embeddedworld2023
Exhibition&Conference
... it´s a smarter world

3D Computer Vision for the Industrial Metaverse

On the potentials of Neural Radiance Fields

Uwe Hahne
Furtwangen University
Faculty Digital Media
Furtwangen, Germany
Uwe.Hahne@hs-furtwangen.de

Sabine Schleise
Furtwangen University
Faculty Digital Media
Furtwangen, Germany
Sabine.Schleise@hs-furtwangen.de

***Abstract*—The industrial metaverse refers to the use of virtual reality (VR) and augmented reality (AR) technologies in the context of industry and manufacturing. It is envisioned as a shared, immersive digital space where people can interact with and manipulate virtual representations of physical objects and processes. The industrial metaverse has the potential to transform the way products are designed, manufactured, and maintained, enabling new levels of collaboration, automation, and innovation.**

It further includes virtual representations of humans, also known as avatars. These avatars can be used to enable remote collaboration and communication between people in the virtual space. In this way, the industrial metaverse can facilitate virtual meetings, trainings, and other interactive experiences that involve human participants.

Neural Radiance Fields (NeRFs) are a powerful tool for synthesizing photorealistic images of 3D objects, including virtual representations of humans known as avatars. In this talk, we will discuss the potential applications of NeRFs in generating high-fidelity objects and avatars for use in the industrial metaverse.

Keywords—3D Vision, industrial metaverse, immersion, neural rendering

I. INTRODUCTION

The academic research field of computer vision has been extremely vibrant in recent years. 2012 saw a revolution in computer vision. An established competition, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [1], which evaluates algorithms for object detection and image classification, was won by a deep learning solution [2] by a wide margin. While the approach was not new, being based on the older ideas of neural networks, its implementation with the heavy use of GPU computing power led to it replacing almost all current methods and rethinking an entire field [3]. In addition, this new methodology has triggered a previously unimaginable surge. More than 1000 (sic!) papers on machine or deep learning are now published every week.

The success of deep learning can be attributed to several factors. First, the explosion of data in recent years, combined with advances in computing power, has made it possible to train large neural networks on massive datasets. Second, the development of novel architectures and optimization techniques has allowed the construction of more powerful and flexible models. Third, the availability of open-source software frameworks and pre-trained models has lowered the barrier to entry and facilitated the adoption of deep learning by researchers and practitioners in various fields.

As a result, deep learning has achieved breakthroughs in a wide range of fields, including computer vision, natural language processing, speech recognition and gaming. These breakthroughs have been driven by the development of neural methods, which have recently become increasingly influential in various applications.

In this article, we will take a closer look at Neural Radiance Fields (NeRFs) as an exemplary neural method and how this method might impact the field of embedded vision. We will discuss some of the challenges and limitations of this method, and explore future directions and potential applications in the industrial metaverse.

We first introduce the different ideas of the industrial metaverse in order to identify the major challenges for research. We then discuss the potential of neural methods to address these challenges. We focus on NeRFs as an exemplary neural method, as we see it as the one with the highest potential. Finally, we conclude this paper with an outlook on the opportunities for the embedded vision industry.

II. INDUSTRIAL METAVERSE

There are various definitions and conceptions of the metaverse. Luis Bravo Martins and Samantha Wolfe [4] point out that "there is no definition for the metaverse". Simple definitions call it "the next generation of the Internet" or "infinite worlds that connect the physical with the digital", while others focus on the aspect of Extended Reality (XR) and its variations from Virtual Reality (VR) to Mixed Reality (MR) to Augmented Reality (AR). Bitkom, Germany's digital industry association, notes that there are currently several factors that have helped the term and concept of the metaverse to attract so much attention [5]. They see the value of content at the heart of the metaverse, which has a direct impact on embedded vision: Content creation is cumbersome, but good 3D reconstruction is key to simplifying

the creation process. The Fraunhofer Society [6] also sees the capture of user states and the real environment as the key action of metaverse applications, alongside virtual immersion. Therefore, embedded vision technologies will be needed to capture the real world and its state.

While the term metaverse is also controversially discussed in the media [7] and its success is on the edge [8], it seems clear that it will play a decisive role in the scope of Industry 4.0, which was introduced by Klaus Schwab in 2016 [9]. Other authors agree and see the current changes as the sixth long wave of innovation [10]. It is therefore not surprising that the concept of the *industrial metaverse*, which refers to the use of XR technologies in the context of industry and manufacturing, appears on the roadmaps of most industrial players.

Microsoft [11] distinguishes three metaverses: Consumer, Commercial and Industrial Metaverse. While the consumer and commercial metaverses are characterized by immersiveness, the industrial metaverse links the physical and digital worlds. It enables process simulation, data analysis, and machine or deep learning reasoning to improve performance and sustainability. Microsoft wants to take it further by involving people to work between the digital and the real. They see three main aspects: Design, Build and Operate.

NVIDIA calls its vision of an industrial metaverse "Omniverse". They define it as a set of tools based on a new data exchange system invented by Pixar: Universal Scene Description (USD) [12]. They describe it as the HTML of 3D virtual worlds. The system includes a file format specification that allows data to be exchanged in real time, and the tools currently available are aimed at artists, or more precisely teams of artists, who are much more productive when they work together. The use of USD enables real-time collaboration that was not possible before. NVIDIA goes on to say [13]: "We believe that anything that will be built will be visualized. Anything that moves will be autonomous. And anything that's autonomous will be simulated."

Embedded vision plays a critical role in transforming the industrial metaverse by enabling the integration of computer vision algorithms and machine learning models directly into sensors, cameras and other embedded devices. This enables real-time analysis of visual data and the creation of intelligent and interactive environments that can adapt to user needs. For example, smart cameras and sensors can be enabled to detect and track objects and people in real time, and then provide feedback to users in the form of augmented reality overlays or haptic feedback. In addition, embedded vision can drive intelligent control systems that can optimize manufacturing processes, reduce waste and energy consumption, and improve the overall efficiency and safety of industrial operations. The combination of embedded vision, deep learning and the industrial metaverse represents a powerful paradigm shift in the way we design, create and interact with products and systems, and has the potential to drive significant advances in a wide range of fields, from manufacturing to healthcare to transportation.

A major challenge for the industrial metaverse is the ability to seamlessly bridge the gap between the physical and virtual worlds. To create an immersive and interactive digital space that accurately reflects the physical environment, it is necessary to be able to track and analyze real-world data in real time and use this information to generate virtual representations that are accurate, up-to-date and responsive. Embedded vision is a key technology for achieving this synchronization, as it enables computer vision and machine learning algorithms to be integrated directly into sensors and cameras embedded in the physical environment. Using embedded vision, it is possible to capture and analyze visual data in real time and use this information to generate virtual representations that are synchronized with the physical environment. This synchronization is critical for a wide range of applications, from remote collaboration and training to virtual prototyping and testing, and requires a combination of high-performance computing, advanced algorithms and real-time processing capabilities. The development of embedded vision technologies that can quickly and dynamically bridge reality and virtuality will be essential to realize the full potential of the industrial metaverse and unlock the benefits of increased collaboration, automation and innovation in manufacturing and beyond.

As mentioned in the introduction, computer vision research has been dominated by deep learning and thus neural methods in the last decade. Some challenges, such as image classification, have been more or less solved for controlled environments, while others, such as 3D reconstruction, remain active topics. We argue that NeRFs have the potential to be a big step towards solving the 3D reconstruction problem for the scope of the industrial metaverse, which is defined by the need to unite the digital and physical world.

III. Neural Radiance Fields

The NeRF algorithm was published by Mildenhall, Srinivasan and Tancik in 2020 as a method that achieves state-of-the-art results for synthesizing novel views of complex scenes [14]. This is achieved by optimizing an underlying continuous volumetric scene function using a sparse set of input views. Given a set of images capturing the same object from multiple angles along with their corresponding poses, a neural network learns to represent the 3D object so that novel views can be synthesized with the training set of views. The neural network is implemented as a simple multilayer perceptron, which means that there is no special architecture, just nine layers of fully connected neurons. The size of the network is also small, containing only a few thousand parameters. This allows for a very compact representation of the 3D scene. As shown by Hornik, Stinchcombe and White in 1989 [15], multilayer networks are universal function approximators. Thus, given a 3D coordinate input together with the viewing direction, resulting in a 5D input, the network gives a color value together with a density value as an output. These form the parameters of a volume renderer that produces color images. Note that the input values can be taken from the continuous 5D space of positions and viewing directions. In other words, it is possible to create new views of 3D scenes, allowing virtual exploration and interaction with the scene, all based on the compact neural network.

Training the network is straightforward because volume rendering is essentially defined as integration along a line of

sight. This can be implemented as a regular point sampling and results in an easily differentiable sum. The loss function is simply the difference between the resulting output images from

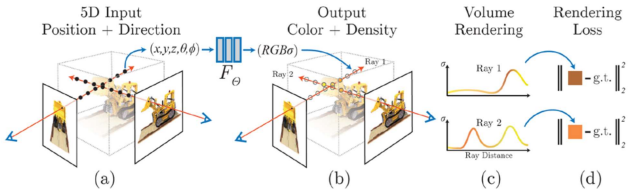

Figure 1: An overview of the NeRF process from [14]. They synthesize images by sampling 5D coordinates (location and viewing direction) along camera rays (a), feeding those locations into an MLP to produce a color and volume density (b), and using volume rendering to composite these values into an image (c). This rendering method is differentiable, so they can use those images in the loss function during training (d).

the mesh and the input images. Figure 1 illustrates the process. The only problem with this approach is that neural networks tend to have a spectral bias towards low frequencies [16]. This is overcome by positional coding, which is similar to a Fourier transform of the coordinates.

The NeRF idea spread very quickly. In less than three years, a huge amount of new ideas and improvements have been published. We refer to the comprehensive review by Gao et al. [17] to get an idea of the impact of NeRFs in the research community. Their review provides a taxonomy for organizing the large number of NeRF-inspired papers and how the method has been improved.

A major drawback of the original solution [14] is performance. Its implementation takes about one to two days to train on a single high-end GPU (NVIDIA A100). The generation of a single image by inference on the resulting model also takes a few minutes.

This drawback has recently been addressed by Müller et al. [18] in a work called *instant-ngp*. The researchers from NVIDIA improved the original idea in three ways. Each improvement sped up the process by a factor of 10, multiplying to a total factor of 1000. Here we describe how this improvement was achieved and what it means for future developments in the field.

First, instead of using TensorFlow, a general-purpose framework for training neural networks, they implemented the whole process in CUDA, the NVIDIA low-level programming language designed to get the most out of its GPUs. Some algorithms from TensorFlow or PyTorch, which are currently the two most popular neural network frameworks, can be easily converted to a hardware-optimized version in CUDA. While CUDA is only available for NVIDIA hardware, other dedicated hardware is becoming available. Google has released its Coral developer board [19], which includes a Tensor Processing Unit (TPU) capable of accelerating computations on neural nets. Smaller companies like Graphcore also offer alternatives [20] to NVIDIA's GPUs, including their own APIs to get the most out of the hardware. It is difficult to predict how much performance gain is possible, but the factor of 10 presented seems reasonable for these other systems as well.

Second, they optimized the sampling scheme by assuming a reasonable distribution. This is possible for real-world objects

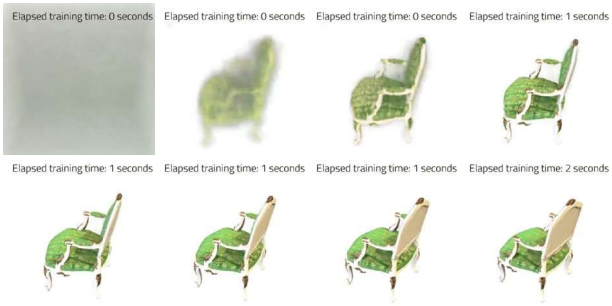

Figure 2: An illustration of the first two seconds of a NeRF training process with *instant-ngp* [18]. After one second the 3D model is clearly recognizable and after less than 2 seconds the model is complete.

and therefore usually applicable in industrial applications where the environment is controlled to some extent.

Third, they improved the positional coding by using a multi-resolution hashing mechanism. The use of an efficient hash function is much faster than a cosine function, so the process could be further improved.

Instant-ngp can produce impressive results even on affordable hardware such as a modern gaming PC, as shown in Figure 2. It also generalizes neural fields to other applications such as Signed Distance Functions (SDF) and gigapixel images, which may also be relevant in the context of the industrial metaverse. Note that NeRF is only one application using neural fields, while other ideas address even more applications, as described in great detail by Xie et al. [21].

As mentioned above, NeRFs have a wide range of potential applications, especially in the context of the industrial metaverse. For example, NeRFs can be used to generate high-fidelity virtual representations of physical objects and processes, which can be used for virtual prototyping and testing, as well as for training and simulation. In addition, NeRFs can be used to create virtual representations of humans [22], as avatars, which can be used for remote collaboration and communication, as well as for training and interactive experiences. By using NeRFs to generate photorealistic images of avatars, it is possible to create immersive and engaging virtual experiences that closely mimic real-world interactions. Overall, NeRFs have the potential to transform the way virtual representations of the physical world are created and used, enabling new levels of collaboration, innovation and automation in the industrial metaverse.

A. What is the connection to embedded vision?

Embedded vision systems can play a key role in supporting the use of Neural Radiance Fields (NeRFs) in the industrial metaverse. One of the challenges in creating photorealistic 3D models and avatars using NeRFs is the need for high quality input data including the camera poses (extrinsics). Current implementations rely on a system like COLMAP [23], [24] that reconstructs the camera extrinsics. These are needed as input for the training process and can be a bottleneck in real-world applications. Industrial cameras usually provide accurate intrinsic data, and the extrinsics can usually be obtained through calibration procedures during initial operation. The required input data can also come from a variety of sources, including images, video, LiDAR scans, and other sensor data. Embedded

vision systems can help capture and process this data to ensure it is of high quality and suitable for use with NeRFs.

In addition, embedded vision systems can help synchronize the virtual and physical worlds in the industrial metaverse. This is essential to enable real-time interaction and collaboration between virtual and physical actors, and to monitor and control physical processes from within the virtual environment. Embedded vision systems can be used to capture and process real-time video and sensor data from the physical world, and to integrate this data in real-time with the virtual environment.

Finally, embedded vision systems can help optimize the use of NeRFs in the industrial metaverse by enabling more efficient and accurate processing of 3D data. By using embedded vision systems to pre-process and filter input data, it is possible to reduce the computational burden on NeRFs and improve their performance and accuracy. This can help enable faster prototyping and testing, as well as more sophisticated and realistic virtual simulations and interactions. Overall, embedded vision systems are a key enabler of the industrial metaverse and the use of advanced technologies such as NeRFs to transform the way we design, manufacture and interact with physical objects and processes.

B. *Further challenges*

The big challenge is interacting with neural representations. They are usually just a large number of weights in a network and are far from human readable. Their output is images or 3D models based on triangles, but it is not possible for a human user to simply edit the network weights to adjust the result. Images can be manipulated with many tools in the spatial, color, or frequency domain, all of which are easily incomprehensible to human users. Triangular 3D models tend to roughly represent the geometry of 3D objects that humans experience in the physical world. Therefore, the mental mapping from model to reality is straightforward and allows for intuitive editing methods. When only the output of the neural network is editable, we rely on existing tools for image editing and 3D modeling, and the remaining benefit is the compact representation of the data. Therefore, there is a strong need for new methods to overcome this challenge.

Xie et al. [21] proposed a classification of editing methods for neural fields into two categories: Coordinate remapping and network parameter editing. While network parameter editing may become increasingly accessible to novice users in the coming decades as neural methods become more ubiquitous, today only data scientists are able to edit network parameters with any prediction of the results of a change.

As supported by Alex Evans in his keynote accompanying the CVPR tutorial [25] that is based on [21], the main and inescapable advantage of neural fields is their compactness, which makes them an ideal transport format. Although processing is challenging, Evans notes that neural fields bring some superpowers that form the basis of new tools and metaphors. They can solve inverse problems, as NeRFs do for photogrammetry, and include optimization-based translation between different scene representations. They also support generalization and generation, as well as the possibility of multi-modal modelling approaches such as using text input, as recently demonstrated in stable diffusion [26] for images.

IV. Conclusion

In this paper, we explored the use of NeRFs for 3D modeling in the context of the industrial metaverse. We described how NeRFs can be used in comparison to traditional methods of 3D modeling to achieve performance improvements in product development, training and education of employees, optimization of production processes, and marketing and sales of products and services. We also examined the efficiency of NeRFs in terms of quality and time, and addresses the challenges and limitations of the technology. In a forthcoming research project, we are working with a sensor manufacturer to design a framework for generating NeRFs in the context of the industrial metaverse. The goal is to integrate neural methods as NeRFs into design processes that support designers in their daily work.

References

[1] O. Russakovsky *et al.*, 'ImageNet Large Scale Visual Recognition Challenge', *Int. J. Comput. Vis. IJCV*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems*, 2012, vol. 25. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[3] Y. Bengio, Y. Lecun, and G. Hinton, 'Deep learning for AI', *Commun. ACM*, vol. 64, no. 7, pp. 58–65, Jun. 2021, doi: 10.1145/3448250.

[4] L. Bravo Martins and S. G. Wolfe, *Metaversed: See beyond the hype*. Standards Information Network, 2022.

[5] S. Klöß, 'A guidebook to the metaverse', *Leitfaden Publikation 2022 | Bitkom e.V.*, 2022. https://www.bitkom.org/EN/Publications/Guidebook-Metaverse (accessed Feb. 21, 2023).

[6] D. Laß, 'Fakt oder Fiktion? Technologien und Use Cases für das (Industrial) Metaverse', Jun. 2022, Accessed: Feb. 22, 2023. [Online]. Available: https://www.iuk.fraunhofer.de/content/dam/iuk/de/Download/Technologien%20und%20Use%20Cases%20f%C3%BCr%20das%20(Industrial)%20Metaverse.pdf

[7] B. Inkster, 'A Teenager's view of the Metaverse: "A term that is made fun of"', *The Time Blawg*, Feb. 19, 2023. https://thetimeblawg.com/2023/02/19/a-teenagers-view-of-the-metaverse/ (accessed Feb. 21, 2023).

[8] '4 tech trends to explore in 2023 | Globant Reports', *Reports*. https://reports.globant.com/en/trends/tech-trends-report-2023/ (accessed Feb. 21, 2023).

[9] K. Schwab, *The Fourth Industrial Revolution*. World Economic Forum, 2016.

[10] D. Reis, 'Learn about long waves to better ride the next one (Part 1)', *Thinkergy*, Nov. 03, 2021. https://www.thinkergy.com/2021/11/04/learn-about-long-waves-to-better-ride-the-next-one-part-1/ (accessed Feb. 21, 2023).

[11] J. Althoff, ‘Microsoft Ignite – The Industrial Metaverse’, *Microsoft Ignite*, Oct. 12, 2022. https://ignite.microsoft.com/en-US/sessions/d12206d9-ee2d-4d99-9dfb-dedd50bf7f0a?source=sessions (accessed Feb. 21, 2023).
[12] ‘Universal Scene Description’, *Universal Scene Description (USD)*, Jan. 23, 2023. https://graphics.pixar.com/usd/release/api/index.html (accessed Feb. 21, 2023).
[13] *The Metaverse Begins: NVIDIA Omniverse and a Future of Shared Worlds*, (Jun. 03, 2021). Accessed: Feb. 21, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=fEm99cwca2k
[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, ‘NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis’. arXiv, Aug. 03, 2020. doi: 10.48550/arXiv.2003.08934.
[15] K. Hornik, M. Stinchcombe, and H. White, ‘Multilayer feedforward networks are universal approximators’, *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989, doi: 10.1016/0893-6080(89)90020-8.
[16] N. Rahaman *et al.*, ‘On the spectral bias of neural networks’, in *Proceedings of the 36th international conference on machine learning*, Jun. 2019, vol. 97, pp. 5301–5310. [Online]. Available: https://proceedings.mlr.press/v97/rahaman19a.html
[17] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li, ‘NeRF: Neural Radiance Field in 3D Vision, A Comprehensive Review’. arXiv, Nov. 08, 2022. doi: 10.48550/arXiv.2210.00379.
[18] T. Müller, A. Evans, C. Schied, and A. Keller, ‘Instant Neural Graphics Primitives with a Multiresolution Hash Encoding’, *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–15, Jul. 2022, doi: 10.1145/3528223.3530127.
[19] ‘Coral Dev Board’, *Coral*, 2020. https://coral.ai/products/dev-board (accessed Feb. 20, 2023).
[20] C. Jin, ‘Graphcore launches C600 PCIe card for AI compute’, 2022. https://www.graphcore.ai/posts/graphcore-launches-c600-pcie-card-for-ai-compute (accessed Feb. 20, 2023).
[21] Y. Xie *et al.*, ‘Neural Fields in Visual Computing and Beyond’. arXiv, Apr. 05, 2022. doi: 10.48550/arXiv.2111.11426.
[22] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman, ‘HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video’. arXiv, Jun. 14, 2022. doi: 10.48550/arXiv.2201.04127.
[23] J. L. Schönberger and J.-M. Frahm, ‘Structure-from-Motion Revisited’, in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[24] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, ‘Pixelwise View Selection for Unstructured Multi-View Stereo’, in *European Conference on Computer Vision (ECCV)*, 2016.
[25] *CVPR 2022 Tutorial on Neural Fields in Computer Vision*, (Sep. 12, 2022). Accessed: Feb. 22, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=PeRRp1cFuH4
[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ‘High-Resolution Image Synthesis with Latent Diffusion Models’. arXiv, Apr. 13, 2022. doi: 10.48550/arXiv.2112.10752.