data

MDPI

Data Descriptor

Dataset of Annotated Virtual Detection Line for Road Traffic Monitoring

Ivars Namatēvs [1,*], Roberts Kadiķis [1], Anatolijs Zencovs [1], Laura Leja [1] and Artis Dobrājs [2]

[1] Institute of Electronics and Computer Science, Dzērbenes Str. 14, LV-1006 Riga, Latvia; roberts.kadikis@edi.lv (R.K.); anatolijs.zencovs@edi.lv (A.Z.); laura.leja@edi.lv (L.L.)
[2] Mondot Ltd., Balsta Dambis 80a, LV-1048 Riga, Latvia; artis.dobrajs@mondot.lv
* Correspondence: ivars.namatevs@edi.lv; Tel.: +371-264-33-567

Abstract: Monitoring, detection, and control of traffic is a serious problem in many cities and on roads around the world and poses a problem for effective and safe control and management of pedestrians with edge devices. Systems using the computer vision approach must ensure the safety of citizens and minimize the risk of traffic collisions. This approach is well suited for multiple object detection by automatic video surveillance cameras on roads, highways, and pedestrian walkways. A new Annotated Virtual Detection Line (AVDL) dataset is presented for multiple object detection, consisting of 74,108 data files and 74,108 manually annotated files divided into six classes: Vehicles, Trucks, Pedestrians, Bicycles, Motorcycles, and Scooters from the video. The data were captured from real road scenes using 50 video cameras from the leading video camera manufacturers at different road locations and under different meteorological conditions. The AVDL dataset consists of two directories, the Data directory and the Labels directory. Both directories provide the data as NumPy arrays. The dataset can be used to train and test deep neural network models for traffic and pedestrian detection, recognition, and counting.

Dataset: https://zenodo.org/record/6274296#.YjGVWJaxVPY.

Keywords: computer vision; multi-object detection; intelligent transportation systems; video surveillance; video analysis

Citation: Namatēvs, I.; Kadiķis, R.; Zencovs, A.; Leja, L.; Dobrājs, A. Dataset of Annotated Virtual Detection Line for Road Traffic Monitoring. *Data* **2022**, *7*, 40. https://doi.org/10.3390/data7040040

Academic Editor: Joaquín Torres-Sospedra

Received: 24 February 2022
Accepted: 29 March 2022
Published: 31 March 2022

1. Summary

Moving object detection using video surveillance systems [1] is an important research area for various computer vision applications, where it plays an important role in intelligent video surveillance [2,3], traffic monitoring [4], and pedestrian detection [5]. Although moving object detection methods have been extensively studied to achieve higher detection performance, i.e., accuracy [6], since videos from surveillance cameras are captured in an uncontrolled environment, the performance may be degraded due to illumination problems, complex backgrounds, occlusions, moving shadows, unpredictable motion, changes in the appearance of moving objects, and camera problems. Another important factor is processing speed, especially in real-time road monitoring with edge devices, where execution time and processing capacity are the primary concerns. If the processing speed is too slow, the performance may decrease, i.e., frame (image) dropouts may occur [7].

Several techniques have been proposed in the context of motion detection algorithms. Recently, Deep Learning (DL) has been proposed to achieve good results in detecting moving objects in the camera environment [8]. For example, Convolution Neural Networks (CNN) models in video processing have provided impressive results in tracking moving objects [9], Recurrent Neural Networks (RNN) models can be applied to various vision

tasks that involve sequential inputs and outputs, such as detecting the activity of an object in time [10], or using You Only Look Once YOLO, a DL-based real-time object detection algorithm [11]. The tremendous success of road traffic monitoring systems has been made possible primarily by improvements in the methodology for moving objects, the availability of appropriate datasets, and the computational gains achieved with GPU cards.

Currently, there are four approaches for detecting objects: background subtraction [12], optical flow [13], frame difference [14], and interframe difference [15]. These approaches are characterized by processing the captured images by manipulating pixels. Therefore, one of the approaches to develop faster object recognition algorithms is to reduce the number of pixels processed [10]. This approach can be improved by processing only a portion of the image or frame. For videos taken with a static camera, a region of interest (ROI) can be defined in the image. This means that only the pixels of ROI are processed, i.e., objects must either be inside ROI or cross ROI box to be detected. In works such as [16,17], ROI consists of multiple lines or one line within an image. The lines run perpendicular to the usual movement of the objects of interest. Such lines are also called virtual lines, detection lines, or virtual loop detectors, see Figure 1.

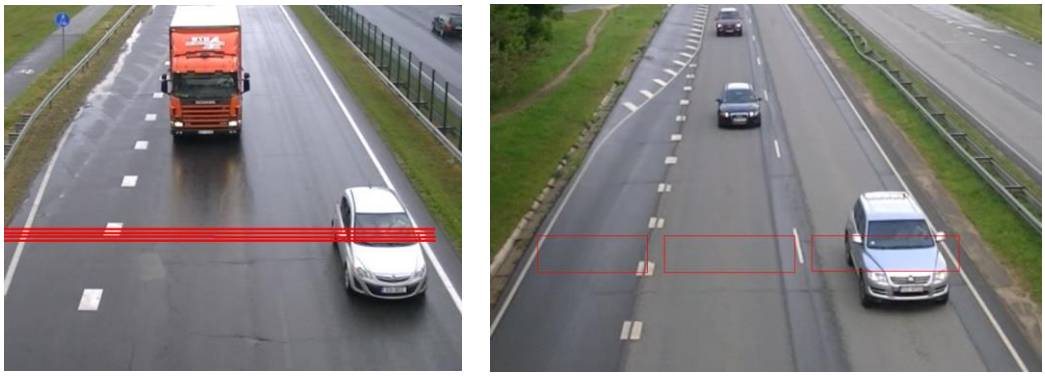

Figure 1. Examples of ROI based object detection.

Based on the detection results, object counting methods have been developed [18,19], where the information about the direction of the object's motion is based on the intersection of a single, virtual line. There is extensive research on the use of virtual line-based recognition methods or spatiotemporal images collected at such a line [10,19,20]. We need to be aware that these methods and datasets used for deep neural models are relevant to computationally limited use cases. For example, traffic monitoring in a city consisting of numerous cheap edge devices lends to such an approach. The dataset presented here is unique and suitable for most approaches used for surveillance-related applications.

Solving moving object detection with DL the important requirement is large video datasets such as Kinetic [21], GRAM Road-Traffic Monitoring [22] or MIT Traffic dataset [23] and DL frameworks such as Tensroflow, Café. For moving object detection using the ROI approach for processing speed, several practical datasets have recently emerged, e.g., the WiseNet dataset, consisting of videos for indoor human activity monitoring with annotations of meta-information for person detection and tracking [24]. The dataset presented here is unique and suitable for most approaches used for surveillance-related applications. To the best of our knowledge, this is the first large dataset to provide a ROI as a virtual pixel with encoding objects which cross it.

The proposed Annotated Virtual Detection Line (AVDL) dataset provides data based on the ROI approach, where objects are detected on the line in a video frame. The virtual detection line (a row of pixels) is placed perpendicularly to the road and covers all the width of the road. All further processing is carried out on the line only, thus making the detection algorithm computationally efficient, see Figure 2.

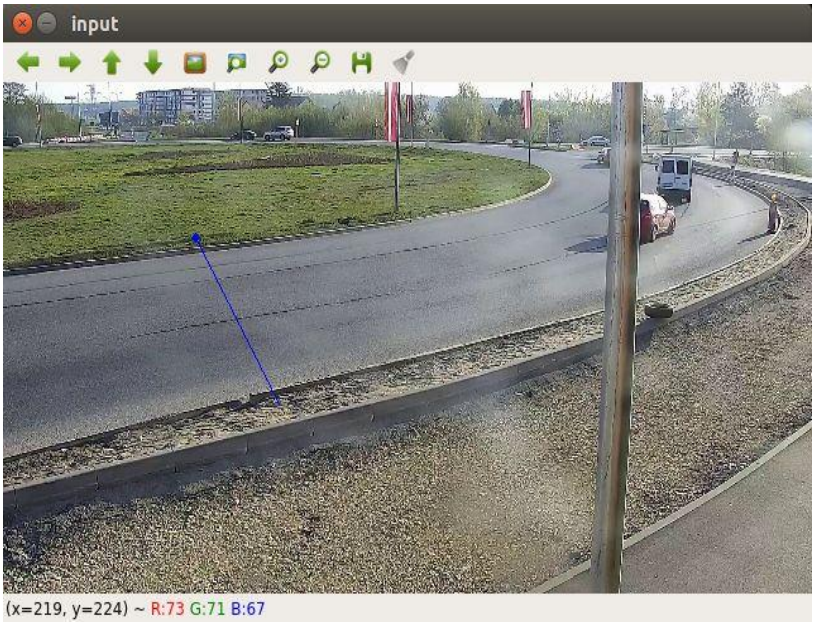

Figure 2. Examples of the positioning of a virtual object detection line on the first video frame.

The AVDL dataset provides the pixel matrices of virtual detection lines with the corresponding annotation matrices in images of moving objects in video data captured by video surveillance cameras. The approach to data collection is based on the annotation of traffic and pedestrians crossing and exiting the virtual detection line, which is perpendicular to the direction of movement of the objects. Instead of the bounding box, the six-element vector defines the correct class. The annotation process is based on the automatic creation of a time–space image, which allows us to see the object and capture the situation on the detection line in the video image. The surveillance camera video file was divided into video fragments consisting of 28 video frames. Each frame is represented by a pair of files: named.npy and name_labels.npy. The first file represents a matrix of pixel values of the detection line. The second file represents an annotated matrix. The dataset contains six manually labelled classes: Vehicles, Trucks, Pedestrians, Bicycles, Motorcycles, and Scooters. The data comes from videos taken by 50 traffic surveillance cameras at various locations in Riga and the suburbs and municipalities around Riga, Latvia. Traffic was recorded under different weather and lighting conditions. The AVDL dataset was created for traffic and pedestrian detection and counting using a deep neural network architecture. The dataset is used to develop multiple object detection and counting algorithms for low-cost edge devices.

2. Data Collection

2.1. Video Data Collection

Detecting and counting objects using computer vision technologies requires extensive video data to train, validate, and test DNN. We considered several principles before collecting video data for multiple object detection and counting from surveillance cameras. These were as follows [10]: (1) the data must be collected from different vendors and a sufficient number of video cameras, (2) the data must be collected from video cameras located at different locations in different traffic environments, from different angles and viewpoints, and (3) the data must be collected considering different environmental factors such as weather conditions and low light.

The first principle eliminates the influence of the technical characteristics and capabilities of individual surveillance camera manufacturers on the quality of the captured video data. All video sequence data in the AVDL dataset was recorded from the video cameras of the following manufacturers: AXIS Communication AB, HikVision Technology, and

DAHUA Security. These three major manufacturers together account for about 70–80% of the global video camera market. The resolution of the cameras was 640 × 480 × 3, 1920 × 1080 × 3, 2688 × 1520 × 3, 3072 × 2048 × 3. Since different video cameras record the videos with different resolutions, the corresponding FPS varies from 15 to 25. A total of 50 cameras were selected for video data collection. This was sufficient for the effective amount of video data and provided enough video material for the annotation. In this way, the technical characteristics of the respective camera manufacturer and the number of video cameras were avoided from influencing the quality of the captured data.

The second principle was to collect video data from different locations, traffic environments, and video camera viewpoints. The video data were collected from real public scenes, including highways, typical public roads, city bypasses, and pedestrian walkways, see Figure 3.

Figure 3. Examples of video images of typical video camera sightings describing the location, environment, and viewing angle of surveillance cameras to be used for data collection.

Video data were collected from real public venues, including highways, typical public roads, urban bypasses, and pedestrian walkways. The cameras we used to collect the data are installed on Riga's streets, roads, and pedestrian walkways. The following images show some of the typical video camera sightings that best describe the location of the video cameras selected for data collection.

Pan-Tilt-Zoom (PTZ) and bullet video cameras were used for data collection. The installation height of the PTZ and bullet video cameras was approximately 2.5 m above the road. Since the tilt angle, focal length, and deflection angle are variable for PTZ cameras, the recording angle can be freely rotated without a pre-set position.

The third principle considers the effects of weather conditions and lighting in obtaining video data. Videos were taken under various climatic conditions, including sunny, rainy, foggy, overcast, and cloudy weather. Lighting conditions varied between strong (day) light, insufficient (night) light, dawn and dusk, see Figure 4.

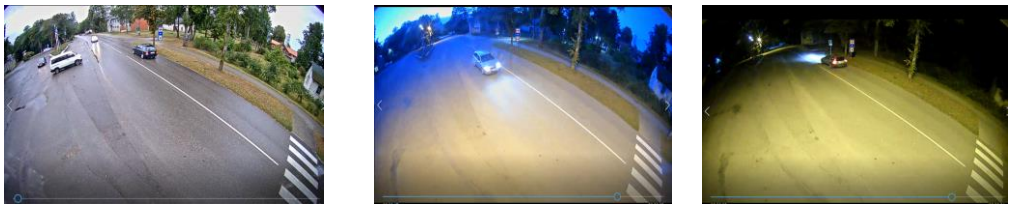

Figure 4. Examples of surveillance camera images in different weather and lighting conditions.

The recorded video streams from the accepted video cameras were processed by Luxriot EVO video management software. A specific video camera, location, and time were considered when selecting the video streams for the AVDL dataset. The video stream (video file) was divided into video fragments. Each video fragment consists of video frames. The length of each video fragment was not allowed to exceed 4 min. The whole process was performed with the help of a Python script.

2.2. Video Data Annotation

The AVDL dataset consists of two linked directories Data.zip, 2.2 GB and Labels.zip, 7 MB. All data in the AVDL dataset is represented as a ndarray for ease of use in Deep Neural Network (DNN) training, validation, and testing. The first directory, Data.zip, contains 74,108 .npy files representing the virtual detection line as a matrix of pixel values. The associated Labels.zip contains the same number of files in the form of an annotated matrix. The video files are broken into short video fragments. Each video fragment has a pair of files: name.npy and labels.npy. For example, name.npy consists of n_frames, n_pixels, n_channels, where n_frames is the number of frames in the video fragment, n_pixels is the number of pixels on the virtual detection line, and n_channels is the number of video channels. For example, name_labels.npy consists of n_frames, n_classes), where n_frames is the frame number in the video fragment and n_classes is the object class number.

For example, the file of the image (see Figure 5, first row) 2019-12-02 09-27-31 Roja-Maxima118_line_x152-706_y116-486_labels.npy is from the video recorded on 12 February 2019. The filename is decoded as follows: the abbreviation 2019-12-02 09-27-31 Roja-Maxima represents the year, day, month and location where the video was recorded. The abbreviation 118_line_x152-706_y116-486_labels.npy means the following: 118_line the length of the annotated line, the _x152-706_y116-486 the coordinates of the virtual line through the image on the x and y axis. The first endpoint is $x_1 = 152$, $y_1 = 706$, the second endpoint $x_2 = 116$, $y_2 = 486$.

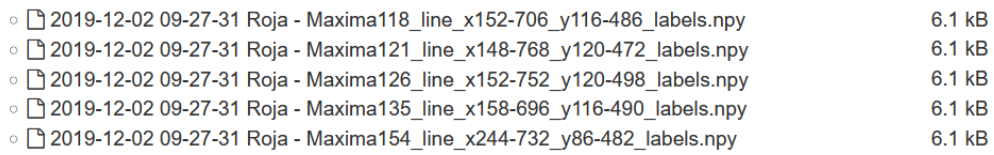

Figure 5. Screenshot from Zenodo (https://zenodo.org/record/6274296#.YjGVWJaxVPY, accessed on 23 February 2022) of the proposed AVDL dataset for road traffic monitoring consisting of the files of labelled matrices.

To facilitate annotation, the video stream was downloaded in 1.5 to 2-min videos. A total of 427 short videos were labelled. The detection line for each short video is manually determined and varies in length. A total of about 900 GB of video data was collected, or about 500 h of video.

2.3. Encoding Object Classes

The encoding consists of an n-element vector for each video frame of video fragment, where n is the number of defined classes, in the AVDL dataset $n = 6$. Each vector element is an integer, starting at 0 and above. The position of the non-zero elements in the vector indicates that at least one object has left the detection line in the frame. The number of the element determines how many objects of a certain class have crossed the detection line. For example, the encoding [0, 0, 0, 0, 0, 0] means that six object classes are possible, but none of the potential objects has cross the detection line in this frame. The encoding [0, 1, 0, 0, 0, 0] means that a single truck has left the virtual detection line in this frame. The coding [0, 3, 1, 0, 0, 0] means that three trucks and one pedestrian have left the line at this time.

After confirming the line position, the frames of the video file are prepared and processed for further visualization. The process of video visualization usually takes more than 30 s. During this process, the annotator should recognize the moment when the object of interest leaves the virtual detection line and make an annotation. Each class of objects represents its label and was coded as follows. In such coding, only one element of the vector has the value 1, while all other elements are zeros. Summary of the AVDL dataset, see Table 1.

Table 1. Summary of the total number of annotations per class and example of encoding.

	Available Classes					
	Vehicle	**Truck**	**Pedestrian**	**Bicycle**	**Motorcycle**	**Scooters**
Example of an encoding vector for the corresponding class when only one object crosses the virtual line.	[1, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0]	[0, 0, 1, 0, 0, 0]	[0, 0, 0, 1, 0, 0]	[0, 0, 0, 0, 1, 0]	[0, 0, 0, 0, 0, 1]
Number of annotations for each class.	56,134	5540	9073	2534	756	71

Note, if multiple objects of the same class left the detection line in one frame, the number of objects is entered into the label code. For instance, the code is [200,000], which means that the detection line was left by two vehicles. If the code is [210,000], it means that different types of objects have left the line in one frame. In this case, two vehicles and a truck.

For several months, EDI and Mondon staff labelled the AVDL dataset with an annotation script. Using an AVDL dataset with annotated video surveillance data, deep learning network models can be trained to detect traffic flows, track pedestrians, and determine vehicle counts. A pixel-level virtual detection line approach can be used in various object detection and tracking systems, such as detecting vehicles on highways, counting people in supermarkets or public events, detecting intruders in secure areas, and counting various objects on a conveyor belt.

Author Contributions: Conceptualization, R.K.; methodology, R.K.; software, R.K.; validation, R.K. and A.Z.; formal analysis, I.N., R.K. and A.Z.; investigation, A.D., and L.L.; resources, A.D.; data curation, L.L.; writing—original draft preparation, I.N.; writing—review and editing, I.N. and R.K.; visualization, I.N., A.D. and L.L.; supervision, R.K.; project administration, R.K. and A.D.; funding acquisition, R.K. and A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ERDF project No. 1.2.1.1/18/A/006 research No. 1.5 "Efficient module for automatic detection of people and transport using surveillance cameras".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in EDI https://zenodo.org/record/6274296#.YjGVWJaxVPY (accessed on 23 February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elharrous, O.; Almaadeed, N.; Al-Maadeed, S. A review of video surveillance systems. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103116. [CrossRef]
2. Meng, Q.; Song, H.; Zhang, Y.; Zhang, X.; Li, G.; Yang, Y. Based on Vehicle Detection and Correlation-Matched Tracking Using Image Data from PTZ Cameras. *Math. Probl. Eng.* **2020**, *9*, 1–16.
3. Kalbo, N.; Mirsky, Y.; Shabtai, A.; Elovici. The Security of IP-Based Video Surveillance Systems. *Sensors* **2020**, *20*, 4806. [CrossRef] [PubMed]
4. Liu, G.; Shi, H.; Kiani, A.; Khreishah, A.; Lee, J.Y.; Ansari, N.; Liu, C.; Yousef, M. Smart Traffic Monitoring System using Computer Vision and Edge Computing. *IEEE Trans. Intell. Transp. Syst.* **2021**, *e-print*. arXiv:2109.03141v1. [CrossRef]
5. Zhao, X.; Ye, M.; Zhu, Y.; Zhong, C.; Zhou, C.; Zhou, J. Real Time ROI Generation for Pedestrian Detection. In Proceedings of the 2009 International Conference on Computational Intelligence and Software Engineering, Wuhan, China, 11–13 December 2009.
6. Zhao, Z.-Q.; Zheng, P.; Xu, S.; Wu, X. Object Detection with Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
7. Wang, W.; Wang, L.; Ge, X.; Li, J.; Yin, B. Pedestrain Detection Based on Two-Stream UDN. *Appl. Sci.* **2020**, *10*, 1866. [CrossRef]

8. Ye, D.H.; Li, J.; Chen, Q.; Wachs, J.; Bouman, C. Deep Learning for Object Detection and Tracking form a Single Camera in Unmanned Aerial Vehicles (UAVs). *Electron. Imaging* **2018**, *10*, 4661–4666. [CrossRef]
9. Zhou, Y.; Maskel, S. Detecting and Tracking Small Moving Objects in Wide Area Motion Imagery (WAMI) Using Convolution Neural Networks (CNNs). In Proceedings of the 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019. arXiv:1911.01727.
10. Kadikis, R. Recurrent neural network based virtual detection line. In Proceedings of the Tenth International Conferences on Machine Vision (ICMV), Vienna, Austria, 13–15 November 2018; Volume 10696.
11. Zhu, J.; Wang, Z.; Wang, S.; Chen, S. Moving Object Detection Based on Background Compensation and Deep Learning. *Symmetry* **2020**, *12*, 1965. [CrossRef]
12. Liu, H.; Hou, X. Moving Detection Research of Background Frame Difference based on Gaussian model, Computer Science and Service System. In Proceedings of the International Conference on Computer Science and Service System, Nanjing, China, 11–13 August 2012; pp. 258–261.
13. Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High Accuracy Optical Flow Estimation Based on the theory for Wrapping. In *Computer Vision—ECCV 2004*; Lecture Notes in Computer Science; Pajdla, T., Matas, J., Eds.; Springer: Berlin, Germany, 2004; Volume 3024, pp. 25–36.
14. Weng, M.; Huang, G.; Da, X. A New Interframe Difference Algorithm for Moving Target Detection. In Proceedings of the 2010 3rd International Congress on Image and Signal Processing, Yantai, China, 16–18 October 2010; Volume 1, pp. 285–289.
15. Nakashima, T.; Yabuta, Y. Object Detection by using Interframe Difference Algorithm. In Proceedings of the 12th France-Japan and 10th Europe-Asia Congress on Mechatronics, Tsu, Japan, 10–12 September 2018; pp. 98–102.
16. An-an, L. Video vehicle detection algorithm based on virtual-line group. In Proceedings of the ASITIS'08, IEEE International Conference on Signal Image Technology and Internet Based Systems, Singapore, 4–7 December 2006; pp. 1148–1151.
17. Lei, M.; Lefloch, D.; Gouton, P.; Madani, K. A video-based real-time vehicle counting system using adaptive background method. In Proceedings of the 2008 IEEE International Conference on Signal Image Technology and Internet Based Systems, Bali, Indonesia, 30 November–3 December 2008; p. 523.
18. Yue, Y. A Traffic-flow Parameters Evaluation Approach based on Urban Road Video. *Int. J. Intell. Eng. Syst.* **2009**, *2*, 33–39. [CrossRef]
19. Kadikis, R.; Freivalds, K. Efficient Video Processing method for Traffic Monitoring Combining Motion Detection and Background Subtraction. In Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP 2012); Springer: New Delhi, India, 2013; pp. 131–141.
20. Zhang, Y.; Zhao, C.; Zhang, Q. Counting Vehicles in Urban Traffic Scenes using Foreground Time-spatial Images. *IET Intell. Transp. Syst.* **2016**, *11*, 61–67. [CrossRef]
21. Kinetic (Kinetic Human Action Video Dataset) Dataset. Available online: https://arxiv.org/abs/1705.06950 (accessed on 23 February 2022).
22. MIT Traffic Dataset. Available online: http://mmlab.ie.cuhk.edu.hk/datasets/mit_traffic/index.html (accessed on 23 February 2022).
23. GRAM Road-Traffic Monitoring Dataset. Available online: https://gram.web.uah.es/data/datasets/rtm/index.html (accessed on 23 February 2022).
24. WiseNet: Multi-Camera Dataset. Available online: https://doi.org/10.4121/uuid:c1fb5962-e939-4c51-bfd5-eac6f2935d44 (accessed on 23 February 2022).