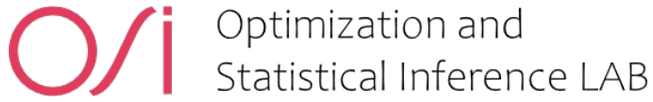
Synergistic Integration of Coordinate Network and Tensorial Feature for Improving Neural Radiance Fields from Sparse Inputs

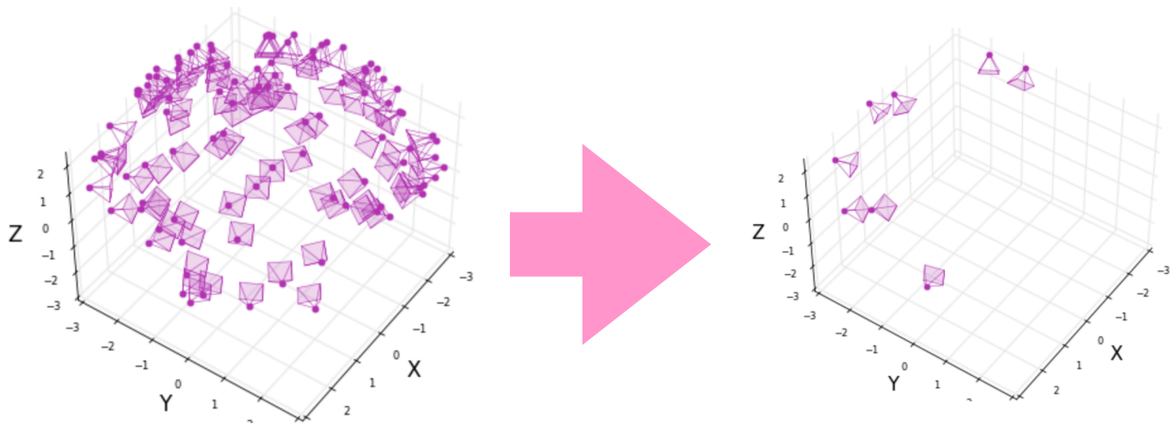
Presented @ ICML2024



NeRFs under Sparse Inputs

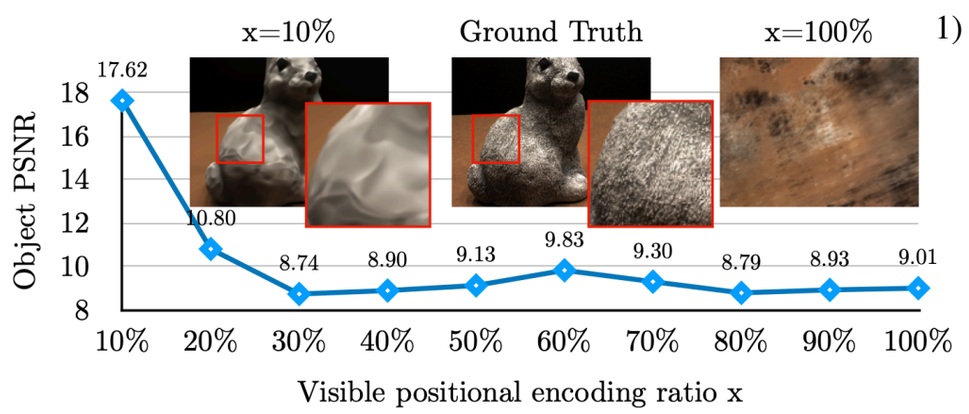
Optimization and Statistical Inference LAB

- The object-centric NeRFs datasets assumes a training set with 100 views. However, this concept doesn't align closely with practical usage.
- In real-world scenarios, assuming 100 views is quite ambitious.
- If we train NeRFs with only 8 views, learning both low/high-frequencies is challenging

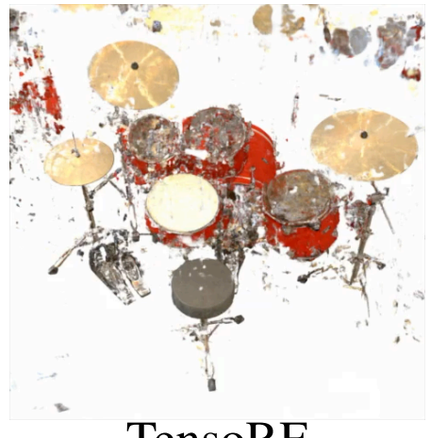


Views : 100 Views : 8

1)



2)



TensoRF

1) Yang, Jiawei, Marco Pavone, and Yue Wang. "Freenerf: Improving few-shot neural rendering with free frequency regularization." *CVPR2023*
2) Chen, Anpei, et al. "Tensorf: Tensorial radiance fields." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

Literature Survey

- Summarize prior research on sparse inputs by manipulating feature fields.

1. **Simplified NeRF**[3]**/HALO**[4]**/FreeNeRF**[1] : Manipulating frequencies in sinusoidal encoding

2. **DietNeRF**[5]**/VisionNeRF**[6] : Fulfill features with pre-trained networks (CLIP or ViT)

3. **DVGO**[7]**/VSOS**[8] : Voxel-grid based NeRF and its emphasize on low resolution feature fields

4. **TensoRF**[2]**/K-Planes**[9]**/HexPlane**[10] : Multi-plane feature grid and its TV regularization

3) Jain, Ajay, Matthew Tancik, and Pieter Abbeel. "Putting nerf on a diet: Semantically consistent few-shot view synthesis." *ICCV2021*
4) Song, Liangchen, et al. "Harnessing low-frequency neural fields for few-shot view synthesis." *Arxiv2023*
5) Lin, Kai-En, et al. "Vision transformer for nerf-based view synthesis from a single input image." *WACV2023*
7) Sun, Cheng, Min Sun, and Hwann-Tzong Chen. "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction." *CVPR2023*
8) Sun, Jiakai, et al. "VGOS: Voxel grid optimization for view synthesis from sparse inputs." *IJCAI2023*
9) Fridovich-Keil, Sara, et al. "K-planes: Explicit radiance fields in space, time, and appearance." *CVPR2023*
10) Cao, Ang, and Justin Johnson. "Hexplane: A fast representation for dynamic scenes." *CVPR2023*

Key Observation: Free-NeRF (1)

Optimization and Statistical Inference LAB

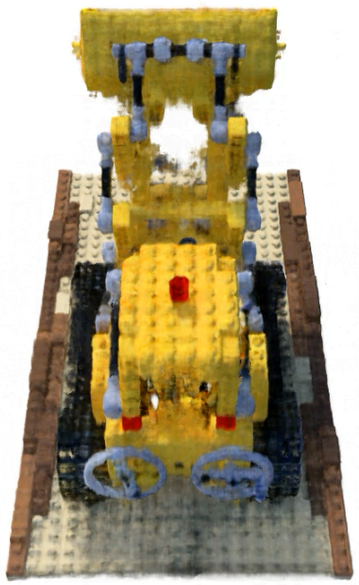
- FreeNeRF, equipped with progressive sinusoidal encoding, is prone to under-fitting issues
  - While it allows robust training and quality, the synthesized images appear blurry.
- It struggles to learning high-frequency details and longer training times
  - Low-frequency sinusoidal encodings learn quickly, whereas high-frequency sinusoidal features struggle to learn as rapidly, requiring at least 5 hours.

Progressive sinusoidal encoding

$$\gamma_L(\mathbf{x}) = \left[\sin(\mathbf{x}), \cos(\mathbf{x}), ..., \sin(2^{L-1}\mathbf{x}), \cos(2^{L-1}\mathbf{x})\right]$$

$$\gamma'_L(t, T; \mathbf{x}) = \gamma_L(\mathbf{x}) \odot \boldsymbol{\alpha}(t, T, L),$$

$$\boldsymbol{\alpha}_i(t, T, L) = \begin{cases} 1 & \text{if } i \leq \frac{t \cdot L}{T} + 3 \\ \frac{t \cdot L}{T} - \lfloor \frac{t \cdot L}{T} \rfloor & \text{if } \frac{t \cdot L}{T} + 3 < i \leq \frac{t \cdot L}{T} + 6 \\ 0 & \text{if } i > \frac{t \cdot L}{T} + 6 \end{cases}$$

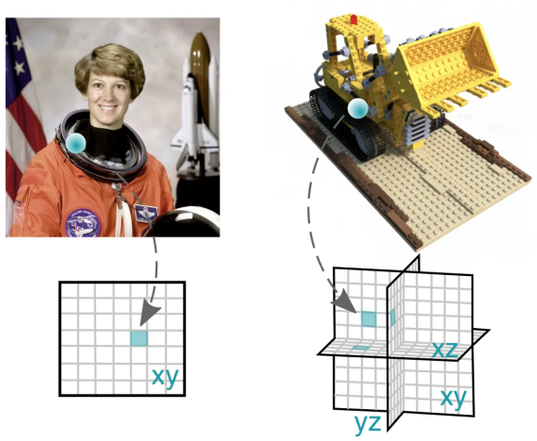


1) Yang, Jiawei, Marco Pavone, and Yue Wang. "Freenerf: Improving few-shot neural rendering with free frequency regularization." *CVPR2023*

Key Observation: K-Planes (2)

Optimization and Statistical Inference LAB

- K-Planes equipped with TV denoising loss is effective to remove floating artifacts.
- Due to the explicit method, this achieves to take 30min. and rendering speed is fast.
- However, it exhibits color distortion that appears authentic but is not present in the training dataset when TV penalty is .
  - Explicit Representation struggles to learn low-frequency details due to their locality



emory: $\mathcal{O}(N^2)$ $\mathcal{O}(3N^2)$

$$\mathcal{L}_{TV}(\mathbf{P}) = \frac{1}{|C|n^2} \sum_{c,i,j} \left(\|\mathbf{P}_c^{i,j} - \mathbf{P}_c^{i-1,j}\|_2^2 + \|\mathbf{P}_c^{i,j} - \mathbf{P}_c^{i,j-1}\|_2^2 \right),$$

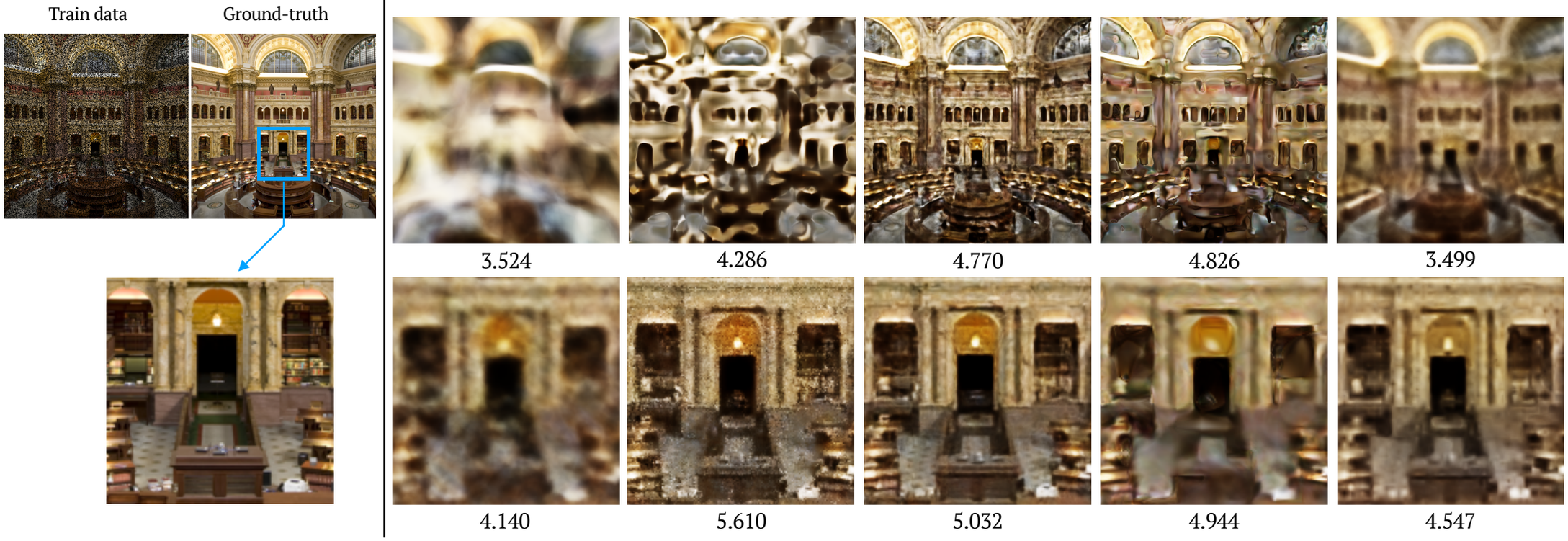
Color Distortion:

9) Fridovich-Keil, Sara, et al. "K-planes: Explicit radiance fields in space, time, and appearance." *CVPR2023*

Intuition

Optimization and Statistical Inference LAB

- The image regression task proves that multi-resolutional tensorial features do not play a role to capture low-frequency details in intended manner.
  - Low-resolution features also tends to learn high-frequency details.
  - iNGP exhibits the widest spectrum among baselines, the image with low resolution features does not adequately capture global reasoning.

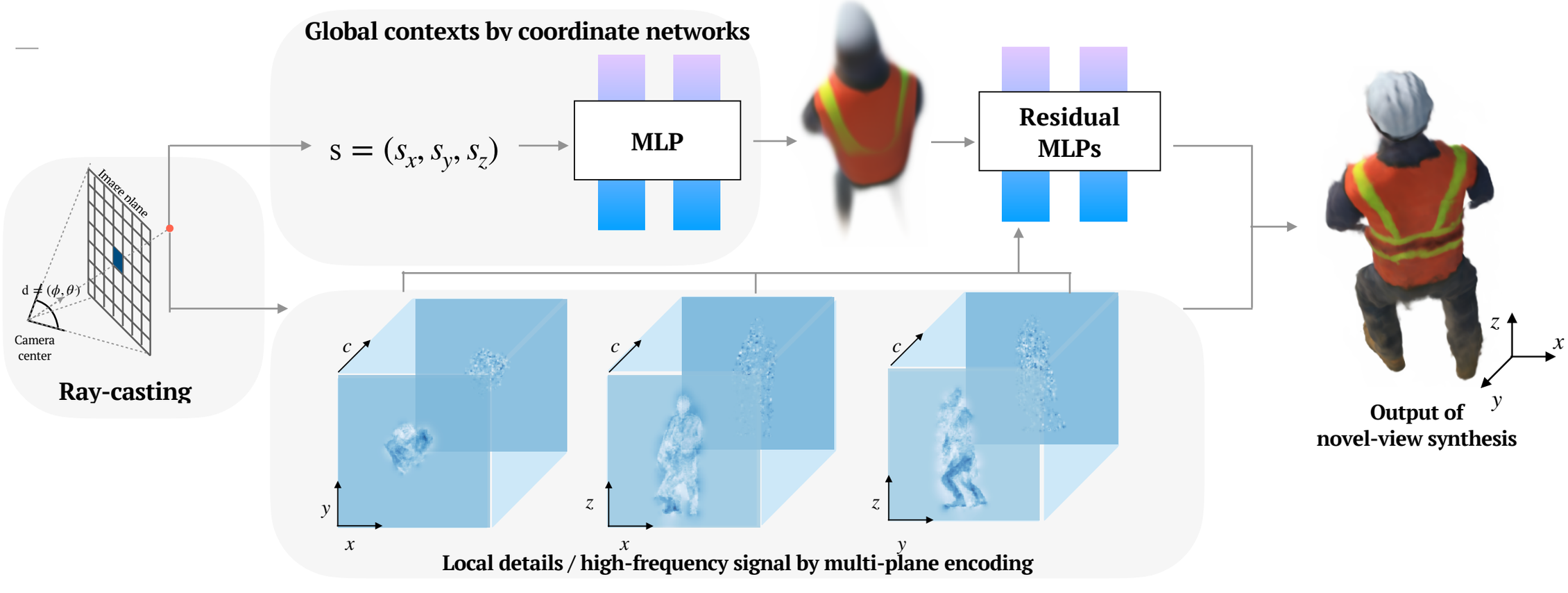


11) Lee, Joo Chan, et al. "Coordinate-aware modulation for neural fields." *ICLR2024*

Proposed Method (1)

Optimization and Statistical Inference LAB

- Residual networks seamlessly incorporates coordinate network and multi-plane features.
- We introduce progressive training strategy:
  - The coordinate network is trained first, and followed by multiple-plane encoding.
    - Coordinate networks: low-frequency details
    - Multi-plane features: high-frequency details.



Proposed Method (2)

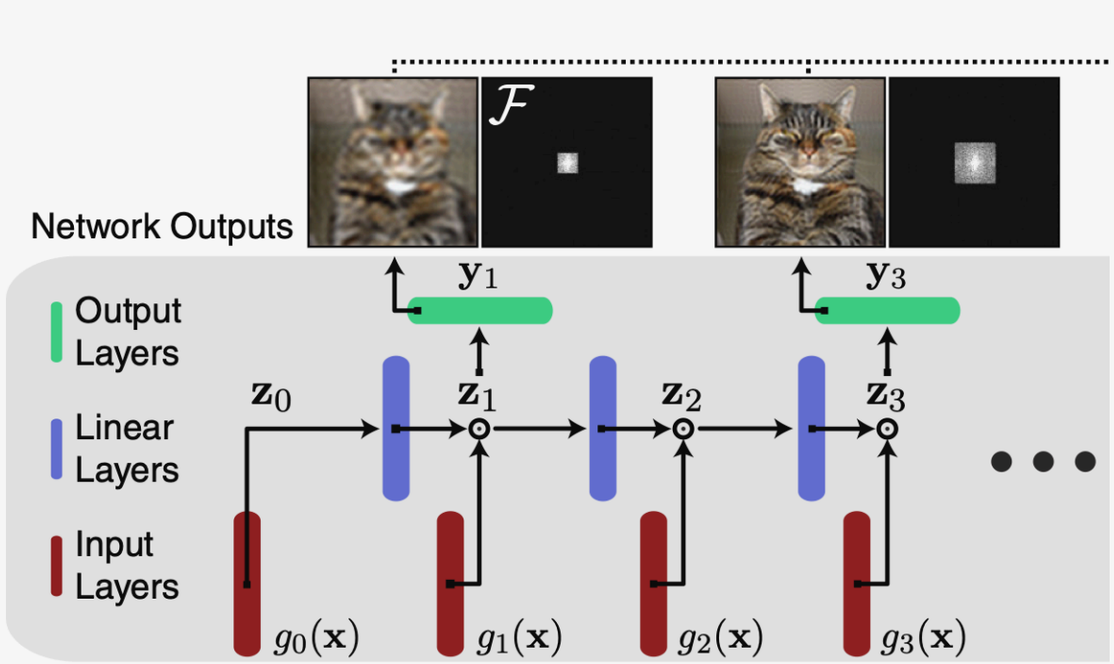
Optimization and Statistical Inference LAB

- The loss function consists of photometric loss, TV regularization and L1 on multi-planes

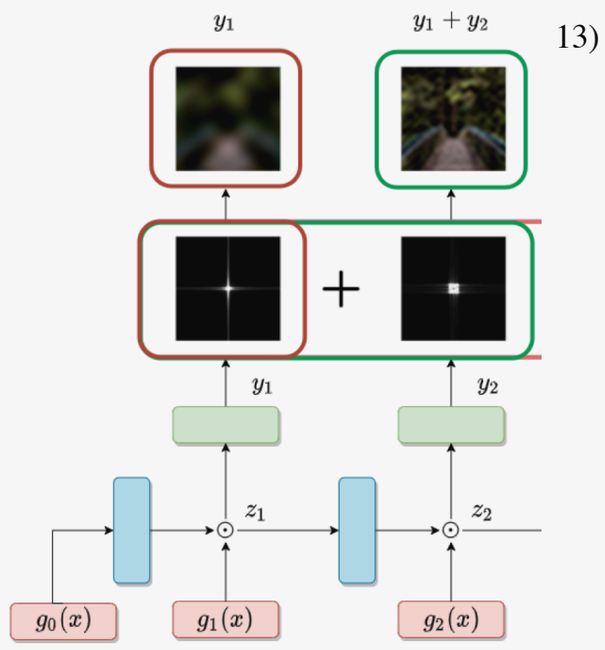
$$\mathscr{L}(\Theta, M) = \sum_r \|\hat{\mathrm{c}}(\mathrm{r}; \Theta, M) - \mathrm{c}\|^2 + \lambda_1 \sum_c \sum_{hw} \left(\left\| M^c_{h+1,w} - M^c_{h,w} \right\|_2^2 + \left\| M^c_{h,w+1} - M^c_{h,w} \right\|_2^2 \right) + \lambda_2 \left(\|M\|_1 \right)$$

- Contributions:
  - Previous band-limited coordinate networks supposed to have homogenous features.
- However, the proposed method allows to incorporate heterogenous different features.

12)



13)



12) Lindell, David B., et al. "Bacon: Band-limited coordinate networks for multiscale scene representation."*CVPR2022*

13) Shekarforoush, Shayan, et al. "Residual multiplicative filter networks for multiscale reconstruction." *NeurIPS2022*

Quantitative Result (1)

Optimization and Statistical Inference LAB

- Dataset : NeRF Synthetic (Training data: 8 Views / Test data : 200 views)
  - Explicit encoding approaches shows competitive performance (Lego, Hotdog, Mic),
  - However, these methods struggle to depict accurately in (Drums, Ship)
  - Nevertheless, `ours` addresses these challenges by incorporating coordinate networks and multi-plane encoding, resulting in improved performance.

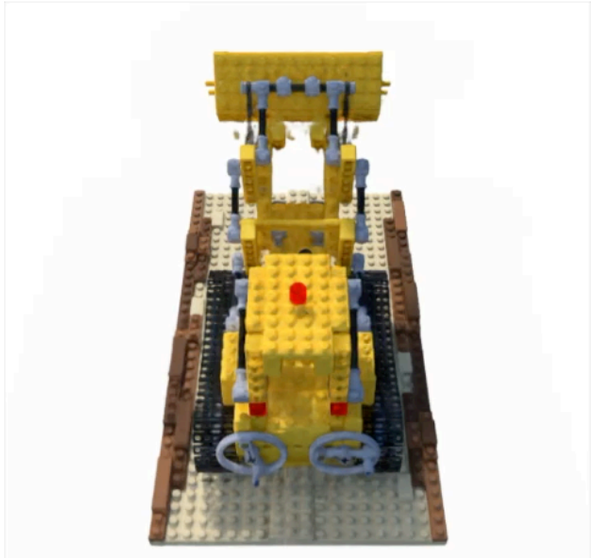
Models	**PSNR ↑**								**Avg. PSNR ↑**	**Avg. SSIM ↑**	**Avg. LPIPS ↓**
	chair	drums	ficus	hotdog	lego	materials	mic	ship			
Simplified_NeRF	20.35	14.19	<u>21.63</u>	22.57	12.45	18.98	24.95	18.65	19.22	0.827	0.265
DietNeRF	21.32	14.16	13.08	11.64	16.12	12.20	24.70	19.34	16.57	0.746	0.333
HALO	24.77	18.67	21.42	10.22	22.41	21.00	24.94	21.67	20.64	0.844	0.200
FreeNeRF	26.08	<u>19.99</u>	18.43	<u>28.91</u>	24.12	<u>21.74</u>	24.89	<u>23.01</u>	23.40	0.877	0.121
DVGO	22.35	16.54	19.03	24.73	20.85	18.50	24.37	18.17	20.57	0.829	0.145
VGOS	22.10	18.57	19.08	24.74	20.90	18.42	24.18	18.16	20.77	0.838	0.143
iNGP	24.76	14.56	20.68	24.11	22.22	15.16	26.19	17.29	20.62	0.828	0.184
TensoRF	26.23	15.94	21.37	28.47	26.28	20.22	26.39	20.29	23.15	0.864	0.129
K-Planes	<u>27.30</u>	**20.43**	**23.82**	27.58	<u>26.52</u>	19.66	**27.30**	21.34	<u>24.24</u>	**0.897**	**0.085**
Ours	**28.02**	19.55	20.30	**29.25**	**26.73**	**21.93**	<u>26.42</u>	**24.27**	**24.56**	<u>0.896</u>	<u>0.092</u>

Results on static NeRF Synthetic

Qualitative Results

Optimization and Statistical Inference LAB

- The proposed method provides qualitatively robust performance across all scenes.



Ours

- When only the coordinate network is engaged, global reasoning is well constructed. Then, tensorial features compliments finest details.

Coordinate Network engaged only

$t = 0$ $t = 0.5$ $t = 1.0$

Progressive manipulation of Tensorial Feature

$t = 0$ $t = 0.5$ $t = 1.0$

Quantitative Result (2) : Stability

Optimization and Statistical Inference LAB

- We define stability as the minimal performance discrepancy between test viewpoints adjacent to and not adjacent to the training views.
  - Variance of PSNR across all test viewpoints in the static NeRF dataset.
- FreeNeRF, which uses MLP and sinusoidal encoding, records the lowest variance among baselines. While K-Planes reduces instability than these methods, its variances still do not reach the level of ours. Quantitatively, ours achieves comparable results to FreeNeRF.

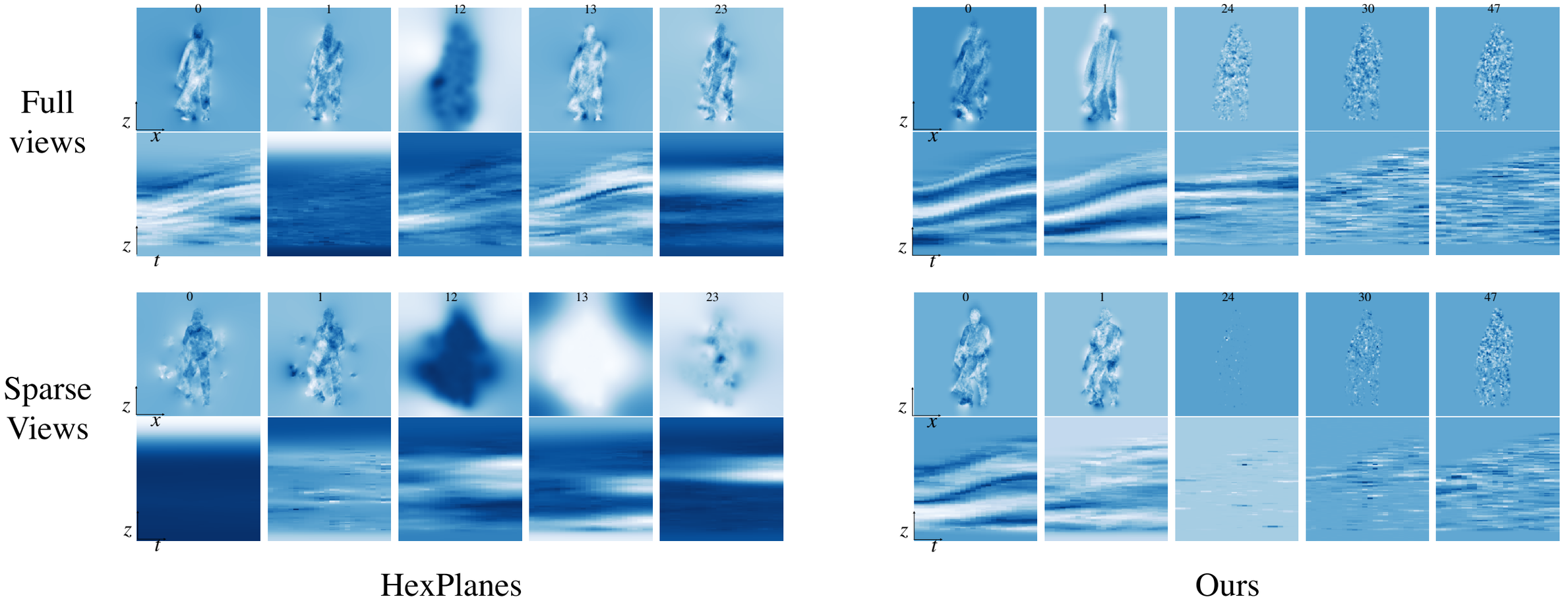
	chair	**lego**	**ship**	**Total**
FreeNeRF	5.07	6.42	6.48	17.31
iNGP	8.43	7.78	6.03	23.95
TensoRF	10.88	10.27	5.71	23.22
K-Planes	10.74	10.76	11.48	19.61
Ours	3.82	8.72	6.01	18.23

Variance of PSNR in each scene

Ablation Study (1)

Optimization and Statistical Inference LAB

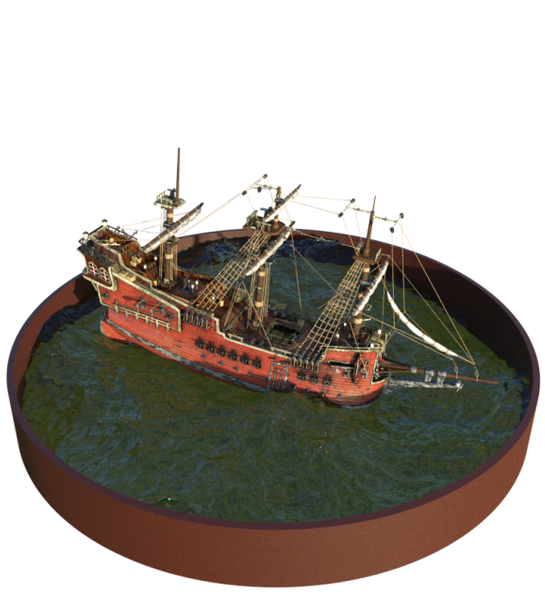
- We provide visualizations of multi-plane features to understand how scenarios with sparse inputs influence the learning of these features.
- Under full viewpoints, `HexPlane` does not exhibit artifacts. However, it suffers from unintended floating artifacts adjacent to human shapes, and a few channels fail to learn standing human features appropriately in the sparse-view.
- `Ours` preserves similar pattern regardless of full and sparse views.



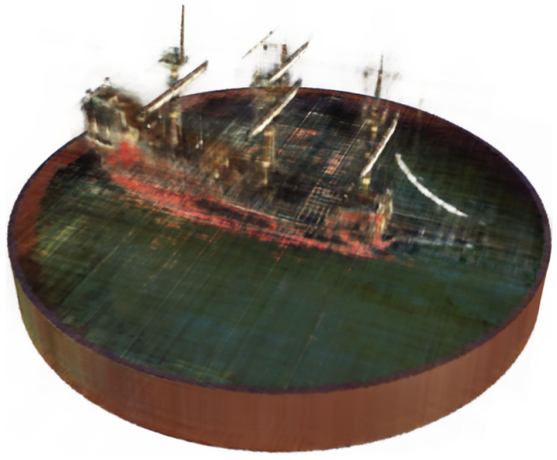
Ablation Study (2)

Optimization and Statistical Inference LAB

- Furthermore, this method is more stable against varying denoising regularization, if this parameter varies, this method less influence compared of TensoRF and K-Planes
  - When excessive denoising weights, the images simply appear faded or less vibrant.



Ground Truth



TensoRF with
$\lambda_1 = 1.0$

K-Planes with
$\lambda_1 = 1.0$

TensorRefine
$\lambda_1 = 1.0$

Quantitative Result (3) : Efficiency

Optimization and Statistical Inference LAB

- The proposed method preserves performance even when the number of parameters is reduced to 1M because the coordinate network supports low-frequency details.
- This efficiency stems from replacing the low-resolutional spatial parameters with the coordinate network.

Model Name	# Params [M]	Avg. PSNR
iNGP (T=19)	11.7M	19.26
iNGP (T=18)	6.4M	19.99
K-Planes (3*16)	17M	23.95
K-Planes (2*16)	4.4M	23.16
TensoRF (64)	17.3M	25.23
TensoRF (20)	6.1M	-
Ours (48)	6.0M	24.36
Ours (24)	3.0M	23.74

Static NeRF dataset

❖ Optimal values for λ_1

Model Name	# Params [M]	Avg. PSNR
K-Planes (3*32)	18.6M	23.85
K-Planes (3*4)	1.9M	23.41
HexPlane (72)	9.7M	24.00
HexPlane (6)	0.8M	22.08
Ours (48)	3.4M	25.17
Ours (12)	1.0M	25.10

Dynamic NeRF dataset

Summary and Limitation

- We have identified that tensorial features are effective for Neural Radiance Fields (NeRFs), but they are quite susceptible to overfitting to training views.

- Integrating the coordinate network with tensorial features significantly improves performance under sparse-input conditions without the need for additional regularization techniques.

- However, due to the per-scene optimization, this approach does not generalize well to novel tasks. The next chapter will explore a strong prior model that, by training on various objects within the same categories, can generalize effectively even in extremely sparse situations.

Q&A