ISSN 0002-306X. Proc. of the RA NAS and NPUA Ser. of tech. sc. 2023. V. LXXVI, N1

UDC 004.832

COMPUTER SCIENCE AND INFORMATICS

DOI: 10.53297/0002306X-2023.v76.1-12

A.A. AVETISYAN, T.B. KHACHATRYAN, M.T. GRIGORYAN

PHOTOREALISTIC AND SYNTHETIC STEREO-DATASET GENERATION METHOD FOR VISUAL ODOMETRY AND DEPTH ESTIMATION

Computer vision is a rapidly developing field in modern computer science that deals with various challenging problems. Both mono and stereo imagery data are widely used for tasks such as depth estimation, visual odometry, and SLAM (Simultaneous Localization and Mapping). To ensure the clean verification and robust performance of the resulting software solutions, datasets should contain precise ground truth data. However, creating a real-world stereo dataset is a costly task as it requires stereo cameras and precise hardware for ground truth measurements (such as lidars, lasers, barometers, accelerometers, etc.). These types of hardware are often expensive and not accessible to intermediate users. An alternative approach is to use synthetic datasets, which are collections of computer-generated data designed to mimic real-world data. Synthetic datasets are used to train AI models when real-world data is not available, or to test the performance of models in simulated environments.

Our method suggests combining real-world data collection with synthetic data generation methods to maintain photorealism while gaining the advantages of synthetic data generation flow.

Keywords: computer vision, stereo dataset, synthetic dataset, simulated environment, photorealism.

Introduction. Stereo datasets are an essential component in various computer vision and autonomous system applications, such as SLAM [1]. They are used to train models for tasks such as semantic segmentation, object detection, and more. Stereo datasets are frequently generated using 3D vehicle simulators like AirSim [2] and Gazebo [3]. These simulators provide a wealth of information that can be extracted, including depth maps, camera positions, and IMU data, among others.

While synthetic dataset production has several advantages over traditional real-world data collection methods, it also has its own set of limitations and weaknesses. One of the main limitations is the cost of obtaining high-quality 3D assets and scenes, which may come with licensing restrictions. Additionally, creating a realistic virtual environment that accurately simulates the real world requires a significant amount of programming and 3D design expertise. The

simulation tools themselves can also be limited and may not accurately reflect the complexity and variability of real-world data, leading to models that perform well in simulation but poorly in real-world scenarios.

Furthermore, synthetic dataset production can also be computationally intensive, requiring significant processing hardware and time to complete. This can be a significant drawback for those with limited resources, as the data collection process can be significantly longer than real-world data collection.

To overcome these limitations, our work describes a novel approach for stereo dataset production that combines real-world data with camera simulation in a 3D virtual environment. This approach provides photorealistic images from the real-world data while also removing hardware errors through ground truth extraction in the simulator. This allows for the generation of high-quality stereo datasets that accurately reflect the real-world, while also reducing the time and computation resources required for data collection.

Moreover, this approach offers greater freedom in terms of camera positioning, lighting, and object placement. Unlike real-world vehicles, virtual environments are not limited by the capabilities of physical vehicles, and cameras in a virtual environment can be programmed to have different intrinsic parameters or lenses. This opens new possibilities for stereo dataset production and allows for the generation of datasets that are not restricted by real-world limitations.

In conclusion, the novel approach to stereo dataset production described in the text provides a compelling solution to the limitations and weaknesses of synthetic dataset production. By combining real-world data with camera simulation in a 3D virtual environment, it offers a cost-effective and efficient method for generating high-quality stereo datasets that accurately reflect the real-world.

Related works. In this section, we study various stereo datasets and recent advancements in stereo dataset synthesis. The advancements in machine learning heavily rely on the quality and quantity of data available. Hence, the current wave of solutions in the field faces challenges related to data size, generation, and ground truth availability. To ensure continued progress and evaluate algorithms, robust and relevant datasets are essential.

Synthetic stereo datasets. To address the data challenges, several synthetic stereo datasets have been created. These datasets provide a platform for machine learning models to learn from and to be evaluated against.

Synthetic dataset performance. One such dataset is the MPI-Sintel dataset [4], which was created in 2012. The dataset is used to evaluate optical flow algorithms and is derived from the animated film "Sintel." The dataset contains 1064 synthesized stereo images and ground truth data for disparity and provides a

wide range of image degradations and effects. The dataset is generated using the Blender 3D generation open-source tool. Despite its popularity, the biggest drawback of the MPI-Sintel dataset is the graphical style of the movie. Models trained on this dataset have difficulties in recognizing real-world depth and are not suitable for practical inference in real-life scenarios.

Another synthetic dataset is the Virtual KITTI dataset [5], created in 2016. The dataset was designed for the evaluation of computer vision models for various video understanding tasks such as object detection, multi-object tracking, semantic segmentation, optical flow, and depth estimation. The dataset includes 50 high-resolution videos, generated from five different virtual worlds in urban settings, under different imaging and weather conditions. The virtual worlds were created using the Unity game engine [6] and a novel real-to-virtual cloning method. Although widely used, the dataset is limited to street views and the camera movements are restricted to the car's forward movement.

The TartanAir dataset [7] is a more recent addition, created in 2020. The aim of this dataset was to expand the limited scenes and camera motions provided by famous real datasets such as KITTI [8] or EuroC [9]. The dataset includes sequences from 30 various simulation environments, including indoor and outdoor scenes, and features balanced camera motion in 6 DoF. The dataset also includes challenging visual effects such as day-night alterations, weather effects, seasons, and others. Although the dataset is diverse, the photorealism is still far from the quality of real video frames.

In conclusion, synthetic stereo datasets have proven to be useful in evaluating optical flow algorithms and training machine learning models. Despite their limitations, they provide a platform for continued progress in the field.

Method. The dataset generation steps are illustrated in Fig. 1. As a first step we are capturing photo images by a DJI Mini SE drone. The images are later used for real-world large scene point cloud creation. The Litchi API [10] application is used to program the drone. The photos are taken at a minimum interval of one second, with a flight speed of 2.5 *km/h.*

To capture an area, the drone is flown in circles with a radius of 50 meters, at a height of 30 meters above the surface. This height was chosen based on considerations of drone safety and the presence of environmental objects, such as trees and buildings.

Several camera views and drone trajectories were tested, and a circle with a 50-meter diameter was found to produce the best results. The image resolution is 4000x3000 and GPS data is attached to each image.

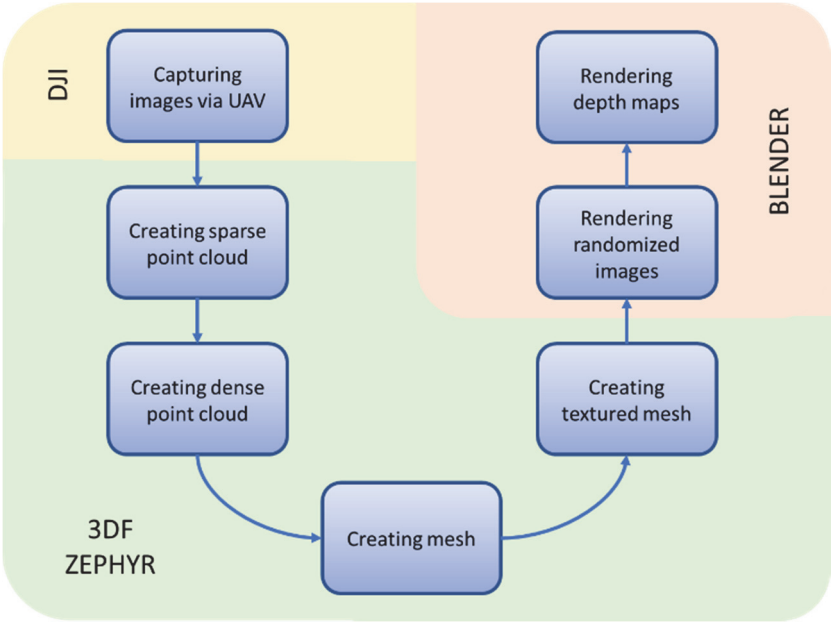

Fig. 1. Dataset generation steps

To enhance the 3D representation of the real-world large scene, the 3DF Zephyr [11] open-source tool is employed to calculate the point cloud (Fig. 2). This tool is known for its ability to create precise and comprehensive 3D models from photographs. The point cloud calculation is implemented by using Multi-View Stereo and Structure from motion algorithms. Multi-View Stereo algorithm consists of two consequent steps. First is depth map computation which is measured via formula (1):

$$E(d) = E\_data(d) + \lambda * E\_smooth(d), \quad (1)$$

where d is the depth value, E_data is the data term that measures the similarity between the reference image and the corresponding pixels in the other images at depth d, E_smooth is the smoothing term that encourages the neighboring pixels to have similar depth values, and λ is a weighting parameter that balances the two terms. The second step is point cloud generation. Once the depth maps have been computed for each image, a dense point cloud can be generated by triangulating the 3D coordinates of corresponding pixels in different views. The formula for triangulating a 3D point X from two camera positions C1 and C2, and corresponding image points x1 and x2, is (2):

$$X = (1/P1) * x1 = (1/P2) * x2, \quad (2)$$

where P1 and P2 are the 3x4 camera matrices for camera positions C1 and C2, respectively.

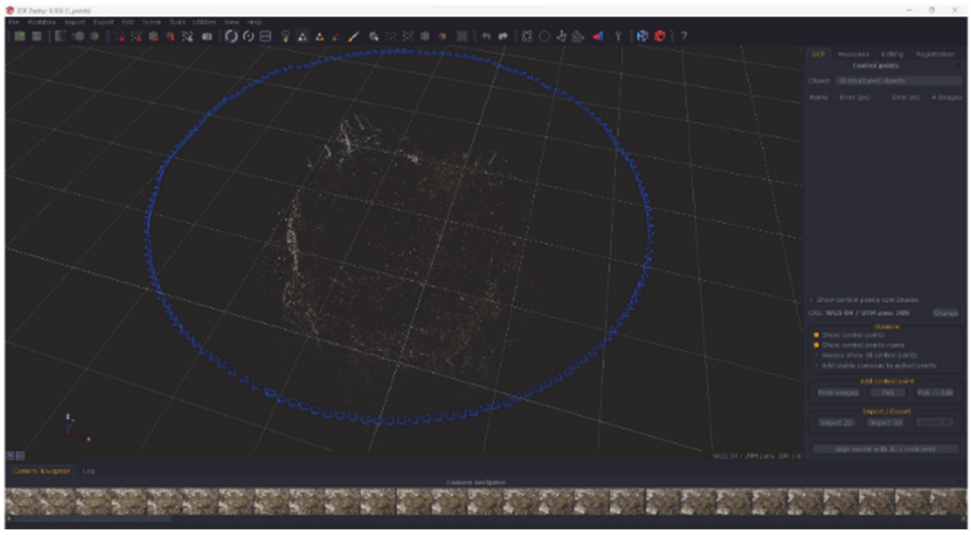

Fig.2. Generating sparse point cloud based on the images and camera positions

Once the point cloud is generated, the next step is to create a mesh using the same tool (Fig. 3). The mesh acts as a framework for the 3D model, providing structure and form. The creation of a mesh from the point cloud allows for a more intuitive and recognizable representation of the real-world scene.

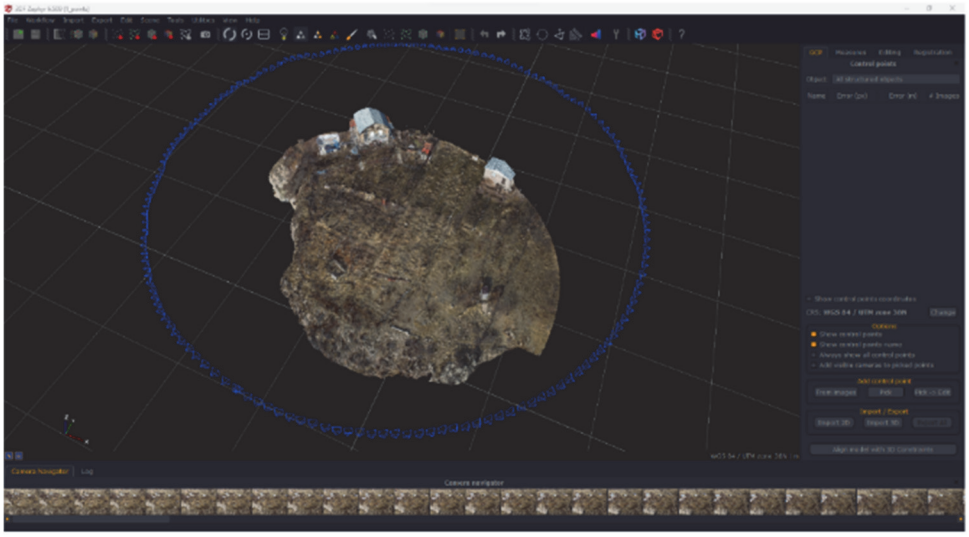

Fig. 3. Generating mesh based on the point cloud

Finally, a texture is added to the mesh, which is created from the images captured by the DJI Mini SE drone. The texture provides visual detail to the 3D model, giving it a more lifelike appearance. The combination of the mesh and the

texture results in a highly detailed and accurate representation of the real-world scene.

Small pieces are joined by using the control point method. The same control points are picked on 2 or more separate meshes, and a unified mesh is created (Fig. 4).

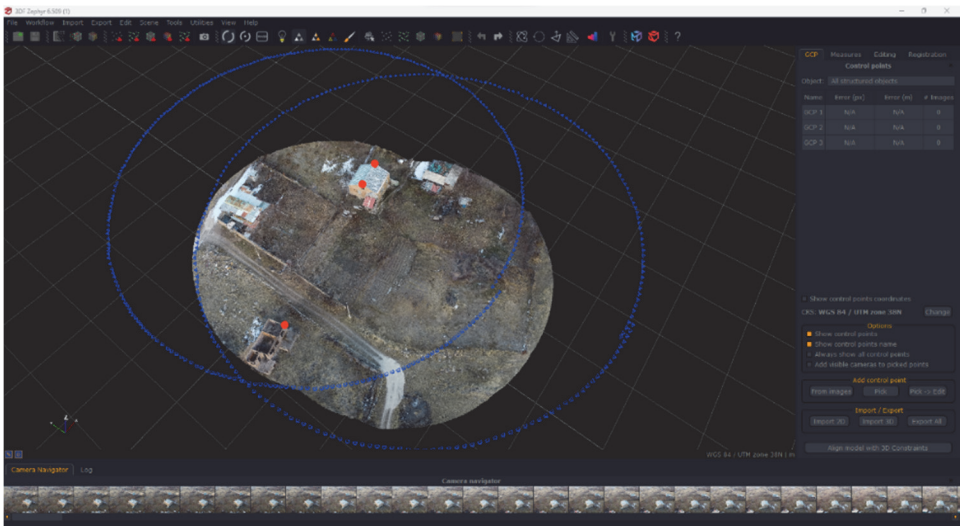

Fig. 4. Merging two small textured meshes

The textured mesh is then uploaded to the Blender 3.4 [12] open-source tool. The Blender platform enables the user to program one or more cameras with randomized positions, allowing for easy extraction of photos and accompanying depth maps. Scripting is achieved using Python 3 and is seamlessly integrated into the Blender environment. It is important to note that the use of two cameras is not a limit - it is possible to render images and depth maps from multiple cameras.

Of course, the photos produced in the virtual environment are not identical to the original photos. This is due to several factors such as the absence of small or intricate objects like trees or chimneys in the virtual environment. Additionally, the scene geometry may be slightly distorted. These differences can be observed in the comparison screenshots (Fig. 5).

The distortion of the 3D environment is evident by comparing the bottom part of the screenshots, as the whole scene is slightly shifted and the same camera position and angles in the virtual environment result in a slightly different viewpoint. The camera positioning used in the virtual environment was sourced from the DJI Mini SE drone flight logs.

Fig. 5. Original photo made by the DJI SE Mini drone (Top). Rendered image from the virtual environment (Bottom)

Blender also allows to generate depth maps for the rendered images (Fig. 6). As rendering happens in simulated environment, depth maps are ideal.

Fig. 6. Rendered image from the virtual environment (Left). Depth map of the rendered image (Right)

Generated dataset. By following the pipeline outlined in methodology section, a dataset of 1,000 images (captured by left and right cameras), corresponding depth maps and camera position ground truths were generated. The structure of the dataset is based on the KITTI visual odometry dataset. Each image and depth map has a resolution of 1280x720.

The generated virtual environment spans an area of 153,000 square meters. The camera positions and angles are randomized, with a height range of 20-100 meters and a horizontal coverage of approximately 38,480 square meters to prevent the capturing of the map edges in the rendered photos. The camera view angles are

also randomized and range from -45 to 45 degrees for X and Y rotations and from 0 to 360 degrees for Z rotation.

Scene lighting is also randomized, utilizing the Sun light feature of Blender 3.4. The script adjusts the light angles (ranging from 30 to 150 degrees for X, Y, and Z rotations) and intensity (from 5 to 10) to simulate different times of day and levels of cloudiness. The photo realism is measured with formula (3):

$$Error = 100 * \frac{\left(\sum \quad |i1(x,y) - i2(x,y)|\right)}{M*N*255*3}, \tag{3}$$

where i1, i2 are the image intensity values, M and N are the image width and height. The mean photo realism error across the image set is 12.49%. As a reference, the famous Virtual KITTI dataset's photo realism error varies from 5 to 20 percents depending on the image sequence.

Conclusion. This work presents a simple and robust flow for producing photorealistic stereo-datasets which do not require expensive hardware and IMUs for ground truth extraction. The flow uses open-source tools and does not require heavy processing for 3D environment synthesis. The other advantage of the presented flow is its flexibility in randomized data generation. That includes generated image resolution, camera positioning, light conditions, etc.

The resulting photo realism error was compared to the widely used Virtual KITTI dataset and can be considered as acceptable.

REFERENCES

1. Exploration: Simultaneous Localization and Mapping (SLAM), Computer Vision: A Reference Guide / **Perera, Samunda Barnes, Dr.Nick, Zelinsky, Dr.Alexander** // Springer US. –2014. -P. 268–275.
2. AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles / **Shital Shah, Debadeepta Dey, Chris Lovett, Ashish Kapoor** // Field and Service Robotics. – 2017.
3. **Open Robotics** https://gazebosim.org/docs.
4. A Naturalistic Open Source Movie for Optical Flow Evaluation / **Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black**.- University of Washington, Seattle, WA, USA, 2012.
5. **Adrien Gaidon, Qiao Wang, Yohann Cabon, Eleonora Vig**. Virtual Worlds as Proxy for Multi-Object Tracking Analysis.- CVPR, Las Vegas, Nevada, USA, 2016.
6. **Unity Technologies** https://docs.unity3d.com/Manual/index.html - 2021.
7. TartanAir: A Dataset to Push the Limits of Visual SLAM / **Wang, Wenshan and Zhu, Delong and Wang, Xiangwei and Hu, Yaoyu and Qiu, Yuheng and Wang, Chen and Hu, Yafei and Kapoor, Ashish and Scherer, Sebastian** // 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). – 2020.

8. **Andreas Geiger and Philip Lenz and Raquel Urtasun.** Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite // Conference on Computer Vision and Pattern Recognition (CVPR). – 2012.
9. The EuRoC micro aerial vehicle datasets / **M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. Achtelik and R. Siegwart** // International Journal of Robotic Research. – 2016.
10. **VC Technology Ltd** Litchi User Guide. – 2023.
11. https://www.3dflow.net – 2017.
12. https://www.blender.org – 2023.

National Polytechnic University of Armenia. The material is received 24.02.2023.

Ա.Ա. ԱՎԵՏԻՍՅԱՆ, Տ.Բ. ԽԱՉԱՏՐՅԱՆ, Մ.Տ. ԳՐԻԳՈՐՅԱՆ

ՖՈՏՈՌԵԱԼԻՍՏԻԿ ԵՎ ԱՐՀԵՍՏԱԿԱՆ ՍՏԵՐԵՈ-ՏՎՅԱԼՆԵՐԻ ՀԱՎԱՔԱԾՈՒՆԵՐԻ ՍՏԵՂԾՄԱՆ ՄԵԹՈԴ՝ ՏԵՍՈՂԱԿԱՆ ՏԵՂՈՐՈՇՄԱՆ ԵՎ ԽՈՐՈՒԹՅԱՆ ՈՐՈՇՄԱՆ ՀԱՄԱՐ

Համակարգչային տեսողությունը ժամանակակից համակարգչային գիտության մեջ արագ զարգացող ոլորտ է, որը զբաղվում է տարատեսակ բարդ խնդիրներով: Ինչպես մոնո, այնպես էլ ստերեո պատկերների տվյալները լայնորեն օգտագործվում են խորության գնահատման, տեսողական տեղորոշման և SLAM-ի (Simultaneous Localization and Mapping) համար: Ստացված ծրագրային լուծումների հստակ ստուգումն ու կայուն կատարումն ապահովելու համար տվյալների հավաքածուները պետք է պարունակեն ճշգրիտ տեղեկություն իրական դիրքի մասին: Այնուամենայնիվ, իրական աշխարհի ստերեո տվյալների հավաքածուի ստեղծումը ծախսատար խնդիր է, քանի որ այն պահանջում է ստերեո տեսախցիկներ և ճշգրիտ սարքեր իրական դիրքի չափումների համար (օրինակ՝ լիդարներ, լազերներ, բարձրաչափեր, արագաչափեր և այլն): Նման սարքավորումները սովորաբար թանկ են և հասանելի չեն հաճախադեպ օգտագործման համար: Այլընտրանքային մոտեցում է արհեստական տվյալների հավաքածուների օգտագործումը, որոնք համակարգչի միջոցով ստեղծված հավաքածուներ են: Դրանք նախատեսված են իրական աշխարհի տվյալների նմանակման համար: Արհեստական տվյալների հավաքածուները սովորաբար օգտագործվում են արհեստական բանականության մոդելների ուսուցման համար, երբ իրական աշխարհի տվյալները հասանելի չեն, ինչպես նաև մոդելավորման միջավայրում մոդելների աշխատանքը փորձարկելու համար:

Առաջարկվում է համատեղել իրական աշխարհի տվյալների հավաքագրումը արհեստական տվյալների ստեղծման մեթոդների հետ՝ պահպանելով ինչպես ֆոտոռեալիզմը, այնպես էլ արհեստական տվյալների ստեղծման ընթացակարգի առավելությունները:

Առանցքային բառեր. համակարգչային տեսողություն, ստերեո տվյալների հավաքածու, արհեստական տվյալների հավաքածու, մոդելավորված միջավայր, ֆոտոռեալիզմ:

А.А. АВЕТИСЯН, Т.Б. ХАЧАТРЯН, М.Т. ГРИГОРЯН

МЕТОД ГЕНЕРАЦИИ ФОТОРЕАЛИСТИЧНЫХ И СИНТЕТИЧЕСКИХ СТЕРЕОНАБОРОВ ДАННЫХ ДЛЯ ВИЗУАЛЬНОЙ ОДОМЕТРИИ И ОЦЕНКИ ГЛУБИНЫ

Компьютерное зрение - это быстроразвивающаяся область в современной компьютерной науке, которая занимается решением различных сложных проблем. Моно- и стереоизображения широко используются для таких задач, как оценка глубины, визуальная одометрия и SLAM (Simultaneous Localization and Mapping). Для обеспечения хорошей проверки и надежной работы полученных программных решений наборы данных должны содержать точную информацию о реальном положении объектов. Однако создание реального набора стереоизображений является затратной задачей, так как требует использования стереокамер и точного оборудования для измерения реального положения объектов (таких как лидары, лазеры, барометры, акселерометры и т.д.). Эти типы оборудования часто являются дорогими и недоступными для рядовых пользователей. Альтернативным подходом является использование синтетических компьютерно-сгенерированных наборов данных, предназначенных для имитации реальных сцен. Синтетические наборы данных часто используются для обучения моделей искусственного интеллекта в случаях, когда реальные данные недоступны, или для тестирования производительности моделей в симулированных средах.

Предлагаемый нами метод предлагает комбинировать сбор реальных данных с методами генерации синтетических данных, чтобы сохранять фотореализм, сохраняя преимущества процедуры генерации синтетических данных.

Ключевые слова: компьютерное зрение, стереонабор данных, синтетический набор данных, симулированная среда, фотореализм.