https://doi.org/10.2352/EI.2024.36.18.3DIA-104

Data Augmentation Based on Depth Information for Neural Radiance Fields

Hamed Razavi Khosroshahi[1], Jaime Sancho[2], Gun Bang[3], Gauthier Lafruit[1], Eduardo Juarez[2], and Mehrdad Teratani[1]

[1]Laboratory of Image Synthesis and Analysis, Universite Libre de Bruxelles; Brussels, Belgium
[2]Research Center on Software Technologies and Multimedia Systems, Universidad Politecnica de Madrid; Madrid, Spain
[3]Electronics, Telecommunication Research Institute; Seoul, South Korea

Abstract

Neural Radiance Fields (NeRF) have attracted particular attention due to their exceptional capability in virtual view generation from a sparse set of input images. However, their scope is constrained by the substantial amount of images required for training. This work introduces a data augmentation methodology to train NeRF using external depth information. The approach entails generating new virtual images at different positions through the utilization of MPEG's reference view synthesizer (RVS) to augment the training image pool for NeRF. Results demonstrate a substantial enhancement in the output quality when employing the generated views in comparison to a scenario where they are omitted.

Introduction

Advancements in neural rendering techniques [1, 2, 3] have increased the capabilities of Neural Radiance Fields (NeRF) [4] in generating realistic scenes from sparse viewpoint data. By utilizing the potential of NeRF, remarkable progress has been made in synthesizing high-quality 3D scenes. However, the effectiveness of NeRF depends largely on the quality and diversity of the input data. To overcome this problem, some authors have considered using depth information as a prior to improve the quality results or to reduce training time [5, 6, 7].

This document follows this lead from the point of view of data augmentation, specifically targeting the enhancement of the quality of rendered images through the proposed approach. Data augmentation techniques can artificially increase the number of training images, allowing the model to be exposed to a wider variety of data without capturing additional real-world images. Our research introduces a novel methodology that integrates depth information and an MPEG (Moving Picture Experts Group) Reference View Synthesizer (RVS) [8] into the NeRF framework. To do this, we synthesize novel views to improve the model quality. This integration is not merely an addition of data; rather, it is a strategic enhancement that leverages depth-aware augmentation to enrich the input dataset. By doing so, we aim to significantly refine the target-rendered images, thereby overcoming some of the limitations posed by sparse data conditions.

Our approach represents an effective step in the evolution of neural rendering, particularly in the context of NeRF-based reconstruction systems. This work presents an in-depth exploration and analysis of our methodology, demonstrating the effectiveness of integrating depth information and the Reference View Synthesizer in advancing NeRF-based rendering. Moreover, the Reference View Synthesizer aids in generating additional viewpoints, further augmenting the dataset and enhancing the robustness of the rendering process. Together, these advancements contribute to achieving higher fidelity and accuracy in the synthesized scenes, marking a significant leap forward in the field of neural rendering. This document not only explains the theory of our approach but also presents empirical evidence to substantiate the efficacy of depth-augmented NeRF in various rendering scenarios.

Methodology

In this section, we discuss a method for augmenting datasets for training with NeRF models where the number of available images is not sufficient. Using this method, virtually synthesized images (known as augmented images) will be added to the training dataset using an augmentation method.

The number of input training images plays a crucial role in the performance of NeRF models. A higher number of diverse training images can lead to a more detailed and accurate representation [9]. Diversity denotes the variety or range of different information present in the dataset. For view synthesis applications, diversity in views would mean having images or observations from a wide range of angles, positions, lighting conditions, and distances, ensuring that the dataset covers a comprehensive representation of the scene. This variety is crucial for training models like NeRF to generalize well across unseen views, leading to more accurate and robust 3D reconstructions and renderings [10, 11]. However, if the training dataset lacks diversity or is limited, the resulting NeRF model may not be able to fully reconstruct the scene.

Synthesizing images using other techniques can augment the training dataset for NeRF, offering some advantages. Here are the steps required for our method (Figure 1):

- In the first step, the original available images must be calibrated, and camera parameters can be extracted using structures-from-motion techniques (SFM) such as COLMAP [12].
- Next, we generate a depth map based on the existing images. In order to generate depth maps, we use MPEG Depth Estimation Reference Software (DERS) [13], a high-quality depth estimation tool (No need for this step if the depth maps are available).
- The MPEG Reference View Synthesizer (RVS), as a tool for synthesizing novel views, is used to synthesize virtual images in the position of missing images based on the original

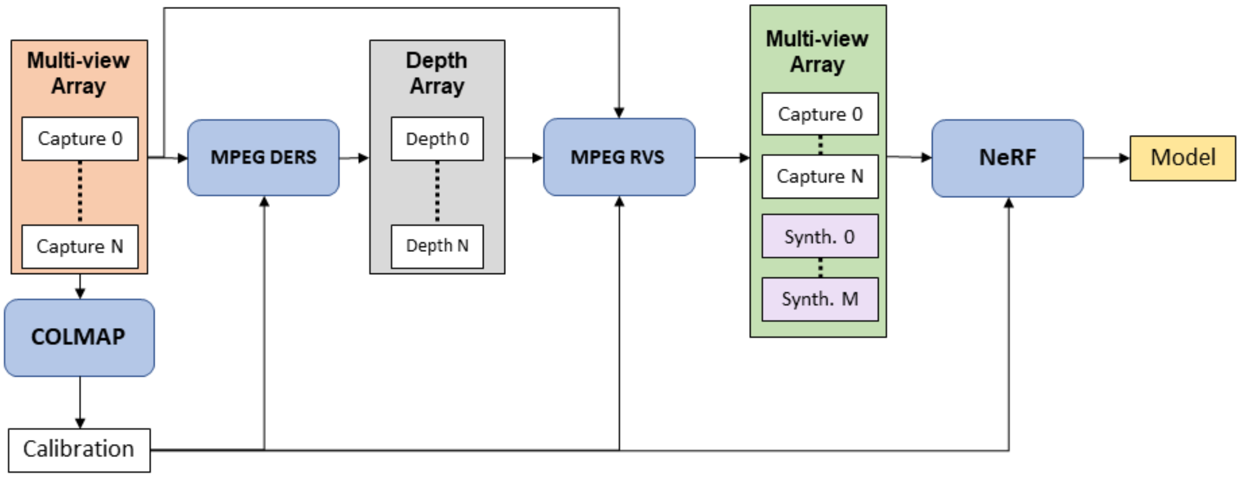

Figure 1. *Proposed pipeline to augment existing data, including calibration, depth map generation, and reference view synthesize*

images and their depth maps and camera parameters.
- The augmented virtual images are added to the training pool of the available images with their corresponding camera parameters.
- New dataset pool is used to train the NeRF model.

Some considerations are important in our pipeline. The virtual images should be positioned in places where they are not out of the original images' field of view (known as their boundaries). In the case that the defined positions are out of the boundary, the augmented images will have occluded areas, which reduces the quality of the trained model. The quality of the depth maps is another important parameter. The quality of the model will be reduced if depth maps are inaccurate and not sharp; otherwise, artifacts will appear in the augmented images.

Experiment condition

We conducted two distinct series of experiments, each using datasets that comprise 5×5 images or a subset of them. The primary objective of these experiments was to strictly evaluate the performance of the Neural Radiance Fields (NeRF) model under varying data availability conditions and adding extra augmented images.

The first series of experiments was dedicated to training the NeRF model exclusively with original images. This process started with the utilization of a complete set of 5×5 original images. Subsequently, in a step-wise manner, we reduced the number of images used for training the models to subsets of 4×4, 3×3, and 2×2 original images. This decremental approach (depicted in Figure 2 - top row), allowed us to systematically analyze the impact of reducing the quantity of training data on the model's performance. This approach was instrumental in understanding the baseline capabilities of the NeRF model when operating with limited data, a common scenario in practical applications.

The second series of experiments was designed to explore the efficacy of incorporating synthesized images alongside the original ones. In this phase, we augmented the datasets by adding synthesized images, which were generated in the previous part, to the positions corresponding to the missing views in the original dataset. This resulted in a comprehensive 5×5 dataset, a mixture of original (denoted as yellow squares) and synthesized (indicated as purple squares) images (illustrated in Figure 2 - bottom row). The purpose of this step was to assess the impact of data augmentation on the NeRF model's ability to render accurate and high-fidelity images. By comparing the outcomes of these two experimental series, we aimed to demonstrate the potential improvements in model performance that can be achieved through the strategic integration of synthesized data, thereby providing valuable insights into the optimization of NeRF models for enhanced neural rendering.

Our experiments were carried out utilizing the nerf-pytorch implementation of NeRF [14]. We conducted each experiment on a 5×5 dataset configuration. To obtain the necessary camera parameters, we employed COLMAP. Depth maps for images were generated using the MPEG's Depth Estimation Reference Software (DERS), taking into account all 5×5 images. In every experiment, the test image was placed in the upper left corner, designated V_0. For the process of augmentation at each stage of the experiment, synthesized images were created using Reference View Synthesizer (RVS). This synthesis was based on the corner images, excluding the test image to maintain the integrity of the experimental conditions.

Datasets

For our experiments, we utilized three datasets. The first is a 5×5 subset extracted from the ULB toys table dataset [15, 16, 17]. This dataset features a baseline distance of 32mm between captured views. The second dataset, also structured as a 5×5 grid, was captured using an Azure Kinect camera mounted on a moving structure at UPM (Universidad Politécnica de Madrid), with a similar baseline of 20mm between the views. The third dataset is a 5×5 subset of the Garage dataset [18], which is a synthetic creation developed using Blender software [19] and provided by ETRI (Electronics and Telecommunications Research Institute). Unlike the first two datasets, the Garage dataset has a larger baseline distance of 60mm between its captured views. These diverse datasets, encompassing both real and synthetic environments, were chosen to test our experiments under varied conditions and baselines.

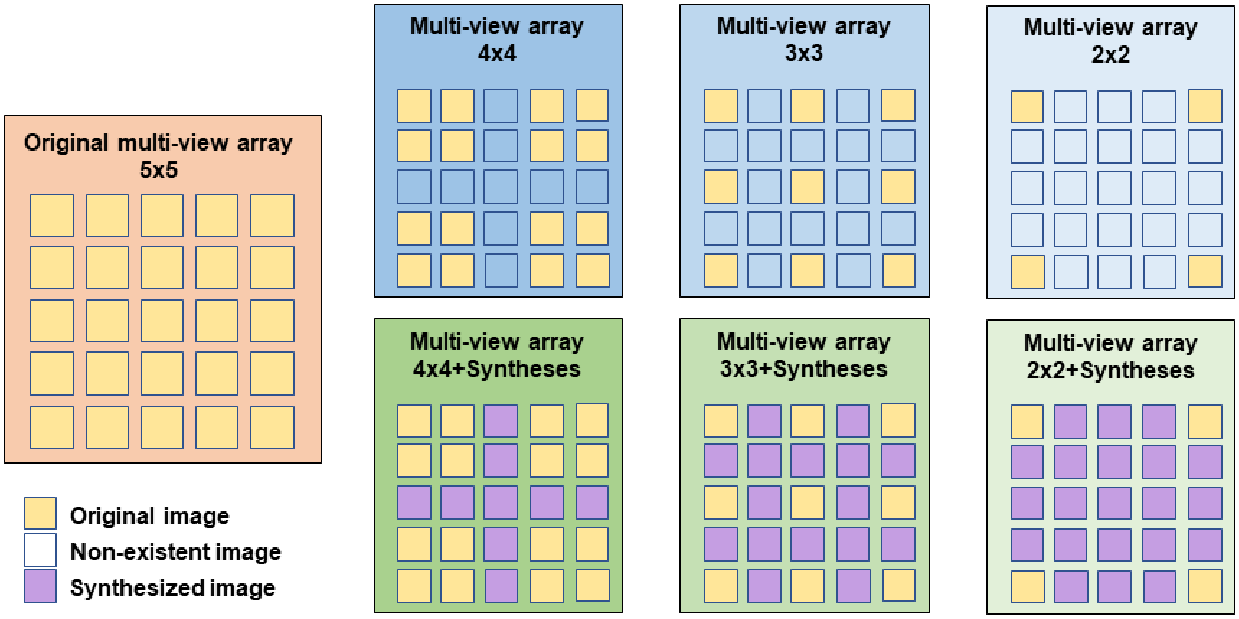
Original multi-view array 5x5
Multi-view array 4x4
Multi-view array 3x3
Multi-view array 2x2
Multi-view array 4x4+Syntheses
Multi-view array 3x3+Syntheses
Multi-view array 2x2+Syntheses
Original image
Non-existent image
Synthesized image

Figure 2. *Configurations for different experiments, training using just original images (top row), training using original and augmented images (bottom row)*

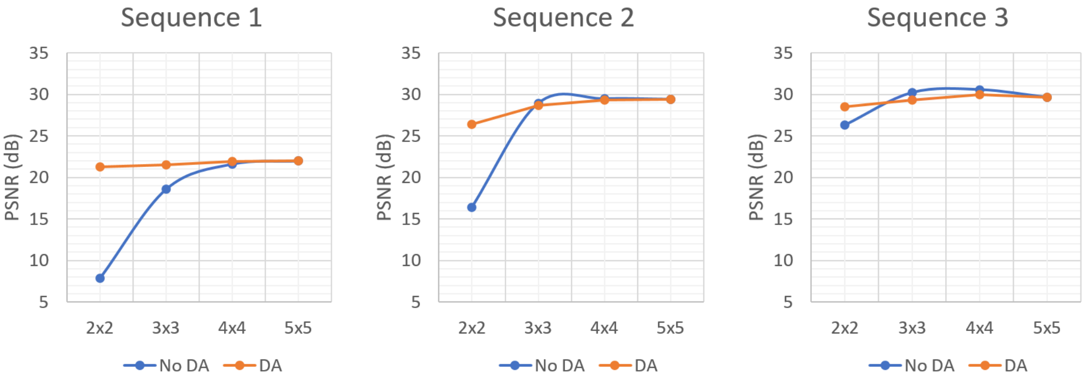

Figure 3. *Objective results for ULB dataset (Sequence 1), UPM dataset (Sequence 2), and ETRI dataset (Sequence 3).*

Results and Discussion

Both objective and subjective findings are discussed here. The objective results, evaluated using the Peak Signal-to-Noise Ratio (PSNR) as a metric, are illustrated in Figure 3. The blue line (without data augmentation) in the figure clearly shows that the objective quality improves as the number of original views increases across all three datasets. More notably, the introduction of synthesized views into the training, indicated by the orange line (with data augmentation), markedly enhances the quality. This improvement is observed across the subsets of the ULB Toys Table and UPM datasets, bringing the results closer to the optimal achievable outcomes. However, the situation differs for the Garage dataset. There, the data augmentation results in a quality increase for the 2 × 2 and augmented subsets, but it does not produce similar improvements for the other experiments. This discrepancy might stem from the lower quality and noisiness of the synthesized images, which in turn could be attributed to the inferior quality of depth maps, particularly given the baseline size of the dataset.

Subjective evaluations also reveal significant insights. When comparing the first set of experiments, which utilized only original images, with the second set, which included both original and synthesized images, we observe noticeable improvements. The augmented images contribute to better synthesized target views, characterized by less blur and more well-defined edges. This is evident in Figure 4, which displays magnified sections of the rendered test image. Consistent with the objective results, we notice quality enhancements in the ULB (Sequence 1) and UPM (Sequence 2) datasets. However, for the Garage dataset (Sequence 3), the subjective results show no substantial differences. This aligns with our objective findings, suggesting that the quality and characteristics of the synthesized images significantly influence the overall performance of the NeRF model under different dataset conditions.

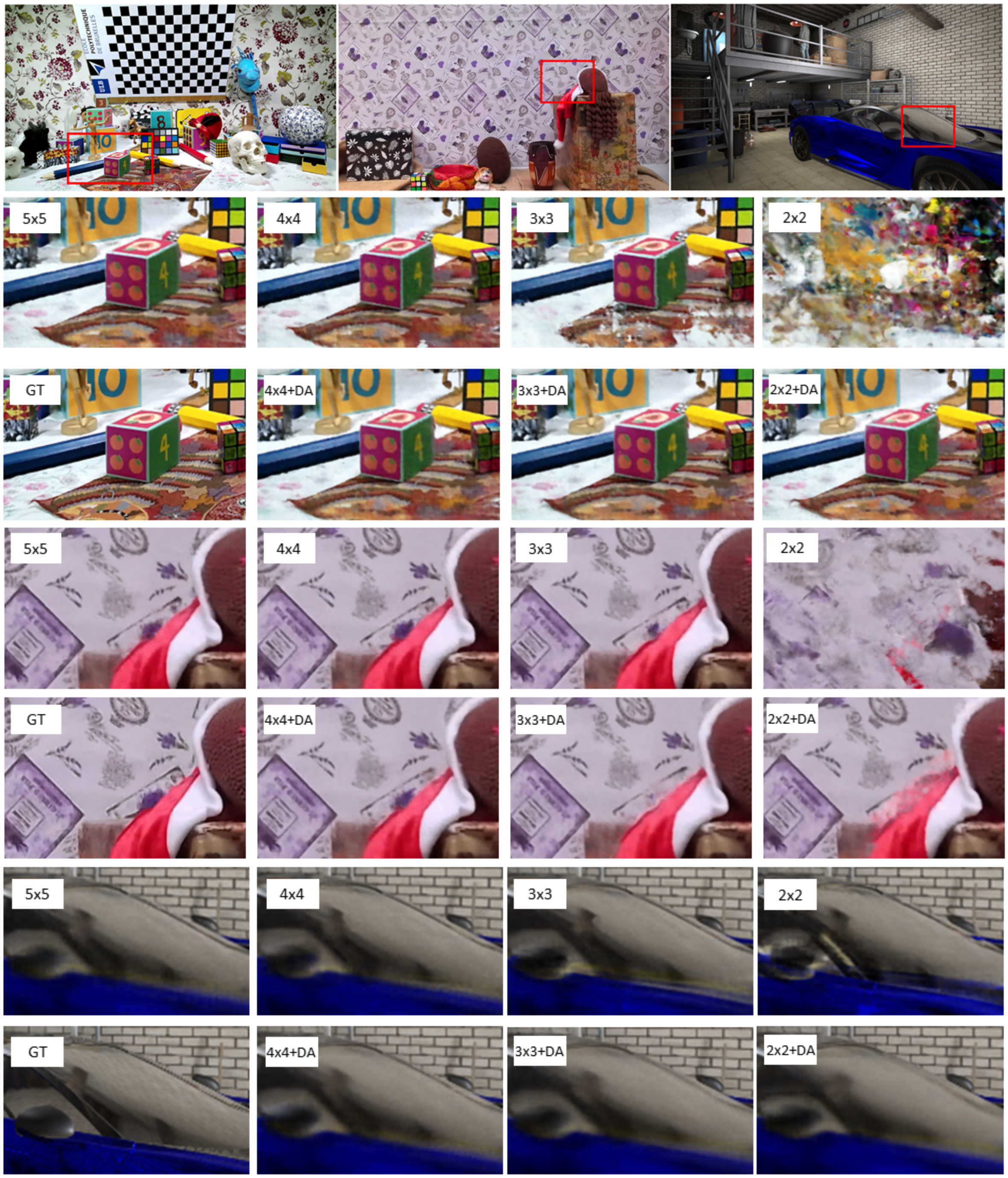

Figure 4. *Magnified part of the rendered image with different configurations for (1) ULB Toys Table (top double rows), (2) UPM dataset (middle double rows), and ETRI Garage dataset (bottom double rows).*

Conclusion

In conclusion, the findings from our objective and subjective analyses strongly suggest that the augmentation of images for training in the NeRF framework, particularly for view synthesis applications, substantially enhances both the model's performance and the quality of the resulting rendered images. This improvement is especially pronounced when there is a limited number of original images available. While this augmentation technique appears to be generally applicable to real-scene datasets, its effectiveness in synthetic datasets warrants further exploration. Apart from the nature of the dataset, another crucial aspect to consider is the baseline distance, which requires additional study. The quality of depth maps emerges as an important factor that varies across different datasets. For a more comprehensive understanding of synthetic datasets, future research should consider utilizing their ground truth depth maps. This approach could provide deeper insights into optimizing the NeRF model for various types of datasets.

Acknowledgement

This work was supported in part by the Service Public de Wallonie Recherche under grant n°2010235 – ARIAC by DIGITALWALLONIA4.AI; in part by the FER 2021 project (n∘ 1060H000066-FAISAN); in part by the Emile DEFAY 2021 project (n∘ 4R00H000236); in part by the FER 2023 project (n∘ 1060H000075), Belgium; Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera media; in part by the project AIMS5.0, which is supported by the Chips Joint Undertaking and its members, including the top-up funding by National Funding Authorities from involved countries under grant agreement no. 101112089; and in part by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00072).

References

[1] Müller, Thomas and Evans, Alex and Schied, Christoph and Keller, Alexander, Instant neural graphics primitives with a multiresolution hash encoding. Association for Computing Machinery (ACM), (2022).

[2] Alex Yu and Ruilong Li and Matthew Tancik and Hao Li and Ren Ng and Angjoo Kanazawa, PlenOctrees for Real-time Rendering of Neural Radiance Fields. International Conference on Computer Vision (ICCV), (2021).

[3] Zhang Chen and Anpei Chen and Guli Zhang and Chengyuan Wang and Yu Ji and Kiriakos N. Kutulakos and Jingyi Yu, A Neural Rendering Framework for Free-Viewpoint Relighting. Conference on Computer Vision and Pattern Recognition (CVPR), (2020).

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi and Ren Ng, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. ECCV, (2020).

[5] Deng, K., Liu, A., Zhu, J. Y., Ramanan, D. (2022). Depth-supervised nerf: Fewer views and faster training for free. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12882-12891).

[6] Johari, M. M., Lepoittevin, Y., Fleuret, F. (2022). Geonerf: Generalizing nerf with geometry priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18365-18375).

[7] Xu, C., Wu, B., Hou, J., Tsai, S., Li, R., Wang, J., ... Tomizuka, M. (2023). Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 23320-23330).

[8] Sarah Fachada, Daniele Bonatto, Mehrdad Teratani, and Gauthier Lafruit, View Synthesis Tool for VR Immersive Video. IntechOpen, (2022).

[9] Gong, Zhiqiang and Zhong, Ping and Hu, Weidong, Diversity in Machine Learning. Institute of Electrical and Electronics Engineers (IEEE), (2019).

[10] M. Goesele, N. Snavely, B. Curless, H. Hoppe and S. M. Seitz, "Multi-View Stereo for Community Photo Collections,"IEEE 11th International Conference on Computer Vision (ICCV), (2007)

[11] Xiaoyan Zhang, Zhengchun Zhou, Ying Han, Hua Meng, Meng Yang, Sutharshan Rajasegarar, Deep learning-based real-time 3D human pose estimation, Engineering Applications of Artificial Intelligence, (2023)

[12] Schonberger, Johannes L., and Frahm, Jan-Michael, Structure-From-Motion Revisited. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016).

[13] Segolene Rogge, Daniele Bonatto, Jaime Sancho, Ruben Salvador, Eduardo Juarez, Adrien Munteanu, Gauthier Lafruit, MPEG-I depth estimation reference software. International Conference on 3D Immersion (IC3D), (2019).

[14] SYen-Chen, Lin, https://github.com/yenchenlin/nerf-pytorch/, (2020).

[15] A. Schenkel, D. Bonatto, S. Fachada, H.-L. Guillaume, et G. Lafruit, « Natural Scenes Datasets for Exploration in 6DOF Navigation », International Conference on 3D Immersion (IC3D), (2018).

[16] Daniele Bonatto, Sarah Fachada, Gauthier Lafruit, "ULB Toys-Table", (2021).

[17] D. Bonatto, A. Schenkel, T. Lenertz, Y. Li, et G. Lafruit, « [MPEG-I Visual] ULB High Density 2D/3D Camera Array data set, version 2 [m41083] », ISO/IEC JTC1/SC29/WG11 MPEG2017/M41083, Torino, Italy, juill. (2017).

[18] https://content.mpeg.expert/data/Explorations/INVR/Garage/

[19] Hess R. Blender Foundations: The Essential Guide to Learning Blender 2.6. Focal Press, (2010).

Author Biography

Hamed Razavi Khosroshahi received his BS in Electronics from the University of Tabriz, (2011) and his M.Sc. in Nano-Electronics from the University of Tabriz (2015). Currently, he is a Ph.D. researcher at Université libre de Bruxelles (ULB). His research interest is in the machine learning methods for light-field cameras, Generative AI, and deep learning for medical images.

Jaime Sancho received his M.Sc. (2018) and Ph.D. (2023, Cum Laude with International Mention) on Systems and Services Engineering for the Information Society in the Universidad Politécnica de Madrid (UPM) Spain. Currently, he is teaching and assistant professor (Ayudante Doctor) in the Telematics and Electronics engineering department (DTE, UPM) and member of the research Center in Software Technologies and Multimedia Systems for Sustainability (CITSEM, UPM), where he develops his research career. His main research interests include biomedical real-time systems, computer vision applied to immersive video, and GPU computing.

Gun Bang received his Ph.D. in Computer Science in 2014 from Korea University in Seoul, Rep. of Korea. He specializes in total varia-

tional optimization for enhancing the visual quality of 3D video synthesis. Since joining ETRI in 2000, he holds the position of Principal Scientist and a Specialized Member of Standard Committee at Electronics and Telecommunications Research Institute, Deajeon, Rep. of Korea. His work currently focuses on standardizing immersive video compression as a co-chair of the MPEG-INVR group.

Gauthier Lafruit is professor Multimedia, with a research focus on Volumetric Reality and Light Field technologies at Université Libre de Bruxelles (ULB). He received his Master and Ph.D. degree in electromechanical engineering with a speciality in electronics from the Vrije Universiteit Brussel (VUB), in 1989 and 1995 respectively. He has worked for 30 years in the domain of visual data analysis and compression, participating to compression standardization committees like CCSDS (space applications), JPEG (still picture coding) and MPEG (moving picture coding). In 2014, he joined the LISA department of ULB, Laboratories of Image Synthesis and Analysis, with a research focus on image synthesis techniques for six degrees of freedom virtual reality using real content, like in the movie Déjà-Vu where highly realistic viewpoints to a scene can be rendered without ever having captured them. This includes depth image-based rendering, immersive video, and point cloud technologies. From 2014 to 2016, he was co-chair of the FTV (Free viewpoint TV) working group in the international MPEG standardization committee. In 2018, LISA's Volumetric Reality research unit (LISA-VR) has actively contributed to part of the MPEG reference software for immersive experiences that will be promoted by Q1-2024 to the MIV standard: "MPEG Immersive Video". Prof. Lafruit teaches 3D graphics and virtual/augmented reality with OpenGL, as well as imaging courses with GPU programming in CUDA.

Eduardo Juárez (PhD, EPFL, 2003) is Associate Professor at UPM. Currently, he is vice-director of the research Center in Software Technologies and Multimedia Systems for Sustainability (CITSEM, UPM). His research activity is mainly focused on (1) hyperspectral imaging for health applications, (2) real-time depth estimation and refinement and (3) heterogeneous high performance computing. He is co-author of one book and author or co-author of more than 100 papers and contributions to technical conferences. He has participated in more than 15 competitive research projects and 20 non-competitive industrial projects.

Mehrdad Teratani received his PhD degree in Information Electronics from Nagoya University, Japan, in 2004. During 2004 to 2019, he was affiliated with: Nagoya University as a researcher for three years; KDDI Research as a researcher for two years; National Institute of Information and Communications Technology (NICT) as a senior researcher for two years; and Nagoya University as an associate professor for nine years, Japan. In January 2020, he joined the École Polytechnique de Bruxelles as a professor at Université Libre de Bruxelles (ULB), Belgium. He has been following and contributing to MPEG standardization since 2009, especially the activities of 3D video processing and immersive video applications. From 2019 to 2021 he was a co-chair of the MPEG Immersive Video (MIV) group, and since 2021 he has been a co-chair of the Lenslet Video Coding (LVC) group in the international MPEG standardization committee. His research interests include 3-D Imaging Systems, Light Field Video Processing and Compression, Intelligent Video System, and Computer Vision. He holds 16 granted patents. He is an IEEE Senior Member, active in IEEE Circuits and Systems and IEEE Signal Processing Societies.