remote sensing

MDPI

Article

WHUVID: A Large-Scale Stereo-IMU Dataset for Visual-Inertial Odometry and Autonomous Driving in Chinese Urban Scenarios

Tianyang Chen, Fangling Pu *, Hongjia Chen and Zhihong Liu

Electronic Information School, Wuhan University, Wuhan 430072, China; tychen@whu.edu.cn (T.C.); chj1997@whu.edu.cn (H.C.); zhihongliu@whu.edu.cn (Z.L.)
* Correspondence: flpu@whu.edu.cn; Tel.: +86-1897-169-2061

Abstract: In this paper, we present a challenging stereo-inertial dataset collected onboard a sports utility vehicle (SUV) for the tasks of visual-inertial odometry (VIO), simultaneous localization and mapping (SLAM), autonomous driving, object detection, and other computer vision techniques. We recorded a large set of time-synchronized stereo image sequences ($2 \times 1280 \times 720$ @ 30 fps RGB) and corresponding inertial measurement unit (IMU) readings (400 Hz) from a Stereolabs ZED2 camera, along with centimeter-level-accurate six-degree-of-freedom ground truth (100 Hz) from a u-blox GNSS-IMU navigation device with real-time kinematic correction signals. The dataset comprises 34 sequences recorded during November 2020 in Wuhan, the largest city of Central China. Further, the dataset contains abundant unique urban scenes and features of a complex modern metropolis, which have rarely appeared in previously released benchmarks. Results from milestone VIO/SLAM algorithms reveal that methods exhibiting excellent performance on established datasets such as KITTI and EuRoC perform unsatisfactorily when moved outside the laboratory to the real world. We expect our dataset to reduce this limitation by providing more challenging and diverse scenarios to the research community. The full dataset with raw and calibrated data is publicly available along with a lightweight MATLAB/Python toolbox for preprocessing and evaluation. The dataset can be downloaded in its entirety from the uniform resource locator (URL) we provide in the main text.

Keywords: dataset; WHUVID; SLAM; VIO; autonomous driving; urban scenarios

Citation: Chen, T.; Pu, F.; Chen, H.; Liu, Z. WHUVID: A Large-Scale Stereo-IMU Dataset for Visual-Inertial Odometry and Autonomous Driving in Chinese Urban Scenarios. *Remote Sens.* **2022**, *14*, 2033. https://doi.org/10.3390/rs14092033

Academic Editor: Andrzej Stateczny

Received: 18 March 2022
Accepted: 21 April 2022
Published: 23 April 2022

1. Introduction

Datasets related to visual-inertial odometry (VIO), simultaneous localization and mapping (SLAM), and autonomous driving released by various research institutions and colleges in the past decade have greatly promoted the development of these technologies, e.g., UTIAS Multi-Robot [1], San Francisco Landmark [2], SeqSLAM [3], CCSAD [4], Cityscapes [5], NCLT [6], MPO-Japan [7], etc. Several notable works have emerged among these datasets. The FORD dataset [8] published two sequences of images and three-dimensional (3D) laser data recorded at a research campus and downtown Dearborn in Michigan, United States with corresponding inertial measurement unit (IMU) data and six-degree-of-freedom (6-DOF) ground truth, claiming to be the first to add visual information to the structure of the environment by fusing image and laser. KITTI [9,10] is the most well-known and widely used large dataset to date that provides benchmarks for various computer vision tasks, including (but not limited to) stereo matching, optical flow, visual odometry, SLAM, depth estimation, and 3D object detection. TUM [11] extends the ability of SLAM algorithms from optical-only to RGB and depth (RGB-D) by collecting accurately calibrated and aligned optical and depth images as well as providing several reasonable evaluation metrics. In addition, datasets collected from other platforms beyond four-wheeled vehicles, such as motorcycles [12], mobile robots [13,14], unmanned aerial vehicles (UAV) [15,16], autonomous underwater vehicles (AUV) [17–19], and handheld

devices [20,21] have been released continuously, which has collectively resulted in new breakthroughs in VIO/SLAM and autonomous driving.

However, the aforementioned datasets have two key issues that cannot be ignored. First, they have shortcomings such as insufficient number of images, low resolution and sampling frequency, and lack of scene dynamics. For example, the FORD dataset [8] contains only around 7000 groups of images with a narrow horizontal perspective smaller than 80°, and its 8 fps image acquisition rate is much less than the real-time sampling rate of 30 Hz. These defects make the dataset unstable at turnings and unable to evaluate the long-term robustness of algorithms. Further, KITTI [9,10] is a large-scale benchmark but does not record IMU information. Moreover, the part prepared for visual odometry (VO) lacks scene complexity and the number of samplings is also limited. Furthermore, numerous datasets are artificially split up into multiple segments, where each segment represents a continuous and smooth driving process at almost constant speed. Such a method facilitates the evaluation of algorithms to a certain extent; however, it does not reflect real-world driving conditions. The second issue is that most of the existing datasets have a consistent style because of being recorded in European and North American cities, where fields of traffic, roads, buildings, and residential areas share similar designs and planning. In other words, these datasets are unable to comprehensively cover all characteristics of modern cities [22,23]. For example, urban landscapes common in China, such as contiguous skyscrapers, wide and crowded roads, complex overpasses, large sections of long tunnels, and huge bridges, are rarely included, introducing obvious flaws in the evaluation of VIO/SLAM and autonomous driving algorithms.

In the present study, compared with previous works as summarized in Table 1, we release an exemplary dataset collected in Chinese urban scenarios—Wuhan Urban Visual-inertial Dataset (WHUVID)—aiming to increase the completeness and richness of urban scenes for evaluation. To the best of our knowledge, this dataset is currently the largest and latest visual-inertial benchmark for VIO [24–26], SLAM [27–29], and autonomous driving recorded in a highly modern Chinese city. WHUVID was recorded using an off-the-shelf Stereolabs ZED2 camera in November 2020, covering three main urban areas of Wuhan, namely, Wuchang, Hankou, and Hanyang, as shown in Figure 1. It contains general traffic situations with many static/dynamic objects in multiple scenes such as campus, municipal roads, viaducts, tunnels, parking lots, and huge bridges. In addition, we label all the data with five time slots (morning, midday, afternoon, sunset, and night) according to the recording time to distinguish different lighting conditions. WHUVID contains 34 publicly downloadable sequences, including more than 336,000 pairs of high-resolution and full-frame-rate binocular RGB images and up to 4.5 million 6-axis (triaxle acceleration and angular velocity) IMU readings, with a cumulative duration of 11,285 s and a travel distance of 82.01 km. We also provide timestamped 100 Hz 6-DOF ground truth collected and calculated by a u-blox C100-F9K GNSS-IMU navigation device in a uniform format (x-y-z for translation and quaternion for rotation). To expand the application scope of the dataset, we manually annotated 1860 images for the object-detection task, obtaining 7485 bounding boxes in four categories (car, person, bike, and label). We further used them as a training set and obtained up to 682,000 annotations by applying YOLOv4 [30] on the entire dataset, in which dynamic targets account for more than 90%. Finally, we employed two milestone algorithms, namely, ORB-SLAM2 [31,32] and VINS-Mono [33,34], to evaluate the performance of WHUVID with five other well-known datasets in three cases: monocular, stereo, and visual-inertial. Results show that algorithms that perform well on other formerly released datasets make more mistakes and cause larger errors on WHUVID, indicating that our dataset poses more challenges and requirements.

Table 1. Comparison of established publicly available benchmarks and datasets recorded onboard vehicles and in outdoor environments (except where specified) with vision and IMU data over the last decade. In the table header, "#Seq" denotes the number of sequences of the corresponding datasets and "Dur", "Len", "Avg", "Spd", and "GT" denote duration, length, average, speed, and ground truth, respectively.

Dataset	Release Year	Position	#Seq	#Frame	Dur/s	Len/km	Avg Spd/(m/s)	Camera Parameter	#Category	#Label	GT Quality
FORD [8]	2011	Dearborn, USA	2	7 k	938	5.4	5.8	omni × 6 RGB 1600 × 600 @ 8 fps	None	None	6-DOF
KITTI [9,10]	2013	Karlsruhe, Germany	22	41 k	4517	39.2	8.7	stereo RGB 1241 × 376 @ 10 fps	5	—	6-DOF
Malaga [35]	2014	Malaga, Spain	15	113 k	5655	36.8	6.5	stereo RGB 1024 × 768 @ 20 fps	None	None	3-DOF
Oxford [1] [36]	2016	Oxford, UK	1	35 k	2455	9.3	3.8	stereo RGB 1280 × 960 @ 16 fps	None	None	6-DOF
EuRoC [2] [15]	2016	Zurich, Switzerland	11	27 k	1373	0.89	0.65	mono GRAY 752 × 480 @ 20 fps	None	None	6-DOF
MVSEC [12]	2018	West Philly, USA	5	37 k	1813	9.6	5.3	stereo GRAY 752 × 480 @ 20 fps	None	None	6-DOF
WHUVID	**2021**	**Wuhan, China**	**34**	**336 k**	**11,285**	**82.0**	**7.2**	**stereo RGB 1280 × 720 @ 30 fps**	**4**	**681 k**	**6-DOF**

[1] The Oxford dataset was collected repeatedly along the same route more than 100 times over the period of a year. Here, we employed data from 28 November 2014. [2] The EuRoC dataset was collected indoors on board a UAV, and it was only involved in the evaluation of the visual-inertial case latter in Section 5.

We expect WHUVID to contribute to further improving the robustness and reliability of VIO/SLAM and autonomous driving algorithms; therefore, we have made it open access under the CC-BY 4.0 license (available at https://github.com/chentianyangWHU/WHUVID accessed on 15 March 2022). For personal and social privacy, relative position data in the metric of the east-north-up (ENU) coordinate system are published instead of raw longitude and latitude readings, and sensitive information that is clearly visible in images, such as license plates, faces, and special signs, is blurred with mosaic. Excluding the above mentioned information, the website contains full information with raw and calibrated data, calibration manuals, demo programs, videos for visual inspection, and a toolbox for evaluation and preprocessing supporting both MATLAB and Python.

The main contributions of this paper can be summarized as follows:

1. We propose WHUVID, the latest and largest calibrated and synchronized visual-inertial dataset collected from Chinese urban scenarios with abundant scenes not previously included, along with high-quality recordings and accurate ground truth.
2. We present a brief review of numerous previously published datasets and conduct a detailed evaluation and comparative experiments between some of them and WHUVID, proposing some original evaluation metrics.

The remainder of this paper is structured as follows. Section 2 describes the sensors used as well as their setup. Section 3 gives a detailed introduction of the composition of the dataset and how it is collected and annotated. Content related to data preprocessing, such as calibration and synchronization, is given in Section 4. Evaluation experiments and discussion from multiple aspects are described in Section 5. Finally, a brief conclusion of our work and suggestions about future research are given in Section 6.

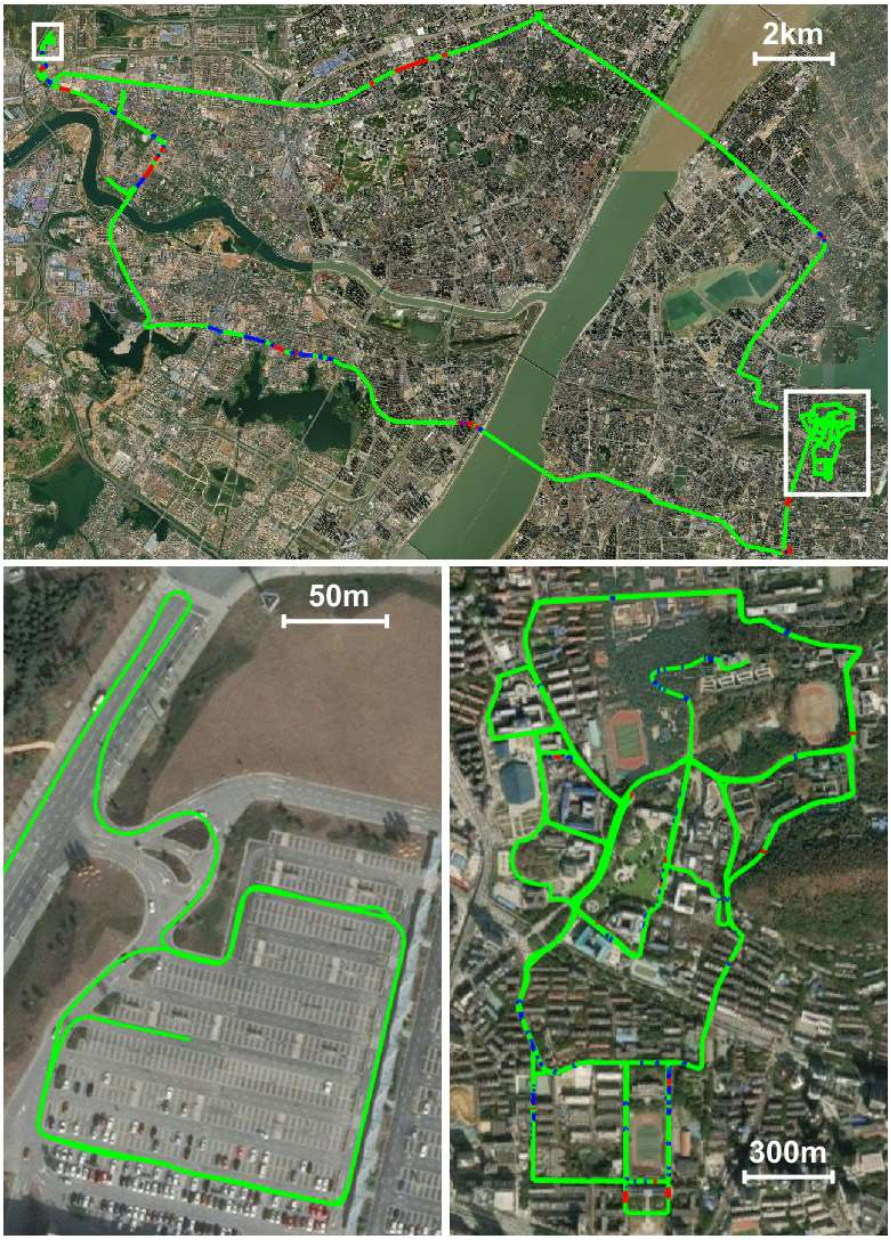

Figure 1. Recording zone. This figure shows the complete trajectory of our dataset (**up**) in the city center of Wuhan, China, together with two enlarged areas marked with white boxes, namely, a parking lot (**bottom left**) and the campus of Wuhan University (**bottom right**). Colors encode the signal quality of the Global Navigation Satellite System (GNSS): green represents EXCELLENT quality with the Position Dilution of Precision (PDOP, a positioning accuracy evaluation indicator; refer to Section 3.1.2 for detailed information) less than 3; blue represents GOOD quality with PDOP between 3 and 7; and red represents BAD quality with PDOP between 7 and 100.

2. Sensor Setup

The sensor setup is illustrated in Figure 2:

1. A Stereolabs ZED2 integrated VI sensor with stereo lenses (1/3″ 4MP CMOS, 2688 × 1520 pixels with each pixel of size 2 × 2 microns, electronic synchronized rolling shutter, baseline: 120 mm, focal length: 2.12 mm, field of view (FOV): 110° horiz. 70° vert.) and a consumer-grade built-in IMU (motion measurement with 6-DOF @ 400 Hz ± 0.4% error, magnetometer with 3-DOF @ 50 Hz ± 1300 μT);

2. A u-blox GNSS-IMU navigation device (184-channel u-blox F9 engine; supporting GPS, GLONASS, BeiDou, Galileo, SBAS, and QZSS; position accuracy < 0.2 m + 1 ppm CEP with real-time kinematic (RTK); and data-update rate up to 30 Hz, with a built-in IMU for a GNSS-denied environment) with a GNSS receiver and a C100-F9K integrated module (IMU inside).

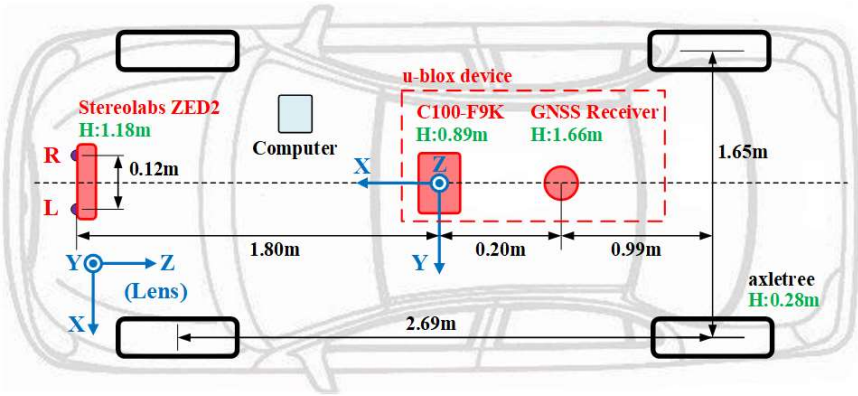

Figure 2. Sensor setup. This figure shows the mounting positions of the sensors (marked in red) with respect to the vehicle body. Heights above ground are marked in green and measured with respect to the road surface. The axes of the lenses of Stereolabs ZED2 and IMU (inside the C100-F9K integrated module) are marked in blue.

As shown in Figure 3, the Stereolabs ZED2 camera is fixed to the front cover of the vehicle using a sports-level suction cup bracket, the GNSS receiver is firmly attached to the top of the vehicle with a magnet, and the C100-F9K integrated module is adhered to the storage table—calibrated using a leveling instrument—in the middle of the two front seats by glue. Distances between the three abovementioned devices are manually measured with centimeter-level accuracy. The Euler angles of the ZED2 camera can be adjusted through screws, but its position and pose as well as those of the other two devices are maintained throughout the recording process. Our sports utility vehicle (SUV) houses a Dell PC with an i7-8750H processor and a 1-TB Western Digital MyPassport SSD, connected to the host computer via a type-C interface. Our computer runs Ubuntu Linux (64 bit) 16.04 and multi-process programs to store stereo images and IMU data from ZED2 and position readings from the u-blox device in real time.

Figure 3. Recording platform. We recorded a large dataset for evaluating VIO/SLAM and autonomous driving using a Stereolabs ZED2 (**top left**) onboard an SUV. The ground truth was collected by a u-blox navigation device, including a GNSS receiver (**bottom left**) and a C100-F9K integrated module (**bottom right**). All these devices were connected to and timestamped by the host computer (**top right**).

3. Dataset

The ZED2 is a USB Video Class video camera with low-level access comprising dual image sensors and motion/environment sensors inside. The u-blox device is an IMU-assisted GNSS navigation instrument designed for vehicles in the case of a GNSS-denied environment. The host computer connects with and collects information from ZED2 and the u-blox device through type-C and USB, respectively, to obtain visual-inertial data and ground truth simultaneously. The dataset contains various scenes including (but not limited to) huge bridge, skyscraper, campus, car park, night, tunnel, and under viaduct. Example frames are illustrated in Figure 4. We picked up 34 sequences that were publicly downloadable and with different challenges in terms of factors such as illumination, travel length, and traffic congestion; their statistical summary is given in Table 2. For each sequence, we provided binocular images, IMU readings, ground truth, and object annotations in the form of two-dimensional (2D) bounding boxes, as illustrated in Figure 5. The recordings were conducted on 13 and 14 November 2020 for the entire day. The total size of the provided data was 570 GB.

Table 2. Statistics of 34 sequences of WHUVID. This table lists several key attributes of each sequence, including scene, time period, duration, frame amount, vehicle speed, object annotation amount, average dynamic amount, and area ratio of each frame. GNSS quality is evaluated using the median PDOP.

Id	Scene [1]	Period	Duration/s Length/m	#Frames	Speed/(m/s)				#Labels	Dynamic [2]		IMU [3]	GNSS	GNSS Quality
					Max	Mid	Mean	Min		Num	Ratio/%			
01	campus	p.m.	301, 1511	9006	12.1	4.9	5.0	0	33,461	3.7	5.0	Y	Y	4.2
02	campus	p.m.	288, 1610	8631	9.9	5.4	5.6	1.4	25,635	2.8	8.0	Y	Y	2.7
03	campus	p.m.	283, 1531	8486	10.4	5.5	5.4	0.4	18,021	2.0	3.2	Y	Y	2.7
04	campus	p.m.	172, 1100	5104	11.5	8.0	6.4	1.5	—	—	—	N	PD	1.4
05	campus	sunset	1188, 4775	35,576	10.0	3.7	4.0	0	—	—	—	N	PD	2.4
06	campus	night	822, 4775	24,618	12.9	5.8	5.8	0.1	—	—	—	Y	Y	1.9
07	campus	night	166, 790	4914	9.3	4.5	4.8	0	—	—	—	Y	Y	1.8
08	campus	night	403, 2474	11,825	11.3	6.4	6.1	1.1	—	—	—	N	Y	2.2
09	campus	night	490, 3221	14,655	10.9	6.9	6.6	1.2	—	—	—	Y	Y	2.8
10	campus	night	131, 887	3866	9.4	7.1	6.8	0.8	—	—	—	Y	Y	2.8
11	campus	night	58, 352	1727	8.4	6.3	6.1	1.4	—	—	—	Y	Y	2.9
12	campus	night	495, 3028	14,784	10.4	6.3	6.1	1.1	—	—	—	Y	Y	2.3
13	campus	a.m.	569, 3891	16,985	11.4	6.9	6.8	2.1	41,410	2.3	4.1	Y	Y	2.0
14	campus	a.m.	797, 4531	23,811	11.3	5.7	5.7	0	52,373	2.0	4.0	Y	Y	2.7
15	campus	a.m.	248, 980	7354	8.8	5.1	5.4	0	11,785	1.5	2.7	Y	Y	2.0
16	urban-RD	midday	608, 4165	18,070	17.1	5.8	6.9	0	79,265	4.1	8.8	Y	Y	1.9
17	urban-BG	midday	621, 8028	18,420	18.9	13.5	12.9	0	83,138	4.1	8.6	Y	Y	1.7
18	urban-RD	midday	267, 2925	8000	17.3	10.6	11.0	2.8	36,472	4.1	7.9	Y	Y	1.9
19	urban-TN	midday	35, 361	1042	13.8	10.3	10.3	7.1	3251	3.1	6.6	Y	Y	100
20	urban-RD	midday	143, 2112	4003	17.9	14.5	14.8	9.9	17,549	3.9	5.4	Y	Y	1.6
21	urban-RD	midday	164, 2902	4903	23.1	17.4	17.7	10.6	8300	1.6	2.3	Y	Y	1.4
22	urban-RD	midday	137, 1137	4107	13.6	9.9	8.3	0	14,414	3.1	7.6	Y	Y	1.6
23	urban-CP	p.m.	80, 206	2381	5.4	2.8	2.6	0	9853	4.1	15.6	Y	Y	1.7
24	urban-CP	p.m.	217, 486	6493	5.7	2.1	2.2	0	28,330	4.2	12.1	Y	Y	1.7
25	urban-RD	p.m.	258, 1696	7713	16.9	5.0	6.6	0	23,055	2.6	8.0	Y	Y	1.6
26	urban-RD	p.m.	219, 585	6478	9.7	1.2	2.7	0	19,727	2.9	9.5	Y	Y	2.1
27	urban-RD	p.m.	239, 1661	7154	16.0	6.7	6.9	0	22,279	2.4	3.7	Y	Y	2.0
28	urban-RD	p.m.	321, 1737	9565	14.5	5.5	5.4	0	14,305	1.4	4.1	Y	Y	100
29	urban-RD	sunset	302, 4437	9045	23.2	14.8	14.7	5.5	19,372	2.0	2.5	Y	Y	1.6
30	urban-RD	sunset	405, 5936	12,027	21.6	14.7	14.7	5.9	34,555	2.7	4.2	Y	Y	2.8
31	urban-BG	sunset	219, 3942	6480	20.2	15.6	18.0	12.7	23,206	3.5	4.3	Y	Y	1.6
32	urban-RD	sunset	72, 108	1864	2.6	1.5	1.5	0.5	11,719	6.3	24.8	Y	Y	1.7
33	urban-RD	sunset	265, 2187	7945	17.3	11.0	8.3	0	27,207	3.0	6.4	Y	Y	2.4
34	campus	sunset	302, 1944	9043	10.3	6.7	6.4	0.9	23,027	2.4	6.0	Y	Y	2.4

[1] In this column, "RD", "BG", "TN", and "CP" are abbreviations of "road", "bridge", "tunnel", and "car park". [2] In this column, we list the average amount of dynamic objects and their area ratio of each frame. [3] In this and the latter column, "Y", "N", and "PD" denote that data are intact, missing, and partially damaged, respectively.

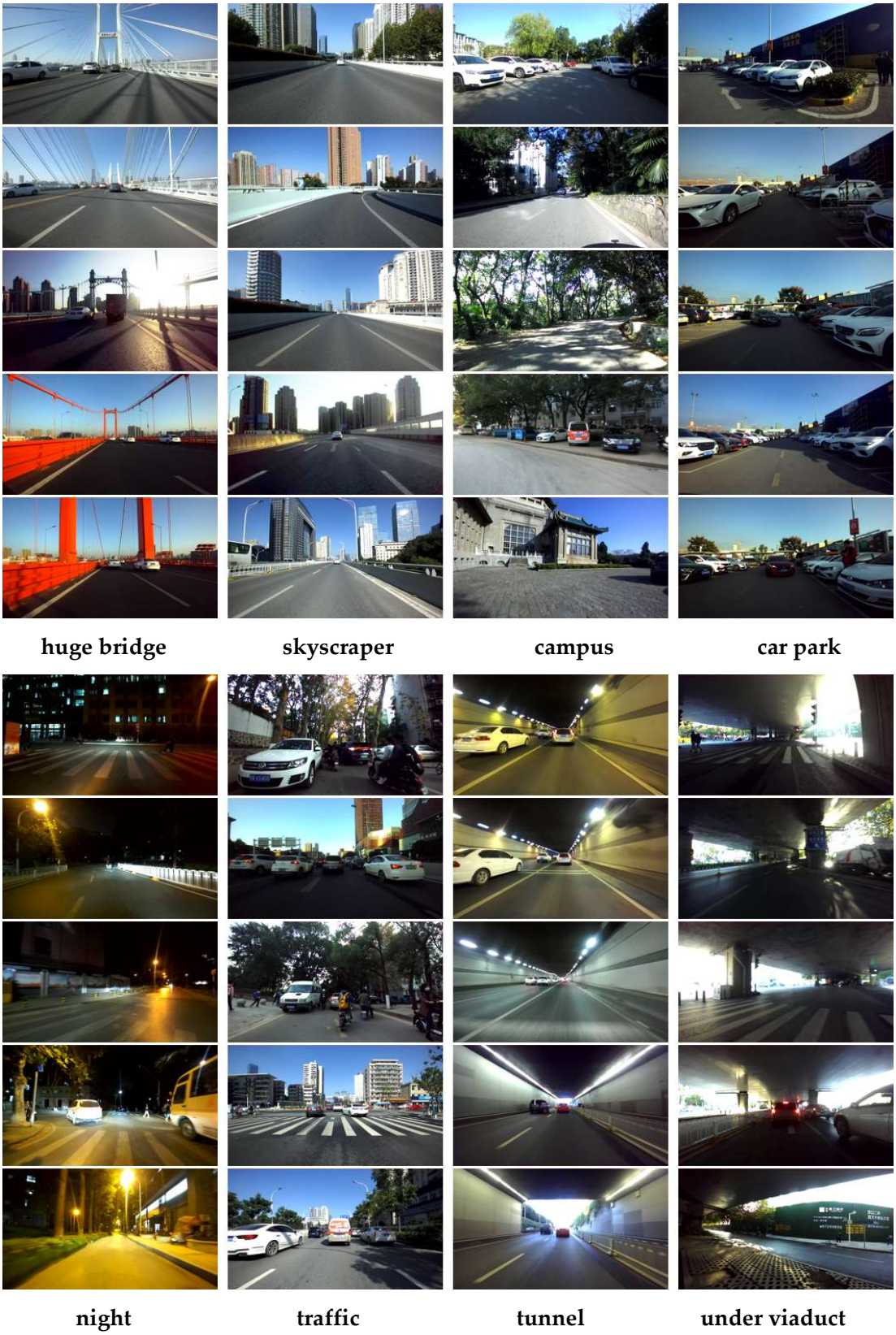

Figure 4. Example frames of WHUVID. This figure contains sample images recorded from the left lens of the Stereolabs ZED2 camera, indicating the diversity and uniqueness of our dataset. Scenes such as huge bridge, skyscraper, tunnel, and under viaduct are rarely seen in previously established datasets.

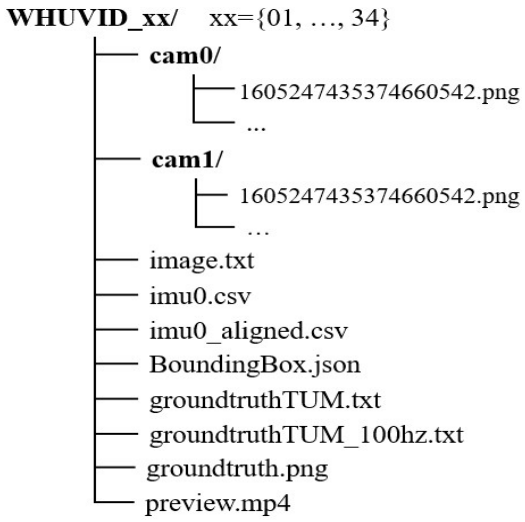

Figure 5. File structure of each provided sequence. Folders cam0 and cam1 store images from the left and right lenses of ZED2, respectively, and the images are named after the Unix timestamp in nanoseconds when they were caught by the host computer. Names of images were listed in the image.txt file for retrieval. Raw and aligned IMU readings were separately saved in the files imu0.csv and imu0_aligned.csv, respectively. BoundingBox.json stores 2D bounding boxes of four types of annotated objects of each image in the current sequence. TUM-style ground truths before and after interpolation were stored in the files groundtruthTUM.txt (10 Hz) and groundtruthTUM_100hz.txt, respectively. The trajectory file groundtruth.png and video preview.mp4 of each sequence are provided for the convenience of checking and usage. More detailed information on each file will be provided later in Sections 4 and 5.

3.1. Data Description

3.1.1. Image and IMU

The output of the image sensor has four modes: Standard Definition (672 × 376 @ 15/30/60/100 fps), High Definition (1280 × 720 @ 15/30/60 fps), Full High Definition (1920 × 1080 @ 15/30 fps), and 2 K (2208 × 1242 @ 15 fps). Considering computing resource limitation and the balance between resolution and update rate, we chose High Definition with 30 fps. We acquired hardware time-synchronized binocular image pairs and 400 Hz 6-DOF IMU readings simultaneously and in real time using a multithread program based on an officially released toolkit *zed-open-capture* [37] by Stereolabs. To ensure real-time performance, images were first written into text files in binary format and then stored with lossless compression using 24-bit PNG files after post-processing. IMU readings were stored in comma-separated value files with seven values: {timestamp, triaxle acceleration (x-y-z), and triaxle angular velocity (x-y-z)}.

Although images and IMU readings were collected by a single program on the same host computer simultaneously, because ZED2 itself does not have a hardware trigger for visual-inertial synchronization, the timestamps of each pair of images and corresponding IMU reading were not strictly aligned, with a time slot up to half of the IMU update period, i.e., 1.25 milliseconds. Unaligned timestamps may cause crashes in some VIO algorithms; to solve this problem, a solution using linear interpolation is proposed in Section 4.2.2.

3.1.2. Ground Truth

In total, 6-DOF ground truths were collected and calculated by the u-blox GNSS-IMU navigation device with RTK correction signals from Qianxun, a high-precision positioning and timing service provider. We first obtained raw positioning data with 13 different fields from the device at 10 Hz: timestamp in seconds; longitude, latitude, and altitude in meters; orientation (i.e., yaw), velocity, and number of satellites used; PDOP (a positioning accuracy evaluation indicator ranging between 0.5 and 100; the smaller the value, the higher the

accuracy); HDOP (horizontal component of PDOP); VDOP (vertical component of PDOP); DGPS (differential GPS, a Boolean variable to indicate RTK correction); and fix type (an enumeration variable to indicate the positioning mode; 3D means only GNSS works, DR, i.e., Dead Reckoning, means only built-in IMU works, and 3D+DR means IMU-assisted GNSS works). The last six indicators mentioned above have different meanings, though they all describe the concept of positioning accuracy from different aspects and are strongly correlated. For example, GNSS and RTK signals were strong during most of our journey with the number of satellites used being more than 12 and PDOP smaller than 3, and HDOP and VDOP were also less than 3. Meanwhile, DGPS would be "YES" and fix type would be "3D+DR". However, when a vehicle is obscured by obstacles such as a dense treetop or just driving underground, the number of satellites used would be less than 6, and PDOP along with HDOP and VDOP would be greater than 7, with a value up to 100. In summary, PDOP was selected for a quantitative description of positioning accuracy after consideration, as shown in Figure 1; the GNSS signal quality of the total trajectory is analyzed in Figure 6.

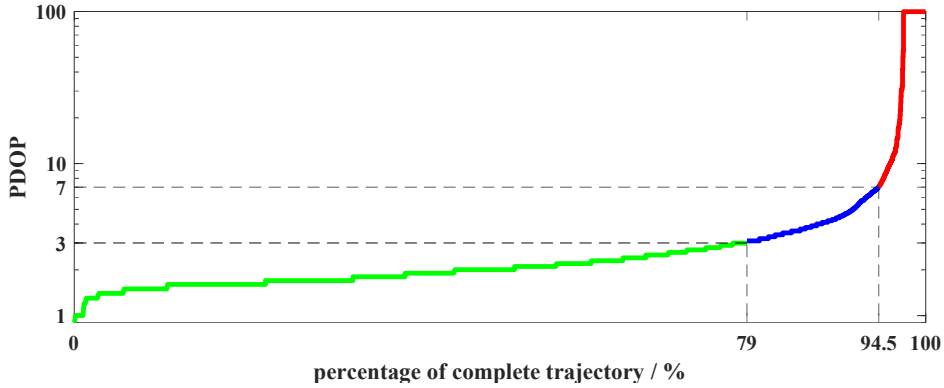

Figure 6. GNSS signal quality analysis of the complete trajectory of WHUVID. The PDOPs of all recorded points in the entire trip are sorted in this figure. As can be seen, 79.0%, 15.5%, and 5.5% of them receive EXCELLENT, GOOD, and BAD GNSS signals. Colors used here (green, blue, and red) have the same meanings as explained in Figure 1.

Among the Euler angles, i.e., roll, pitch, and yaw, only yaw could be acquired directly; therefore, the pitch had to be estimated from the change in altitude and the roll was simply set to zero since the vehicle traveled on straight city roads all the way. During this process, filtering the original positioning data was necessary for suppressing sudden changes in altitude and avoiding incorrect pitch calculation results. Subsequently, quaternion could be calculated. The metric ground truth under the ENU coordinate system was transformed from latitude and longitude using the Mercator projection [38], with the earth radius being 6371 km. Finally, for the convenience of evaluation, we extended the original ground truth from 10 to 100 Hz through linear interpolation and thereby obtained a text file of ground truth in the TUM [11] format: {timestamp, ENU_x, ENU_y, ENU_z, q_x, q_y, q_z, q_w}.

3.1.3. Sequences

Several key statistics including scene, time period, and duration are provided for most of the sequences in Table 2, except for individual ones that lack partial attributes because of severe occlusion, program bug, or other reasons. Sequences 04 and 05 have no IMU readings and partial positioning data because of program error at a certain moment, and the same problem also appears in sequence 08. Median PDOPs of sequences 19 and 28 are 100 because the vehicles were driving in tunnels and under viaducts, respectively, and neither could receive GNSS signals. In this condition, the u-blox navigation device still works by relying on its built-in IMU with the fix mode being "DR". For a more comprehensive assessment of the vehicle driving condition, we provide four types of speed data for each sequence: maximum, minimum, mean, and median. To make the ground truth as reliable as possible, our vehicle ran along digit "8" for 15 min each time before the u-blox device

was restarted to guarantee full self-calibration of GNSS and RTK signals. For each sequence, a folder is provided with its file structure, as shown in Figure 5.

To make working with this dataset more convenient, the trajectories of 32 sequences with intact positioning data are illustrated in Figure 7, and a brief description of their contents is given below:

Figure 7. *Cont.*

Figure 7. Summary of 32 sequences. Each trajectory is shown together with aerial urban imagery (the base map was provided by Google Earth for free) for reference except sequences 04 and 05, owing to partially damaged GNSS signals. Trace colors used here (green, blue, and red) have the same meanings as explained in Figure 1.

(01) Through avenue, some traffic. (02) With several right-angle turns, some traffic. (03) Uphill through trail, little traffic. (04) Through avenue, no GNSS signals on more than 3/4 of the journey. (05) Dense traffic with several loop closures, some GNSS signals lost. (06) A huge loop closure and a smaller one, some traffic. (07) A medium loop closure

with poor illumination. (08) Two loop closures of different sizes. (09) Wide variance in traffic and lighting conditions. (10) Polyline turn and straight path. (11) Mountain road, little traffic. (12) A large loop closure of uphill and downhill. (13) Loop closure with fine illumination. (14) Loop closures connected by a lane, some traffic. (15) Stay still for a long time. (16) At downtown, dense traffic and pedestrians. (17) Cross the Yangtze River, busy traffic. (18) Three-quarters of a turn at an overpass. (19) Straight path in a long tunnel, some traffic. (20) On the viaduct, surrounded by skyscrapers. (21) Long straight path on the viaduct. (22) Branch of a busy expressway. (23) Direct sun conditions in a parking lot. (24) Direct sun conditions in a parking lot with loop closure. (25) Along the light rail transit, busy traffic. (26) Turn left after a prolonged standstill. (27) Half along the light rail transit, half under viaduct. (28) Loop closure under the viaduct. (29) Cross the Hanjiang River, direct sun conditions. (30) A long trajectory with diverse city view. (31) Cross the Yangtze River from west to east. (32) Traffic congestion with very slow speed. (33) Cross a short tunnel, on and down from viaduct. (34) Direct sun conditions on campus, little traffic.

3.2. Annotations

For dynamic and static objects of interest within the camera's FOV, we provided annotations in the form of 2D bounding boxes. We defined the classes "car", "person", "bike" (a bicycle or electric-aided bike being ridden), and "label" (traffic signs). We acquired updated data at 30 fps with similar content between adjacent frames. Meanwhile, consecutive images collected from a moving vehicle are semantically continuous to a certain extent. Therefore, manually annotating all images is not necessary after the rapid development of artificial intelligence and object-detection algorithms. After weighing the accuracy requirements and time cost of this work, we decided to manually annotate one frame every 3–5 s and use them to train an object-detection neural network and apply it to the remaining images. YOLOv4 [30] was chosen as the detection algorithm. We divided 1860 manually annotated images into training and validation sets in the ratio of 4:1; the corresponding results are given in Table 3. The number of tags annotated manually and by neural networks and their distributions in 34 sequences are presented in Table 4 and Figure 8, respectively.

Table 3. Detection results of YOLOv4 on the validation set of manually annotated images.

Class	Car	Person	Bike	Label	In Total
AP or mAP/%	97.47	91.35	94.77	96.41	95.00

Table 4. Number of manually annotated tags and all tags (tags automatically annotated by YOLOv4 are included) of four selected categories and their proportion.

Category	Manually		All	
	Number	Percent/%	Number	Percent/%
car	5396	72.09	528,048	77.46
person	761	10.17	67,885	9.96
bike	474	6.33	39,505	5.79
label	854	11.41	46,271	6.79
In total	7485	100	681,709	100

A single type of object would have obviously different features under different lighting conditions, such as poor or uneven illumination, thus leading to misrecognition or other difficulties in object detection. Therefore, we did not annotate targets on images obtained at night to avoid confusion and thus maintained the recognition accuracy of YOLOv4. To acquire a brief knowledge of the traffic situations of each sequence, we simply defined the classes "car", "person", and "bike" as dynamic objects and calculated the frame-average

number and area ratio (calculated in pixels) of dynamic objects (Table 2). Notably, although the approach described above is not rigorous enough because of a few static objects of these three classes such as stationary cars in a parking lot, it is trustworthy as an auxiliary means in most traffic scenarios.

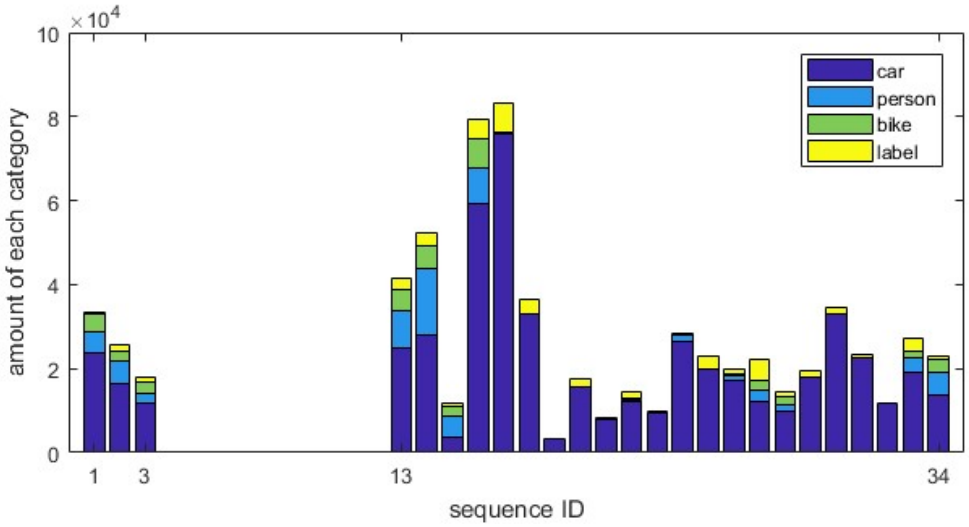

Figure 8. Number of tags of four selected categories (car, person, bike, and label) and their distributions in 34 sequences. Note that not a single tag is annotated in sequences 04–12 because of the following reasons: (1) lost IMU data and damaged GNSS signals of sequence 04, 05, and 08 and (2) adverse effects of different lighting conditions on the accuracy of object detection (driving at night of sequence 06–12).

4. Calibration and Synchronization

All the components of our system, i.e., the stereo camera, built-in IMU, and GNSS-IMU navigation device, required intrinsic and extrinsic calibration. In addition, the timestamp of the sensor messages needed to be synchronized and augmented for the convenience of evaluation. All collected data, calibration results, and intermediate files during this progress are publicly available for download.

4.1. Stereolabs ZED2 Calibration

The Stereolabs ZED2 camera contains two optical lenses and a built-in IMU, which need to be calibrated in sequence. We used the officially released open-source toolbox *zed-ros-wrapper* to obtain low-level access to the device. Some other frequently used systems and tools were also employed for further calibration. A part of the key results is summarized in Table 5 for quick reference.

Table 5. Calibration results of Stereolabs ZED2 for IMU, stereo lenses, and visual-IMU.

IMU	
accelerometer_noise_density	2.499898×10^{-2}
accelerometer_random_walk	3.833771×10^{-4}
gyroscope_noise_density	2.143949×10^{-3}
gyroscope_random_walk	1.716940×10^{-5}

Table 5. *Cont.*

Stereo		
cam0	intrinsic [1]	526.83, 529.30, 638.38, 362.98
	distortion [2]	−0.0615, 0.0148, −0.0000423, −0.00470
cam1	intrinsic	524.90, 529.95, 656.21, 343.63
	distortion	−0.0385, −0.00112, 0.000166, −0.00572
baseline	rotation [3]	−0.00355, 0.00117, 0.000204, 1.0
	translation	−0.120, −0.000202, 0.00207
Visual-IMU		
T_ic (cam0 to imu)	rotation matrix	[0.00815, 1.0, 0.00726; 1.0, −0.00797, −0.0248; −0.0247, 0.00746, −1.0]
	translation	0.00184, −0.0226, −0.0225

[1] The following data are arranged in the order of fx, fy, cx, and cy. [2] The following data are arranged in the order of k1, k2, p1, and p2. [3] Here, "rotation" is represented by quaternions with the order of qx, qy, qz, and qw.

4.1.1. IMU

The calibration of IMU mainly refers to estimating the gyroscope and accelerometer noise model parameters related to a standard inertial sensor noise model. These parameters are usually written as σ_a, σ_ba, σ_g, and σ_bg, denoting accelerometer noise density, accelerometer random walk, gyroscope noise density, and gyroscope random walk, respectively. However, these parameters reflect only the stochastic errors in the inertial data; thus, they should be obtained from an IMU at rest. In this work, the ZED2 camera, together with its built-in IMU, was placed on a horizontal desktop for 2 h, and a data package of around 1.7 GB was recorded by the robot operating system. Subsequently, the four aforementioned parameters could be calculated conveniently using the open-source toolbox *imu_utils* [39].

4.1.2. Stereo

The goal of stereo camera calibration is to obtain intrinsic and extrinsic parameters as well as distortion coefficients. The intrinsic parameters include focal length (fx, fy) and optical center (cx, cy), and the extrinsic parameters include rotation and translation from one lens to the other. The distortion coefficients include radial (k1, k2) and tangential (p1, p2) distortion. By repeatedly moving the camera on multiple axes at the front of a chessboard [40], these parameters can be calculated by the open-source toolbox *kalibr* [41–45].

4.1.3. Visual-IMU

The goal of visual-IMU calibration is to determine the spatial relationship between the camera and IMU. We defined the coordinate of the built-in IMU as the body coordinate of ZED2 and calculated the transformations from the left and right lenses to it. This task was also completed using *kalibr*.

4.2. GNSS Interpolation and IMU Timestamp Alignment

4.2.1. GNSS Interpolation

The output frequency of RTK-GNSS data itself is 1 Hz, and by fusing data of the built-in IMU, the C100-F9K integrated module outputs spatial positioning data for 10 Hz, which are the original data that can be obtained directly. In general, a sudden change in position and posture would not occur during normal driving; therefore, 10 Hz is enough to describe vehicles in motion. However, some currently popular evaluation tools have higher requirements on the data-update rate of ground truth. Therefore, we expanded the original data to 100 Hz through linear interpolation. At the same time, to avoid abrupt turning points caused by interpolation, we performed mean filtering on the interpolated data. The selected filtering radius was not large; consequently, the difference before and

after filtering was normally as low as centimeter level, which did not affect its application. Trajectories before and after modification were stored in the files groundtruthTUM.txt and groundtruthTUM_100hz.txt, respectively, as shown in Figure 5.

4.2.2. IMU Timestamp Alignment

As described in Section 3.1, the ZED2 camera itself does not have a trigger for visual-inertial synchronization at the hardware level; consequently, the timestamps of each pair of images and corresponding IMU reading are not strictly aligned. To compensate for this defect, we performed linear interpolation for original IMU readings and changed the update rate from 400 to 300 Hz, exactly 10 times the image frequency. The detailed operations and results are shown in Figure 9.

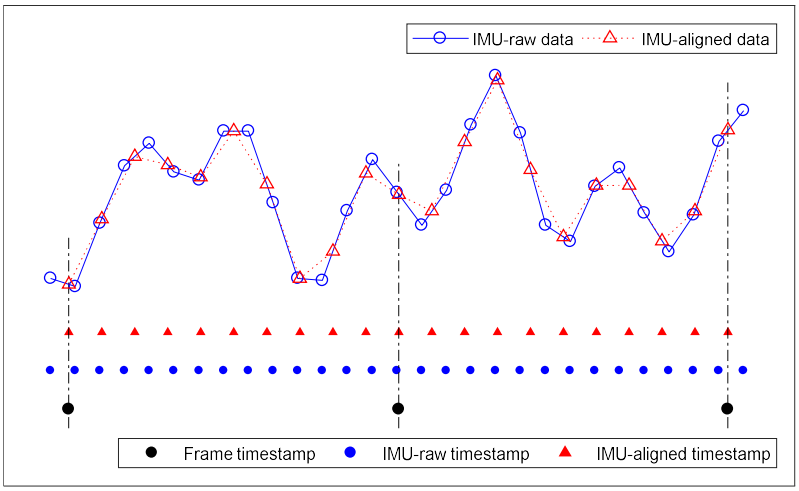

Figure 9. Visual-inertial timestamp alignment and data adjustment. The frequencies of recorded images and IMU readings are 30 fps and 400 Hz, respectively. We obtained strictly synchronized visual-inertial timestamps by linearly interpolating IMU data, thus making the frequency 300 Hz. In this figure, the solid black and blue dots represent the timestamps of images and raw IMU readings, respectively, and the solid red triangles represent aligned IMU timestamps. The hollow blue circles represent raw IMU data, while the hollow red triangles represent the adjusted IMU data after linear interpolation.

5. Evaluation and Discussion

In this work, we employed two milestone algorithms of VIO and SLAM, namely, ORB-SLAM2 and VINS-Mono, to evaluate the performance of our dataset in comparison with five other well-known datasets in three cases: monocular, stereo, and visual-inertial. Because partial data of some sequences of our dataset were missing, we selected 28 intact sequences for a complete comparison of the three aforementioned cases. The *evo* [46] toolbox was employed to implement the evaluation with two widely used metrics: absolute pose error (APE) and relative pose error (RPE). All results and corresponding files are available for download for further checking and analysis.

5.1. Evaluation Metric

APE and RPE give quantitative errors between a section of calculated trajectory and the corresponding ground truth. However, they cannot comprehensively describe the performance of relevant algorithms. We found in experiments that it is common for VIO and SLAM algorithms to be interrupted when tracking a long section of data due to problems such as lack of texture, camera shake, and sudden turnings. When an interruption occurs, algorithms will terminate immediately and restart from the next frame. In this case, one long section would thus be divided into some shorter subsegments, and the smaller the number of subsegments, the stronger the robustness of the algorithms. Therefore, to describe the positioning accuracy of a long section of data comprehensively, we took the

weighted average of APE and RPE of the subsegments as indicators of the long section to which they belonged. In summary, five novel metrics are proposed in this chapter: number of subsegments (NS), average tracked frames of subsegments (ATF), weighted average of the APE of subsegments (WAPE), weighted average of the RPE of subsegments (WRPE), and tracking rate of the entire section (TR). The higher the metrics ATF and TR, the better it is. Moreover, the lower the metrics NS, WAPE, and WRPE, the better it is.

5.2. Results and Discussion

To evaluate the feasibility and complexity of WHUVID accurately, we selected VIO and SLAM algorithms from classic works instead of the latest ones in light of the compatibility with previously published benchmarks [47–60]. Considering the innovation and their popularity in the community, we finally selected ORB-SLAM2 and VINS-Mono. ORB-SLAM2 is the second and latest version (at the time of conducting our experiment) of its series and the most classic and representative implementation of visual SLAM. The third version released later was updated, though it is only slightly modified with some added functions [61]. VINS-Mono is a milestone in the field of VIO and SLAM, and we employed its visual-inertial fusion function. Both algorithms were proposed in 2017 and have enjoyed great popularity among researchers and engineers to date. Despite our best efforts to compare WHUVID with formerly released datasets as comprehensively as possible, such work is obviously not sufficient compared with dozens of benchmarks and algorithms. Therefore, another purpose of choosing classic algorithms familiar to the community is to encourage more readers to participate in and expand our work, so as to further promote the development of this technology and its actualization together.

In the following paragraphs, we evaluate monocular and stereo with ORB-SLAM2 and visual-inertial with VINS-Mono. The experimental results are given in Table 6. All algorithms are comparable as none of them use loop-closure information.

Table 6. Evaluation of experimental results of WHUVID and five other previously published datasets with ORB-SLAM2 and VINS-Mono in three cases: monocular, stereo, and visual-inertial. Considering that some sequences (04, 05, 08, and 26–28) of WHUVID suffer from missing partial data or weak positioning signals, we selected the left 28 ones for a complete comparison.

Dataset		**Monocular**					**Stereo**					**Visual-Inertial**				
Name	**Seq Id**	**NS**	**ATF**	**WAPE**	**WRPE**	**TR/%**	**NS**	**ATF**	**WAPE**	**WRPE**	**TR/%**	**NS**	**ATF**	**WAPE**	**WRPE**	**TR/%**
FORD	1	**9**	276	4.78	4.30	81.3	-	-	-	-	-	-	-	-	-	-
	2	**6**	481	61.47	7.04	61.1	-	-	-	-	-	-	-	-	-	-
Oxford		**9**	3657	19.62	2.27	93.1	**15**	2312	39.77	0.50	98.1	-	-	-	-	-
MVSEC	day 1	**2**	2220	6.12	2.03	84.8	**3**	1684	6.05	0.60	96.5	**3**	1701	23.78	0.42	97.4
	day 2	**1**	12,069	6.33	2.98	92.4	**1**	13,063	25.01	1.51	100	**7**	1638	61.59	0.47	87.7
EuRoC	MH01	-	-	-	-	-	-	-	-	-	-	**1**	3682	0.13	0.0030	100
	MH02	-	-	-	-	-	-	-	-	-	-	**1**	3038	0.13	0.0024	99.9
	MH03	-	-	-	-	-	-	-	-	-	-	**1**	2698	0.17	0.0049	99.9
	MH04	-	-	-	-	-	-	-	-	-	-	**1**	2031	0.30	0.0046	99.9
	MH05	-	-	-	-	-	-	-	-	-	-	**1**	2271	0.30	0.0051	99.9
KITTI	00	**1**	4538	6.67	0.16	99.9	**1**	4541	0.86	0.02	100	-	-	-	-	-
	01	**1**	1062	458.9	10.53	96.5	**1**	1101	8.68	0.04	100	-	-	-	-	-
	02	**1**	4658	22.29	0.22	99.9	**1**	4661	5.42	0.02	100	-	-	-	-	-
	03	**1**	798	0.66	0.05	99.6	**1**	801	0.26	0.02	100	-	-	-	-	-
	04	**1**	268	1.26	0.09	98.9	**1**	271	0.55	0.02	100	-	-	-	-	-
	05	**1**	2650	8.21	0.23	96.0	**1**	2761	0.49	0.01	100	-	-	-	-	-
	06	**1**	1098	11.98	0.29	99.7	**1**	1101	0.47	0.01	100	-	-	-	-	-
	07	**1**	1094	2.89	0.12	99.4	**1**	1101	0.42	0.01	100	-	-	-	-	-
	08	**1**	4067	50.24	0.73	99.9	**1**	4071	2.89	0.03	100	-	-	-	-	-
	09	**1**	1586	42.51	0.78	99.7	**1**	1591	2.40	0.02	100	-	-	-	-	-
	10	**1**	1168	6.08	0.12	97.3	**1**	1201	0.82	0.01	100	-	-	-	-	-

Table 6. *Cont.*

Dataset		Monocular					Stereo					Visual-Inertial				
Name	**Seq Id**	**NS**	**ATF**	**WAPE**	**WRPE**	**TR/%**	**NS**	**ATF**	**WAPE**	**WRPE**	**TR/%**	**NS**	**ATF**	**WAPE**	**WRPE**	**TR/%**
WHUVID	01	**4**	2181	83.57	2.74	96.9	**10**	536	6.60	1.53	59.5	**1**	8834	20.89	0.18	98.1
	02	**2**	4111	115.11	3.00	95.3	**6**	811	4.12	0.56	56.4	**5**	1681	36.99	0.32	97.4
	03	**2**	4230	116.43	2.81	99.7	**6**	718	5.21	0.91	50.8	**5**	1652	46.89	0.27	97.3
	06	**9**	2675	100.59	3.29	97.8	**2**	12,303	82.01	2.73	100	**11**	2154	54.28	0.30	96.2
	07	**3**	1632	50.71	3.11	99.7	**1**	4905	37.28	2.24	99.8	**3**	1392	31.13	0.25	85.0
	09	**7**	2048	99.01	3.68	97.8	**3**	4849	187.46	3.11	99.3	**7**	1939	73.71	0.33	92.6
	10	**2**	1886	76.88	4.65	97.6	**16**	225	16.75	3.22	93.1	**3**	1010	37.91	0.45	78.4
	11	**1**	1107	28.89	2.52	64.1	**2**	843	20.57	2.79	97.6	**1**	1208	20.32	0.27	69.9
	12	**5**	2885	109.55	3.64	97.6	**2**	7278	195.97	2.89	98.5	**5**	2855	105.23	0.32	96.6
	13	**7**	2330	75.67	3.29	96.0	**9**	1273	12.33	1.95	67.4	**7**	2334	81.54	0.32	96.2
	14	**9**	2146	67.18	3.00	81.1	**13**	1309	32.44	2.00	71.5	**10**	2290	84.49	0.22	96.2
	15	**2**	2705	80.33	2.53	73.6	**4**	1494	10.49	1.86	81.3	**2**	2417	65.43	0.32	65.7
	16	**4**	4284	179.72	5.23	94.8	**7**	2122	26.51	2.64	82.2	**8**	1947	61.90	0.43	86.2
	17	**3**	5606	538.88	9.50	91.3	**5**	3478	583.01	6.67	94.4	**10**	1763	61.53	0.71	95.7
	18	**3**	2620	321.37	6.76	98.2	**2**	3772	75.25	5.11	94.3	**5**	1502	70.05	0.63	93.9
	19	**1**	1031	48.38	6.51	98.9	**1**	1033	11.11	4.76	99.1	**1**	990	17.50	0.47	95.0
	20	**3**	1286	196.98	9.06	96.4	**1**	3993	330.95	7.05	99.8	**2**	1693	75.11	0.86	84.6
	21	**2**	2429	357.05	12.38	99.1	**1**	3745	330.79	8.07	76.4	**3**	1235	87.53	1.14	75.6
	22	**2**	1970	73.91	4.56	95.9	**1**	3657	38.81	3.78	89.0	**3**	1122	40.90	0.52	82.0
	23	**1**	2369	36.30	1.70	99.5	**1**	2371	11.36	1.22	99.6	**2**	1138	11.41	0.12	95.5
	24	**2**	3237	33.94	1.77	99.7	**1**	6483	25.80	1.06	99.8	**3**	1877	13.40	0.13	86.7
	25	**3**	2560	73.52	4.93	99.6	**1**	7383	292.80	3.07	95.7	**3**	1641	67.57	0.52	63.8
	29	**6**	1406	182.90	13.20	93.3	**4**	2080	180.47	7.04	92.0	**6**	1431	105.03	0.92	94.9
	30	**5**	2393	242.01	7.76	99.5	**4**	1434	56.71	6.73	47.7	**9**	1160	92.40	0.89	86.8
	31	**4**	1572	602.26	13.80	97.0	**8**	741	181.64	7.82	91.5	**4**	1462	81.36	0.89	90.2
	32	**1**	1828	3.92	3.34	98.1	**1**	1855	11.18	0.78	99.5	**1**	1751	5.80	0.09	93.9
	33	**2**	3887	196.06	7.93	97.8	**1**	7935	174.40	3.93	99.9	**4**	1728	39.99	0.47	87.0
	34	**3**	2878	94.23	2.39	95.5	**1**	9033	132.31	-	99.9	**5**	1746	54.59	0.27	96.5

The results clearly reveal that FORD performs worst and KITTI and EuRoC perform best among all comparative datasets. On the one hand, FORD encounters too many interruptions, resulting in far smaller ATFs compared with others, mainly because its monocular images themselves have heavy distortion and narrow FOV (<80°). Such images are usually not suitable for visual SLAM systems. On the other hand, algorithms achieve minimum trajectory errors and maximum tracking rates in both monocular and stereo cases on KITTI and in the visual-inertial case on EuRoC. One notable fact is that no interruption occurred on the two abovementioned datasets during the entire trip. This is probably because both KITTI and EuRoC have good-quality data particularly suitable for testing, and algorithms along with related hyperparameters are carefully tuned and optimized in advance.

In contrast to KITTI and EuRoC, errors and number of interruptions on WHUVID are higher, indicating the increased level of difficulty of our real-world dataset. According to the phenomena observed in experiments, we conclude that most interruptions come from nonstationary driving situations including (but not limited to) turning at large angles, long-time low-speed movement, being blocked by surrounding vehicles, and frequent start and stop. Such driving conditions are rarely seen in comparative datasets, whose sequences are usually carefully split up into pieces of continuous driving segments with steady speeds. As analyzed before, although such splits make subsequent evaluations convenient, they do change the real-world driving condition. WHUVID does not perform meticulous splits, and thus, different types of real-world challenges continually pose additional difficulties to relevant algorithms from multiple aspects. Apart from that, wide roads, dense pedestrian and traffic flow, and other complex environments also present great challenges. Nevertheless, the fact that the ATFs of almost every sequence of WHUVID are universally as high as thousands of frames proves the reliability of our dataset. One interesting fact to be noticed is that the positioning accuracy of the stereo and visual-inertial

cases is not designed to be significantly better than that of the monocular case, which may be caused by problems such as insufficient baseline length, high random noise of the consumer-grade built-in IMU, and the lack of hardware-level visual-inertial synchronization.

6. Conclusions

In this paper, we proposed a large-scale dataset with a diverse set of sequences representing complex urban scenarios for the evaluation of VIO/SLAM, autonomous driving, and object detection. It contains well-calibrated high-resolution binocular RGB image pairs and high-frequency 3-axis accelerometer and gyroscope measurements, together with accurate 6-DOF ground truths. We believe that our work will complement other works by helping to reduce overfitting to existing datasets and will contribute to the practical progress of these technologies. Furthermore, we plan to offer a more comprehensive benchmark to facilitate an evaluation service for more computer vision tasks such as semantic segmentation, target tracking, scene recognition, and 3D object detection in the near future.

Author Contributions: Conceptualization, T.C. and F.P.; methodology, T.C.; software, T.C.; validation, Z.L.; formal analysis, T.C.; investigation, Z.L.; resources, H.C.; data curation, T.C.; writing—original draft preparation, T.C.; writing—review and editing, H.C. and F.P.; visualization, T.C.; supervision, F.P.; project administration, F.P.; funding acquisition, F.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Key Research and Development Program of China, grant number 2018YFB2100503.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/chentianyangWHU/WHUVID accessed on 15 March 2022.

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

Abbreviations

The following abbreviations are used in this manuscript:

WHUVID	Wuhan Urban Visual-inertial Dataset
VIO	visual-inertial odometry
SLAM	simultaneous localization and mapping
GNSS	Global Navigation Satellite System
IMU	inertial measurement unit
DOF	degree of freedom
VO	visual odometry
RGB-D	RGB-depth
UAV	unmanned aerial vehicle
AUV	autonomous underwater vehicles
ENU	east-north-up
PDOP	position dilution of precision
FOV	field of view
GPS	Global Positioning System
SBAS	Satellite-Based Augmentation System
QZSS	Quasi-Zenith Satellite System
CEP	circular error probable
RTK	real-time kinematic
SUV	sports utility vehicle
HDOP	horizontal component of position dilution of precision
VDOP	vertical component of position dilution of precision
DGPS	Differential Global Positioning System
DR	dead reckoning
APE	absolute pose error
RPE	relative pose error
NS	the number of subsegments

ATF	average tracked frames of subsegments
WAPE	weighted average of absolute pose error of subsequences
WRPE	weighted average of relative pose error of subsequences
TR	tracking rate of the whole section

References

1. Leung, K.Y.; Halpern, Y.; Barfoot, T.D.; Liu, H.H. The UTIAS multi-robot cooperative localization and mapping dataset. *Int. J. Rob. Res.* **2011**, *30*, 969–974. [CrossRef]
2. Chen, D.M.; Baatz, G.; Köser, K.; Tsai, S.S.; Vedantham, R.; Pylvänäinen, T.; Roimela, K.; Chen, X.; Bach, J.; Pollefeys, M.; et al. City-scale landmark identification on mobile devices. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 737–744.
3. Milford, M.J.; Wyeth, G.F. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Guangzhou, China, 11–14 December 2012; pp. 1643–1649.
4. Guzmán, R.; Hayet, J.B.; Klette, R. Towards ubiquitous autonomous driving: The CCSAD dataset. In *International Conference on Computer Analysis of Images and Patterns*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 582–593.
5. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
6. Carlevaris-Bianco, N.; Ushani, A.K.; Eustice, R.M. University of Michigan North Campus long-term vision and lidar dataset. *Int. J. Rob. Res.* **2016**, *35*, 1023–1035. [CrossRef]
7. Jung, H.; Oto, Y.; Mozos, O.M.; Iwashita, Y.; Kurazume, R. Multi-modal panoramic 3D outdoor datasets for place categorization. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 4545–4550.
8. Pandey, G.; McBride, J.R.; Eustice, R.M. Ford campus vision and lidar data set. *Int. J. Rob. Res.* **2011**, *30*, 1543–1552. [CrossRef]
9. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.* **2013**, *32*, 1231–1237. [CrossRef]
10. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
11. Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. In Proceedings of the 2012 IEEE/RSJ international conference on intelligent robots and systems, Faro, Portugal, 7–12 October 2012; pp. 573–580.
12. Zhu, A.Z.; Thakur, D.; Özaslan, T.; Pfrommer, B.; Kumar, V.; Daniilidis, K. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2032–2039. [CrossRef]
13. Chebrolu, N.; Lottes, P.; Schaefer, A.; Winterhalter, W.; Burgard, W.; Stachniss, C. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *Int. J. Rob. Res.* **2017**, *36*, 1045–1052. [CrossRef]
14. Hewitt, R.A.; Boukas, E.; Azkarate, M.; Pagnamenta, M.; Marshall, J.A.; Gasteratos, A.; Visentin, G. The Katwijk beach planetary rover dataset. *Int. J. Rob. Res.* **2018**, *37*, 3–12. [CrossRef]
15. Burri, M.; Nikolic, J.; Gohl, P.; Schneider, T.; Rehder, J.; Omari, S.; Achtelik, M.W.; Siegwart, R. The EuRoC micro aerial vehicle datasets. *Int. J. Rob. Res.* **2016**, *35*, 1157–1163. [CrossRef]
16. Majdik, A.L.; Till, C.; Scaramuzza, D. The Zurich urban micro aerial vehicle dataset. *Int. J. Rob. Res.* **2017**, *36*, 269–273. [CrossRef]
17. Ferrera, M.; Creuze, V.; Moras, J.; Trouvé-Peloux, P. AQUALOC: An underwater dataset for visual–inertial–pressure localization. *Int. J. Rob. Res.* **2019**, *38*, 1549–1559. [CrossRef]
18. Mallios, A.; Vidal, E.; Campos, R.; Carreras, M. Underwater caves sonar data set. *Int. J. Rob. Res.* **2017**, *36*, 1247–1251. [CrossRef]
19. Bender, A.; Williams, S.B.; Pizarro, O. Autonomous exploration of large-scale benthic environments. In Proceedings of the 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 6–10 May 2013; pp. 390–396.
20. Golodetz, S.; Cavallari, T.; Lord, N.A.; Prisacariu, V.A.; Murray, D.W.; Torr, P.H. Collaborative large-scale dense 3d reconstruction with online inter-agent pose optimisation. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 2895–2905. [CrossRef] [PubMed]
21. Schubert, D.; Goll, T.; Demmel, N.; Usenko, V.; Stückler, J.; Cremers, D. The TUM VI benchmark for evaluating visual-inertial odometry. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1680–1687.
22. Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J.S.; An, K.; Kweon, I.S. KAIST multi-spectral day/night data set for autonomous and assisted driving. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 934–948. [CrossRef]
23. Jeong, J.; Cho, Y.; Shin, Y.S.; Roh, H.; Kim, A. Complex urban dataset with multi-level sensors from highly diverse urban environments. *Int. J. Rob. Res.* **2019**, *38*, 642–657. [CrossRef]
24. Fraundorfer, F.; Scaramuzza, D. Visual odometry: Part i: The first 30 years and fundamentals. *IEEE Robot. Autom. Mag.* **2011**, *18*, 80–92.
25. Fraundorfer, F.; Scaramuzza, D. Visual odometry: Part ii: Matching, robustness, optimization, and applications. *IEEE Robot. Autom. Mag.* **2012**, *19*, 78–90. [CrossRef]

26. Aqel, M.O.; Marhaban, M.H.; Saripan, M.I.; Ismail, N.B. Review of visual odometry: Types, approaches, challenges, and applications. *Springerplus* **2016**, *5*, 1897. [CrossRef]
27. Durrant-Whyte, H.; Bailey, T. Simultaneous localization and mapping: Part I. *IEEE Robot. Autom. Mag.* **2006**, *13*, 99–110. [CrossRef]
28. Bailey, T.; Durrant-Whyte, H. Simultaneous localization and mapping (SLAM): Part II. *IEEE Robot. Autom. Mag.* **2006**, *13*, 108–117. [CrossRef]
29. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, D.; Neira, J.; Reid, I.; Leonard, J.J. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robot.* **2016**, *32*, 1309–1332. [CrossRef]
30. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
31. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [CrossRef]
32. Mur-Artal, R.; Tardós, J.D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
33. Qin, T.; Li, P.; Shen, S. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [CrossRef]
34. Qin, T.; Shen, S. Online temporal calibration for monocular visual-inertial systems. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 3662–3669.
35. Blanco-Claraco, J.L.; Moreno-Duenas, F.A.; González-Jiménez, J. The Málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario. *Int. J. Rob. Res.* **2014**, *33*, 207–214. [CrossRef]
36. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 year, 1000 km: The Oxford RobotCar dataset. *Int. J. Rob. Res.* **2017**, *36*, 3–15. [CrossRef]
37. Walter, L.; Oliver, B.; Oliver, B.; Karol, P.; Pierre, Y.; Max, P. A Platform-Agnostic Camera and Sensor Capture API for the ZED Stereo Camera Family. 2020. Available online: https://github.com/stereolabs/zed-open-capture (accessed on 22 May 2018).
38. Grafarend, E. The optimal universal transverse Mercator projection. In *Geodetic Theory Today*; Springer: Berlin/Heidelberg, Germany, 1995; p. 51.
39. Woodman, O.J. An introduction to inertial navigation. In *Technical Report UCAM-CL-TR-696*; University of Cambridge, Computer Laboratory: Cambridge, UK, 2007. [CrossRef]
40. Zhang, Z. Flexible camera calibration by viewing a plane from unknown orientations. In Proceedings of the seventh IEEE international conference on computer vision, Washington, DC, USA, 20–25 September 1999; Volume 1, pp. 666–673.
41. Rehder, J.; Nikolic, J.; Schneider, T.; Hinzmann, T.; Siegwart, R. Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 4304–4311.
42. Furgale, P.; Rehder, J.; Siegwart, R. Unified temporal and spatial calibration for multi-sensor systems. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; pp. 1280–1286.
43. Furgale, P.; Barfoot, T.D.; Sibley, G. Continuous-time batch estimation using temporal basis functions. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Guangzhou, China, 11–14 December 2012; pp. 2088–2095.
44. Maye, J.; Furgale, P.; Siegwart, R. Self-supervised calibration for robotic systems. In Proceedings of the 2013 IEEE Intelligent Vehicles Symposium (IV), Gold Coast City, Australia, 23–26 June 2013; pp. 473–480.
45. Oth, L.; Furgale, P.; Kneip, L.; Siegwart, R. Rolling shutter camera calibration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1360–1367.
46. Grupp, M. Evo: Python Package for the Evaluation of Odometry and SLAM. 2017. Available online: https://github.com/MichaelGrupp/evo (accessed on 14 September 2017).
47. Klein, G.; Murray, D. Parallel tracking and mapping for small AR workspaces. In Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, Japan, 13–16 November 2007; pp. 225–234.
48. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef]
49. Newcombe, R.A.; Lovegrove, S.J.; Davison, A.J. DTAM: Dense tracking and mapping in real-time. In Proceedings of the 2011 IEEE international conference on computer vision, Barcelona, Spain, 6–13 November 2011; pp. 2320–2327.
50. Kerl, C.; Sturm, J.; Cremers, D. Dense visual SLAM for RGB-D cameras. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; pp. 2100–2106.
51. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 834–849.
52. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. In Proceedings of the 2014 IEEE international conference on robotics and automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 15–22.
53. Leutenegger, S.; Lynen, S.; Bosse, M.; Siegwart, R.; Furgale, P. Keyframe-based visual–inertial odometry using nonlinear optimization. *Int. J. Rob. Res.* **2015**, *34*, 314–334. [CrossRef]
54. Bloesch, M.; Omari, S.; Hutter, M.; Siegwart, R. Robust visual inertial odometry using a direct EKF-based approach. In Proceedings of the 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS), Hamburg, Germany, 28 September–3 October 2015; pp. 298–304.

55. Zihao Zhu, A.; Atanasov, N.; Daniilidis, K. Event-based visual inertial odometry. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5391–5399.
56. Yu, C.; Liu, Z.; Liu, X.J.; Xie, F.; Yang, Y.; Wei, Q.; Fei, Q. DS-SLAM: A semantic visual SLAM towards dynamic environments. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1168–1174.
57. Bescos, B.; Fácil, J.M.; Civera, J.; Neira, J. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robot. Autom. Lett.* **2018**, *3*, 4076–4083. [CrossRef]
58. Henein, M.; Zhang, J.; Mahony, R.; Ila, V. Dynamic SLAM: The need for speed. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 2123–2129.
59. Nair, G.B.; Daga, S.; Sajnani, R.; Ramesh, A.; Ansari, J.A.; Jatavallabhula, K.M.; Krishna, K.M. Multi-object monocular SLAM for dynamic environments. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 651–657.
60. Zhang, J.; Henein, M.; Mahony, R.; Ila, V. VDO-SLAM: A visual dynamic object-aware SLAM system. *arXiv* **2020**, arXiv:2005.11052.
61. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.; Tardós, J.D. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [CrossRef]