MultiNeRF: Multiple Watermark Embedding for Neural Radiance Fields

Yash Kulthe[1], Andrew Gilbert[2], John Collomosse[1,2]

[1]Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey

[2]Adobe Research

Motivation

1. NeRFs are revolutionizing 3D content generation, from immersive VR to product modelling.
2. But this also opens up vulnerabilities- NeRF models are:
   a. Expensive to create
   b. Easy to copy or leak
3. Existing watermark methods for NeRFs:
   a. Embed just a single watermark
   b. Offer low payload capacity (~48bits)

→ We need a robust, high-capacity watermarking framework that works natively in 3D and supports multiple identities/watermarks.

Contributions

1. Introduced a dedicated **watermark grid** in NeRF to separate watermark from appearance content.
2. Enable **multiple conditional watermarks** using FiLM[1] -based modulation.
3. Achieve state-of-the-art performance on both single and multi-watermark tasks with minimal visual artifacts.

Applications

3D Content Attribution

- Track ownership in 3D assets and environments

IP protection & Licensing

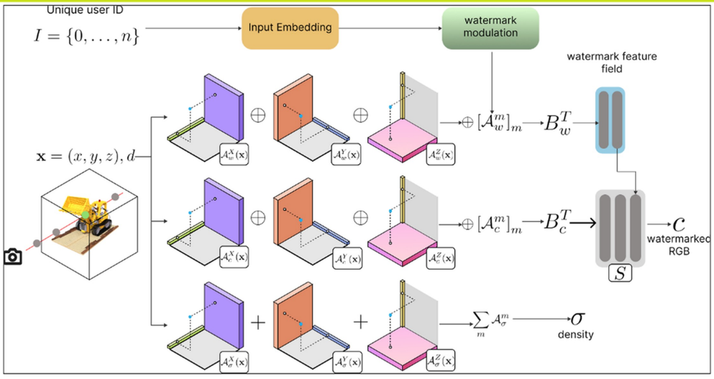
- Multiple IDs for different collaborators
- Or encode long payloads (eg, URLs) via segmented short watermarks

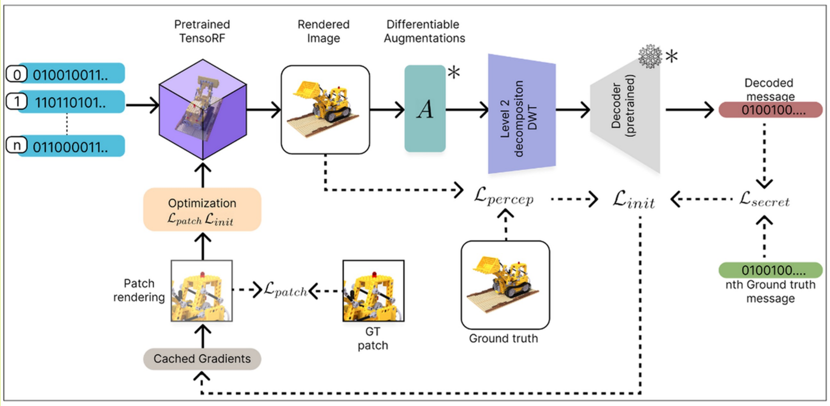
Watermarking module

1. We extend the TensoRF[2] NeRF framework by adding an additional watermark grid along with the appearance and geometry grids.
2. For each watermark ID:
   a. Embedding layer maps watermark ID → embedding
   b. Modulation layer creates scaling & shifting vectors
   c. These vectors modulate the Watermark Grid; which is converter to watermark features using a basis Matrix (B^T_w)
   d. These watermark features are then injected to the decoding MLP of TensoRF.

→ No model retraining needed per watermark!

→ All the watermarks persists across all the views





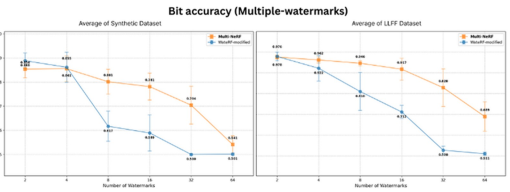
Training Framework

We begin by training a TensoRF NeRF model, using the geometry and appearance grids to initialize those parts of MultiNeRF.

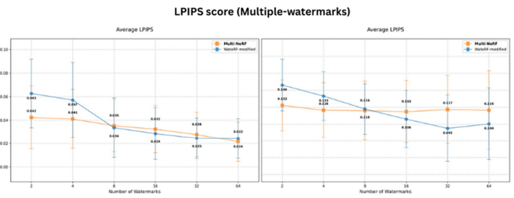
1. Phase 1:
   a. A HiDDeN[3] decoder is trained using full-resolution images
   b. Each image is then rendered from MultiNeRF and is decomposed using 2-level DWT and the LL2 sub-band is used as input to decoder.
   c. Objective: Minimize BCE loss between GT and decoded message. A Watson-VGG perceptual loss ensures visual fidelity
2. Phase 2:
   a. Use of patch-wise rendering to save memory.
   b. Introduce Differentiable Augmentations (for MultiNeRF-noised) to boost the robustness.
   c. Losses: RGB loss + SSIM loss+ Total Variation regularization

Results

Bit accuracy (Multiple-watermarks)



LPIPS score (Multiple-watermarks)



Bit accuracy (Single-watermark)

Method (on SYN)	Avg.	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
WateRF [17]	91.51	98.31	92.19	79.83	96.21	93.16	82.33	95.92	94.10
NeRFProtector [31]	90.81	96.41	89.73	-	93.47	90.12	84.05	90.39	91.54
MultiNeRF (ours)	93.18	98.35	95.14	83.06	96.97	94.86	85.16	96.89	95.03
MultiNeRF-Noised (ours)	89.70	92.60	93.61	78.60	94.36	92.49	83.54	89.72	92.65
Method (on LLFF)	**Avg.**	**Fern**	**Flower**	**Fortress**	**Horns**	**Leaves**	**Orchids**	**Room**	**Trex**
WateRF [17]	99.32	99.75	99.56	99.95	99.92	99.53	96.07	99.89	99.91
NeRFProtector [31]	95.73	94.68	-	99.58	98.77	-	82.23	99.73	99.37
MultiNeRF (ours)	99.23	99.39	99.48	99.82	99.87	99.68	95.92	99.77	99.88
MultiNeRF-Noised (ours)	98.55	99.04	99.05	99.90	99.86	99.28	91.81	99.65	99.81

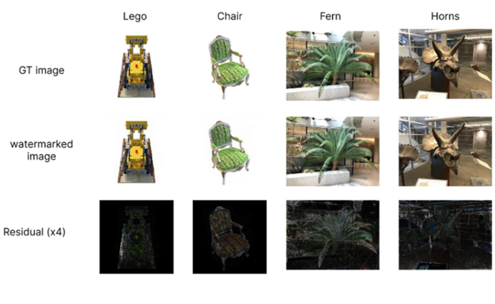
LPIPS score (Single-watermark)

Method (on SYN)	Avg.	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
WateRF	0.04	0.02	0.05	0.02	0.03	0.02	0.04	0.02	0.08
NeRFProtector	0.08	0.04	0.07	-	0.08	0.03	0.08	0.05	0.19
MultiNeRF (ours)	0.04	0.02	0.05	0.02	0.04	0.02	0.04	0.02	0.08
MultiNeRF-Noised (ours)	0.04	0.02	0.06	0.03	0.04	0.02	0.04	0.03	0.09
Method (on LLFF)	Avg.	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	Trex
WateRF	0.10	0.13	0.09	0.07	0.08	0.12	0.17	0.06	0.06
NeRFProtector	0.07	0.10	-	0.07	0.15	-	0.08	0.05	0.06
MultiNeRF	0.09	0.14	0.09	0.06	0.08	0.12	0.18	0.05	0.07
MultiNeRF-Noised (ours)	0.10	0.14	0.09	0.07	0.08	0.12	0.17	0.08	0.06

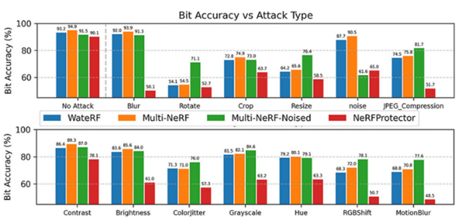
We evaluate MultiNeRF across a range of conditions:

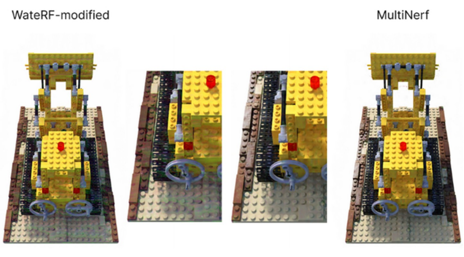
- Single watermarking on SYN and LLFF datasets, showing high bit acc and minimal visual degradation.
- Multi-watermarking with 'n' unique watermarks embedded into a single model
- Robustness tests with common image transformation and regeneration attacks with MultiNeRF-noised performing best.

→ Across all experiments, MultiNeRF consistently delivers higher accuracy, greater robustness and stronger scalability than prior NeRF watermarking methods.



	No Attack	Diffusion [43]	VAE-Cheng [8]	VAE-Bmshj [3]
WateRF	89.98	64.68	56.78	56.96
NeRFProtector	78.59	51.56	63.35	63.00
MN	91.51	67.53	56.26	56.48
MN-noised	**92.52**	**80.48**	**73.94**	**74.72**





References

[1] Perez, Ethan, et al. "Film: Visual reasoning with a general conditioning layer." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.

[2] Chen, Anpei, et al. "Tensorf: Tensorial radiance fields." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

[3] Zhu, Jiren, et al. "Hidden: Hiding data with deep networks." Proceedings of the European conference on computer vision (ECCV). 2018.

Funding Acknowledgement

This research was supported in part by DECaDE: under EPSRC Grant EP/T022485/1