**3GPP TSG SA WG4#129-e S4-241518**

**Online, 19th - 23rd August 2024**

|  |
| --- |
| *CR-Form-v12.2* |
| **PSEUDO CHANGE REQUEST** |
|  |
|  | **26.956** | **CR** | **pseudo** | **rev** | **-** | **Current version:** | **0.0.3** |  |
|  |
| *For* ***[HE](http://www.3gpp.org/3G_Specs/CRs.htm%22%20%5Cl%20%22_blank)******[LP](http://www.3gpp.org/3G_Specs/CRs.htm%22%20%5Cl%20%22_blank)*** *on using this form: comprehensive instructions can be found at <http://www.3gpp.org/Change-Requests>.* |
|  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Proposed change affects:*** | UICC apps |  | ME | **X** | Radio Access Network |  | Core Network | **X** |

|  |
| --- |
|  |
| ***Title:***  | [FS\_Beyond2D] Representation Format - Neural Radiance Fields (NeRF) |
|  |  |
| ***Source to WG:*** | China Mobile Com. Corporation, Qualcomm Incorporated |
| ***Source to TSG:*** |  |
|  |  |
| ***Work item code:*** | FS\_Beyond2D |  | ***Date:*** | 2024-08-08 |
|  |  |  |  |  |
| ***Category:*** | **B** |  | ***Release:*** | Rel-19 |
|  | *Use one of the following categories:****F*** *(correction)****A*** *(mirror corresponding to a change in an earlier release)****B*** *(addition of feature),* ***C*** *(functional modification of feature)****D*** *(editorial modification)*Detailed explanations of the above categories canbe found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | *Use one of the following releases:Rel-8 (Release 8)Rel-9 (Release 9)Rel-10 (Release 10)Rel-11 (Release 11)…Rel-16 (Release 16)Rel-17 (Release 17)Rel-18 (Release 18)Rel-19 (Release 19)* |
|  |  |
| ***Reason for change:*** | The study item description in SP-240479 addresses the following objectives1. Identify and document beyond 2D formats, that are market-relevant within the next years, generated from established and emerging capturing systems (including cameras for spatial video capturing), contribution, and usable on display technologies (smartphones, VR HMDs, AR glasses, autostereoscopic and multiscopic displays).

During SA4#128, several scenarios were defined, that are considered to address the distribution scenarios and evaluation frameworks. However, some of the scenarios already assume a specific Representation Format that seems to be of less relevance initially.The evaluation framework is important, once Representation formats are defined.  |
|  |  |
| ***Summary of change:*** | This document focuses on Neural Radiance Fields (NeRF) . It is a starting point. |
|  |  |
| ***Consequences if not approved:*** |  |
|  |  |
| ***Clauses affected:*** | 4.3 |
|  |  |
|  | **Y** | **N** |  |  |
| ***Other specs*** |  | **X** |  Other core specifications  | TS/TR ... CR ...  |
| ***affected:*** |  | **X** |  Test specifications | TS/TR ... CR ...  |
| ***(show related CRs)*** |  | **X** |  O&M Specifications | TS/TR ... CR ...  |
|  |  |
| ***Other comments:*** |  |
|  |  |
| ***This CR's revision history:*** |  |

## ===== CHANGE ===== (add to References)

[1] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: representing scenes as neural radiance fields for view synthesis. Commun. ACM 65, 1 (January 2022), 99–106. https://doi.org/10.1145/3503250

[2] Gao, Kyle et al. “NeRF: Neural Radiance Field in 3D Vision, A Comprehensive Review.” (2022).

[5] Li, Sicheng et al. “NeRFCodec: Neural Feature Compression Meets Neural Radiance Fields for Memory-Efficient Scene Representation.” ArXiv abs/2404.02185 (2024): n. pag.

[6] Dong-Ha Kim, Jun Young Jeong, Gwangsoon Lee, and Jae-Gon Kim "Compression method of NeRF model using NNC and VVC", Proc. SPIE 13164, International Workshop on Advanced Imaging Technology (IWAIT) 2024, 131642V (2 May 2024); https://doi.org/10.1117/12.3019533

[7] G. Lafruit, Y. Liao, and G. Bang, “AhG on Implicit Neural Video Representations (INVR),” ISO/IEC JTC1/SC 29/WG04, M60641, Oct. 2022.G. Lafruit, Y. Liao, and G. Bang, “AhG on Implicit Neural Video Representations (INVR),” ISO/IEC JTC1/SC 29/WG04, M60641, Oct. 2022

[8] RABBY, AKM SHAHARIAR AZAD and Chengcui Zhang. “BeyondPixels: A Comprehensive Review of the Evolution of Neural Radiance Fields.” ArXiv abs/2306.03000 (2023): n. pag.

[X1] Daniel Duckworth, Peter Hedman, Christian Reiser, Peter Zhizhin, Jean-François Thibert, Mario Lučić, Richard Szeliski, and Jonathan T. Barron. 2024. SMERF: Streamable Memory Efficient Radiance Fields for Real-Time Large-Scene Exploration. ACM Trans. Graph. 43, 4, Article 63 (July 2024), 13 pages. https://doi.org/10.1145/3658193

[X2] Müller, T., Evans, A., Schied, C., & Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG), 41(4), 1-15..

## ===== CHANGE =====

### 4.3.X Formats under Research

Editor’s Note: Formats in that section will not be part of the evaluation framework of release 19, due to their maturity status, or complexity. However, it is recommended that 3GPP follows the research work on NERF, INVR and GS and awaits stabilization in the industry to commonly agreed formats.

#### 4.3.X.1 Neural Radiance Fields

##### 4.3.X.1.1 Introduction

Neural Radiance Field (NeRF) is a technology at the intersection of Artificial Intelligence (AI) and 3D graphics, and has gained interest based on remarkable progress in computer vision, neural processing units and graphics processing. NeRF was an important research area over the last few years, but recently the interest in NeRF has declined and more attention is given to other formats documented in the remainder of this clause 4.3.X. The documentation reflects the state of the art at the time of writing, but the technology has reached a level of maturity.

##### 4.3.X.1.2 Definition

NeRF is the implicit representation of a 3D scene or object using a fully-connected (non-convolutional) deep network, whose input is a single continuous 5D coordinate (spatial location (x, y, z) and viewing direction (Θ,Φ)) and whose output is the volume density (α) and view-dependent emitted radiance (r, g, b) at that spatial location [1].



**Figure.4.3.X.1.1-1 NeRF representation[1]**

The key idea behind NeRF is to represent the appearance of a scene as a function of 3D position and viewing direction, known as the radiance field. The radiance field describes how light travels through the scene and interacts with its surfaces and can be used to generate images from arbitrary viewpoints [8].

The following is an overview pipeline for NeRF:



**Figure.4.3.X.1.1-2 NeRF pipeline (source: https://docs.nerf.studio/nerfology/methods/nerf.html )**

**Field representation:** For each point in space the NeRF represents a view dependent radiance.

**Positional encoding:** The input coordinates (x,y,z,θ,ϕ) need to be encoded to a higher dimensional space prior to being input into the network.

**Rendering**: NeRF rely on classic volumetric rendering techniques to composite the points into a predicted color.

**Sampling:** NeRF use a hierarchical sampling scheme that first uses a uniform sampler and is followed by a PDF sampler.

##### 4.3.X.1.3 Production and Capturing Systems

Mobile apps such as NeRFCapture (https://github.com/jc211/NeRFCapture), Spectacular AI (https://github.com/SpectacularAI), or Record3D (<https://record3d.app/>) are available to capture NeRFs.

A tutorial for capturing NeRFs is provided here: https://github.com/NVlabs/instant-ngp/blob/master/docs/nerf\_dataset\_tips.md.

The NeRFCapture app allows any iPhone™ or iPad™ to quickly collect or stream posed images to InstantNGP. If your device has a LiDAR, the depth images will be saved/streamed as well. It has two modes: Offline and Online. In Offline mode, the dataset is saved to the device and can be accessed in the Files App in the NeRFCapture folder. Online mode uses CycloneDDS to publish the posed images on the network. A Python script then collects the images and provides them to InstantNGP.

The Spectacular AI SDK and apps can be used to capture data from various devices:

- iPhones (with LiDAR)

- OAK-D cameras

- RealSense D455/D435i

- Azure Kinect DK

The Record3D can create a dataset with an iPhone 12 Pro or newer (based on ARKit), a python code is needed to convert the captured data to NeRF (https://github.com/NVlabs/instant-ngp/blob/master/scripts/record3d2nerf.py)

The state-of-art of NeRF at the time of writing includes:

- SMERF (Streamable Memory Efficient Radiance Fields for Real-Time Large-Scene Exploration) is a view synthesis approach that achieves state-of-the-art accuracy among real-time methods on large scenes with footprints up to 300 m2 at a volumetric resolution of 3.5 mm3 [X1] . It enables fully 6DoF navigation within a web browser, and renders real-time on smartphones and laptops.

- Instant Neural Graphics Primitives (Instant-NGP) using multi-resolution hash encoding to split the processing into multiple chunks and using parallel processing using cuda software to effectively change run time from hours to seconds [X2]. Instant-NGP is a method that uses hash-grid and a shallow MLP to accelerate training and rendering. This method reaches speedups of 1000x and train very fast (~6 min) and renders also fast ~3 FPS.

- [NerfStudio](https://docs.nerf.studio/%22%20%5Ct%20%22https%3A//medium.com/%40heyulei/_blank) (https://docs.nerf.studio/), which is open-source and combines many radiance fields methods, and supports the storage of NeRF data in a structured format, which includes key elements as follows. An example is attached to the zip file:

Camera intrinsics:

{

 "camera\_model": "OPENCV\_FISHEYE", // camera model type [OPENCV, OPENCV\_FISHEYE]

 "fl\_x": 1072.0, // focal length x

 "fl\_y": 1068.0, // focal length y

 "cx": 1504.0, // principal point x

 "cy": 1000.0, // principal point y

 "w": 3008, // image width

 "h": 2000, // image height

 "k1": 0.0312, // first radial distortion parameter, used by [OPENCV, OPENCV\_FISHEYE]

 "k2": 0.0051, // second radial distortion parameter, used by [OPENCV, OPENCV\_FISHEYE]

 "k3": 0.0006, // third radial distortion parameter, used by [OPENCV\_FISHEYE]

 "k4": 0.0001, // fourth radial distortion parameter, used by [OPENCV\_FISHEYE]

 "p1": -6.47e-5, // first tangential distortion parameter, used by [OPENCV]

 "p2": -1.37e-7, // second tangential distortion parameter, used by [OPENCV]

 "frames": // ... per-frame intrinsics and extrinsics parameters

}

Camera extrinsics:

{

 // ...

 "frames": [

 {

 "file\_path": "images/frame\_00001.jpeg",

 "transform\_matrix": [

 // [+X0 +Y0 +Z0 X]

 // [+X1 +Y1 +Z1 Y]

 // [+X2 +Y2 +Z2 Z]

 // [0.0 0.0 0.0 1]

 [1.0, 0.0, 0.0, 0.0],

 [0.0, 1.0, 0.0, 0.0],

 [0.0, 0.0, 1.0, 0.0],

 [0.0, 0.0, 0.0, 1.0]

 ]

 // Additional per-frame info

 }

 ]

}

Depth images:

{

 "frames": [

 {

 // ...

 "depth\_file\_path": "depth/0001.png"

 }

 ]

}

Masks:

{

 "frames": [

 {

 // ...

 "mask\_path": "masks/mask.jpeg"

 }

 ]

}

##### 4.3.X.1.4 Rendering and Display Systems

NeRF heavily relies on the volumetric rendering process to obtain rendered pixels. This rendering function is differentiable, so scene representation can be optimized by minimizing the residual between synthesized and ground truth observed images. The rendering process requires sampling tens to hundreds of points along each ray and inputting them into the neural network to produce the final imaging result. Consequently, rendering a single 1080p image necessitates on the order of 108 neural network forward passes, which often takes several seconds [2].

Display System: VR HMD, mobile devices.

##### 4.3.X.1.5 Supporting Information

- Typical quality criteria for evaluating the format

- Evaluation metrics such as PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), and LPIPS (LearnedPerceptual Image Patch Similarity)

- Training iteration, training time, inference speed.

- Conversion from other formats (lossless, lossy)

- Meshes, point clouds

- Uncompressed data size

 The original NeRF model has 8 fully connected layers, with a layer width of 256, and each pixel is synthesized based on 128 samplings along the ray. The standard NeRF model demands an impractical 5,600 Terabytes cache size.

- Known compression technologies:

Early research on NeRF compression is ongoing. The MPEG established the ad-hoc group called Implicit Neural Visual Representation (INVR) and is currently exploring the potential standardization of 6 Degree of Freedom (6DoF) video compression using NeRF-based technologies [7]. The following methods are applied in current research for NeRF compression and encoding:

- Parameter quantization techniques, transform coding, and entropy coding [5]

- VVC and NNC [6]

- Extensibility of the format

- Mip-NeRF, Point-NeRF, KiloNeRF, Mega-NeRF and etc [8].

##### 4.3.X.1.6 Benefits and Limitations

###### 4.3.X.1.6.1 Benefits

- High-quality 3D representation: NeRF can create photo-realistic 3D reconstructions of complex scenes, including fine surface details, reflections and realistic lighting effects.

- Improved view synthesis capabilities: NeRF can synthesize novel views of a scene or object from a small number of input images, allowing rendering from any viewpoint.

- Flexibility: NeRF can handle non-rigid and dynamic scenes, adapting well to varying spatial conditions and changes over time.

- Unsupervised training: NeRF can learn to reconstruct a scene or object without explicit supervision.

###### 4.3.X.1.6.2 Limitations

- More computationally demanding and slower to render compared to photogrammetry and 3D Gaussian Splatting.

- Not reductionistic: The entire scene is encoded in a single NeRF function, which makes it challenging to segment the scene into parts, edit individual objects within the scene, or combine different NeRF scenes into one.

- Currently, NeRF representation formats do not seem to effectively handle dynamic content within 3D scenes.