Appendix

1. Ablation Studies on Tensor Factorization Strategies

	# Comp	PSNR ↑	SSIM ↑	# Param.(M) ↓
Multi(0.6, 0.3, 0.15)	24	**33.24**	**0.963**	**7.07**
Single(0.3)	96	33.02	0.963	9.15
VM-Cloud (0.3)	6	32.59	0.959	11.36
VM-Cloud (0.3)	12	32.99	0.962	21.64

Table 1: (a) Comparisons on our method pairing with different factorization strategies, e.g., CP decomposition and vector-matrix (VM) decomposition (row 2 vs 3,4). The local tensors' edge lengths are all set as 0.3. (b) We also compare a single-scale model with a multi-scale model (row 1 vs 2). We evaluate these settings on the NeRF Synthetic dataset [8] and evaluate them with both rendering quality and model capacity (#Param. denotes the number of parameters).

Other than CP decomposition, TensoRF [2] also proposes vector-matrix (VM) decomposition, which factorizes a 3D tensor as the summation of vector-matrix bases. Each basis is the outer product of a matrix along a plane, e.g., the XY plane, and a vector along an orthogonal direction, e.g., the Z axis. For comparison, we also explore to replace our tri-vector representation with the vector-matrix representation for each local tensor. Tab. 1 shows that the single-scale tri-vector cloud can outperform the vector-matrix cloud representation with less model capacity.

It is not a surprise that our tri-vector cloud representation achieves more compactness. It applies more compression by factorizing each component of a 3D tensor, with a space complexity of $O(IJK)$, into three vectors, with a space complexity of $O(I + J + K)$. On the other hand, vector-matrix cloud representation factorizes it into three vectors and three matrices, which have a space complexity of $O(IJ + JK + IK)$. Even if we reduce the number of components, the vector-matrix clouds still require more space than our tri-vector representations.

In terms of quality, since our method exploits the spatial sparsity of natural scenes, we only need to factorize each local space independently instead of the entire scene together. The more compact tri-vector representation can benefit from the appearance coherence in local space and result in better performance. In TensoRF [2], since the entire space is factorized all at once, the radiance information is, in general, less coherent across locations and the CP decomposition will lead to a shortage of rank.

2. Ablation Studies on Multi-scale Models

In Tab.1, we also compare our multi-scale tri-vector radiance fields with the single-scale strategy. In our default model, we have three scales, composed of tensors with lengths 0.15, 0.3, and 0.6, respectively. Similar to the findings in iNGP [9], our multi-scale models provide more smoothness and lead to a better rendering quality than their single-scale counterparts. The multi-scale model with 24 components (row 1) can already outperform the single-scale model (row 2), which has more parameters.

3. Ablation Studies on the Number of Tensor Components

We conduct experiments on the NeRF Synthetic dataset [8] to show the relationship between rendering performance and the number of tensor components. In Tab.2, we compare our multi-scale models with 12, 24, 48, and 96 appearance components, respectively. In general, more tensor components will lead to better performance. We also observe that the benefit of adding more components becomes marginal when the number reaches 48. We speculate that it is harder to learn high-frequency details even though the model's capacity can hold high-rank information. Improvement in this aspect can be a promising future direction.

4. Ablation Studies on Initial Geometry

We emphasize that our superior quality stems from our novel scene representation rather than the initial geometry. The initial geometry is simply acquired from a low-res RGBA volume reconstruction, which is coarse and only used to roughly prune empty space.

We show in Fig. 1 that our approach performs robustly with various choices of these geometry structures and consistently achieves high PSNRs, even with a much worse early-stopped RGBA reconstruction. This showcases the key to our superior quality is our Strivec model itself.
In particular, the self-bootstrap geometry is generated

	PSNR↑	SSIM↑	LPIPS$_{Vgg}$ ↓	LPIPS$_{Alex}$ ↓	# Param.(M)↓
Ours-12	32.94	0.961	0.049	0.028	**4.87**
Ours-24	33.24	0.963	0.046	0.026	7.07
Ours-48	33.55	0.965	0.044	0.025	13.52
Ours-96	**33.59**	**0.965**	**0.043**	**0.024**	21.01

Table 2: Ablation study on the number of tensor components. We use the same setting as our default model but only change the number of components in each variant. These variants are evaluated on the NeRF Synthetic dataset [8].

purely from our own model with 8 coarse tri-vectors without existing modules in previous work. Moreover, we can also further prune unoccupied tensors during training but we find this leads to similar quality (0.03db difference) and unnecessary extra (+22%) training time. We instead choose to use one single initial geometry to prune empty space in implementation for its simplicity and efficiency.

Figure 1: Our quality with initial geometry by different methods.

5. Speed v.s. Performance

Though speed is not our focus, here, if we reduce the number of scales from 3 to 2 and TopK from 4 to 2 (i.e., Multi(0.6, 0.3) with TopK=2), and Strivec becomes faster than CP and close to VM, while still having competitive quality (see Ours-48(fast) in Tab.3). The fewer ranks of our tensor and the less number of TopK to be find for each sample point along a ray lead to less computation, and thus, acceleration. To conclude, Strivec is capable to improve quality, training time and compactness all together with proper hyper-parameters.

	Train(s)↓	Inference(s/it)↓	#Params.(M)↓	PSNR↑
TensoRF-CP	1914	2.01	0.98	31.56
TensoRF-VM	915	1.60	17.95	33.14
Ours-48(fast)	959	1.67	6.20	33.09

Table 3: Comparison on NeRF Synthetic dataset [8]. We compare the average training time (s), inference time (s/it), the number of parameters (M) and PSNR.

6. Per-scene Breakdown Results of the NeRF Synthetic Dataset

We show the per-scene detailed quantitative results of the comparisons on the NeRF Synthetic dataset [8] in Tab. 6 and qualitative comparisons in our video. With compact model capacity, our method outperforms state-of-the-art methods [8, 9, 11, 2] and achieves the best PSNRs, and LPIPSs in most of the scenes. We report two versions of

	garden	room	Model Size(avg)
DVGO	24.32	28.35	5.1GB
Ours-48	24.13	28.11	12.6MB

Table 4: Results on the Mip-NeRF 360 dataset.

iNGP [9]. Specifically, iNGP-dark$_{100k}$ is reported in the original paper. According to issue #745 in iNGP's official repo, the method uses a random color background in training and dark background in testing. The number of iterations, 100k, is referenced to its initial code base release. We also refer to the results reported in [3] as iNGP-white$_{30k}$, since the authors use a white background in both training and testing for 30k iterations, which has the same setting as ours and many other compared methods. Please refer to issue #745 and #1266 in iNGP's official repo for more details.

7. The Tanks and Temples Dataset

We show the qualitative comparison between our Strivec and TensoRF-VM [2] on the Tanks and Temples dataset [5] in Fig.2. Similar to the procedures on the NeRF Synthetic dataset, we build the coarse scene geometry within 30 seconds to place our local tensors. The quantitative results are reported in Tab.5.

8. Mip-NeRF360 Dataset

We evaluate our method on two scenes (one indoor scene and one outdoor scene) of Mip-NeRF360 dataset [1]. Note that we only use the scene warping scheme the same as DVGO [10] and Mip-NeRF360 [1] and keeping other components (i.e., positional encoding, point sampling, etc.) the same as TensoRF [2]. The qualitative and quantitative results are shown in Fig. and Tab. , respectively. Here, we use only two scales in implementation to show our compactness and scalability.

	Tanks & Tamples					
	Ignatius	Truck	Barn	Caterpillar	Family	Mean
			PSNR ↑			
NV [7]	26.54	21.71	20.82	20.71	28.72	23.70
NeRF [8]	25.43	25.36	24.05	23.75	30.29	25.78
NSVF [6]	27.91	26.92	27.16	26.44	33.58	28.40
TensoRF-CP[2]	27.86	26.25	26.74	24.73	32.39	27.59
TensoRF-VM[2]	28.34	27.14	27.22	26.19	**33.92**	28.56
Ours-48	**28.39**	**27.32**	**28.09**	**26.58**	33.13	**28.70**
			SSIM ↑			
NV [7]	0.992	0.793	0.721	0.819	0.916	0.848
NeRF [8]	0.920	0.860	0.750	0.860	0.932	0.864
NSVF [6]	0.930	0.895	0.823	0.900	0.954	0.900
TensoRF-CP[2]	0.934	0.885	0.839	0.879	0.948	0.897
TensoRF-VM[2]	0.948	0.914	0.864	0.912	**0.965**	0.920
Ours-48	**0.948**	**0.915**	**0.884**	**0.917**	0.957	**0.924**
			LPIPS$_{Alex}$ ↓			
NV [7]	0.117	0.312	0.479	0.280	0.111	0.260
NeRF [8]	0.111	0.192	0.395	0.196	0.098	0.198
NSVF [6]	0.106	0.148	0.307	0.141	0.063	0.153
TensoRF-CP[2]	0.089	0.154	0.237	0.176	0.063	0.144
TensoRF-VM[2]	**0.081**	0.129	0.217	0.139	**0.057**	0.125
Ours-48	0.083	**0.123**	**0.167**	**0.125**	0.065	**0.113**
			LPIPS$_{Vgg}$ ↓			
TensoRF-CP[2]	0.106	0.202	0.283	0.227	0.088	0.181
TensoRF-VM[2]	**0.078**	**0.145**	0.252	0.159	**0.064**	0.140
Ours-48	0.083	0.150	**0.216**	**0.154**	0.078	**0.136**

Table 5: Quantity comparison on five scenes in the Tanks and Temples dataset [5] selected in NSVF [6]. NV, NeRF, and NSVF have not reported their LPIPS$_{Vgg}$

	NeRF Synthetic							
	Chair	Drums	Lego	Mic	Materials	Ship	Hotdog	Ficus
				PSNR↑				
NeRF [8]	33.00	25.01	32.54	32.91	29.62	28.65	36.18	30.13
NSVF [6]	33.19	25.18	32.54	34.27	**32.68**	27.93	37.14	31.23
Point-NeRF$_{20k}$ [11]	32.50	25.03	32.40	32.31	28.11	28.13	34.53	32.67
Point-NeRF$_{200k}$ [11]	35.40	26.06	35.04	35.95	29.61	30.97	37.30	36.13
iNGP-dark$_{100k}$ [9]	35.00	26.02	36.39	36.22	29.78	31.10	37.40	33.51
iNGP-white$_{30k}$ [9, 4]	35.42	24.24	34.82	35.98	28.99	30.72	37.45	32.09
TensoRF-CP [2]-384$_{30k}$	33.60	25.17	34.05	33.77	30.10	28.84	36.24	30.72
TensoRF-VM [2]-192$_{30k}$	35.76	26.01	36.46	34.61	30.12	30.77	37.41	33.99
Ours-12$_{30k}$	35.21	25.96	35.60	35.29	29.54	30.64	37.03	34.21
Ours-24$_{30k}$	35.60	26.16	36.05	35.81	29.79	30.89	37.24	34.37
Ours-48$_{30k}$	**35.88**	**26.20**	**36.52**	**36.65**	29.90	**31.13**	**37.63**	34.47
				SSIM↑				
NeRF	0.967	0.925	0.961	0.980	0.949	0.856	0.974	0.964
NSVF	0.968	0.931	0.960	0.987	**0.973**	0.854	0.980	0.973
Point-NeRF$_{20k}$	0.981	0.944	0.980	0.986	0.959	0.916	0.983	0.986
Point-NeRF$_{200k}$	**0.991**	**0.954**	**0.988**	**0.994**	0.971	**0.942**	**0.991**	**0.993**
iNGP-white$_{30k}$	0.985	0.924	0.979	0.990	0.945	0.892	0.982	0.977
TensoRF-CP-384$_{30k}$	0.973	0.921	0.971	0.983	0.950	0.857	0.975	0.965
TensoRF-VM-192$_{30k}$	0.985	0.937	0.983	0.988	0.952	0.895	0.982	0.982
Ours-12$_{30k}$	0.983	0.937	0.980	0.989	0.948	0.888	0.981	0.983
Ours-24$_{30k}$	0.984	0.940	0.982	0.990	0.952	0.893	0.982	0.984
Ours-48$_{30k}$	0.985	0.940	0.984	0.992	0.953	0.899	0.983	0.985
				LPIPS$_{Vgg}$ ↓				
NeRF	0.046	0.091	0.050	0.028	0.063	0.206	0.121	0.044
Point-NeRF$_{20k}$	0.051	0.103	0.054	0.039	0.102	0.181	0.074	0.043
Point-NeRF$_{200k}$	0.023	0.078	0.024	0.014	0.072	**0.124**	0.037	0.022
iNGP-white$_{30k}$	0.022	0.092	0.025	0.017	0.069	0.137	0.037	0.026
TensoRF-CP-384$_{30k}$	0.044	0.114	0.038	0.035	0.068	0.196	0.052	0.058
TensoRF-VM-192$_{30k}$	0.022	0.073	0.018	0.015	0.058	0.138	0.032	0.022
Ours-12$_{30k}$	0.025	0.070	0.022	0.015	0.062	0.145	0.033	0.022
Ours-24$_{30k}$	0.022	0.067	0.020	0.013	0.058	0.141	0.031	0.021
Ours-48$_{30k}$	**0.021**	**0.064**	**0.017**	**0.011**	**0.056**	0.138	**0.029**	**0.018**
				LPIPS$_{Alex}$ ↓				
NSVF	0.043	0.069	0.029	0.010	**0.021**	0.162	0.025	0.017
Point-NeRF$_{20k}$	0.027	0.057	0.022	0.024	0.076	0.127	0.044	0.022
Point-NeRF$_{200k}$	0.010	0.055	0.011	0.007	0.041	**0.070**	0.016	**0.009**
iNGP-white$_{30k}$	0.022	0.093	0.025	0.017	0.069	0.140	0.037	0.026
TensoRF-CP-384$_{30k}$	0.022	0.069	0.014	0.018	0.031	0.130	0.024	0.024
TensoRF-VM-192$_{30k}$	0.010	0.051	0.007	0.009	0.026	0.085	0.013	0.012
Ours-12$_{30k}$	0.011	0.051	0.009	0.007	0.027	0.092	0.015	0.013
Ours-24$_{30k}$	0.010	0.049	0.008	0.006	0.024	0.087	0.014	0.012
Ours-48$_{30k}$	**0.009**	**0.048**	**0.007**	**0.005**	0.023	0.086	**0.012**	0.011

Table 6: Detailed breakdown of quantitative metrics on individual scenes in the NeRF Synthetic [8] for our method and baselines. All scores are averaged over the testing images. The subscripts are the number of iterations of the models. NeRF only [8] reports the LPIPS$_{Vgg}$ [12] while NSVF only reports LPIPS$_{Alex}$.

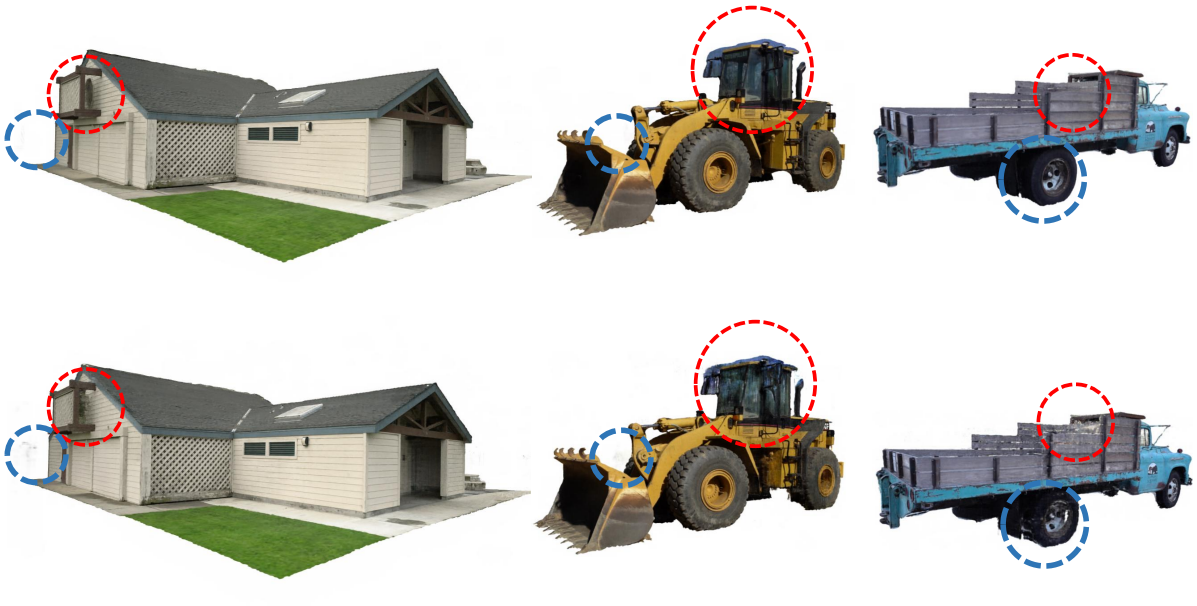

Figure 2: Qualitative comparison on the Tanks and Temples dataset. Top: ours. Bottom: TensoRF-VM.

Figure 3: Qualitative results on Mip-NeRF360 dataset.

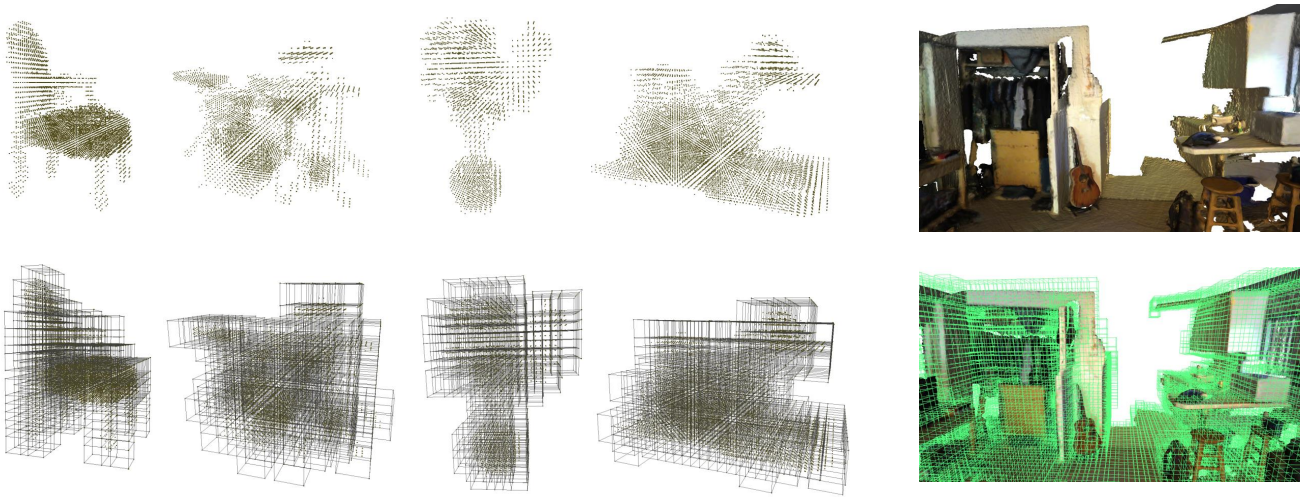

Figure 4: Visualization of local tensors (single scale) on initial geometry.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. 2

[2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, pages 333–350. Springer, 2022. 1, 2, 3, 4

[3] Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger. Factor fields: A unified framework for neural fields and beyond, 2023. 2

[4] Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger. Factor fields: A unified framework for neural fields and beyond. *arXiv preprint arXiv:2302.01226*, 2023. 4

[5] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 2, 3

[6] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *arXiv preprint arXiv:2007.11571*, 2020. 3, 4

[7] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 3

[8] 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. In *European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 1, 2, 3, 4

[9] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022. 1, 2, 4

[10] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5459–5469, 2022. 2

[11] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5438–5448, 2022. 2, 4

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4