

TUMTraf Event: Calibration and Fusion Resulting in a Dataset for Roadside Event-Based and RGB Cameras

Christian Creß*

Walter Zimmer

Venkatnarayanan Lakshminarasimhan

Nils Purschke

Leah Strand

Bach Ngoc Doan

Alois C. Knoll

Abstract—Event-based cameras are predestined for Intelligent Transportation Systems (ITS). They provide very high temporal resolution and dynamic range, which can eliminate motion blur and make objects easier to recognize at night. However, event-based images lack color and texture compared to images from a conventional rgb camera. Considering that, data fusion between event-based and conventional cameras can combine the strengths of both modalities. For this purpose, extrinsic calibration is necessary. To the best of our knowledge, no targetless calibration between event-based and rgb cameras can handle multiple moving objects, nor data fusion optimized for the domain of roadside ITS exists, nor synchronized event-based and rgb camera datasets in the field of ITS are known. To fill these research gaps, based on our previous work, we extend our targetless calibration approach with clustering methods to handle multiple moving objects. Furthermore, we develop an early fusion, simple late fusion, and a novel spatiotemporal late fusion method. Lastly, we publish the TUMTraf Event Dataset, which contains more than 4k synchronized event-based and rgb images with 21.9k labeled 2D boxes. During our extensive experiments, we verified the effectiveness of our calibration method with multiple moving objects. Furthermore, compared to a single rgb camera, we increased the detection performance of up to +16% mAP in the day and up to +12% mAP in the challenging night with our presented event-based sensor fusion methods. The TUMTraf Event Dataset is available at <https://innovation-mobility.com/tumtraf-dataset>.

Index Terms—Event-Based Cameras, RGB Cameras, Sensor Fusion, Targetless Calibration, Multi-modal Dataset, Intelligent Transportation Systems

I. INTRODUCTION

The principle of event-based cameras is the recognition of changes in the brightness of each pixel asynchronously. This technique results in a very high temporal resolution and a very high dynamic range [1], [2]. Therefore, an event-based camera is predestined for Intelligent Transportation Systems (ITS). Even in difficult visibility conditions, these systems require robust and accurate perception of traffic participants. Here, event-based cameras can achieve significant improvements, e.g., at night, with unclear visibility or even fast-moving objects, leading to motion blur when using conventional cameras. However, the disadvantages of these novel sensors are the lack

All authors are with the Chair of Robotics, Artificial Intelligence and Real-time Systems, TUM School of Computation, Information and Technology, Technical University of Munich, Munich, Germany.
E-mail: christian.cress@tum.de,
knoll@in.tum.de

* Corresponding author.

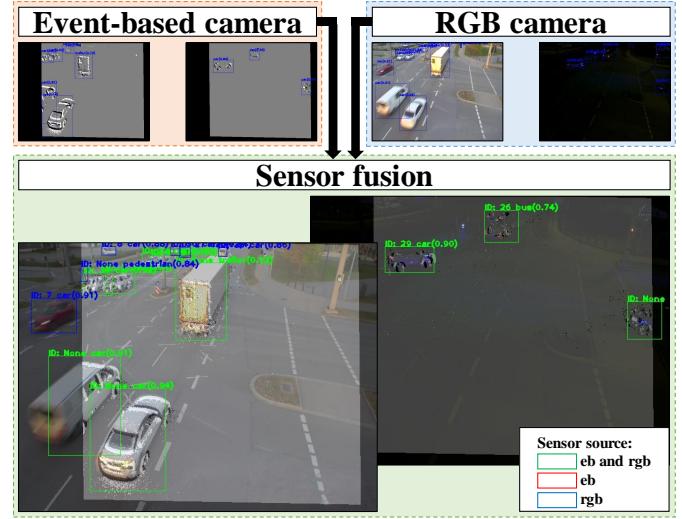


Fig. 1: This figure shows the sensor fusion between event-based and rgb cameras and its impact during a sunny day and a night in slet. The blue bounding boxes in the event-based respectively rgb camera represent detections in the sensor. In the sensor fusion component, a green bounding box indicates that the object was detected in the event-based and the rgb camera. However, a blue bounding box indicates an object detection exclusively in the rgb camera, and a red bounding box exclusively in the event-based camera (not available here). As soon as objects have been detected in several frames, a unique track ID is assigned.

of color and texture information compared to conventional cameras and the fact that the event-based camera only detects moving objects. An optimized combination of roadside event-based and conventional rgb cameras offers advantages from both modalities. So far, ITS perception mainly relies on conventional cameras, Radars, and Lidars. Event-based cameras are yet to be widespread but are slowly being established there [3]. For this reason, an investigation into calibration, detection, and data fusion between roadside event-based and conventional cameras is necessary.

Detection and tracking with a stationary event-based camera mounted on roadside ITS was performed for the first time we know by [4]. Here, clustering and tracking methods achieved sufficient performance. In addition, [5] used a stationary event-based camera for pedestrian detection with a convolutional

neural network (CNN). For data fusion, the authors [6]–[9] carried out fusion on the feature level from a perspective with ego-motion. Furthermore, [9] performed early fusion, and [10] late fusion. Nevertheless, a fusion between event-based and rgb cameras is a developing field [9]; more knowledge about this topic in the area of stationary sensors in ITS is desirable. In [11]–[15] the authors provided datasets from an ego-motion perspective. Unfortunately, data with ego-motion significantly differs from data from stationary cameras. Simulators [13], [16] could tackle this problem, but suffer from the sim2real gap [9]. Another possibility is using pseudo-labels based on rgb camera detections [17]. This approach promises sufficient results, but an extrinsic calibration between event-based and rgb cameras is required. To calculate the intrinsic camera matrix, the authors [18]–[24] used classical checkerboards, which were moved in front of the camera, or used checkerboards, which emitted changes in brightness. However, calibration patterns are impractical on an ITS. Therefore, in our previous work [25], we presented a novel targetless extrinsic calibration method between event-based and rgb cameras. The approach produced adequate results but needed to be more robust in situations of multiple moving objects (e.g., several cars or shadows), which were not equally imaged by both cameras. These gaps is intended to be closed with this work.

To the best of our knowledge, there exists neither a targetless calibration approach that can handle multiple moving objects nor a data fusion between event-based and rgb cameras that takes the unique characteristics of a stationary camera setup during day and night in the domain of roadside ITS into account. Furthermore, there is still a lack of datasets in the mentioned domain.

For this reason, to increase the practicability, we improve in this work our previously presented targetless calibration approach between event-based and rgb camera [25] to better handle multiple moving objects. Furthermore, to combine the advantages of both sensor modalities, we provide three fusion approaches between event-based and conventional cameras: Early fusion, simple late fusion, and spatiotemporal late fusion based on SORT [26] tracking. Here, we demonstrate the effectiveness of our calibration and fusion methods with comprehensive experiments based on real data. For the experiments with sensor fusion, we analyzed the combination of event-based and rgb cameras during a sunny day and a night in slet on our ITS [27], see Figure 1. Lastly, we would like to share our novel TUMTraf Event Dataset. It contains synchronized event-based and rgb images, which show a complex road intersection with several traffic scenarios during the day and night. The dataset labels for training and validation are based on partially optimized pseudo-labels extracted from the YoloV7 [28]–[30] detector using an extrinsic calibration matrix obtained by our targetless calibration tool. The test dataset is carefully labeled and allows analysis with accurate ground truth data.

In summary, the main contributions of this work are:

- Based on our previous work [25], an improved targetless calibration between event-based and rgb cameras, which can handle multiple moving objects.

- An early fusion, a simple late fusion, and a spatiotemporal late fusion between event-based and rgb cameras to profit from both sensor modalities' advantages and reduce their limitations.
- A novel TUMTraf Event Dataset, which contains spatiotemporal calibrated event-based and rgb images. The dataset shows the possibility of robust detection of traffic participants with an event-based camera during day and night in the domain of ITS.

II. RELATED WORK

First, we give an overview of event-based cameras and their usage in roadside ITS. Furthermore, we briefly summarize existing calibration algorithms for multi-sensor setups and state-of-the-art detection and fusion between event-based and conventional cameras.

A. Event-based cameras in roadside ITS

Event-based cameras recognize changes in the brightness of each pixel asynchronously. As explained in our previous work [25], which refers to [2], each pixel responds independently to brightness changes in the continuous log brightness signal $L(\mathbf{u}_k, t)$. Here, an event $e_k = (\mathbf{u}_k, t_k, p_k)$ is triggered at pixel $\mathbf{u}_k = (x_k, y_k)^T$ at time t_k when the brightness change $\Delta L(\mathbf{u}_k, t_k)$ since the last event at the same pixel Δt_k reaches a threshold C :

$$\Delta L(\mathbf{u}_k, t_k) = L(\mathbf{u}_k, t_k) - L(\mathbf{u}_k, t_k - \Delta t_k) = p_k C, \quad (1)$$

where $C > 0$. The event polarity is the sign of the brightness change $p_k \in \{+1, -1\}$. Visually, we can interpret an event as motion. With this, event-based cameras have a very high temporal resolution and a dynamic range of 140 dB, far more than conventional cameras (60 dB) [2]. So, the perception system of an ITS could benefit from low energy consumption, low latency, and higher detection performance in challenging conditions (night vision, no motion blur from high-speed vehicles) [1]. Because of these advantages, event-based cameras are slowly being established in roadside ITS [3]. The authors of [4] presented the first and the only one we know detection and tracking approach using a stationary event-based camera, which was mounted on roadside infrastructure. The approach contains for object detection several clustering (e.g., DBSCAN [31]) and tracking methods (e.g., SORT [26]). They achieved sufficient detection performance with a more than 110 Hz frame rate. However, the authors noted a lack of datasets in the ITS domain. Over and beyond, we recognize in this scope research gaps in using state-of-the-art CNNs, analysis of performance in different lighting conditions (e.g., day or night), and the fusion with other sensor systems in the domain of roadside ITS.

B. Calibration

Data fusion between event-based and conventional cameras requires an accurate calibration. We can generally distinguish between target-based (e.g., with a checkerboard pattern) and

targetless methods. [18], [19] used a classical checkerboard, which was moved in front of the camera. [20] also calibrated an event-based camera with a checkerboard, whose lighting was changed using a flashlight to trigger events. Furthermore, [21], [22] apply a flashing LED grid pattern and [23], [24] apply a flashing screen with a shown checkerboard. The usage of these targets in the domain of roadside ITS is impracticable. Consequently, a targetless method is required.

In general, extrinsic calibration algorithms for multi-sensor systems are based on the principle of registering similarities. Due to thermal factors, wind gusts, or other uncontrolled movements, [32] performed a multi-camera autocalibration to improve the calibration accuracy during runtime. [33] calibrated targetless cameras, thermal cameras, and laser sensors, based on key points extraction in the images using SIFT [34] and point cloud registration between the modalities using ICP [35]. The extrinsic calibration approach between camera and Lidar of [36] used the assignment of the point cloud segmentation map and the semantic segmentation map of the camera image. Another approach for camera and Lidar in unstructured environment using Kalman Filter was developed by [37]. The authors [38] performed extrinsic targetless calibration between stationary camera and Radar based on point cloud clustering with DBSCAN [31]. Each Radar object cluster is tracked first, then assigned with the camera detections. Similar to this, [39] also have developed a calibration for the camera and Radar using the Hungarian Method to find an association between the Radar clusters, generated via DBSCAN [31], and the camera bounding boxes, generated via YoloV3 [40]. Interestingly, [39] provided an automatically labeled radar dataset based on the rgb detections.

In our previous work [25], we developed a targetless calibration method. In this work, we noticed that established image registration methods (e.g., SIFT [34]) cannot be applied to event-based images. Therefore, we extracted the edges of moving objects and registered them to each other. Our approach provided sufficient calibration accuracy if the same objects were presented in the cameras. The main limitation was the robustness against disturbances (e.g., shadow) or multiple moving objects not equally imaged by both cameras. In these cases, the algorithm was unable to produce accurate results. To the best of our knowledge, we are unaware of other targetless calibration methods for stationary event-based cameras.

C. Detection

Object detection with event-based cameras can be realized using an unsupervised learning method (e.g., clustering), spiking neural network (SNN), or a convolutional neural network (CNN). Here, the weights of the neural networks can be initialized with pre-trained knowledge from a different domain [1]. Based on these methods, robust object detection or several fusion approaches can be implemented. As already mentioned, [4] used several clustering methods (e.g., DBSCAN [31]) for object detection with event-based cameras. In Contrast, [5] applied a YoloV3 [40] object detector for pedestrian detection with a stationary event-based camera.

A primary factor for high-quality detections is a broad dataset. Therefore, [11]–[15] provided data from an event-

based camera from an ego-motion perspective (e.g., recorded from a vehicle). Unfortunately, there are significant differences in the data between an event-based camera with ego-motion and a stationary-mounted event-based camera on a roadside ITS: In addition to the difference in the general perspective, the representation of non-moving objects is logically not displayed in recordings of stationary event-based cameras. Although many datasets have been published, there is still a lack of realistic stationary event-based camera datasets from an ITS roadside perspective [1], [5]. To tackle this problem, the simulators [13], [16] could create synthetic datasets. However, particularly for event-based cameras, [9] mentioned the existing sim2real gap. Another way to get around the lack of labeled datasets is the approach of [17], which is generated in a multi-sensor setup using pseudo-labels. With this, the authors transformed the detections of an rgb camera into the domain of an event-based camera and used them as labels. This approach promises sufficient results for training CNNs with event-based images and, therefore, is also used in our work.

D. Fusion

In addition to the previously mentioned strengths of event-based cameras, there are limitations such as the lack of color and texture information [6], [7]. Nevertheless, these systems can complement frame-based rgb cameras with intelligent data fusion [9]. However, compared to the fusion of Radar, camera, and Lidar, the fusion with event-based cameras is a relatively nascent field [9]. According to [41], we can distinct between the fusion levels “Data level (early)”, “Feature level (middle)”, and “Decision level (late)”. On an early level, raw data is fused. This approach is useful when the data of the different modalities are comparable and compatible. Fusion on the feature level extracts features from each modality and combines them. The output can be used in learning algorithms. In contrast, late fusion considers the detections of each sensor modality and combines them to produce a final decision. This optimal assignment between the object detections can, e.g., be implemented using a modified Jonker-Volgenant algorithm [42], as in the late fusion for camera and Lidar of [43].

Several data fusion approaches for event-based cameras have already been developed. The authors of [6]–[9] fused the separately extracted features of event-based and rgb images, which are generated from a perspective with ego-motion. Furthermore, [9] chose a voxel grid for the representation of event-based data and extracted features from them. After applying the homography transformation from event-based to the rgb camera, the features were fused and fed as input to a RetinaNet [44] for object detection. The dataset was also recorded from a moving vehicle. In addition, [9] experimented with image reconstruction using the method proposed in [45] with the result of high computation costs. They also performed an early fusion and detection by combining the rgb and event voxels. Last but not least, a late fusion between event-based and rgb cameras was developed by [10]. The authors used DBSCAN clustering [31] as a detector for the event-based camera and RetinaNet for the rgb camera. Nevertheless, there is still a lack of knowledge about the detection performance

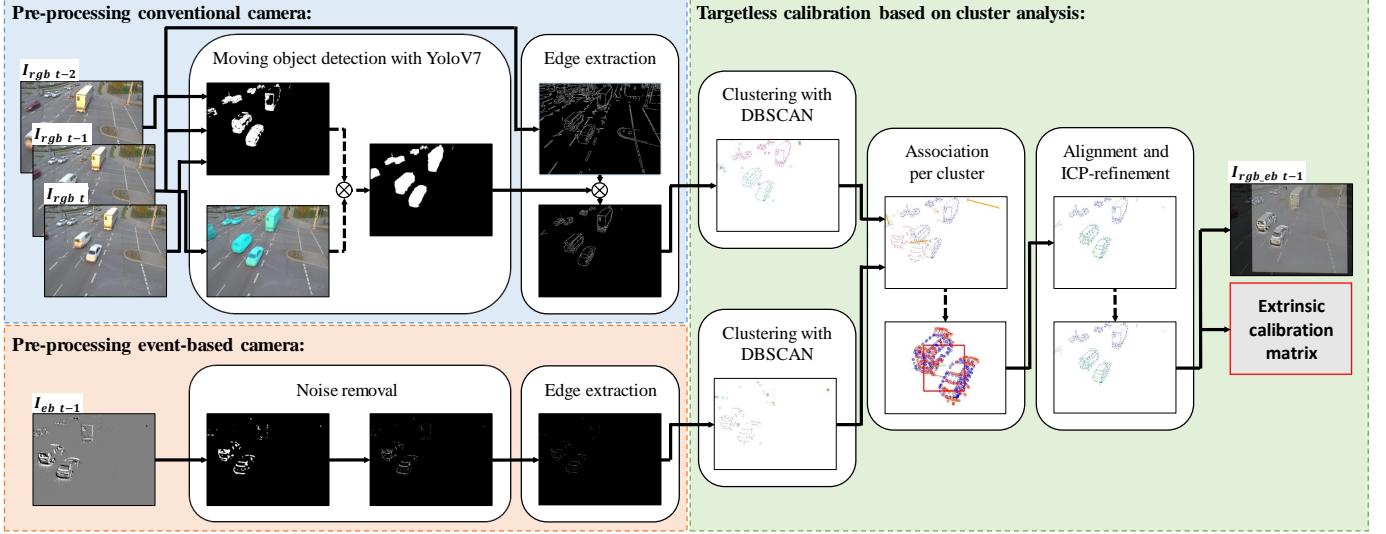


Fig. 2: The main components of our targetless extrinsic calibration algorithm are “pre-processing conventional camera”, “pre-processing event-based camera”, and “targetless calibration based on cluster analysis”. The event-based camera naturally indicates moving image areas. However, we identify such areas in the rgb camera by analyzing the last three images. We extended our previous work [25] with DBSCAN [31] and can handle multiple moving objects. The fundamental goal is to find associations per cluster pair to calculate a global transformation matrix. This approach allows us to calibrate even in more complex traffic scenarios.

of early and late fusion between a stationary event-based and rgb camera in the domain of ITS with real data under adverse weather conditions.

III. METHODOLOGY

In this section, we describe our targetless extrinsic calibration approach, which is an essential improvement of our previous work [25]. Furthermore, we present our method to generate a synchronized event-based and conventional camera dataset, and the YoloV7 object detector that we used to train on our dataset. Lastly, we close this section with our developed early fusion, simple late fusion, and novel spatiotemporal late fusion algorithms between event-based and rgb cameras optimized for usage at stationary roadside ITS.

A. Targetless calibration

As mentioned in our previous work, the image content of event-based cameras is indicated by motion in the scene. The optical flow represents the motion in images from conventional cameras. For calculating the extrinsic calibration matrix, this approach also aims to accurately match the detected motion between event-based and rgb cameras to find image correspondences between both sensor modalities. The calibration algorithm can be divided into “Pre-processing conventional camera”, “Pre-processing event-based camera”, and “Targetless calibration based on cluster analysis”. With this cluster analysis, we can tackle the main weakness of our previous works [25] that only one dynamic object can be in the field of view of all cameras. The other assumptions are still valid: We need a time-synchronized event-based and conventional camera setup for correct targetless calibration that recognizes the same objects from almost the same perspective. In addition,

the objects must have a sufficient distance from the camera to be considered as a planar plane. Figure 2 gives an overview of our extrinsic calibration approach based on multiple objects.

In the first step, as in our previous work [25], our algorithm accurately extracts the edges of moving objects in the event-based camera image. This grayscale image contains the brightness changes accumulated over the last 5000 μ s. White areas indicate event polarity of +1, and black areas of -1. Gray areas show no motion. For simplification, we ignore the polarity of the events, and therefore, convert the image into a black-and-white image: Black defines static image areas and white dynamic image areas. In contrast to [25], we apply directly a dilation operation on the image and use a kernel size of $ksize = 3 \times 3$. Then, a median filter with $ksize = 3$ removes noise from the binary image. The faster an object moves, the larger the white area in the processed event-based image. As previously shown in [25], to enhance an edge image $E \in \mathbb{R}^2$, we apply efficient morphological hit-miss operations using a combination of structuring elements (kernels $K \in \mathbb{R}^3$) in vertical, horizontal and diagonal directions, as follows:

$$\begin{aligned} K_{verti} &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{bmatrix}, K_{horiz} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \\ K_{diag1} &= \begin{bmatrix} 0 & -1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, K_{diag2} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned} \quad (2)$$

Then, we combine the edge images, based on the kernels mentioned, with $E = E_{verti} + E_{horiz} + E_{diag1} + E_{diag2}$.

In the second step, we detect the edges of multiple moving objects in the conventional camera. To enable more accurate

motion detection, inspired by [46], unlike [25], we extract motion $M_t \in \mathbb{R}^2$ with the last three grayscale images $I_t \in \mathbb{R}^2$, I_{t-1} , and I_{t-2} and a binary threshold with $T \in \mathbb{Z}$, e.g., $T = 10$:

$$\begin{aligned} D1_t &= |I_t - I_{t-1}|, \quad D2_t = |I_{t-1} - I_{t-2}|, \\ T1_t &= \begin{cases} 255 & \text{if } D1_t > T \\ 0 & \text{otherwise} \end{cases}, \quad T2_t = \begin{cases} 255 & \text{if } D2_t > T \\ 0 & \text{otherwise} \end{cases}, \\ M_t &= T1_t \wedge T2_t. \end{aligned} \quad (3)$$

Nevertheless, as in [25], we only consider motion caused by moving objects, not environmental influences, e.g., camera vibrations due to wind gusts. Therefore, inspired by [47], we also analyze the optical flow based on the methods Good Features To Track by J. Shi and Tomasi [48] and Lucas-Kanade optical flow in pyramids [49]: A flow vector $v \in \mathbb{R}^2$ with a specific length l is assigned to camera motion if

$$v_l < (m + C), \quad (4)$$

with m as the median of the length of all optical flow vectors and $C \in \mathbb{R}$ as a constant value, e.g., $C = 0.5$. The other flow vectors indicate motion by moving objects. Furthermore, we also apply a KNN background subtractor [50] to receive the motion from the image sequence. To consider the edge extraction of the complete texture of a moving object, we use deep-learning-based instance segmentation provided by YoloV7 [28]–[30], pretrained on the MS Coco dataset [51]. In contrast to our previous work [25], we don't just consider the object containing the most movement. Instead, for calculation M_{yolo_t} , we consider all detected instance segmentation masks, where the ratio of the motion r_{motion} is greater than $C_{motion} \in \mathbb{R}$, e.g. $C_{motion} = 0.2$, and the ratio to the total image r_{total} is greater than $C_{total} \in \mathbb{R}$, e.g. $C_{total} = 0.002$, as follows:

$$r_{motion} = \frac{m_i}{d_i}, \quad r_{total} = \frac{d_i}{S} \quad \text{for } i = 0, \dots, n \quad (5)$$

with the number of pixels m in the motion mask and the number of pixels d in the instance segmentation mask of each detected object i . Here, n is the total number of detected objects, and S is the total number of pixels in the image.

Then, we combine the motion mask M_t , described above, and the motion mask M_{yolo_t} with a logical OR operation. With this procedure, we obtain motion, which includes moving traffic participants and background, e.g., shadows or blowing trees in the wind. To receive the edges of these moving objects, we first apply Canny edge detection [52] on the conventional camera image, and then we combine it with the extracted motion mask via a bitwise AND operation.

At this point, similar edge images, including moving objects from the event-based and conventional camera, are available. To deal with multiple moving objects, we divide the edge images into several clusters using DBSCAN [31]. We want to emphasize the importance of dividing the event-based image into clusters, which has to be similar to the division of the conventional camera. After clustering, we determine the median centroids of each cluster. An optimal assignment

between the clusters from event-based and conventional cameras can be found with these positions. For this purpose, the linear sum assignment problem will be solved based on the implementation of a modified Jonker-Volgenant algorithm [42].

Next, we search for an association inside each cluster pair. For this, we have to align the event-based 2D point cloud optimally with the conventional camera 2D point cloud. To be robust against outliers, we create an imaginary rectangle, see red rectangle in Figure 2, between the 13th and 87th percentile for each event-based respectively conventional camera cluster. With these two rectangles r_{eb} and r_{rgb} for each cluster pair, we can, similar to our previous work [25], find a suitable scaling $s \in \mathbb{R}^2$ and displacement $t \in \mathbb{R}^2$ to transform each cluster of the event-based camera with $T_{coarse} \in \mathbb{R}^3$.

$$\begin{aligned} r_{rgb_w} &= r_{rgb}_{x_{0.87}} - r_{rgb}_{x_{0.13}}, \\ r_{rgb_h} &= r_{rgb}_{y_{0.87}} - r_{rgb}_{y_{0.13}}, \\ r_{eb_w} &= r_{eb}_{x_{0.87}} - r_{eb}_{x_{0.13}}, \\ r_{eb_h} &= r_{eb}_{y_{0.87}} - r_{eb}_{y_{0.13}}, \\ s_x &= \frac{r_{rgb_w}}{r_{eb_w}}, \quad s_y = \frac{r_{rgb_h}}{r_{eb_h}}, \end{aligned} \quad (6)$$

$$\begin{aligned} r_{eb_{scaled_x}} &= s_x \cdot r_{eb}_{x_{0.13}}, \quad r_{eb_{scaled_y}} = s_y \cdot r_{eb}_{y_{0.13}}, \\ t_x &= (-1 \cdot r_{eb_{scaled_x}}) + r_{rgb}_{x_{0.13}}, \\ t_y &= (-1 \cdot r_{eb_{scaled_y}}) + r_{rgb}_{y_{0.13}}, \\ \Rightarrow T_{coarse} &= \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned} \quad (7)$$

After translation and scaling, an optimal point-to-point assignment with the modified Jonker-Volgenant algorithm [42] inside of each cluster can be found. To filter outliers from the point assignments, we calculate the length of each point assignment i and determine the length l of the 70th percentile. If the assignment length l is greater than the length l of the 70th percentile, we calculate the extension factor $f \in \mathbb{R}$. If the factor f is greater than a threshold $C \in \mathbb{R}$, e.g., $C = 0.5$, the point-to-point assignment i is defined as outliers:

$$f_i = \frac{l_i - l_{p=0.70}}{l_i}, \quad (8)$$

$$f_i > C, \quad (9)$$

After filtering the outliers, we consider only cluster pairs, where the minimum number of assignments $M \in \mathbb{Z}$, e.g., $M = 1$, is achieved. Based on the remaining point-to-point assignments between event-based and conventional cameras, which we determined for each cluster pair, we calculate the coarse alignment as optimal affine transformation using RANSAC [53]. In the last step, we refine our coarse estimation between the point clouds of event-based and conventional cameras using point-to-point ICP [35].

TABLE I: The novel TUMTraf Event Dataset is designed for training an event-based detector from a roadside perspective. The training and validation set are generated via pseudo-labeling. It consists of labels for the event-based and rgb cameras. The test set was labeled manually to ensure accurate evaluation. Here, we split the test set into the subsets “day”(D), “night with street lights on” (N1), and “night with street lights off” (N2). Unfortunately, we noticed a lack of pedestrians, bicyclists, and motorcycles in our recordings, particularly at night. For the sake of completeness, we have listed these classes too.

ID	Class	Train EB	Val EB	Test EB (D, N1, N2)			Test RGB (D, N1, N2)		
0	Pedestrian	739	61	163	0	0	1196	0	0
1	Bicycle	17	0	38	0	0	109	0	0
2	Car	8620	1624	1211	302	44	3717	620	74
3	Motorcycle	2	0	17	0	0	21	0	0
4	Bus	352	16	37	<u>22</u>	0	49	<u>42</u>	<u>1</u>
5	Truck	<u>754</u>	<u>127</u>	155	0	0	400	2	0
6	Trailer	698	112	<u>225</u>	0	0	366	0	<u>1</u>

B. Dataset generation

According to [5], a lack of datasets exists in the domain of stationary roadside event-based cameras. For this reason, we present our approach to creating our stationary and synchronized roadside event-based and conventional camera dataset.

First, we generate spatiotemporal synchronized frame triplets of event-based, rgb, and rgb-event-based combined frames. Then, to receive accurate detections from the conventional camera, we train the YoloV7 object detector [28]–[30] on the nuImages dataset [54], which counts 93,000 images, including rain, snow, and night from the ego perspective of a vehicle [54]. To consider the roadside perspective from a stationary camera on a gantry bridge, we perform transfer learning with the 2D annotations of our TUMTraf dataset family [55], [56], which also includes snow and night images. With this robust object detector for the conventional camera and the previously calculated extrinsic calibration matrix, we generate, similar to [17], pseudo-labels using the confidence threshold $C = 0.80$. In this way, we can obtain any number of accurate pseudo-labels, which enables robust object detection for the event-based camera.

To enhance the dataset’s quality, we control roughly the training and validation set variant and exclude frame triplets where the pseudo labels are obviously incorrect. This procedure results in an optimized dataset with 2,538 frame triplets, including synchronized event-based, rgb, and rgb-event-based combined frames for the training set, 580 frame triplets for the validation set, and 915 frames for the test set. True to the motto “Train during the day, detect at night,” we intentionally select only images from the scenario day for the training and validation set. This choice enables us to achieve the best possible quality when generating the pseudo-labels. Image triples from the night scenarios are only used in the test set. This procedure is possible since the event-based camera has the same image content day and night. We also performed meticulous manual fine-tuning of each label on the test set to enable an accurate evaluation.

Table I shows the number of labels per object class. At this point, we would like to mention that the labels in the training and validation set only include moving objects so that the event-based camera can, in principle, capture them. On the other hand, we provide two categories of labels in the test set: One category is used to test the event-based camera

and includes the labels of moving objects. The other category consists of all objects that can, in principle, be recognized by the rgb camera, including not-moving objects. The labels are available in OpenLABEL format [57].

C. Detection and fusion

In this section, we describe our sensor fusion between event-based and conventional cameras, which combines the strengths of both sensor modalities mentioned above. In detail, we explain our early fusion, simple late fusion, and novel spatiotemporal late fusion approach.

As a pre-processing step for the conventional camera, we calculate a motion mask using the grayscaled rgb images I_t and I_{t-1} . Then, we can determine the per-element absolute difference between both images and, thereby, we get the motion mask M_t with the application of a binary threshold function with threshold T :

$$D_t = |I_t - I_{t-1}|,$$

$$M_t = \begin{cases} 255 & \text{if } D_t > T \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Similar to our calibration approach, we also use a KNN background subtractor [50] for refinement. The calculated motion mask allows the fusion component to distinguish between static and moving objects.

Event-based cameras accurately recognize very rapid changes in brightness. If street lights are on, the accumulated

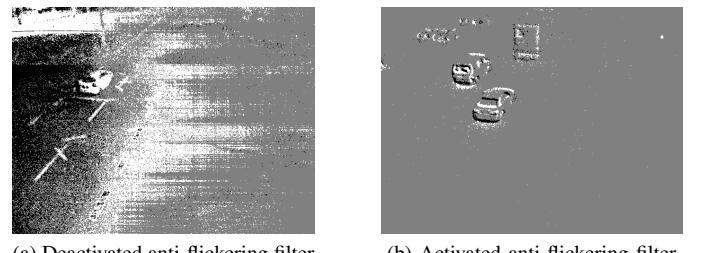


Fig. 3: Street lights cause significant noise at night, which must be eliminated with an anti-flickering filter. This phenomenon is due to the lamps operated with 50 Hz alternating current or pulse width modulation.

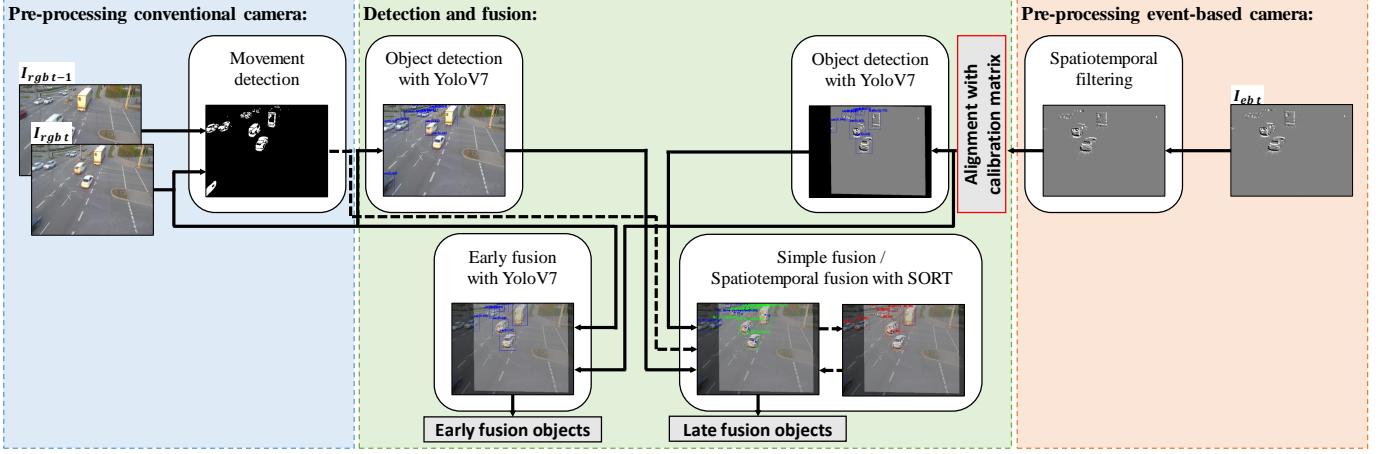


Fig. 4: This figure illustrates the logic of early fusion, simple late fusion, and spatiotemporal late fusion between event-based and conventional cameras. The early fusion uses the raw images from the rgb and event-based cameras. On the other hand, the late fusion approaches operate on the detections based on the individual raw images and the motion mask of the rgb camera. In addition, the spatiotemporal late fusion utilizes tracking information of each object for fusion decision.

grayscale image shows flickering. This phenomenon is due to the lamps operated with 50 Hz alternating current or pulse width modulation. For this reason, in the very first step in the processing pipeline of the event-based camera, the flickering is removed by the camera driver, see Figure 3. Then, inspired by [58], a spatiotemporal filter is applied as a preprocessing step for data fusion. Here, for each event $e_k = (\mathbf{u}_k, t_k, p_k)$ at pixel position $\mathbf{u}_k = (x_k, y_k)^T$ at time t_k , we investigate the spatial-temporal neighborhood $N = (N_x, N_y, N_t)$, where

$$\begin{aligned} N_x &= [x_k - r_x, x_k + r_x], \\ N_y &= [y_k - r_y, y_k + r_y], \\ N_t &= [t_k - r_t, t_k], \end{aligned} \quad (11)$$

and r_x, r_y, r_t are the sizes of the neighborhood. The amount of events in the neighborhood N is defined as noise if the total number of events in the neighborhood does not achieve the threshold $E \in \mathbb{Z}$, e.g., $E = 30$. This efficient noise suppression ensures that only moving objects are in the image of the event-based camera.

The next step after pre-processing the event-based and conventional camera is the detection and fusion. Here, we develop the methods of early fusion, simple late fusion, and a novel spatiotemporal late fusion. For early fusion, we simply blend the event-based image I_{eb} and the rgb image I_{rgb} with

$$I_{eb_rgb} = (1 - \alpha)I_{eb} + \alpha I_{rgb}, \quad (12)$$

and $\alpha \in \mathbb{R}$, e.g., $\alpha = 0.5$. Then, we train the YoloV7 detector [28]–[30] on the fused image I_{eb_rgb} .

For the simple late fusion and spatiotemporal late fusion, the YoloV7 detector [28]–[30] for the rgb camera is applied to the rgb images, and the detector for the event-based camera is applied to the aligned event-based images. In the second step, we use the previously calculated motion mask M_t and, thus, we find an optimal assignment between detected objects from the event-based camera and moving objects from the conventional camera using the modified Jonker-Volgenant algorithm [42]. If the Euclidean distance between

the objects of each pair is greater than a threshold $L \in \mathbb{R}$, e.g., $L = 50.0$, we reject the assignment. Otherwise, we create a fused object O_f and declare it as the output object of the fusion component. In principle, the rgb camera provides more texture information and is more precise in determining the object class. Consequently, the fused object receives the following properties with the weight $\alpha \in \mathbb{R}$, e.g., $\alpha = 0.4$:

$$\begin{aligned} O_{f\text{class}} &= O_{rgb\text{class}} \\ O_{f\text{position}} &= (1 - \alpha)O_{rgb\text{position}} + \alpha O_{eb\text{position}}, \\ O_{f\text{size}} &= (1 - \alpha)O_{rgb\text{size}} + \alpha O_{eb\text{size}}. \end{aligned} \quad (13)$$

Then, in the third step, we take all fused and unfused objects of the event-based and conventional camera as output objects for the simple late fusion so that moving and non-moving objects are considered. Here, we noticed that in difficult visibility conditions, e.g., night, the false positive rate for the conventional camera detector increases noticeably, even with confidence thresholds $C = 0.70$. Therefore, the YoloV7 detector [28]–[30] for the rgb images has to operate with a relatively high confidence threshold. On the other hand, due to the noise-free grayscale mask, the detector of the event-based camera produces significantly fewer to no false positives, even in night images with confidence threshold $C = 0.30$, see Figure 7.

For this reason, we develop a novel spatiotemporal late fusion, where we track each fused object and potential output object with SORT [26] and assign a unique tracking ID. So, we can identify these objects in multiple frames over time. If an object was previously detected by the event-based camera or by a combination of event-based and conventional cameras, we classify this object as trustworthy. Subsequently, only object detections from the conventional camera with a confidence threshold, e.g., greater than $C = 0.77$, or trustworthy detected objects are considered as output objects of the fusion component. This filter can allow us to significantly reduce the number of false positives.

IV. EVALUATION

In this section, we present the results of our improved targetless extrinsic calibration method based on our previous work [25]. Furthermore, we evaluate the performance of the developed event-based and rgb camera object detector. Lastly, we analyze the strengths and limitations of our presented fusion approaches: early fusion, simple late fusion, and spatiotemporal late fusion.

A. Sensor setup

Our experiments apply our new TUMTraf Event Dataset, which was recorded with our ITS from the Providentia++ project [59], [60]. As illustrated in Figure 5, roadside sensors were set up on a gantry at a height of 7 m located at an intersection in Garching near Munich, Germany. The TUMTraf Event Dataset consists of time-synchronized images from an event-based and a conventional rgb camera with the following specifications:

- **Conventional camera:** Basler ace acA1920-50gc, 1920 × 1200, Sony IMX174, global shutter, color, GigE, with 8 mm lens.
- **Event-based camera:** Imago VisionCam EB, 640 × 480, > 120 dB dynamic range, 30 000 000 events/s, with 8 mm lens.

Furthermore, to be comparable with our previous work, we also applied our targetless extrinsic calibration approach on the generated data of [25], which were recorded with identical cameras but a 16 mm lens. The intrinsic parameters and the distortion models of the event-based and rgb cameras have been calibrated beforehand.

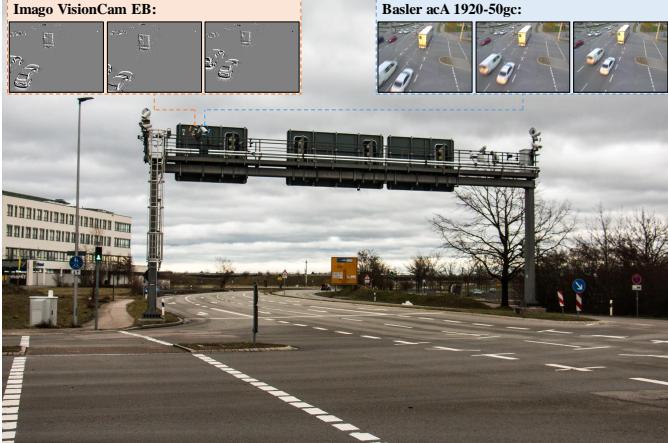


Fig. 5: We recorded the novel TUMTraf Event Dataset at this intersection in Garching near Munich. In addition to event-based and RGB cameras, the gantry contains numerous other sensors, e.g., Lidar, which form the basis for the TUMTraf Dataset family.

B. Extrinsic Calibration

To evaluate the accuracy of our improved method, we calculated the reprojection of our calibration approach based on ground truth data, which we manually created with the

same tool as in our previous work [25]. To create comparability, the test sequences 1–3 are identical to our previous work. Here, sequences 1 and 3 contain a single moving car; sequence 2 includes a crossing van with oncoming traffic very slowly. Furthermore, fast-moving traffic participants with different vehicle classes in two directions are included in test sequence 4, which is part of the test set in the TUMTraf Event Dataset. We would like to mention that our previous approach, fortunately, did not detect the very slow movement of oncoming traffic in most frames in sequence 2. However, once this multiple motion was detected, our previous approach could no longer determine a meaningful transformation matrix due to its coarse alignment procedure.

TABLE II: We measure the accuracy of our calculated extrinsic calibration and the manually created ground truth using the reprojection error in pixels. We also analyzed sequences 1–3 from our previous work for comparison purposes. Sequence 4 is part of the TUMTraf Dataset and contains numerous small self-moving objects. We can achieve similar accuracies in all sequences compared to our previous work, now also in complex traffic scenarios.

Sequence	# Frame	Rep. Err.	Rep. Err. GT
1	1	10.16	1.15
	2	10.76	1.97
	3	3.37	1.65
	4	13.62	2.76
	5	6.05	1.57
	6	5.41	1.79
2	1	8.69	3.30
	2	11.09	2.09
	3	14.49	1.73
	4	8.67	5.16
	5	13.51	2.14
	6	12.69	2.17
3	1	4.76	1.55
	2	6.42	1.42
	3	8.20	1.99
	4	6.69	1.11
	5	12.49	1.84
	6	7.85	1.69
4	1	17.41	1.62
	2	15.60	1.36
	3	7.79	1.90
	4	6.72	2.54
	5	7.70	2.84
	6	8.18	3.40

Table II shows the reprojection error of the ground truth data and our improved extrinsic calibration. In general, the results are roughly comparable to our previous work. In the sequences of a single moving vehicle, we achieve a reprojection error of up to $3.37px$. In sequence 4 from the TUMTraf Event test set, which contains several small, independently moving objects, our targetless calibration method achieved a reprojection error of up to $6.72px$. The main advantage of our improved targetless calibration approach between event-based and rgb cameras is the ability to handle multiple moving objects. Since the result from frame #3 with a reprojection error of $7.79px$ provides the subjectively best extrinsic calibration result, all further data fusion experiments were carried out with this calculated transformation. Figure 6 shows the effectiveness of our targetless calibration method, even with independently

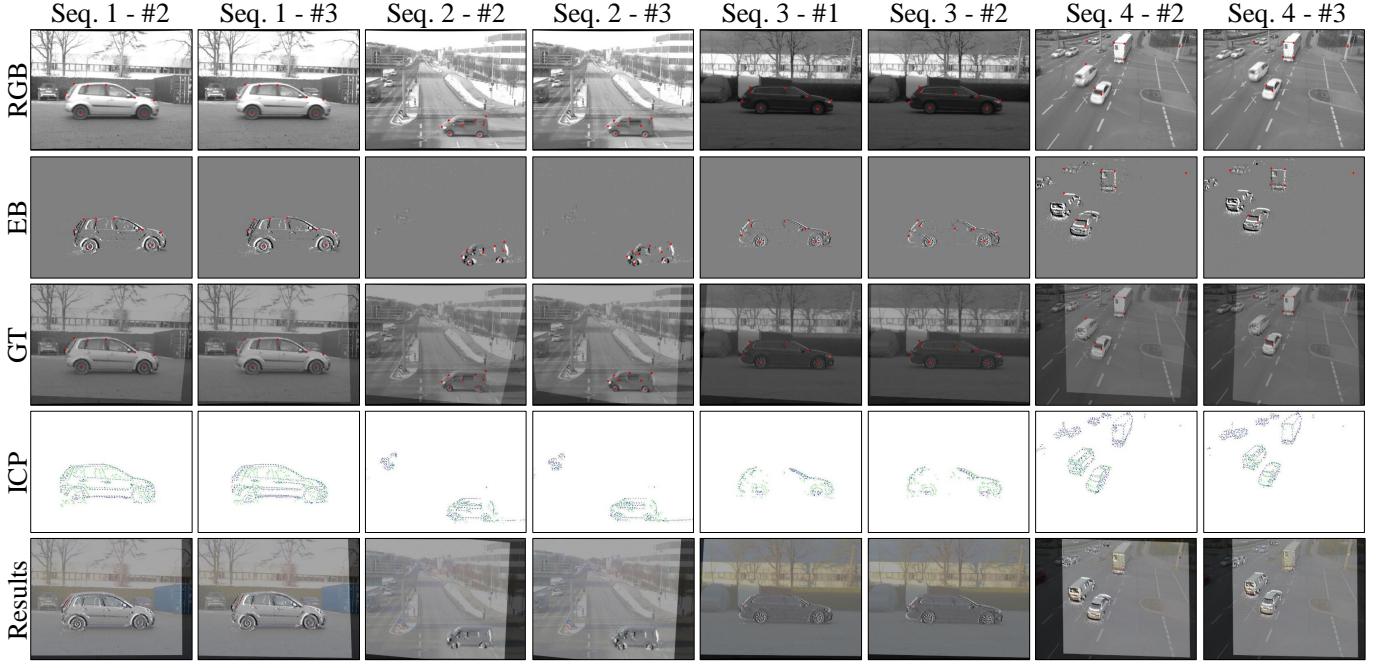


Fig. 6: The main characteristic of targetless extrinsic calibration is the association of features, e.g., edges, in both modalities. By manually selecting keypoints, we marked common image points in the images from the event-based and rgb cameras and thus created a ground truth. As can be seen, our improved targetless calibration is not only convincing for single-moving objects, the approach can also carry out valid camera calibration in complex traffic scenarios with multiple-moving objects.

moving small objects in complex traffic scenarios. Therefore, our improvement increases the flexibility of our previous work [25].

C. Detection and fusion

The detector for the rgb camera forms the basis for our pseudo-labels for the event-based sensor and is also the data input for our presented late fusion approaches. Therefore, a very high detection performance of the detector is essential for us. As described above, we first trained the YoloV7 CNN [28]–[30] on the nuImages [54] dataset. Since the TUMTraf dataset family considers roadside perspective, we finetuned the CNN to our TUMTraf Highway [55] and TUMTraf Intersection Dataset [56] using transfer learning. Here, we achieved a significant increase in detection performance from 0.36 mAP to 0.85 mAP on our combined TUMTraf test set. Thus, our rgb detector allows us to generate pseudo-labels accurately.

Based on the robust rgb detector mentioned above and the extrinsic calibration, we generated pseudo-labels to train the event-based object detectors for early and late fusion. For the late fusion approach, we trained with the event-based frames; for the early fusion approach, we trained with the rgb-event-based combined frames. We would like to emphasize that our training is based on the optimized variant of our training and validation set. Then, we evaluated early fusion, simple late fusion, spatiotemporal late fusion, and the rgb baseline on our carefully manually corrected test set, which includes the scenarios day, night with street lights on, and night with street lights off. In this way, we ensured correct ground truth during the evaluation. The performance results of all detectors, as

well as early fusion, simple late fusion, and spatiotemporal late fusion, are shown in Table III.

In the “day” scenario, we have recognized a significantly worse performance of the event-based camera with optically small objects. The event-based detector, e.g., can’t detect pedestrians or bicycles. However, the rgb camera achieves there an AP of 0.64 and 0.73. We assume that these small moving objects produce less edge information, which the CNN can use. Compared to small objects, we can achieve significantly better performance with the event-based camera with larger object classes, e.g., car (0.44 AP) or bus (0.92 AP). Suppose the objects are shown in the image sufficiently large, providing sufficient edge information. In that case, they can be recognized equally well by the rgb and the event-based cameras; see Figure 7. Surprisingly, the early fusion approach does not produce satisfactory results. We suspect that the neural network has dominantly learned the edges from the even-based camera on the day dataset and, therefore, cannot recognize objects that do not contain event-based edges. In the qualitative investigation in Figure 7, we notice that objects that are not in the field of view of both sensors cannot be detected.

The scenario “night with street lights on” includes the object classes car and bus. Unfortunately, early fusion no longer enables the detection of traffic objects in this scenario. We think due to the training data being based purely on daytime images, the dark background at night no longer allows meaningful recognition. However, the event-based and rgb camera detectors still deliver satisfactory results. We can detect a car with the event-based camera with 0.65 AP and the rgb camera with 0.71 AP. Interestingly, false positive detections occur significantly more often with the rgb detector under

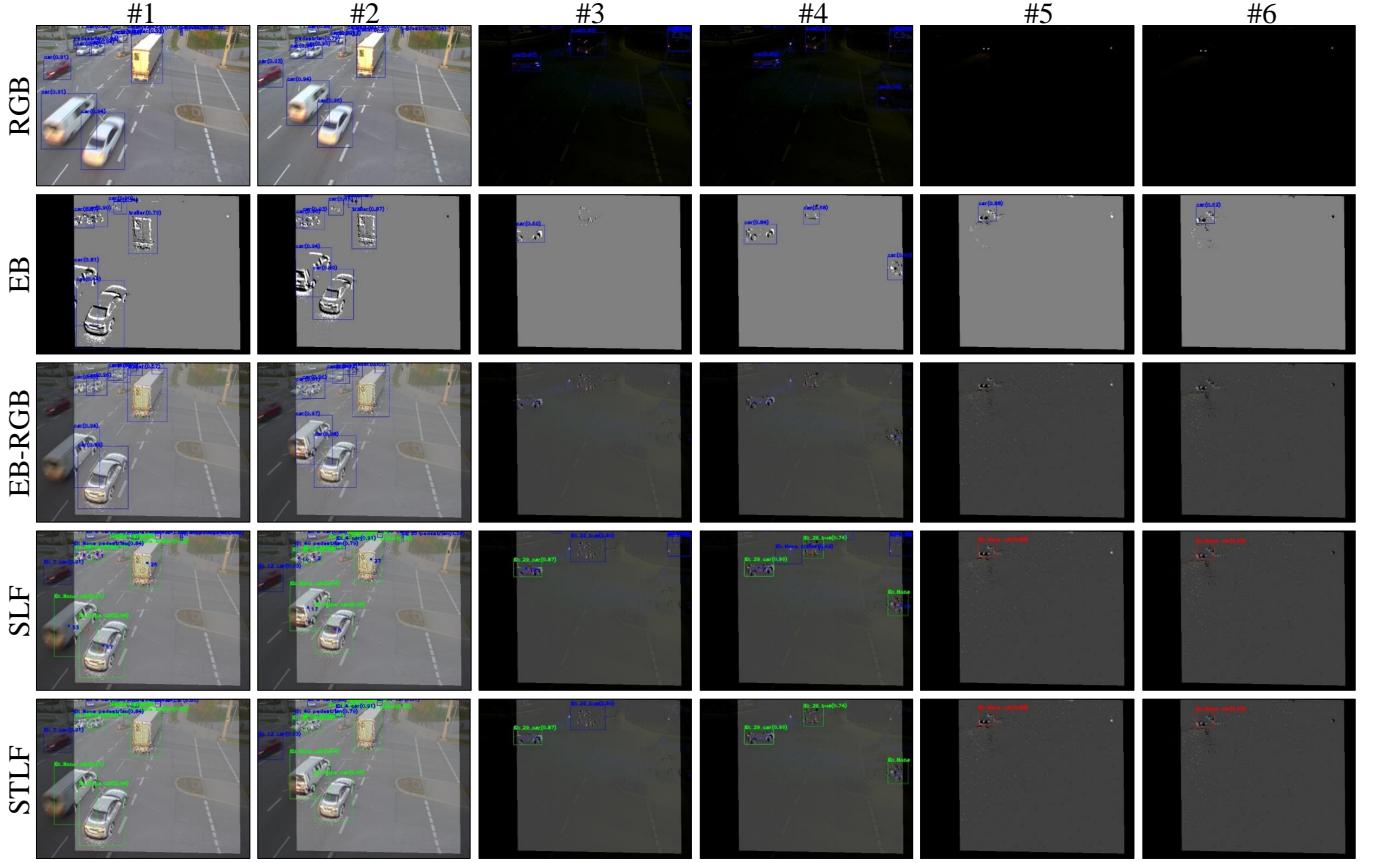


Fig. 7: This figure shows the qualitative results of the event-based (EB) and rgb (RGB) camera detectors. Furthermore, it shows the performance of the early fusion (EB-RGB), simple late fusion (SLF), and spatiotemporal late fusion (STLF) during the scenario “day” (#1–#2), “night with street lights on” (#3–#4), and “night with street lights off” (#5–#6). The blue bounding boxes in RGB, EB, and EB-RGB represent the detections of the YoloV7 detector. In the fusion components SLF and STLF, a green bounding box indicates that the object was detected in the event-based and the rgb camera. However, a blue bounding box indicates object detection exclusively in the rgb camera, and a red bounding box indicates object detection exclusively in the event-based camera. Unfortunately, the early fusion does not deliver any valuable results at night. We recognize the successful filtering of false positives in the STLF caused by the RGB camera in frames #3 and #4 in spatiotemporal late fusion and the advantages of the event-based camera in absolute darkness in frames #5 and #6. The confidence threshold was 0.3; the IoU threshold was 0.45.

these brightness conditions, as shown in Figure 7. As seen in Table III, using a higher confidence threshold of the rgb detector could solve this problem, but it would also result in significantly fewer objects being detected. At this point, the advantage of the event-based camera stands out clearly: Since the event-based camera only provides a clear image mask, false positives in the background area are significantly reduced. This effect can also be seen in the stable high precision value of the event-based camera in Table III.

In the evaluation scenario “night with streets lights off,” the high dynamic range of the event-based camera significantly improves the quality of detection: In the example shown in Figure 7, the rgb camera is no longer able to detect the car due to the extremely challenging lighting conditions. However, detection with satisfactory precision values can be achieved with the event-based camera. Nevertheless, the recall of the class car compared to the day scenario drops significantly from 0.61 to 0.34 in the event-based detector and 0.95 to 0.39

in the rgb camera, which shows the difficulties in detection in absolute darkness. For this reason, the sensor fusion of an event-based and an RGB camera can improve the overall result.

Multi-modal sensor fusion ideally achieves more robust detection results by combining two sensor systems’ strengths while eliminating their weaknesses. We have achieved this goal with our fusion. Since the sensor fusion is potentially calculated for every frame on an ITS, we would first like to examine the runtime performance. Table IV shows the runtime in the detection and fusion program flow. Early fusion requires 15 ms, while late fusion requires 43.6 ms. Interestingly, most of it requires pre-processing with 101 ms, which mainly includes the event-based camera filtering method and the application of the intrinsics and extrinsics calibration matrices. The measurements were on an Intel Core i9-9900KF CPU, NVIDIA GeForce RTX 2080 SUPER - 8GB VRAM, and 32 GB RAM. Here, GPU parallelization could significantly

TABLE III: We evaluated our detectors and fusion methods on the three subsets “day,” “night with street lights on,” and “night with street lights off” of the TUMTraf Event test set. We used precision, recall, and average precision (P, R, AP) as metrics for the event-based (EB) and rgb (RGB) detectors, as well as the early fusion (EB-RGB). The metric to measure the performance of our simple late fusion (SLF) and spatiotemporal late fusion (STLF) was the average precision (AP). The confidence threshold was 0.3; the IoU threshold was 0.45. In addition, the performance of the rgb detector, with a confidence threshold of 0.8, was also examined. The early fusion’s poor performance and the event-based camera’s high precision values are noteworthy. The drop in performance with spatiotemporal late fusion is due to the more strict spatiotemporal late fusion logic conditions. The simple late fusion method significantly outperforms the rgb detector in the subsets “day,” and “night with street lights off.” In the table, we highlighted the best AP values.

Subset	Class	EB ¹			EB-RGB ²			RGB ²			RGB ² (conf.=0.8)			SLF ²	STLF ²
Day	Pedestrian	0.10	0.08	0.02	0.01	0.00	0.00	0.91	0.66	0.64	0.97	0.17	0.17	0.84	0.33
	Bicycle	0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.77	<u>0.73</u>	0.98	0.38	0.38	0.82	0.61
	Car	0.61	0.61	0.44	0.37	0.15	0.10	0.95	0.95	<u>0.94</u>	0.99	0.83	0.83	0.98	0.88
	Motorcycle	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.48	0.37	0.44	0.19	0.20	0.85	<u>0.62</u>
	Bus	0.73	0.95	0.92	0.82	0.61	0.61	1.00	0.98	0.91	0.91	<u>0.92</u>	0.98	0.98	
	Truck	0.29	0.13	0.10	0.25	0.08	0.05	0.48	0.76	<u>0.58</u>	0.76	0.39	0.36	0.69	0.53
	Trailer	0.83	0.37	0.35	0.65	0.33	0.26	0.69	0.74	<u>0.60</u>	0.80	0.28	0.24	0.78	0.58
	\emptyset Total	0.37	0.31	0.26	0.30	0.17	0.15	0.72	0.77	0.69	0.84	0.45	0.44	0.85	0.65
Night w. st. lights on	Car	0.84	0.74	0.65	0.31	0.08	0.03	0.91	0.56	<u>0.71</u>	0.98	0.22	0.22	0.75	0.49
	Bus	0.91	0.46	0.44	1.00	0.10	0.10	0.62	0.83	0.80	0.96	0.57	0.57	<u>0.60</u>	0.60
	\emptyset Total	0.88	0.60	0.55	0.66	0.09	0.07	0.77	0.70	0.76	0.97	0.40	0.40	<u>0.67</u>	0.54
Night w. st. lights off	Car	0.88	0.34	0.34	0.25	0.05	0.01	0.85	0.39	<u>0.44</u>	1.00	0.13	0.13	0.56	0.26
	\emptyset Total	0.88	0.34	0.34	0.25	0.05	0.01	0.85	0.39	<u>0.44</u>	1.00	0.13	0.13	0.56	0.26

¹Evaluated on the test set with ground truth labels from the event-based camera.

²Evaluated on the test set with ground truth labels from the rgb camera.

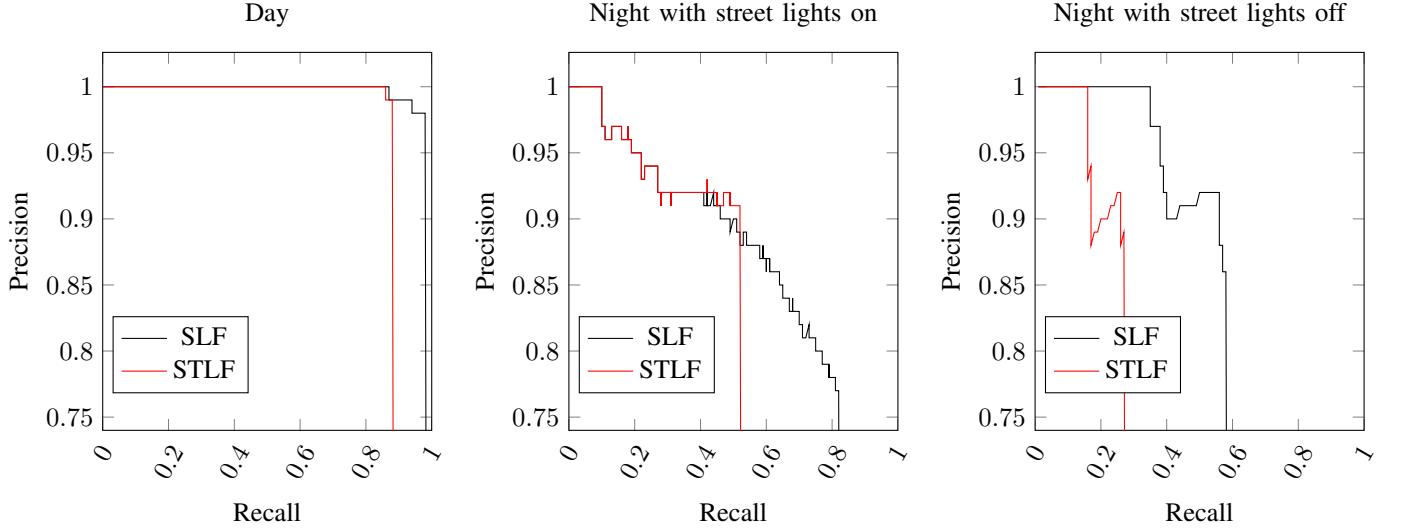


Fig. 8: The spatiotemporal late fusion (STLF) effect becomes clear on the precision-recall curve for the object class “car”: Because this fusion approach gives preferential treatment to the event-based camera, higher precision values are guaranteed. Since not every non-moving object is recognized in difficult visibility conditions, e.g., at night, the recall drops in such situations. On the other hand, simple late fusion (SLF) enables higher recall values even in difficult visibility conditions with the compromise of lower precision values. Since the object class “car” is present in all three subsets, we chose the precision-recall curve for this class to enable comparability under different visibility conditions.

accelerate the pre-processing overhead in future work.

In the next step, we analyze the detection performance of the individual fusion algorithms in the scenarios “day,” “night with street lights on,” and “night with street lights off.” Here, we used the toolkits of [29] and [61]. As shown in Table III, in the scenario “day,” we achieved with our simple late fusion an mAP of 0.85. With it, this fusion approach outperforms the rgb camera detector (0.69 mAP). Furthermore, we achieve

with the spatiotemporal late fusion an mAP of 0.65, which is, therefore, comparable to the rgb camera detector.

The sensor fusion in “night with street lights on” scenario also produces satisfactory results. Here, our simple late fusion achieves 0.67 mAP, spatiotemporal late fusion achieves 0.54 mAP. Although quantitatively both sensor fusion approaches deliver a slightly worse detection performance than the rgb camera (0.76 mAP), we were able to eliminate false positives

TABLE IV: Runtime analysis of the detection and fusion.

Component	Runtime
Preprocessing images	101 ms
Motion calculation	10 ms
CNN YoloV7 for rgb	17 ms
CNN YoloV7 for eb	15 ms
CNN YoloV7 for early	15 ms
Late fusion	1.6 ms
Total	159.6 ms

very effectively with the spatiotemporal late fusion, see Figure 7.

The simple late fusion outperforms significantly at “night with street lights off” with an mAP of 0.56 the rgb camera detector (0.44 mAP), due to the high dynamic range of the event-based camera. The spatiotemporal late fusion achieves a lower mAP value of 0.26. Interestingly, in all scenarios, the performance of the spatiotemporal late fusion is slightly weaker compared to the simple late fusion approach. This effect is because objects are only accepted by the spatiotemporal late fusion logic that either have a high confidence threshold in the rgb camera, e.g., of $C = 0.77$ or have been recognized over time by the event-based camera. In challenging brightness conditions, e.g., night, this leads to predominantly moving objects being considered for the fusion. In Figure 8, we can recognize this effect: Cars are slightly less detected in the spatiotemporal fusion, but higher precision values are guaranteed, so false positive detections are prevented. This is particularly noticeable in “Night with street lights on.” Here, the minimum possible precision value achieves 0.76, whereas the spatiotemporal late fusion achieves 0.90.

In summary, the sensor fusion of event-based and rgb cameras offers comparable or significantly more stable detection performance than a single sensor. It can, therefore, in particular in difficult brightness conditions, bring more stability into object detection for an intelligent transportation system.

V. CONCLUSION

This work has improved our previous approach [25] for targetless extrinsic calibration between event-based and rgb cameras. Here, we have extended the corresponding matching algorithm with DBSCAN clustering [31] to enable the handling of multiple moving objects. Furthermore, we have developed an early fusion, simple late fusion, and spatiotemporal late fusion based on SORT [26] tracking method to combine the strengths of event-based and rgb cameras with parallel reduction of their limitations. Last but not least, we published the novel TUMTraf Event Dataset, which contains spatiotemporal calibrated event-based and rgb images from stationary roadside cameras in the domain of intelligent transportation systems.

We showed a significant increase in the flexibility of our calibration approach, as it can now also be used in more complex traffic scenarios. In addition, we have analyzed the advantages and disadvantages of event-based and rgb cameras in roadside ITS during the day and at night. Here, we were able to show that our presented sensor fusion algorithms can

significantly increase the detection performance of up to +16% mAP compared to a rgb camera in the day and up to +12% mAP in the night.

For future work, we propose to significantly expand the database of object detectors for the event-based and rgb cameras, taking additional object classes into account so that more accurate object detection is possible during the day and at night. This goal can be achieved in a highly scalable manner using so-called pseudo-labeling. The approach necessary for this pseudo-labeling is the accurate extrinsic calibration between event-based and rgb cameras, which we provide with this work.

ACKNOWLEDGMENT

This research was funded by the Federal Ministry of Education and Research of Germany in the project AUTOtech.agil, FKZ: 01IS22088U. We would like to express our gratitude for making this paper possible.

REFERENCES

- [1] G. Chen, H. Cao, J. Conradt, H. Tang, F. Röhrbein, and A. Knoll, “Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception,” *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 34–49, 2020.
- [2] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, “Event-based vision: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2022.
- [3] C. Creß, Z. Bing, and A. C. Knoll, “Intelligent transportation systems using roadside infrastructure: A literature survey,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–0, 2023.
- [4] G. Chen, H. Cao, M. Aafaque, J. Chen, C. Ye, F. Röhrbein, J. Conradt, K. Chen, Z. Bing, X. Liu, G. Hinz, W. Stechele, and A. Knoll, “Neuromorphic vision based multivehicle detection and tracking for intelligent transportation system,” *Journal of Advanced Transportation*, vol. 2018, pp. 1–13, 2018.
- [5] G. Chen, H. Cao, C. Ye, Z. Zhang, X. Liu, X. Mo, Z. Qu, J. Conradt, F. Röhrbein, and A. Knoll, “Multi-cue event information fusion for pedestrian detection with neuromorphic vision sensors,” *Frontiers in neurorobotics*, vol. 13, p. 10, 2019.
- [6] 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 Journal*, vol. 21, no. 21, pp. 24 540–24 548, 2021.
- [7] J. Zhang, Y. Wang, W. Liu, M. Li, J. Bai, B. Yin, and X. Yang, “Frame-event alignment and fusion network for high frame rate tracking,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6/17/2023 - 6/24/2023, pp. 9781–9790.
- [8] Z. Zhou, Z. Wu, R. Boutteau, F. Yang, C. Demonceaux, and D. Ginhac, “Rgb-event fusion for moving object detection in autonomous driving,” in *ICRA 2023*, M. O’Malley, Ed. Piscataway, NJ: IEEE, 2023, pp. 7808–7815.
- [9] A. Tomy, A. Paigwar, K. S. Mann, A. Renzaglia, and C. Laugier, “Fusing event-based and rgb camera for robust object detection in adverse conditions,” in *2022 IEEE International Conference on Robotics and Automation (ICRA)*, M. O’Malley, Ed. IEEE: IEEE, 2022, pp. 933–939.
- [10] H. Sun and V. Fremont, “Object tracking with a fusion of event-based camera and frame-based camera,” in *Intelligent Systems and Applications*, ser. Lecture Notes in Networks and Systems Ser, K. Arai, Ed. Cham: Springer International Publishing AG, 2023, vol. 543, pp. 250–264.
- [11] A. Z. Zhu, D. Thakur, T. Ozaslan, B. Pfommer, V. Kumar, and K. Daniilidis, “The multivehicle stereo event camera dataset: An event camera dataset for 3d perception,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018.
- [12] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, “Hats: Histograms of averaged time surfaces for robust event-based object classification,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2018, pp. 1731–1740.

- [13] E. Mueggler, H. Rebecq, G. Gallego, T. Delbrück, and D. Scaramuzza, “The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam,” *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 142–149, 2017.
- [14] W. Cheng, H. Luo, W. Yang, L. Yu, S. Chen, and W. Li, “Det: A high-resolution dys dataset for lane extraction,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*. Piscataway, NJ: IEEE, 2019, pp. 1666–1675.
- [15] J. Binas, D. Neil, S.-C. Liu, and T. Delbrück, “Ddd17: End-to-end davis driving dataset.” [Online]. Available: <https://arxiv.org/pdf/1711.01458.pdf>
- [16] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza, “Esim: an open event camera simulator,” *Conference on Robot Learning*, pp. 969–982, 2018. [Online]. Available: <https://proceedings.mlr.press/v87/rebecq18a.html>
- [17] N. F. Y. Chen, “Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*. Piscataway, NJ: IEEE, 2018, pp. 757–75709.
- [18] M. Muglikar, M. Gehrig, D. Gehrig, and D. Scaramuzza, “How to calibrate your event camera,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2021.
- [19] K. Huang, Y. Wang, and L. Kneip, “Dynamic event camera calibration,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021.
- [20] Qifan Zhang, Jinwei Ye, Philip Osteen, and S Susan Young, *Co-Calibration and Registration of Color and Event Cameras*, 2020. [Online]. Available: <https://apps.dtic.mil/sti/citations/AD1116489>
- [21] M. J. Dominguez-Morales, A. Jimenez-Fernandez, G. Jimenez-Moreno, C. Conde, E. Cabello, and A. Linares-Barranco, “Bio-inspired stereo vision calibration for dynamic vision sensors,” *IEEE Access*, vol. 7, pp. 138 415–138 425, 2019.
- [22] E. Mueggler, N. Baumli, F. Fontana, and D. Scaramuzza, “Towards evasive maneuvers with quadrotors using dynamic vision sensors,” in *2015 European Conference on Mobile Robots (ECMR)*, 2015, pp. 1–8.
- [23] C. Plasberg, M. G. Besselmann, A. Roennau, and R. Dillmann, “Intrinsic and extrinsic calibration method for a trinocular multimodal camera setup,” in *2022 25th International Conference on Information Fusion (FUSION)*. IEEE, 7/4/2022 - 7/7/2022, pp. 1–6.
- [24] E. Mueggler, B. Huber, and D. Scaramuzza, “Event-based, 6-dof pose tracking for high-speed maneuvers,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 2761–2768.
- [25] C. Creß, E. Schütz, B. L. Žagar, and A. C. Knoll, “Targetless extrinsic calibration between event-based and rgb camera for intelligent transportation systems,” in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 6/4/2023 - 6/7/2023, pp. 1–8.
- [26] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in *2016 IEEE International Conference on Image Processing*. Piscataway, NJ: IEEE, 2016, pp. 3464–3468.
- [27] innovation mobility.com, “Providentia++: A9 testfeld für autonome fahren und digitale erkennung von fahrzeugen.” 18/08/2022. [Online]. Available: <https://innovation-mobility.com/projekt-providentia/>
- [28] arXiv.org, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” 07/07/2022.
- [29] W. Kin-Yiu, “Implementation of paper - yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” 14/01/2023. [Online]. Available: <https://github.com/WongKinYiu/yolov7/tree/mask>
- [30] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, “Blendmask: Top-down meets bottom-up for instance segmentation.” [Online]. Available: <https://arxiv.org/pdf/2001.00309>
- [31] Ester, Martin and Kriegel, Hans-Peter and Sander, Jörg and Xu, Xiaowei, “A density-based algorithm for discovering clusters in large spatial databases with noise,” *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.
- [32] G. Simarro, D. Calvete, and P. Souto, “Ucalib: Cameras autocalibration on coastal video monitoring systems,” *Remote Sensing*, vol. 13, no. 14, p. 2795, 2021.
- [33] T. Fu, H. Yu, W. Yang, Y. Hu, and S. Scherer, “Targetless extrinsic calibration of stereo cameras, thermal cameras, and laser sensors in the wild.” [Online]. Available: <https://arxiv.org/pdf/2109.13414>
- [34] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: <https://link.springer.com/article/10.1023/B:VISI.0000029664.99615.94>
- [35] P. J. Besl and N. D. McKay, “A method for registration of 3-d shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [36] A. Tsaregorodtsev, J. Müller, J. Strohbeck, M. Herrmann, M. Buchholz, and V. Belagiannis, “Extrinsic camera calibration with semantic segmentation.” [Online]. Available: <http://arxiv.org/pdf/2208.03949v1>
- [37] M. A. Munoz-Banon, F. A. Candelas, and F. Torres, “Targetless camera-lidar calibration in unstructured environments,” *IEEE Access*, vol. 8, pp. 143 692–143 705, 2020.
- [38] L. Wang, Z. Zhang, X. Di, and J. Tian, “A roadside camera-radar sensing fusion system for intelligent transportation,” in *2020 17th European Radar Conference*. Piscataway, NJ: IEEE, 2021, pp. 282–285.
- [39] A. Sengupta, A. Yoshizawa, and S. Cao, “Automatic radar-camera dataset generation for sensor-fusion applications,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2875–2882, 2022.
- [40] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *29th IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2016, pp. 779–788.
- [41] M. Brenner, N. H. Reyes, T. Susnjak, and A. L. C. Barczak, “Rgb-d and thermal sensor fusion: A systematic literature review,” *IEEE Access*, vol. 11, pp. 82 410–82 442, 2023.
- [42] D. F. Crouse, “On implementing 2d rectangular assignment algorithms,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 4, pp. 1679–1696, 2016.
- [43] W. Zimmer, J. Birkner, M. Brucker, H. Tung Nguyen, S. Petrovski, B. Wang, and A. C. Knoll, “Infradet3d: Multi-modal 3d object detection based on roadside infrastructure camera and lidar sensors,” in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 6/4/2023 - 6/7/2023, pp. 1–8.
- [44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *2017 IEEE International Conference on Computer Vision*, ser. IEEE Xplore Digital Library. Piscataway, NJ: IEEE, 2017, pp. 2999–3007.
- [45] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, “High speed and high dynamic range video with an event camera,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 1964–1980, 2021.
- [46] M. Ross, *Modellfreie, statistische Objektverfolgung in Farbbildsequenzen*. Zugl.: Koblenz-Landau, Univ., Diss., 2007. Tönning: Der Andere Verlag, 2007.
- [47] J. Kim, X. Wang, H. Wang, C. Zhu, and D. Kim, “Fast moving object detection with non-stationary background,” *Multimedia Tools and Applications*, vol. 67, no. 1, pp. 311–335, 2013. [Online]. Available: <https://link.springer.com/article/10.1007/s11042-012-1075-3>
- [48] Jianbo Shi and Tomasi, “Good features to track,” in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.
- [49] J. Bouguet, “Pyramidal implementation of the lucas kanade feature tracker,” 1999. [Online]. Available: <https://www.semanticscholar.org/paper/Pyramidal-implementation-of-the-lucas-kanade-Bouguet/a9972b40c0f8e20b07e02d1fd320bc7ebadfffc7?pdf>
- [50] Z. Zivkovic and F. van der Heijden, “Efficient adaptive density estimation per image pixel for the task of background subtraction,” *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.
- [51] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft coco: Common objects in context.” [Online]. Available: <http://arxiv.org/pdf/1405.0312.pdf>
- [52] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.
- [53] M. A. Fischler and R. C. Bolles, “Random sample consensus,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [54] Motional, “nuimages,” 08/08/2023. [Online]. Available: <https://www.nusenes.org/nuimages>
- [55] Creß, Christian and Zimmer, Walter and Strand, Leah and Fortkord, Maximilian and Dai, Siyi and Lakshminarasimhan, Venkatnarayanan and Knoll, Alois, “A9-dataset: Multi-sensor infrastructure-based dataset for mobility research,” in *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022, pp. 965–970.
- [56] W. Zimmer, C. Creß, H. T. Nguyen, and A. C. Knoll, “Tumtraf intersection dataset: All you need for urban 3d camera-lidar roadside perception,” in *2023 IEEE Intelligent Transportation Systems (ITSC)*. IEEE, 2023.

- [57] “Openlabel concept paper,” 31/12/2023. [Online]. Available: <https://www.asam.net/index.php?eID=download&t=f&f=3876&token=413e8c85031ae64cc35cf42d0768627514868b2f>
- [58] C. Yang, P. Liu, G. Chen, Z. Liu, Y. Wu, and A. Knoll, “Event-based driver distraction detection and action recognition,” in *2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. Piscataway, NJ: IEEE, 2022, pp. 1–7.
- [59] A. Krämer, C. Schöller, D. Gulati, V. Lakshminarasimhan, F. Kurz, D. Rosenbaum, C. Lenz, and A. Knoll, “Providentia – a large-scale sensor system for the assistance of autonomous vehicles and its evaluation,” *Field Robotics*, vol. 2, no. 1, pp. 1156–1176, 2022.
- [60] PROVIDENTIA – A9 Testfeld für autonomes Fahren und digitale Erkennung von Fahrzeugen, “Providentia++: Bmvdprojekt schafft digitalen zwilling,” 15/02/2023. [Online]. Available: <https://innovation-mobility.com/projekt-providentia/>
- [61] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. Da Silva, “A comparative analysis of object detection metrics with a companion open-source toolkit,” *Electronics*, vol. 10, no. 3, p. 279, 2021.



Christian Creß joined the Chair of Robotics, Artificial Intelligence, and Real-time Systems at the Technical University of Munich (TUM), Germany, in 2020 as a Research Assistant and Ph.D. student, where he is currently researching computer vision and data fusion for multi-modal sensor systems. He completed his M.Sc. in Applied Computer Science at the University of Applied Sciences Kempten in 2016. The Master’s thesis was in computer vision and machine learning. His further research interests are artificial intelligence and software architecture.



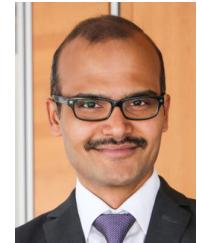
Walter Zimmer is currently a Ph.D. candidate and Research Assistant at the Chair of Robotics, Artificial Intelligence, and Real-time Systems of the Technical University of Munich (TUM). He received his M.Sc. degree in Computer Science from the Technical University of Munich in 2018. During his studies, he stayed abroad at the Technical University of Delft (Netherlands) and the University of California, San Diego (UCSD), where he was involved in developing perception and autonomous driving algorithms. His research interests are mainly 3D perception, simulation, and autonomous driving.



Nils Purschke is a PhD student at the Technical University of Munich in Germany. He started his PhD in 2023 at the Chair of Robotics, Artificial Intelligence and Real-time Systems and conducts research in the fields of autonomous driving, vehicle software architecture and artificial intelligence. He previously completed both his Bachelor’s degree (2021) and his Master’s degree (2023) at the Technical University of Darmstadt. In his master’s thesis, he focused on the creation of three-dimensional digital twins. Parallel to his scientific work, he has been working as a software developer in the field of IT security since 2020.



Bach Ngoc Doan began her B.Sc. in Informatics at the Technical University of Munich (TUM), Germany, in 2020. She is currently working on her Bachelor’s thesis at the Chair of Robotics, Artificial Intelligence, and Real-time Systems at TUM. Her thesis is in the field of computer vision and focuses on monocular 3D perception for autonomous driving.



Venkatnarayanan Lakshminarasimhan is a Ph.D. candidate and Research Assistant at the Chair of Robotics, Artificial Intelligence and Real-time Systems, TUM School of Computation, Information and Technology, Technical University of Munich. His current research interests include V2X communication in intelligent transportation systems.



Leah Strand joined the Chair of Robotics, Artificial Intelligence, and Real-time Systems at the TUM as a research assistant and a PhD candidate in March 2020. She received her Master’s degree in Mechatronics and Information Technology from the TUM after finishing her Bachelor’s degree in Engineering Science at the TUM. Her research interests are autonomous driving, multi-object tracking, sensor calibration, and fusion.



Alois Knoll (Senior Member) received the diploma (M.Sc.) degree in electrical/communications engineering from the University of Stuttgart, Stuttgart, Germany, in 1985, and the Ph.D. (summa cum laude) degree in computer science from the Technical University of Berlin (TU Berlin), Berlin, Germany, in 1988. He served on the faculty of the Computer Science Department at TU Berlin until 1993. He joined the University of Bielefeld, Germany, as a Full Professor and served as the Director of the Technical Informatics Research Group until 2001. Since 2001, he has been a Professor at the Department of Informatics, Technical University of Munich (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. Between 2007 and 2009, he was a member of the EU’s highest advisory board on information technology, the Information Society Technology Advisory Group, and a member of its subgroup on Future and Emerging Technologies (FET). In this capacity, he was actively involved in developing the concept of the EU’s FET Flagship projects. 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.