IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON POSITIONING AND NAVIGATION IN CHALLENGING ENVIRONMENTS

Received 31 May 2022, accepted 21 June 2022, date of publication 29 June 2022, date of current version 5 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3187209

APPLIED RESEARCH

Visual Odometry in Challenging Environments: An Urban Underground Railway Scenario Case

MIKEL ETXEBERRIA-GARCIA[1], MAIDER ZAMALLOA[1], NESTOR ARANA-AREXOLALEIBA[2,3], AND MIKEL LABAYEN[4,5]

[1]Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BRTA), 20500 Arrasate/Mondragón, Spain
[2]MGEP, Mondragon Unibertsitatea, Loramendi Kalea, Arrasate/Mondragón, 20500 Gipuzkoa, Spain
[3]Department of Materials and Production, Aalborg University, 9220 Aalborg, Denmark
[4]CAF Signaling, Donostia, 20018 Gipuzkoa, Spain
[5]Faculty of Informatics, UPV/EHU, Manuel Lardizabal Ibilbidea, Donostia, 20018 Gipuzkoa, Spain

Corresponding author: Mikel Etxeberria-Garcia (mikel.etxeberria@ikerlan.es)

This work was supported in part by the Basque Government through BIKAINTEK2018 Program and CAF Signaling in Collaboration Between Ikerlan, CAF Signaling, and Mondragon Unibertsitatea.

ABSTRACT Localization is one of the most critical tasks for an autonomous vehicle, as position information is required to understand its surroundings and move accordingly. Visual Odometry (VO) has shown promising results in the last years. However, VO algorithms are usually evaluated in outdoor street scenarios and do not consider underground railway scenarios, with low lighting conditions in tunnels and significant lighting changes between tunnels and railway platforms. Besides, there is a lack of GPS, and it is not easy to access such infrastructures. This research proposes a method to create a ground truth of images and poses in underground railway scenarios. Second, the EnlightenGAN algorithm is proposed to face challenging lighting conditions, which can be coupled with any state-of-the-art VO techniques. Finally, the obtained ground truth and the EnlightenGAN have been tested in a real scenario. Two different VO approaches have been used: ORB-SLAM2 and DF-VO. The results show that the EnlightenGAN enhancement improves the performance of both approaches.

INDEX TERMS Visual Odometry, autonomous vehicles, computer vision, data enhancement, simultaneous localization and mapping, image processing, railway domain.

I. INTRODUCTION

Visual Odometry (VO) is a particular case of odometry based on Computer Vision (CV), where the position and motion information are acquired through camera images [1]. VO algorithms aiming to derive localization data through visual sensors are usually evaluated and compared by reference standard datasets such as KITTI [2], [3] and EuRoC-MAV [4]. This situation leads solutions adapted to the visual characteristics contained on those scenarios with adequate lighting conditions (good illumination and similar lighting conditions in subsequent frames), relatively sufficient textures and Lambertian surfaces. However, few algorithms, datasets, and benchmarks can be found in challenging scenarios with varying light conditions, low illumination, low textures, or non-Lambertian surfaces.

For instance, one of the latest benchmark challenges in visually challenging odometry is the Subterranean Challenge (SubT), organized by the Defense Advanced Research Projects Agency (DARPA). Perceptually challenging scenarios and tasks were stated in this challenge, such as navigation through tunnel systems, cave networks, or urban underground environments. The participating teams presented several approaches [5]–[8] to study the robotics autonomy in underground scenarios exploration and navigation. These works emphasize the complexity of localization and navigation in underground environments due to their perceptually-degraded conditions. They also emphasize on the importance of field testing.

The associate editor coordinating the review of this manuscript and approving it for publication was Kegen Yu.

The railway domain is also moving towards the *Intelligent Transportation Systems* (ITS) and the *Advanced Driving Assistance Systems* (ADAS) industry. A train that implements autonomous operations requires accurate localization estimation to carry out operations as precise stop operation or coupling successfully. Algorithms applied in urban underground railway scenarios must deal with significant light changes from tunnel areas to platforms, with insufficient illumination and low textures in tunnels.

In this context, the application of state of the VO algorithms and data enhancement techniques was analyzed in a perceptually challenging driving car scenario [9]. The results showed that the Generative Adversarial Network (GAN)-based image enhancement methods can improve the performance achieved by state-of-the-art VO solutions.

In this paper, an analysis of state-of-art VO algorithms is performed and the use of a data enhancement method in underground railway VO solutions is evaluated. Algorithms applied in these scenarios must deal with significant light changes from tunnel areas to platforms, with insufficient illumination and low textures in tunnels. Therefore, an image enlightening technique is integrated to improve the results of state-of-the-art VO algorithms.

A dataset with challenging characteristics is really needed in order to evaluate VO performance in such scenarios. From an analysis of datasets used in CV for localization (datasets labeled with 6-DoF pose), no standard dataset of the railway domain was found; hence, an ad-hoc underground railway dataset generation was pursued.

The following section (II) includes a literature review of the main VO algorithms, a description of the applied enlightening data enhancement technique, and a list of reference VO datasets. Section III depicts the urban underground railway dataset generation process. Then, the results of state-of-art VO algorithms in the underground railway dataset and the influence of an enlightening technique are shown in sections IV and V, respectively. Finally, some conclusions are drawn in section VI.

II. LITERATURE REVIEW

A. VISUAL ODOMETRY

The term Visual Odometry was first introduced by Niester *et al.* [10] proposing a technique to estimate camera motion using RANSAC [11] outlier refinement method and tracking extracted features across the frames. Previously, feature matching was done just in consecutive frames. Later works have shown that VO methods might perform as well as wheel odometry while the cost of cameras is much lower compared to wheel sensors [1].

The VO research community started from the robotics domain to, later, focus on the localization in other sub-domains. In this context, different types of vehicles from distinct sub-domains and diverse characteristics have been studied, such as, cars [12], [13], trains [14], or lately UAVs [15].

Depending on the algorithm used to estimate odometry data, VO techniques can be classified as learning-based and geometry-based [16], [17]. *Geometry-based VO* is usually divided into appearance-based VO (also referred to as direct), feature-based VO, and a hybrid approach that mixes the two of them.

Direct VO techniques operate directly on intensity values. In feature-based VO methods features are extracted from the image and a tracking-matching process is done. Feature-based methods have good accuracy, are robust in dynamic scenes, and can deal with variances in viewpoint [18]; however, in contrast to direct methods, feature-based techniques are inadequate in low texture areas. However, the performance of direct VO algorithms degrades if the dataset is not photometrically calibrated and is sensitive to geometric distortions as those induced by the camera speed [19]. Furthermore, as mentioned in [20], direct methods require a constant irradiation appearance between matched pixels, which hinders its application in some scenarios.

Geometry-based VO approaches rely on image geometric characteristics and camera model to reconstruct the ego-motion between consecutive frames. One of the most standard geometric VO approach is ORB-SLAM2 [21]. It is based on the ORB [22] feature matching and a bundle adjustment algorithm. It is the reference geometric solution in the VO community [19], [23]–[28].

Geometry-based VO is reliable and accurate under favorable conditions, when there are enough illumination and textures to make the feature matching among consecutive frames. As stated in [29], monocular VO experiences a scale drift issue and global bundle adjustment algorithms needs to be applied. Furthermore, monocular VO algorithms have a depth-translation scale ambiguity issue [30].

Stereo geometry-based VO works have been also targeted lately. Semi-direct visual odometry (SVO) [31] is one of the most predominant approaches among direct monocular and stereo VO algorithms. It uses a probabilistic mapping method to estimate ego-motion and explicitly models outlier measurements. In 2017, Wang *et al.* presented Stereo Direct Sparse Odometry (Stereo DSO) [19], a method for VO estimation from stereo cameras based on the previously proposed monocular DSO algorithm [32]. Lately, Koestler *et al.* presented TANDEM [33], a SLAM system that estimates ego-motion based on a direct VO pipeline and deep multi-view stereo.

The expansion of Deep Learning-based Computer Vision techniques carried the emergence of *Deep Learning-based VO* solutions. Learning-based VO/vSLAM algorithms usually rely on learning parts of a standard VO/vSLAM pipeline or designing end-to-end trainable algorithms for ego-motion estimation.

One of the first and most relevant learning-based VO algorithms was PoseNet proposed by [34] Kendall *et al.*, a robust and real-time monocular re-localization system based on an end-to-end trained CNN. This approach was later improved by introducing loss functions based on geometry and scene

reprojection error [35]. Following this end-to-end pose estimation networks, DeepVO [36] was published, a solution that infers camera poses directly in an end-to-end manner from a sequence of RGB frames through a supervised Deep Recurrent Convolutional Neural Network (RCNN).

Some research works have tried to adapt traditional non-learning approaches into Deep Learning pipelines. Brachmann *et al.* introduced DSAC (Differentiable Sample Consensus) [37] algorithm based on previously proposed RANSAC [11]. They applied DSAC in a camera localization solution, learning an end-to-end camera localization pipeline.

However, most of the research works from the literature emphasize the importance of an accurate depth and flow estimation for VO/vSLAM. Depth information is crucial for the localization as it enables the inference of the scene geometry from 2D images. Moreover, it allows scale recovery [38] and the distinction of foreground and background points, allowing a better environment understanding. Together with depth estimation, the optical flow estimation is also a critical component of some VO/vSLAM algorithms as it models the motion between consecutive images. Therefore, most of learning-based VO/vSLAM algorithms have focused on learning depth and flow estimation for the pose inference process.

Following this research line, several works have focused the depth estimation [39], [40], [41]. In 2018, Zhan *et al.* presented Depth-VO-Feat [42], where stereo training was introduced to reduce the spatial and temporal photometric error. At the same time, DVSO was presented by Yang *et al.* [29], introducing deep depth predictions in Direct Sparse Odometry (DSO). D3VO [43] algorithm was also proposed in this direction, including the uncertainty estimation with camera pose and depth.

Zhan *et al.* proposed the unsupervised VO algorithm DF-VO [17]. This algorithm applies a deep learning-based depth and flow estimation, and, geometric image information to estimate the camera pose. As shown in [17], DF-VO outperforms most learning-based state-of-the-art algorithms in standard datasets.

Some works have proposed loss functions to handle challenging scenario characteristics. Yin *et al.* proposed GeoNet [44], to increase robustness towards outliers and non-Lambertian surfaces. After GeoNet, more works were proposed in this direction [45], [46].

However, as mentioned in [47], literature VO solutions have limitations in challenging scenarios that contain insufficient illumination and textures, or, variable lighting conditions. Literature VO solutions, as they are adapted to the characteristics of standard datasets, require sufficient illumination and enough textured surfaces for a correct feature matching. A good illumination allows motion extraction from images, as pixel displacement can not be accurately estimated otherwise. Therefore, the lighting issue needs to be handled in scenarios that contain low illumination or varying illumination conditions. These are the conditions that face the urban underground railway scenario.

DF-VO and *ORB-SLAM2* have been selected from the literature review as reference VO algorithms. As stated before, the DF-VO algorithm outperforms most learning-based state-of-the-art algorithms, while ORB-SLAM2 is the most referenced geometric algorithm. Moreover, these algorithms represent two distinct types of VO algorithms (learning-based and geometric). Both solutions can use mono-vision or stereo-vision camera frames as input. The stereo-vision input was chosen for the analysis, as stereo-vision solutions keep the real-world scale, i.e. the predictions are directly aligned to a real-world scale.

B. DATA ENHANCEMENT FOR VISUAL ODOMETRY IN CHALLENGING ENVIRONMENTS

In order to afford the scenario limitations of VO in challenging environments, the application of a data enhancement technique was considered. In this work, the data enhancement process is dedicated to the lighting limitations of the target domain. It aims to reduce the impact of the drastic lighting conditions found in the underground railway scenario.

In this paper, the work published in [9] is extended. In the previous work the application of *EnlightenGAN* [48] data enhancement approach in an outdoor driving car scenario with varying lighting conditions was evaluated. This previous research was focused on a driving car scenario where the lighting conditions of the underground railway domain where replicated driving by night. The results showed that the performance of DF-VO algorithm is improved when EnlightenGAN is applied in the recorded frames.

EnlightenGAN is based on machine learning models proposed by Ian Goodfellow *et al.* [49]. The algorithm uses an unsupervised Generative Adversarial Network (GAN) pretrained on the ImageNet dataset [50] and then trained on several datasets [51]–[54] to improve input image lighting.

EnlightenGAN was previously used for several tasks such as image reconstruction [55], photo exposure correction [56], image quality assessment [57] or illumination enhancement [58]. However, to our knowledge, the use of data enhancement methods to handle specific problems of VO methods in such challenging scenarios has not been researched yet.

In this paper, the application of EnlightenGAN in the underground railway domain when using geometric and hybrid VO solutions is evaluated. The study aims to explore if EnlightenGAN technique can afford the lighting limitations of reference VO approaches (DF-VO and ORB-SLAM2). The evaluation procedure and results are detailed in section V.

C. DATASETS FOR UNDERGROUND RAILWAY VISUAL ODOMETRY

In this work, a propietary dataset is generated as no standard or reference railway dataset fitted to the underground railway scenario was identified. Table 1 resumes the reference datasets used by starte-of-the art VO approaches.

Most state-of-the-art VO approaches are evaluated in the standard KITTI [2], [3] vision benchmark [17], [29], [36], [42], [43], [81]. This benchmark includes several datasets for

TABLE 1. Referenced datasets for Computer Vision-based VO approaches application and evaluation ordered by domain or motion type.

Dataset	Domain	Sensor configuration	Pose ground truth	Environment
Cambridge Landmarks [34]	Handheld sensor	Monocular	SfM	outdoors
7-scenes [59]	Handheld sensor	RGB-D	MoCap	indoors
BigSFM [60]	Handheld sensor	Monocular	GPS	outdoors
ICL-NUIM [61]	Handheld sensor	RGB-D	SLAM	indoors
ADVIO [62]	Handheld sensor	Stereo/IMU	IMU	in/outdoors
OIVIO [63]	Handheld sensor	Stereo/IMU	Total station	in/outdoors
Rawseeds [64]	Robot	Stereo/IMU	GPS	in/outdoors
SUN3D [65]	Robot	RGB-D	SfM	indoors
TUM-VI [66]	Robot	Stereo/IMU	MoCap	in/outdoors
TUM-RGB-D SLAM [67]	Robot	RGB-D	MoCap	indoors
TUM-Monocular VO [68]	Robot	Monocular	LSD-SLAM/MoCap	in/outdoors
NavVis [69]	Robot	Monocular	GPS	indoors
MIT Stata [70]	Robot	Stereo/RGB-D/Laser	Laser	indoors
The Wean Hall [71]	Robot	Stereo/IMU/Laser/Wheel odometry	GPS	in/outdoors
RGB-D SLAM [67]	Robot	RGB-D	MoCap	indoors
ETH3D [72]	Robot	Stereo/RGB-D/Laser/IMU	MoCap/SfM/LIDAR	in/outdoors
NCLT [73]	Segway	Stereo/IMU/Laser	GPS/IMU/Laser	in/outdoors
KITTI [2, 3]	Car	Stereo/IMU/Laser	GPS/IMU	outdoors
Málaga Urban [74]	Car	Stereo/IMU/Laser	GPS	outdoors
Oxford RobotCar [75]	Car	Stereo/Laser	GPS	outdoors
Ford Campus [76]	Car	Stereo/Laser/IMU	GPS	outdoors
KAIST Urban [77]	Car	Stereo/IMU	GPS/Laser	outdoors
Nordland [78]	**Railway**	**Monocular**	**GPS**	**outdoors**
Zurich Urban [79]	MAV	Monocular/IMU	GPS	outdoors
EuroC/MAV [4]	MAV	Stereo/IMU	MoCap/Laser	indoors
MVSEC [80]	Multi Vehicle	Stereo/IMU/Laser	GPS/MoCap/Laser	in/outdoors

* MoCap=Motion Capture System. SfM=Structure From Motion

tasks like VO, optical flow estimation, 3D object detection, or 3D tracking. The data is captured from a moving car in outdoor urban scenarios, and they provide datasets and evaluation metrics for each task. However, as the KITTI odometry dataset contains images from an outdoor environment with good lighting conditions, it is not adequate to evaluate the VO algorithms in the pursued scenario. Among the other analyzed datasets, it should be noted that only one database (Norland [78]) covers the railway domain; however it only covers outdoor scenarios, which is also out of the scope of this research work. Searching for a publicly available VO dataset from an indoor urban railway domain, no dataset was found. Following the idea that the evaluation of the VO approaches that have previously been evaluated in standard datasets is essential to adapt the algorithms to other industrial scenarios. Therefore, the generation of a proprietary database was considered.

The data for a proprietary dataset can be collected from different sources: from real scenarios or simulated environments. Real environment datasets are based on real-world scenarios, and therefore, the performance of algorithms can be effectively evaluated in the target scenario. However, the database generation in real-world scenarios increases recording and processing time, effort, and cost. In addition, it also depends on the access and permission to make the recordings in the target scenario.

Simulated environments can overcome these problems. The drawback of simulated environments is that it can not be assured that an algorithm trained and validated in a simulated environment will perform the same way in a real-world scenario. As stated in [82], all the challenging conditions inherent to underground environments can not be recreated in virtual scenarios.

Consequently, and as a real-world underground railway scenario was accessible, a proprietary dataset was generated from a real underground railway scenario. The definition, generation and validation processes of the proprietary *CAF* dataset is explained in the next section III.

III. URBAN UNDERGROUND RAILWAY DATASET GENERATION

The proprietary (*CAF*) was generated for the evaluation of VO algorithms in underground railway scenarios. The sensor set validation and camera calibration procedure was done by generating a complementary dataset (*CarDriving*) in an urban driving car domain. *CarDriving* dataset generation is described in [9].

The *CAF* dataset was recorded in an underground scenario in the railway *Line 3* of Euskotren-Bilbao. The line is composed by seven stations from Matiko to Kukullaga and it has a whole track length of 5.8km. It contains poor lighting conditions in tunnel areas and significant light changes in platform areas. Furthermore, the images captured in the tunnels contain repetitive and light dependent textures, and therefore, they are challenging for feature extraction algorithms. Figure 1 shows two frames of this scenario: (a) tunnel frame and (b) platform frame.

The camera was placed in the front of the train, inside the driving cabin according to the safety requirements of the railway domain. Figure 2 shows the camera placement in the active cabin.

FIGURE 1. The *CAF* dataset's tunnel and platform areas where the poor light conditions and textureless areas can be appreciated.

FIGURE 2. Camera setup for *CAF* dataset, placed in the cabin of a train moving through an underground urban railway scenario.

The recording camera is a ZED Stereo Camera. The image's resolution is 1280 × 720 pixels at 30 Hz with an electronic synchronized rolling shutter, automatic gain and a lens aperture of F2.0.

A. CAF DATASET

The dataset is composed by 19 sequences captured in the two directions of the rail Matiko-Kukullaga. A sequence is a record that begins at one station and ends in the stations the train stops. A 6-DoF pose is estimated for each captured frame. The dataset format follows the standard KITTI odometry dataset format and naming convention. The frames are rectified RGB color images stored with lossless compression using 8-bit PNG files.

The camera calibration parameters and the poses are stored in files specified by the KITTI format [3]. Each row of the pose file contains the first three rows of a 4×4 homogeneous pose matrix flattened into one line. The homogeneous pose matrix p_n can be represented as:

$$p_n = [r_n|\mathrm{tr}_n] = \begin{bmatrix} r_{11} & r_{12} & r_{13} & x_n \\ r_{21} & r_{22} & r_{23} & y_n \\ r_{31} & r_{32} & r_{33} & z_n \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

TABLE 2. *CAF* dataset resume with recorded sequences, the direction of the sequences, arriving station for each sequence, frame quantity, and sequence length.

Direction	Arrival station	Sequence	Frames	Length (m)
Matiko	Otxakoaga	01_50	3048	1420
	Txurdinaga	01_53	1977	699
	Zurbaranbarri	01_54	2663	1029
		03_49	6700	3148
	Zazpikaleak	02_22	3260	1011
	Uribarri	01_15	2724	903
		02_25	2639	
		03_54	5904	1913
	Matiko	01_17	2532	505
		02_27	2505	
Kukullaga	Uribarri	01_31	2140	524
	Zazpikaleak	01_33	2830	979
	Zurbaranbarri	01_35	2494	1007
		03_36	6560	2449
	Txurdinaga	01_37	2550	1032
	Otxarkoaga	01_39	2126	695
		03_36	4493	1729
	Kukullaga	01_40	4095	1405
		03_44	4144	
TOTAL			65384	23261

where r_n and tr_n are the rotation matrix and the translation matrix of the n-th frame, respectively. The translation component of the pose matrix follows the right-hand rule when defining axes in a 3D space (x-axis forward, y-axis right and z-axis up).

The dataset generated in this domain is represented in table 2 where the recorded sequences, recording direction, the arrival station for each sequence, the number of frames, and the track length of each sequence are depicted. The entire set of sequences yields 65.384 frames, with varying speed and length.

B. GROUND TRUTH GENERATION ALGORITHM DATA SOURCES

In general, the ground truth of VO datasets is generated using a GPS sensor [3], [74]–[77] (refer to Table 1). But, the GPS signal is unavailable in underground zones like the urban underground railway domain. Thus, a method that computes the 6-DoF pose of each frame from the train ERTMS/ETCS ATP data, geodetic map coordinates, and railway infrastructure gradient profile data was defined and implemented (see figure 3).

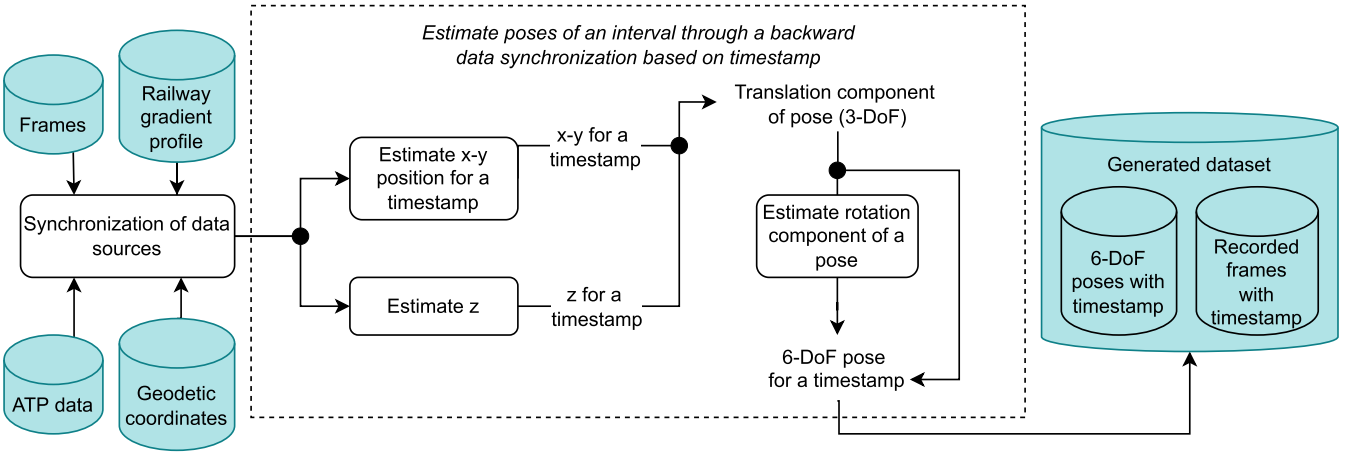


FIGURE 3. Diagram of the algorithm processes, with the data sources and the outputs.

The algorithm first estimates (x,y) positions based on geodetic coordinates, then z is added through the gradient profile. Afterwards the (x,y,z) translation data is estimated for each frame by using ERTMS ATP data, and, finally, the rotation data of each pose is calculated.

1) GEODETIC COORDINATES

The geodetic coordinates are represented by a pair (ϕ, λ) expressing *Latitude (Lat.)* and *Longitude (Lon.)* in decimal degrees. These coordinates use an ellipsoid to approximate the the earth's surface locations [83].

In this research, the geodetic coordinates define the coordinates followed by the trains in the target railway and have been extracted from a Geomap called ÖPNVKarte [84]. This Geomap contains public data that includes worldwide public transport facilities on a uniform map with information concerning several transport methods such as train, railway, ferry or bus. It is derived from OpenStreetMap [85], an initiative to create and provide accessible geographic data (i.e. street maps, etc.). It also contains railway-related information, such as platforms, stop positions, and routes.

The entire trajectory of an underground train in L3 extracted from ÖPNVKarte is shown in figure 4. As stated before, the trajectory of L3 is made up of seven stations in the route Kukullaga - Matiko, where some route positions, the station entrances, and train stop positions of each station are known in geodetic coordinates. However, the frequency of the camera is higher than the geodetic coordinates defined in the Geomap, and, therefore, a method based on ERTMS ATP data has been designed and implemented in order to generate the poses of the frames that were recorded between the geodetic coordinates.

The geodetic coordinates must be transformed from 3D plane to a 2D plane to assign an equal-area (x,y) position to each geodetic coordinate. Figure 5 shows a trajectory sample in geodetic coordinates and the generated equal-area (x,y) coordinates. In the ground truth generation algorithm, an equal-area (x,y) coordinate refers to tr_x and tr_y components of a 6-DoF pose.

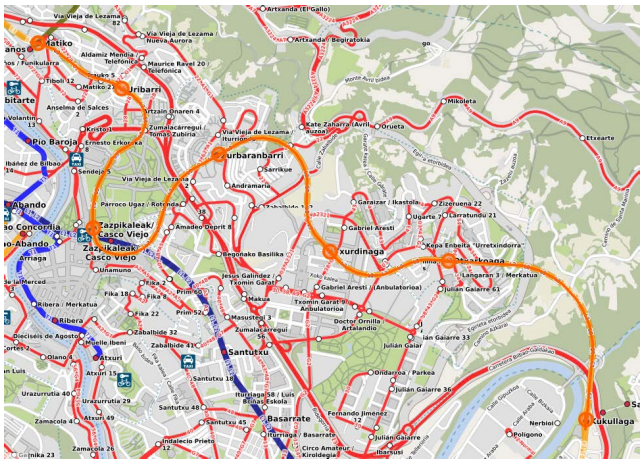


FIGURE 4. Line 3 railway extracted from ÖPNVKarte map [84]. Each circle represents one station from Line 3.

2) RAILWAY GRADIENT PROFILE

The railway gradient profile provided by the railway infrastructure managers, defines how the slope of the railway varies in predefined sections and allows the estimation of the height (z) for each 6-DoF pose. For that, a height profile can be constructed with this gradient profile. The initial height is initialized as 0, and then the height for each 1m section is calculated using the Equation 1.

$$h(d_n) = h(d_{n-1}) + (0.01 \cdot \text{grad}_n)\,, \tag{1}$$

where h refers to height, d_n refers to 1m railway sections and grad_n is the gradient value corresponding to that section from the gradient profile. Figure 6 shows the obtained railway gradient profile of the whole L3 railway.

3) ATP DATA: TRAIN's DYNAMICS AND SPEED DATA

The ERTMS/ETCS ATP train speed estimation process is based on redundant wheel encoder and radar sensor in order to get a safe and accurate estimation. By using these sensors, the ATP subsystem embedded in the train estimates the train position in the track, i.e. the distance traveled from an station or a beacon of the track. Track beacon position or

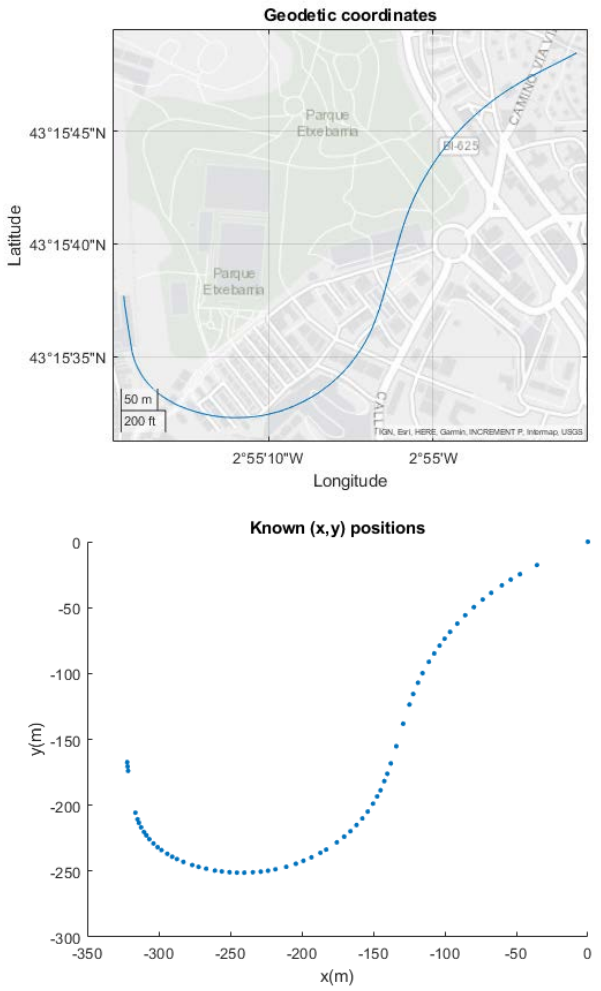


FIGURE 5. Transformation of a given sequence from L3 railway defined by geodetic coordinates into equal-area (x,y) positions.

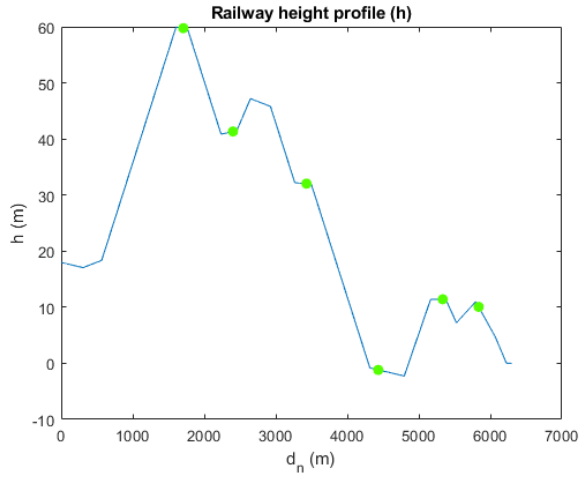


FIGURE 6. Results of height generation process. Height profile (*h*) is generated from gradient profile provided by railway constructor. The green circles represent the stations.

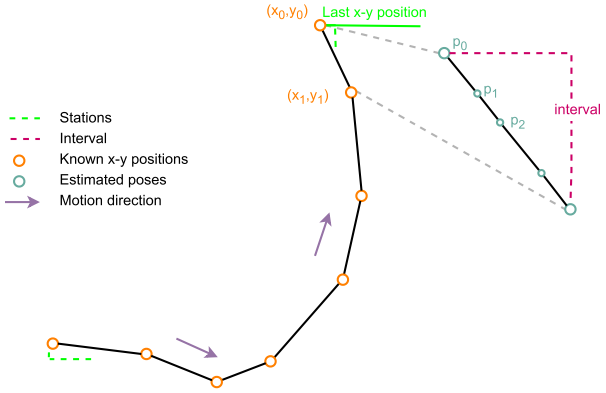


FIGURE 7. Railway with the known (x,y) positions, the intervals and estimated poses.

inter-beacon distance is predefined and known by railway infrastructure managers, even by the ATP subsystem, and therefore the ATP train position is re-adjusted when a beacon signal is received obtaining a precise estimation. The 6-DoF pose estimation of each frame is made by synchronizing the ATP system monitoring process with the image recording process as both are installed in the train. The objective of this process is to obtain a synchronized train position information for each frame. The data monitored from the ATP system is the following one:

- *timestamp* (s): time measured in the Coordinated Universal Time (UTC) standard read from the train's internal clock.
- *linear position estimation* (cm): distance traveled by the train from a previous station.
- *train speed* (m/s): train speed calculated by ATP.
- *train acceleration* (cm/s^2): train acceleration calculated by ATP.
- *train stopped*: boolean reflecting whether the train has reached stopping point or not.

All those variables are extracted from a ATP monitoring proprietary application that monitors ATP data with a frequency of 128000 Hz. The data acquisition frequency higher than the camera frequency (30Hz), and, consequently, they have been synchronized and a pose estimated for each frame.

C. ESTIMATE POSES OF AN INTERVAL THROUGH A BACKWARD DATA SYNCHRONIZATION BASED ON TIMESTAMP

The main idea of the synchronization algorithm is the estimation of poses in the trajectory sections between the known (x,y) positions obtained by transforming the known geodetic coordinates. These known (x,y) positions define the trajectory, but they are not enough for camera frequency and, therefore, more poses must be estimated between them. The *interval* has been defined to represent the idea of the trajectory sections, and it is a straight line between two consecutive known (x,y) positions. The estimated poses are located in the intervals. Figure 7 represents the intervals, known (x,y) positions and estimated poses in the railway. The main concepts of the ground truth generation algorithm are described in 1.

As the data sources are synchronized at the sequence ending, from now on, the ground truth generation is done in a

Algorithm 1 Ground Truth Data Generation Algorithm

Input: Given an *interval* (i) defined as a straight line between two known (x,y) positions

Phase 1 - Synchronize last (x,y) position, last image and ATP data of an interval

1: **if** $i = 0$ **then** ▷ First interval
2: Last image $\leftarrow$ $SSIM > threshold$ ▷ SSIM [86]
3: Last (x_i, y_i) position $\leftarrow$ given in the interval definition
4: ATP data $\leftarrow$ $train\_stopped = 1$
5: **else** ▷ Following intervals
6: Last image, last (x_i, y_i) position and ATP data $\leftarrow$ taken from $i - 1$
7: **end if**

Phase 2 - Estimate poses on an interval through a backward data synchronization based on timestamps

Input: V_n: train speed, a_n: train acceleration, t: timestamp, h: height profile, d_n: linear position estimation
8: Estimate translation component of poses (tr_n)
a: $(x_n, y_n) \leftarrow f(v_n, a_n, t)$ ▷ Eqn. 2
b: $z_n \leftarrow h(d_n)$ ▷ Eqn. 1
9: Estimate rotation component of poses (r_n)
a: $r_n \leftarrow g(tr_{n-1}, tr_n)$ ▷ Eqn. 3, 4, 5

backward data synchronization process of an *interval* based on the images timestamps. The last (x,y) position, last image and ATP data are taken for a given interval and the poses for all timestamps in that interval are estimated. Then, the poses of the following interval are estimated by taking the last (x,y) position and the last image of the previous interval as the initial position.

However, the train speed is variable and, therefore, the distribution of these poses can not be linear in different intervals. The total number of poses within the whole sequence should match the record frame amount.

1) SYNCHRONIZE LAST (x,y) POSITION, LAST IMAGE AND ATP DATA

The first step is to synchronize the different data sources using the last (x,y) position, last image and ATP data. The algorithm generates ground-truth poses for each recorded sequence using the position where the train has stopped as origin. For that, first the image where train stops (last image of the sequence) must be estimated. When there is motion, the similarity between consecutive frames is very low, however the similarity increases when the train has stopped. Due to the similarity of the frames corresponding to the train stopping point, the last frame is selected using the Structural Similarity Index (SSIM) [86]. SSIM is one of the most standard algorithms for image quality assessment [57], and therefore, for image similarity measure. It has shown that can outperform other common image similarity measurements as MSE [87] and has been previously referenced [88]. The SSIM measures the luminance, contrast, and structure of two given images and returns a similarity value between them.

Also, it only requires a starting optimization phase where the threshold is selected. Furthermore, the index was used to find just the first image within the threshold in each sequence, which gives a little number of results totally. Although SSIM is sensitive to image distortions, the environment being static, and the view fixed enables the SSIM application in underground railway scenarios.

The threshold was selected by exploratory testing. A predefined threshold was stated and iterated it until a SSIM threshold that best fitted to the lighting conditions of the scenario was identified. In this case a $SSIM > 0.965$ has been used as similarity threshold at the train stopping point.

The last (x,y) coordinates refer to the train stopping position; therefore, this coordinate pair and the last image are already synchronized. Finally, ATP monitored data is synchronized using the *train stopped* variable.

2) ESTIMATE POSES OF AN INTERVAL THROUGH A BACKWARD DATA SYNCHRONIZATION BASED ON TIMESTAMPS

A ground truth pose is generated for each recorded image in an interval using a backward synchronization process based on the timestamp. This process has two steps; first, the translation component is estimated, and then, the rotation is calculated from that translation.

a: ESTIMATION OF TRANSLATION COMPONENT

Translation component $T = \{tr_0, tr_1, \ldots, tr_m\}$ is defined as a set containing all the 3-DoF poses ($tr_n = [x_n, y_n, z_n]$) of an interval where n is the pose number ($0 \leq n \leq m$) and m is the total number of poses for that interval. For the translation component of a pose, first, the (x,y) position is estimated, and then the height (z) is added. The translation is estimated by taking an initial (x,y) position and calculating the motion to the next one using the ATP data *train speed* and *train acceleration*. The translation between two consecutive (x,y) positions in a straight line that forms the interval can be calculated using *Uniformly Accelerated Motion (UAM)* equations. This estimation is possible because it is considered that the poses follow a motion in a straight line and with a constant acceleration between them. Equation 2 shows the application of UAM equations in this case.

$$d_n = v_{n-1}t + \frac{1}{2}a_{n-1}t^2\,, \tag{2}$$

where t refers to the timestamp, v_n and a_n refer to ATP data train speed and acceleration respectively. The initial (x,y) translation component is set as [0, 0].

After calculating the (x,y) positions, the z or height is estimated using the height profile estimated from the gradient profile and ATP data. The railway height profile can be synchronized with the train stopping point, and therefore, with the first (x,y) position.

Then, previously calculated (x,y) positions can be used to extract the Euclidean distance traveled from position to position. Each pose's height (z) is calculated using traveled distances and the height profile. Therefore, after height estimation, the translation component of a pose has been estimated with respect to a timestamp.

b: ESTIMATION OF ROTATION COMPONENT

Rotation component $R = \{r_0, r_1, \ldots, r_m\}$ is defined as a set containing all the rotation matrices (r_n) within an interval where n is the pose number ($0 \leq n \leq m$) and m is the total number of poses for that interval calculated in the previous steps.

To calculate the rotation component r_n for each translation tr_n the transformation between two consecutive orientation vectors or_{n-1} and or_n is estimated. or_n defines the orientation of the train in tr_n and represents the vector between consecutive translations tr_{n-1} and tr_n. It is calculated as shown in 3:

$$\mathrm{or_n}(\mathrm{tr}_{n-1}, \mathrm{tr}_n) = (x_n - x_{n-1}, y_n - y_{n-1}, z_n - z_{n-1}), \quad (3)$$

where x, y and z represent the translation components of tr_{n-1} and tr_n. Then, using the axis-angle representation, the transformation between consecutive orientation vectors or_{n-1} and or_n can be calculated. For that, first the orientation vectors are normalized by dividing their value with the Euclidean norm (vector magnitude) $\|or_n\|$ of each vector (Eqn. 4) to align them at the same origin. The Euclidean norm can also be defined as the Euclidean distance of a vector from the origin to a point.

$$\mathrm{normalize}(\mathrm{or}_n) = \frac{\mathrm{or}_n}{\|\mathrm{or}_n\|}, \quad (4)$$

Then, the Euclidean norm of the cross product between the normalized consecutive orientations is estimated to get the axis. Finally, the rotation component is estimated using the inverse tangent function as shown in equation 5, where the angle between the orientations vectors is calculated trough the dot product:

$$r_n = \arccos\left(\frac{\|\mathrm{or}_n \times \mathrm{or}_{n-1}\|}{\mathrm{or}_n \cdot \mathrm{or}_{n-1}}\right), \quad (5)$$

where arccos refers to the inverse cosine function and or_{n-1} and or_n to two consecutive orientation vectors. This rotation estimation method accumulates an error relative to the previous estimations. However, as the train is tied to the rails, the trains' orientation is always fixed, and the orientation estimation is not critical.

The previously calculated translation component is added to the newly calculated rotation component to obtain the target 6-DoF ground truth pose. This is done by following the representation in equation 1.

Once all the poses from a given interval have been estimated, the next interval is taken and the process is repeated until all the intervals of a sequence have been covered.

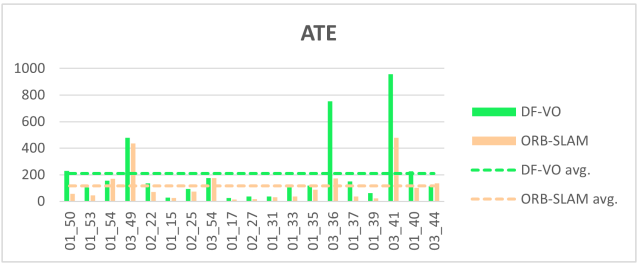


FIGURE 8. ATE of DF-VO and ORB-SLAM2 application on the generated *CAF* dataset.

IV. VO APPLICATION IN URBAN UNDERGROUND RAILWAY ENVIRONMENT

In this section the application of DF-VO and ORB-SLAM2 in the CAF dataset is evaluated.

In the following subsection, the standard VO evaluation metrics are explained. Then, the experimentation setup is described. Finally, the experimental results are discussed.

A. VO EVALUATION METRICS

The metrics used to evaluate the performance of the experiments are the following: Absolute Trajectory Error – *ATE* [67], Relative Pose Error – *RPE* [67], Average Translational Error – t_{err} and Average Rotational Error – r_{err}.

All the sequences were transformed with a 6-DoF Umeyama alignment [89], a standard alignment method used in most VO and SLAM evaluation benchmarks. [2]. A 6-DoF alignment is recommended to evaluate shape similarities of trajectories [90].

Given this transformation, ATE evaluates the global consistency of an estimated trajectory compared to the ground-truth trajectory. The RPE measures the drift error for each pose of the trajectory and the rotation and the translation components are calculated separately.

Finally, following KITTI evaluation benchmark criteria, the Average Translational Error (t_{err}) and the Average Rotational Error (r_{err}) are calculated on sub-sequences of different lengths. These errors measure the average relative pose error at a fixed distance. The sub-sequences length in meters is (100,200,...,800) because the error for smaller sub-sequences was large and hence biased the evaluation results.

B. EXPERIMENTATION SETUP

These experiments extend the evaluation done at [9], where ORB-SLAM2 and DF-VO were evaluated in an outdoor urban car driving scenario. In those experiments, the bad lighting conditions were replicated by car driving recordings in the night.

DF-VO implementation [91] flow-weights and depth estimation deep models were selected from the authors' trained models. The flow model is trained by the authors in the synthetic dataset Scene Flow [92].

To handle the non-deterministic nature of the ORB-SLAM2 algorithm, each sequence is run five times, and the

TABLE 3. DF-VO and ORB-SLAM2 application evaluation using standard VO evaluation metrics: Average Translational Error (t_{err}), Average Rotational Error (r_{err}), ATE and RPE. The sequences are organized by the direction they are recorded. The average errors for all 19 sequences are calculated, and the best result is in bold.

Algorithm	Record	14_11_2021 (->Matiko)									
	Seq	01_50	01_53	01_54	03_49	02_22	01_15	02_25	03_54	01_17	02_27
DF-VO	t_{err} (%)	80	52.97	85.74	86.27	77.94	64.09	94.52	111.55	56.61	55.69
	r_{err} (º/100m)	13.21	21.25	29.92	18.48	35.88	24.43	33.5	41.71	21.34	21.33
	ATE	230.66	106.38	157.46	478.76	135.36	29.11	94.27	175.64	26.1	36.39
	RPE (m)	0.402	0.232	0.354	0.423	0.269	0.236	0.314	0.34	0.132	0.133
	RPE (º)	0.156	0.135	0.156	0.176	0.124	0.13	0.157	0.143	0.057	0.062
ORB-SLAM2	t_{err} (%)	68.79	54.93	177.16	125.85	80.28	55.03	94.26	136.06	51.89	53.88
	r_{err} (º/100m)	5.95	19.19	50.42	17.2	25.6	20.17	31.97	39.34	15.21	14.24
	ATE	56.58	44.98	169.39	435.36	72.35	26.71	74.61	177.23	15.61	17.1
	RPE (m)	0.34	0.192	0.641	0.646	0.277	0.204	0.302	0.371	0.116	0.122
	RPE (º)	0.081	0.092	0.429	0.264	0.125	0.104	0.11	0.121	0.065	0.064

Algorithm	Record	14_11_2021 (->Kukullga)								Avg. Err.
	Seq	01_31	01_33	01_35	03_36	01_37	01_39	03_41	01_40	03_44	
DF-VO	t_{err} (%)	63.64	135.76	175.28	148.11	136.82	58.1	110	93.32	99.36	**93.9879**
	r_{err} (º/100m)	17.81	28.84	29.64	23.51	27.19	21.01	22.21	31.19	41.2	26.50789
	ATE	38.38	104.98	119.08	754.45	150.64	62.88	957.69	226.86	114.75	210.5179
	RPE (m)	0.183	0.467	0.583	0.541	0.594	0.192	0.365	0.332	0.35	0.35389
	RPE (º)	0.088	0.13	0.132	0.116	0.166	0.103	0.111	0.103	0.147	0.125895
ORB-SLAM2	t_{err} (%)	58.49	145.36	185.8	101.98	155.7	51.05	125.01	94.92	96.31	100.6711
	r_{err} (º/100m)	16.71	22.53	24.31	20.11	17.51	17.15	18.82	10.41	10.41	**20.9079**
	ATE	30.67	38.47	88.96	172.66	36.31	22.28	478.87	103.74	137.45	**115.754**
	RPE (m)	0.167	0.498	0.649	0.388	0.672	0.154	0.362	0.311	0.312	**339053**
	RPE (º)	0.09	0.099	0.113	0.11	0.113	0.079	0.097	0.07	0.07	**0.12084**

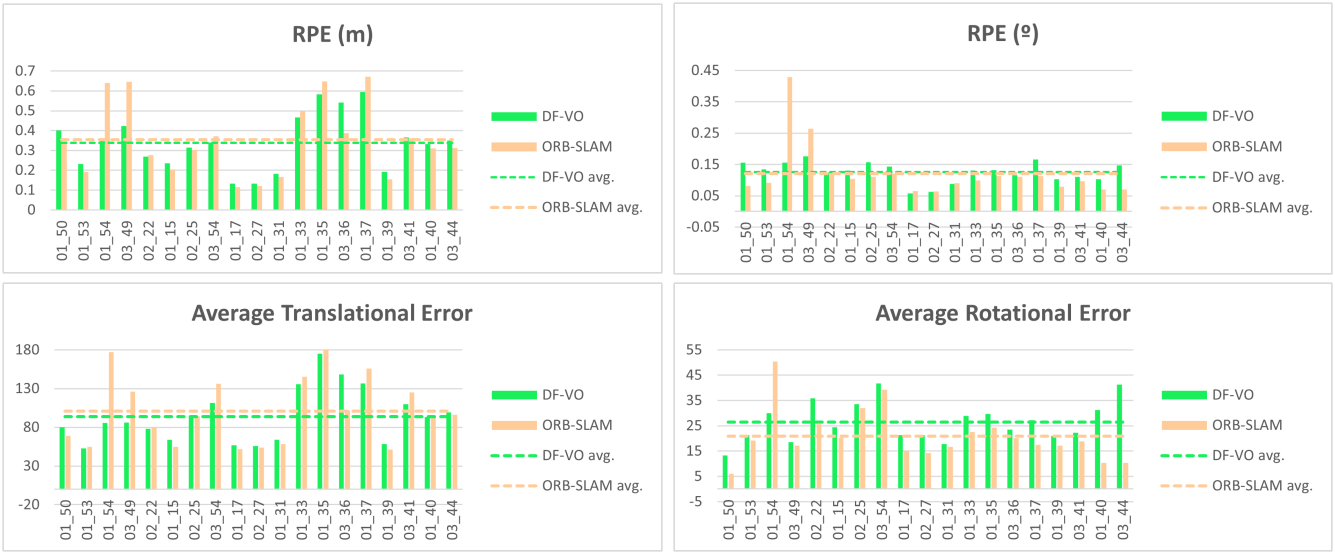


FIGURE 9. Comparison of relative VO evaluation metrics when applying DF-VO and ORB-SLAM2 algorithms in CAF datasets. Translational and rotational components of relative errors are shown separately.

median accuracy is evaluated as proposed by authors in [21]. The VO evaluation is done using the *KITTI Odometry Evaluation Toolbox* [17].

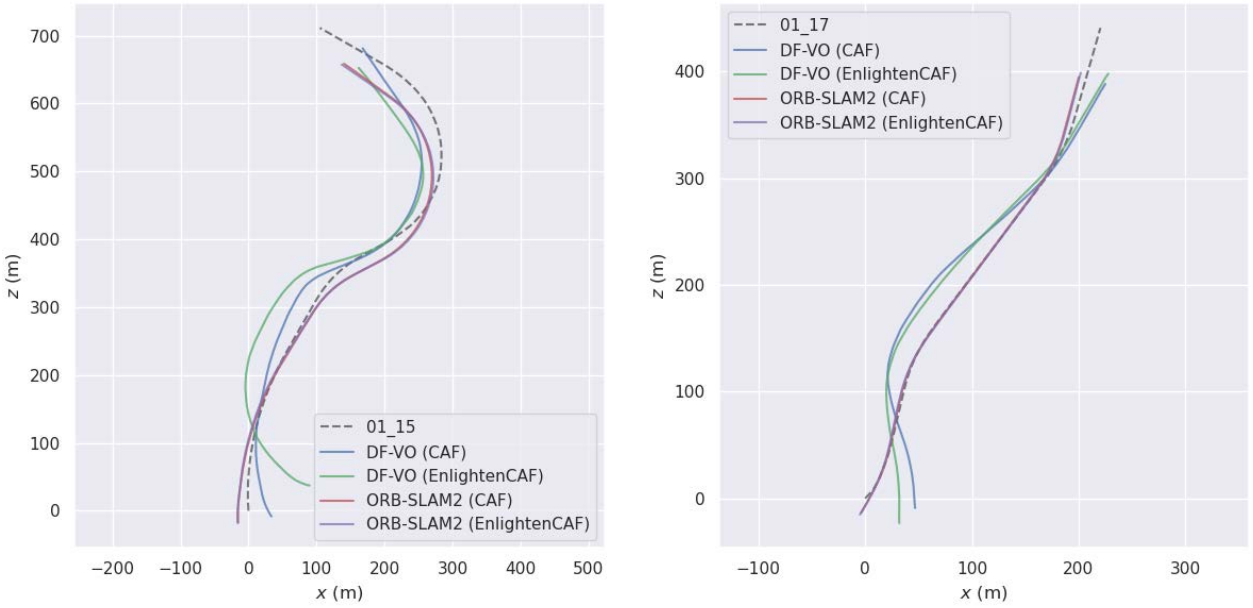
C. VO RESULTS IN CAF DATASET

Table 3 shows the results of DF-VO and ORB-SLAM2 in the CAF dataset. Figures 8 and 9 represent the results depicted in table 3. The visual representation can be found in Figure 9.

Previously, DF-VO and ORB-SLAM2 were evaluated in the KITTI Odometry dataset; however, KITTI does not contain those perception challenges as it contains considerably different properties related to the sequence length and visual characteristics. Results in *CAF* dataset show that the errors of both algorithms are higher than those found in the KITTI dataset. The RPE for DF-VO is 0.038 and 0.339 in KITTI dataset and *CAF* dataset, respectively. While for ORB-SLAM2, RPE measures are 0.130 and 0.353.

In the case of the ATE, the error of DF-VO in KITTI dataset is 6.344 while in the *CAF* dataset is 210.517. For ORB-SLAM2, the ATE is 26.48 and 115.754 in KITTI dataset and *CAF* dataset, respectively.

It can be seen that ORB-SLAM2 outperforms DF-VO in this challenging scenario, where the sequences are longer



(a) Sequence 01_15 (b) Sequence 01_17

FIGURE 10. Comparison of ORB-SLAM2 and DF-VO application on two sample sequences in both CAF and EnlightenCAF datasets and the ground truth for each trajectory.

TABLE 4. Average standard VO errors in *CAF* dataset when reducing the sequences to platform areas without lighting constraints.

Algorithm	Metric	Avg. Err
DF-VO	t_{err} (%)	18.520
	r_{err} (°/100m)	6.975
	ATE	2.298
	RPE (m)	0.049
	RPE (°)	0.037
ORB-SLAM2	t_{err} (%)	19.484
	r_{err} (°/100m)	14.681
	ATE	4.113
	RPE (m)	0.0798
	RPE (°)	0.126

FIGURE 11. A frame from the *CAF* dataset enhanced by EnlightenGAN.

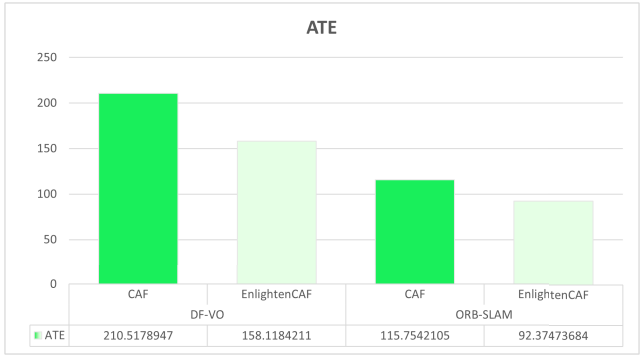


FIGURE 12. Comparative of ATE when applying DF-VO and ORB-SLAM2 algorithms in CAF and EnlightenCAF datasets.

than the standard KITTI dataset. If the *CAF* sequences are shortened to just platform areas where the lighting challenges are more limited, and more similar to the lighting conditions of the KITTI dataset, the errors are reduced to similar values (see Table 4) of executing DF-VO, and ORB-SLAM2 in KITTI dataset [17], [21]. For instance, DF-VO achieves an RPE (m) of 0.027 in KITTI dataset and 0.049 in shortened *CAF* dataset. ORB-SLAM2 achieves an ATE of 9.464 in KITTI dataset while 4.113 is achieved in shortened *CAF* dataset. Furthermore, the same behavior as in KITTI dataset is observed: DF-VO performance is higher than ORB-SLAM2. These results seem to support that the challenging scene conditions hinder the application of VO algorithms in such scenarios.

Results are visually shown in figure 10. In the case of DF-VO, a scale misalignment can be appreciated as the shape of most estimated trajectories is similar to the ground truth shape, but a dimensionality error appears.

As mentioned in [17], geometry-based VO algorithms as ORB-SLAM2 suffer from a scale drift when ideal visual conditions are not met. In the case of DF-VO, being a

FIGURE 13. Comparison of relative VO evaluation metrics when applying DF-VO and ORB-SLAM2 algorithms in CAF and EnlightenCAF datasets. Translational and rotational components of relative errors are shown separately.

hybrid algorithm, the scale may be wrongly estimated due to issues related to the geometric characteristics of the underground visual domain or deep-learning training process. The estimation error of the learning part of the algorithm could be reduced by training the deep models in the target scenario.

Nevertheless, these results require an adaptation of reference VO solutions to increase the performance in the underground railway domain. Image enhancement techniques or solutions based on the fusion of different odometry sensors could provide the precision required by autonomous train operations.

V. ENLIGHTENGAN IN VO APPLICATION

This section explores the application of the image enhancement technique EnlightenGAN in ORB-SLAM2 and DF-VO algorithms.

VO algorithms are based on minimizing the reprojection error of consecutive frames captured by the camera. The error is estimated by solving the essential matrix, which depends on the intrinsic camera parameters, and assuming the camera satisfies the pinhole camera model. In a previous work, the enhanced images calibration procedure was pursued to assess the EnglithenGAN architecture's effect on the camera's calibration. The experimental results showed that the GAN architecture did not significantly disturb the camera calibration parameters. Therefore, it was concluded that VO algorithms could be applied directly to the dataset enhanced by EnlightenGAN.

In the following section the enhanced dataset generation, the experimental configuration, and, finally, the results are explained.

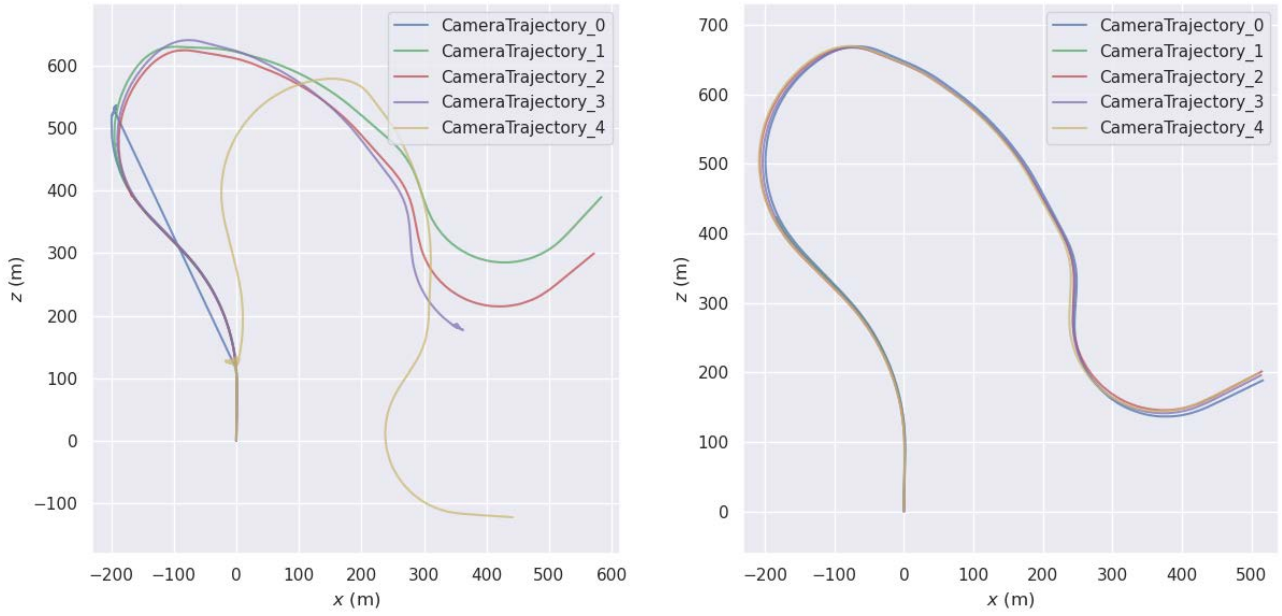
A. ENHANCED DATASET GENERATION: EnlightenCAF

The CAF dataset enhanced by EnlightenGAN is named *EnlightenCAF*. Figure 11 shows the result of the enhancement in the same tunnel zone frame as in figure 1.

The same algorithm configuration from *CAF* dataset experimentation has been used. The enhancing inference model is composed of pretrained weights from original authors.

B. RESULTS IN EnlightenCAF

In the previous work [9], the experimental results showed that *EnlightenGAN* improves the DF-VO performance in low-light car scenarios. In this case, the same behavior was confirmed: quantitative results show that EnlightenGAN reduces the VO errors for both algorithms. Figure 12 shows the reduction in the mean ATE and mean RPE of both algorithms for all the sequences in *EnlightenCAF*.



(a) *CAF* dataset

(b) *EnlightenCAF* dataset

FIGURE 14. Pose dispersion analysis on sample trajectory *01_54*. ORB-SLAM2 algorithm is executed five times on each dataset.

A relative ATE reduction of 24.89% and 20.20% is observed, respectively, when DF-VO and ORB-SLAM2 are applied in the enhanced sequences. Figure 13 shows RPE, t_{err} and r_{err} evaluation metrics in EnlightenCAF dataset.

In the case of RPE, DF-VO algorithm obtains a relative improvement of 1.97% and 4.74% for translation and rotation components, respectively. ORB-SLAM2 gets a relative improvement of 14.59% for the RPE translation component and a relative improvement of 18.55% for the rotation component. t_{err} and r_{err} present a relative reduction of 0.22% and 4.16% when applying DF-VO, and a relative reduction of 3.63% and 9.31% when applying ORB-SLAM2.

Figure 10 shows a result comparison of DF-VO and ORB-SLAM2 in the sequences of CAF and EnlightenGAF. As in *CAF* dataset, it can be seen that the algorithms can estimate the shape of the *EnlightenCAF* trajectories. However, a scale underestimation problem appears again. Furthermore, DF-VO results show that the rotation estimation is affected in the *EnlightenCAF* dataset.

The results demonstrate that EnlightenGAN improves VO algorithms performance in the underground railway domain. Furthermore, the relative error is reduced more for the geometric-based VO algorithm, while absolute error is reduced more in the learning-based algorithm.

However, as in the *CAF* dataset, the errors continue being higher than the results obtained by the algorithms in the KITTI dataset. Therefore, an affection of lighting conditions of the scenario can still be appreciated. This affection could be related to scale underestimation problems found in both algorithms, especially in the hybrid DF-VO.

Additionally, when evaluating the VO algorithms, it has been seen that the dispersion of the poses estimated by ORB-SLAM2 in different runs is reduced when enhancing the frames with EnlightenGAN.

The dispersion of poses among different executions of ORB-SLAM2 has been evaluated using standard metrics [93]. These metrics include the *variance* (σ^2) and the *Coefficient of Variation* (*cv*).

The evaluation procedure has been to run ORB-SLAM2 five times in each dataset, the original *CAF* and the enhanced *EnlightenCAF*. Figure 14 shows the results of applying ORB-SLAM2 five times for a given sequence (*01_54*) in the *CAF* and the enhanced *EnlightenCAF* datasets. It can be seen that the distribution of the poses through the trajectory is more constant in the enlightened dataset.

From the results, it can be seen that enlightening the datasets with *EnlightenGAN* increases the VO performance and tends to reduce ORB-SLAM2 dispersion. An analysis of the trouble spots in the dispersion results could better understand the high dispersion in such frames and detect further possible improvements for VO algorithms in such scenarios.

VI. CONCLUSION

This paper has presented a method to create a ground truth database for underground railway scenarios, where the GPS is unavailable, or the access to the infrastructure is not easily granted. The ground truth data generation is based on camera frames, ERTMS/ETCS ATP data, the railway gradient profile map, and geodetic coordinates of the target railway. Second,

it has proposed to enhance image lighting conditions with EnlightenGAN, which can be used with any state-of-the-art VO. Finally, it has presented the result of the experiment performed within a real urban underground railway scenario. The scenario was characterized by varying lighting conditions (tunnel vs. platform), low illumination (in tunnels), or textureless areas that challenged the state-of-the-art VO algorithms. The experiments were performed using two VO approaches: geometric (ORB-SLAM) and hybrid (DF-VO). The results show that the data enhancement increases the performance of both VO algorithms, reducing the translational error by at least 18%.

Future research proposes to apply the proposed dataset generation method and image enhancement algorithm in more underground railway scenarios. Sensor fusion is also a promising research direction. It is expected that the inclusion of new sensors will reduce uncertainty and increase accuracy, which will be welcome for autonomous train operations requiring higher localization accuracy (e.g., precise train stop operation).

ACKNOWLEDGMENT

The authors would like to thank Euskotren and Eusko Trenbide Sarea.

REFERENCES

[1] K. Yousif, A. Bab-Hadiashar, and R. Hoseinnezhad, "An overview to visual odometry and visual SLAM: Applications to mobile robotics," *Intell. Ind. Syst.*, vol. 1, no. 4, pp. 289–311, Dec. 2015.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[4] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.

[5] C. G. Atkeson, P. W. Benzun, N. Banerjee, D. Berenson, C. P. Bove, X. Cui, M. DeDonato, R. Du, S. Feng, P. Franklin, and M. A. Gennert, "Achieving reliable humanoid robot operations in the DARPA robotics challenge: Team WPI-CMU's approach," in *Proc. Int. Conf. Modeling Simul. Auton. Syst.* Cham, Switzerland: Springer, 2018, pp. 271–307.

[6] T. Rouček, M. Pecka, P. Cížek, T. Petrícek, J. Bayer, V. Šalanský, D. Hert, M. Petrlík, T. Báca, V. Spurný, and F. Pomerleau, "Darpa subterranean challenge: Multi-robotic exploration of underground environments," in *Proc. Int. Conf. Modeling Simul. Auto. Syst.* Cham, Switzerland: Springer, 2019, pp. 274–290.

[7] K. Ebadi, Y. Chang, M. Palieri, A. Stephens, A. Hatteland, E. Heiden, A. Thakur, N. Funabiki, B. Morrell, S. Wood, L. Carlone, and A.-A. Agha-mohammadi, "LAMP: Large-scale autonomous mapping and positioning for exploration of perceptually-degraded subterranean environments," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2020, pp. 80–86.

[8] A. Agha, K. Otsu, B. Morrell, D. D. Fan, R. Thakker, A. Santamaria-Navarro, S.-K. Kim, A. Bouman, X. Lei, J. Edlund, and M. F. Ginting, "NeBula: Quest for robotic autonomy in challenging environments; TEAM CoSTAR at the DARPA subterranean challenge," 2021, *arXiv:2103.11470*.

[9] J. Z. Ansorregi, M. E. Garcia, M. Z. Akizu, and N. A. Arexolaleiba, "Image enhancement using GANs for monocular visual odometry," in *Proc. IEEE Int. Workshop Electron., Control, Meas., Signals Appl. Mechatronics (ECMSM)*, Jun. 2021, pp. 1–6.

[10] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2004, p. 1.

[11] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[12] G. P. Stein, O. Mano, and A. Shashua, "A robust method for computing vehicle ego-motion," in *Proc. IEEE Intell. Vehicles Symp.*, Oct. 2000, pp. 362–368.

[13] K. Yamaguchi, T. Kato, and Y. Ninomiya, "Vehicle ego-motion estimation and moving object detection using a monocular camera," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2006, pp. 610–613.

[14] F. Tschopp, T. Schneider, A. W. Palmer, N. Nourani-Vatani, C. Cadena, R. Siegwart, and J. Nieto, "Experimental comparison of visual-aided odometry methods for rail vehicles," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1815–1822, Apr. 2019, doi: 10.1109/LRA.2019.2897169.

[15] V. Grabe, H. H. Bűlthoff, D. Scaramuzza, and P. R. Giordano, "Nonlinear ego-motion estimation from optical flow for online control of a quadrotor UAV," *Int. J. Robot. Res.*, vol. 34, no. 8, pp. 1114–1135, 2015.

[16] Y. Zou, P. Ji, Q.-H. Tran, J.-B. Huang, and M. Chandraker, "Learning monocular visual odometry via self-supervised long-term modeling," in *Computer Vision*. Glasgow, U.K.: Springer, Aug. 2020, pp. 710–727.

[17] H. Zhan, C. S. Weerasekera, J.-W. Bian, R. Garg, and I. Reid, "DF-VO: What should be learnt for visual odometry?" 2021, *arXiv:2103.00933*.

[18] H. Gaoussou and P. Dewei, "Evaluation of the visual odometry methods for semi-dense real-time," *Adv. Comput., Int. J.*, vol. 9, no. 2, pp. 01–14, Mar. 2018, doi: 10.5121/acij.2018.9201.

[19] R. Wang, M. Schworer, and D. Cremers, "Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3903–3911.

[20] H. Alismail, M. Kaess, B. Browning, and S. Lucey, "Direct visual odometry in low light using binary descriptors," *IEEE Robot. Automat. Lett.*, vol. 2, no. 2, pp. 444–451, Apr. 2016.

[21] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[22] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[23] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robot. Automat. Lett.*, vol. 2, no. 2, pp. 796–803, Apr. 2017.

[24] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual SLAM algorithms: A survey from 2010 to 2016," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, no. 1, pp. 1–11, Dec. 2017.

[25] J. Delmerico and D. Scaramuzza, "A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2502–2509.

[26] B. Bescos, J. M. Fácil, J. Civera, and J. L. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4076–4083, Oct. 2018.

[27] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "DS-SLAM: A semantic visual SLAM towards dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1168–1174.

[28] F.-A. Moreno, D. Zuñiga-Noël, D. Scaramuzza, and J. Gonzalez-Jimenez, "PL-SLAM: A stereo SLAM system through the combination of points and line segments," *IEEE Trans. Robot.*, vol. 35, no. 3, pp. 734–746, Jun. 2019.

[29] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 817–833.

[30] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2020, pp. 4203–4210.

[31] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2014, pp. 15–22.

[32] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2017.

[33] L. Koestler, N. Yang, N. Zeller, and D. Cremers, "Tandem: Tracking and dense mapping in real-time using deep multi-view stereo," in *Proc. Conf. Robot Learn.*, 2022, pp. 34–45.
[34] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2938–2946, doi: 10.1001/jama.284.15.1980.
[35] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6555–6564, doi: 10.1109/CVPR.2017.694.
[36] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2017, pp. 2043–2050.
[37] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "DSAC-differentiable ransac for camera localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6684–6692.
[38] S. J. Lee, H. Choi, and S. S. Hwang, "Real-time depth estimation using recurrent CNN with sparse depth cues for SLAM system," *Int. J. Control, Autom. Syst.*, vol. 18, no. 1, pp. 206–216, Jan. 2020.
[39] G. Costante and T. A. Ciarfuglia, "LS-VO: Learning dense optical subspace for robust visual odometry estimation," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1735–1742, Jul. 2018.
[40] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "DeMoN: Depth and motion network for learning monocular stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, p. 50.8–5047.
[41] K. Tateno, F. Tombari, I. Laina, and N. Navab, "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6243–6252.
[42] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. M. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 340–349.
[43] N. Yang, L. von Stumberg, R. Wang, and D. Cremers, "D3 VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry," 2020, *arXiv:2003.01060*.
[44] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1983–1992.
[45] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5667–5675.
[46] T. Shen, Z. Luo, L. Zhou, H. Deng, R. Zhang, T. Fang, and L. Quan, "Beyond photometric loss for self-supervised ego-motion estimation," in *Proc. Int. Conf. Robot. Automat. (ICRA)*, May 2019, pp. 6359–6365.
[47] M. O. A. Aqel, M. H. Marhaban, M. I. Saripan, and N. B. Ismail, "Review of visual odometry: Types, approaches, challenges, and applications," *SpringerPlus*, vol. 5, no. 1, pp. 1–26, Dec. 2016.
[48] Y. Jiang, X. Gong, D. Liu, Y. Cheng, and C. Fang, "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.
[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[51] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "RAISE: A raw images dataset for digital image forensics," in *Proc. 6th ACM Multimedia Syst. Conf.*, Mar. 2015, pp. 219–224.
[52] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–144, 2017.
[53] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.
[54] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.
[55] Y. Gong, P. Liao, X. Zhang, L. Zhang, G. Chen, K. Zhu, X. Tan, and Z. Lv, "Enlighten-GAN for super resolution reconstruction in mid-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 6, p. 1104, Mar. 2021.
[56] M. Afifi, K. G. Derpanis, B. Ommer, and M. S. Brown, "Learning multi-scale photo exposure correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9157–9167.
[57] Z. Ni, W. Yang, S. Wang, L. Ma, and S. Kwong, "Towards unsupervised deep image enhancement with generative adversarial network," *IEEE Trans. Image Process.*, vol. 29, pp. 9140–9151, 2020.
[58] W. Xiong, D. Liu, X. Shen, C. Fang, and J. Luo, "Unsupervised low-light image enhancement with decoupled networks," 2020, *arXiv:2005.02818*.
[59] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi, "Real-time RGB-D camera relocalization," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2013, pp. 173–179.
[60] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 791–804.
[61] A. Handa, T. Whelan, J. McDonald, and A. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. IEEE Intl. Conf. Robot. Automat. (ICRA)*, Hong Kong, May 2014, pp. 1524–1531.
[62] S. Cortés, A. Solin, E. Rahtu, and J. Kannala, "ADVIO: An authentic dataset for visual-inertial odometry," in *Computer Vision*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 425–440.
[63] D. Zuñiga-Noël, A. Jaenal, R. Gomez-Ojeda, and J. Gonzalez-Jimenez, "The UMA-VI dataset: Visual–inertial odometry in low-textured and dynamic illumination environments," *Int. J. Robot. Res.*, vol. 39, no. 9, pp. 1052–1060, Aug. 2020, doi: 10.1177/0278364920938439.
[64] S. Ceriani, G. Fontana, A. Giusti, D. Marzorati, M. Matteucci, D. Migliore, D. Rizzi, D. G. Sorrenti, and P. Taddei, "Rawseeds ground truth collection systems for indoor self-localization and mapping," *Auton. Robots*, vol. 27, no. 4, pp. 353–371, 2009. [Online]. Available: http://dblp.uni-trier.de/db/journals/arobots/arobots27.html#CerianiFGMM%MRST09
[65] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using sfm and object labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1625–1632.
[66] S. Klenk, J. Chui, N. Demmel, and D. Cremers, "TUM-VIE: The TUM stereo visual-inertial event dataset," in *Proc. Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 8601–8608.
[67] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 573–580.
[68] J. Engel, V. Usenko, and D. Cremers, "A photometrically calibrated benchmark for monocular visual odometry," 2016, *arXiv:1607.02555*.
[69] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using LSTMs for structured feature correlation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 627–637.
[70] M. Fallon, H. Johannsson, M. Kaess, and J. J. Leonard, "The MIT stata center dataset," *Int. J. Robot. Res.*, vol. 32, no. 14, pp. 1695–1699, 2013.
[71] H. Alismail, B. Browning, and M. B. Dias, "Evaluating pose estimation methods for stereo visual odometry on robots," in *Proc. 11th Int. Conf. Intell. Auton. Syst. (IAS)*, vol. 3, 2010, p. 2.
[72] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3260–3269.
[73] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of Michigan north campus long-term vision and lidar dataset," *Int. J. Robot. Res.*, vol. 35, no. 9, pp. 1023–1035, Aug. 2015.
[74] J.-L. Blanco-Claraco, F.-Á. Moreno-Dueñas, and J. González-Jiménez, "The Málaga urban dataset: high-rate stereo and LiDAR in a realistic urban scenario," *Int. J. Robot. Res.*, vol. 33, no. 2, pp. 207–214, Feb. 2014, doi: 10.1177/0278364913507326.
[75] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford robotcar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2017.
[76] G. Pandey, J. R. McBride, and R. M. Eustice, "Ford campus vision and LiDAR data set," *Int. J. Robot. Res.*, vol. 30, no. 13, pp. 1543–1552, 2011.
[77] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban dataset with multi-level sensors from highly diverse urban environments," *Int. J. Robot. Res.*, vol. 38, no. 6, pp. 642–657, May 2019.

[78] D. Olid, J. M. Fácil, and J. Civera, "Single-view place recognition under seasonal changes," in *Proc. PPNIV Workshop IROS*, 2018, pp. 1–6.
[79] A. L. Majdik, C. Till, and D. Scaramuzza, "The Zurich urban micro aerial vehicle dataset," *Int. J. Robot. Res.*, vol. 36, no. 3, pp. 269–273, 2017, doi: 10.1177/0278364917702237.
[80] A. Z. Zhu, D. Thakur, T. Ozaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3D perception," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2032–2039, Jul. 2018, doi: 10.1109/LRA.2018.2800793.
[81] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.
[82] M. T. Ohradzansky, E. R. Rush, D. G. Riley, A. B. Mills, S. Ahmad, S. McGuire, H. Biggie, K. Harlow, M. J. Miles, E. W. Frew, C. Heckman, and J. S. Humbert, "Multi-agent autonomy: Advancements and challenges in subterranean exploration," 2021, *arXiv:2110.04390*.
[83] H. Zhao, B. Zhang, C. Wu, Z. Zuo, and Z. Chen, "Development of a Coordinate Transformation method for direct georeferencing in map projection frames," *ISPRS J. Photogramm. Remote Sens.*, vol. 77, pp. 94–103, Mar. 2013.
[84] (2017). ÖPNVKarte Map. *Planet Dump*. [Online]. Available: https://planet.osm.org
[85] (2017). OpenStreetMap Contributors. *Planet Dump*. [Online]. Available: https://planet.osm.org
[86] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[87] Y. Gao, A. Rehman, and Z. Wang, "CW-SSIM based image classification," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 1249–1252.
[88] J. Søgaard, L. Krasula, M. Shahid, D. Temel, K. Brunnström, and M. Razaak, "Applicability of existing objective metrics of perceptual quality for adaptive video streaming," *Electron. Imag.*, vol. 28, no. 13, pp. 1–7, Feb. 2016.
[89] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 376–380, Apr. 1991.
[90] (2017). Michael Grupp. *EVO: Python Package for the Evaluation of Odometry and Slam*. [Online]. Available: https://github.com/MichaelGrupp/evo
[91] H. Zhan, C. S. Weerasekera, J. Bian, and I. Reid. (2021). *DF-VO*. [Online]. Available: https://github.com/Huangying-Zhan/DF-VO
[92] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
[93] C. S. Rayat, "Measures of dispersion," in *Statistical Methods in Medical Research*. Singapore: Springer, 2018, pp. 47–60.

MIKEL ETXEBERRIA-GARCIA received the B.S. degree in computer engineering and the M.S. degree in advanced computer systems from the University of the Basque Country(UPV/EHU), Donostia, Spain, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree in applied engineering with Mondragon Unibertsitatea, Arrasate-Mondragón, Spain.

From 2016 to 2018, he was a System and Database Administrator at LKS, Donostia. Since 2018, he has been a Ph.D. student with the Ikerlan Technology Research Centre, Arrasate-Mondragón. His research interests include the application of deep learning techniques in railway domain, more specifically, on autonomous train navigation related tasks and on computer vision.

MAIDER ZAMALLOA received the degree in computer science engineering from the University of the Basque Country (UPV/EHU), Donostia, Spain, in 2003, and the Ph.D. degree in physical engineering from the University of the Basque Country (UPV/EHU), Bilbao, Spain, in 2010, in the field of pattern classification for speech technologies.

She is a Certified TUV Functional Safety Engineer for ISO 26262 (#13278/16). She is currently a Researcher with the Ikerlan Technological Research Centre, Arrasate-Mondragon, Spain, since 2009. She is also a Senior Researcher in the field of dependable embedded systems with more than ten years of experiences in testing of safety critical software for railway signaling systems (SIL4). Her research interests include dependable software and machine learning for computer vision.

NESTOR ARANA-AREXOLALEIBA received the Ph.D. degree from the CNRS-LAAS (Robotics and AI Research Group), and the Ph.D. degree in automatic system form the INSAT, Toulouse, in 2020. Since 2002, he has been working with Mondragon University. From 2018 to 2019, he was a Guest Researcher with the Robotics and Automation Research Group, Aalborg University, Denmark, where he was researching on reinforcement learning strategies for human–robot collaboration. He is currently working with the Robotics and Automation Research Group, Mondragon University. He has more than 50 publications (three books, two book chapters, five magazines, and more than 40 conference papers). His research interests include machine learning and image processing for human–robot collaboration and autonomous vehicles. Besides, he teaches on robotics and automation master (ROS, AI-based control and mobile robotics). For more information visit the link (https://www.researchgate.net/profile/Nestor_Arana-Arexolaleiba).

MIKEL LABAYEN received the degree in technical telecommunication engineering from the Public University of Navarre, in 2005, with a focus on image and sound, and the degree from the Faculty of Telecommunication Engineering, Public University of Navarre, in 2007. He is currently pursuing the Ph.D. degree in computer vision and machine learning fields. He completed his undergraduate dissertation at the Electronic Engineering Department, University of Surrey, U.K., in the audio and speech signal processing area. He completed his master's thesis at the Vicomtech Research Center, Digital Television and Multimedia Services Department.

From 2007 to 2012, he started his professional carrier as a Staff Researcher in the field of computer vision—multimedia content analysis with the Vicomtech Research Center. In 2012, he started a new career as the Co-Founder and the Research and Development Project Manager at Smowltech start-up (spin-off of Vicomtech), where he researches in the automatic facial and voice recognition area developing applications for online user authentication based on human biometrics. In the same period, he was also an Associate Teacher with the Electronic Technology Department, University of the Basque Country. He is currently working in the field of intelligent transport systems (ITS) designing computer vision and machine learning-based solutions for autonomous train operations in the railway sector, within the CAF Group. His work as a Researcher includes a number of publications and four patents.