into the model, the vehicle detector achieves better performance compared with the method that only takes grayscale frames or events as input. The experimental results on the labeled DDD17 dataset indicate that our fusion method is effective. Compared with the traditional framed-based vision, the dataset of event camera is scarce. In the following work, we will collect a multilmodal dataset to promote the research on the fusion of the event signal and other modal information. Since event-based research is still in its infancy, we will try to explore the application of event camera in more fields, such as object tracking, segmentation, etc.

References

[1] C. Urmson *et al.*, "Autonomous driving in urban environments: Boss and the urban challenge," *J. Field Robot.*, vol. 25, no. 8, pp. 425–466, 2008.

[2] 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.

[3] G. Chen, C. Kai, Z. Lijun, Z. Liming, and A. Knoll, "VCANet: Vanishing-point-guided context-aware network for small road object detection," *Automot. Innov.*, pp. 2522–8765, Sep. 2021.

[4] G. Chen *et al.*, "Pole-curb fusion based robust and efficient autonomous vehicle localization system with branch-and-bound global optimization and local grid map method," *IEEE Trans. Veh. Technol.*, Sep. 2021.

[5] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, vol. 9905. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.

[6] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 2016, *arXiv:1612.08242*. [Online]. Available: https://arxiv.org/abs/1612.08242

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 91–99.

[8] G. Gallego *et al.*, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 10, 2020, doi: 10.1109/TPAMI.2020.3008413.

[9] S.-C. Liu, B. Rueckauer, E. Ceolini, A. Huber, and T. Delbruck, "Event-driven sensing for efficient perception: Vision and audition algorithms," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 29–37, Nov. 2019.

[10] G. Chen *et al.*, "A novel visible light positioning system with event-based neuromorphic vision sensor," *IEEE Sensors J.*, vol. 20, no. 17, p. 10211–10219, Sep. 2020.

[11] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "EV-FlowNet: Self-supervised optical flow estimation for event-based cameras," in *Robotics: Science and System*. Pittsburgh, PA, USA: Carnegie Mellon Univ., Jun. 2018, pp. 1–9.

[12] A. Z. Zhu, D. Thakur, T. Ozaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3D perception," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2032–2039, Jul. 2018.

[13] I. Alonso and A. C. Murillo, "EV-SegNet: Semantic segmentation for event-based cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1624–1633.

[14] A. I. Maqueda, A. Loquercio, G. Gallego, N. Garcia, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5419–5427.

[15] G. Chen, L. Hong, J. Dong, P. Liu, J. Conradt, and A. Knoll, "EDDD: Event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor," *IEEE Sensors J.*, vol. 20, no. 11, pp. 6170–6181, Jun. 2020.

[16] G. Chen *et al.*, "NeuroIV: Neuromorphic vision meets intelligent vehicle towards safe driving with a new database and baseline evaluations," *IEEE Trans. Intell. Transp. Syst.*, early access, Oct. 20, 2020.

[17] N. F. Y. Chen, "Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 644–653.

[18] G. Chen *et al.*, "Multi-cue event information fusion for pedestrian detection with neuromorphic vision sensors," *Frontiers Neurorobot.*, vol. 13, p. 10, Apr. 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbot.2019.00010

[19] Z. Jiang *et al.*, "Mixed frame-/event-driven fast pedestrian detection," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8332–8338.

[20] J. Binas, D. Neil, S. Liu, and T. Delbrück, "DDD17: End-to-end DAVIS driving dataset," in *Proc. ICML*, 2017.

[21] J. Li, S. Dong, Z. Yu, Y. Tian, and T. Huang, "Event-based vision enhanced: A joint detection framework in autonomous driving," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1396–1401.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[23] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[25] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6568–6577.

[26] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 19, 2020, doi: 10.1109/TPAMI.2020.3032166.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.

[28] W. Li, H. Cao, J. Liao, J. Xia, L. Cao, and A. Knoll, "Parking slot detection on around-view images using DCNN," *Frontiers Neurorobot.*, vol. 14, p. 46, Jul. 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbot.2020.00046

[29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[30] 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. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbot.2020.00051

[31] G. Chen *et al.*, "Neuromorphic vision based multivehicle detection and tracking for intelligent transportation system," *J. Adv. Transp.*, vol. 2018, pp. 1–13, Dec. 2018.

[32] A. Zanardi, A. Aumiller, J. Zilly, A. Censi, and E. Frazzoli, "Cross-modal learning filters for RGB-neuromorphic wormhole learning," in *Robotics: Science and System*. Freiburg im Breisgau, Germany: Univ. of Freiburg, Jun. 2019.

[33] Y. Hu, T. Delbruck, and S.-C. Liu, "Learning to exploit multiple vision modalities by using grafted networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 85–101.

[34] E. Perot, P. de Tournemire, D. Nitti, J. Masci, and A. Sironi, "Learning to detect objects with a 1 megapixel event camera," in *Proc. NeurIPS Conf.*, 2020, pp 16639–16652.

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

[36] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," 2019, *arXiv:1903.06586*. [Online]. Available: https://arxiv.org/abs/1903.06586

[37] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," in *Proc. IMIDL Conf.*, 2018.

[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[39] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[42] A. N. Burkitt, "A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input," *Biol. Cybern.*, vol. 95, no. 1, pp. 1–19, Jul. 2006.

[43] E. Mueggler, C. Bartolozzi, and D. Scaramuzza, "Fast event-based corner detection," in *Proc. Brit. Mach. Vis. Conf.*, 2017.

Hu Cao received the M.Eng. degree in vehicle engineering from Hunan University, China, in 2019. He is currently pursuing the Ph.D. degree in computer science as a member of the Informatics-6, Chair of Robotics, Artificial Intelligence and Real-Time Systems, Technische Universität München. His research interests include computer vision, neuromorphic engineering, robotics, and deep learning.

Guang Chen (Member, IEEE) received the B.S. and M.Eng. degrees in mechanical engineering from Hunan University, China, and the Ph.D. degree from the Faculty of Informatics, Technical University of Munich, Germany. He is a Research Professor with Tongji University and a Senior Research Associate (Guest) with the Technical University of Munich. He is leading the Intelligent Sensing, Perception and Computing Group, Tongji University. 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. His research interests include computer vision, image processing and machine learning, and the bio-inspired vision with applications in robotics, and autonomous vehicle. He was awarded the Program of Tongji Hundred Talent Research Professor 2018.

Jiahao Xia received the B.E. degree from Wuhan University of Technology in 2017 and the M.E. degree from Hunan University in 2020. He is currently pursuing the Ph.D. degree with the University of Technology, Sydney. His research interests include computer vision, face alignment, and object detection.

Genghang Zhuang received the B.Eng. and M.Eng. degrees in software engineering from Sun Yat-sen University, China, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Chair of Robotics, Artificial Intelligence and Real-Time Systems, Department of Informatics, Technical University of Munich, Germany. His research interests include perception and planning in autonomous driving, especially with LiDAR sensors and biologically inspired methods.

Alois Knoll (Senior Member, IEEE) received the Diploma (M.Sc.) degree in electrical/communications engineering from the University of Stuttgart, Germany, in 1985, and the Ph.D. *(summa cum laude)* degree in computer science from the Technical University of Berlin, Germany, in 1988. He served on the Faculty of the Computer Science Department, TU Berlin, until 1993. He joined the University of Bielefeld, 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, TU München. 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, multi-agent 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.