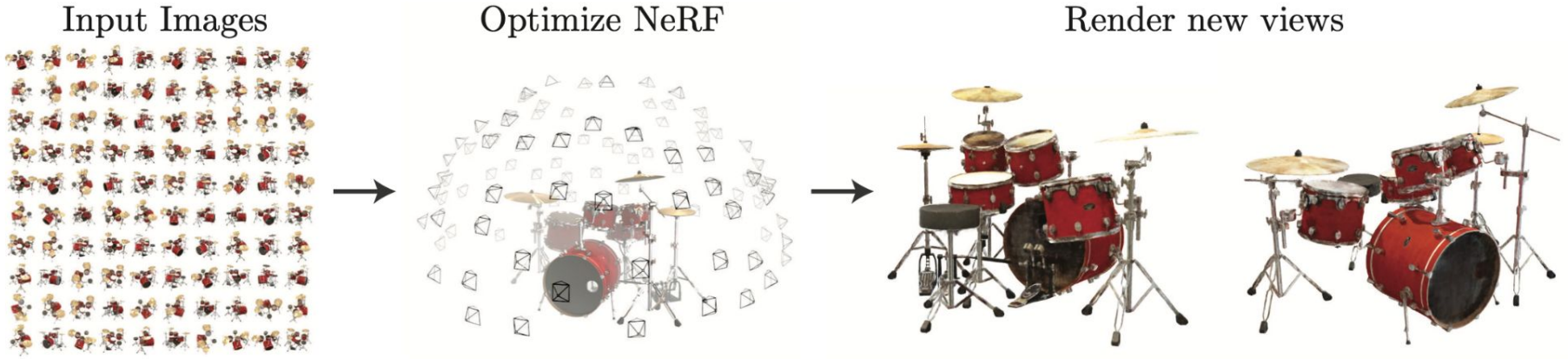
NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis

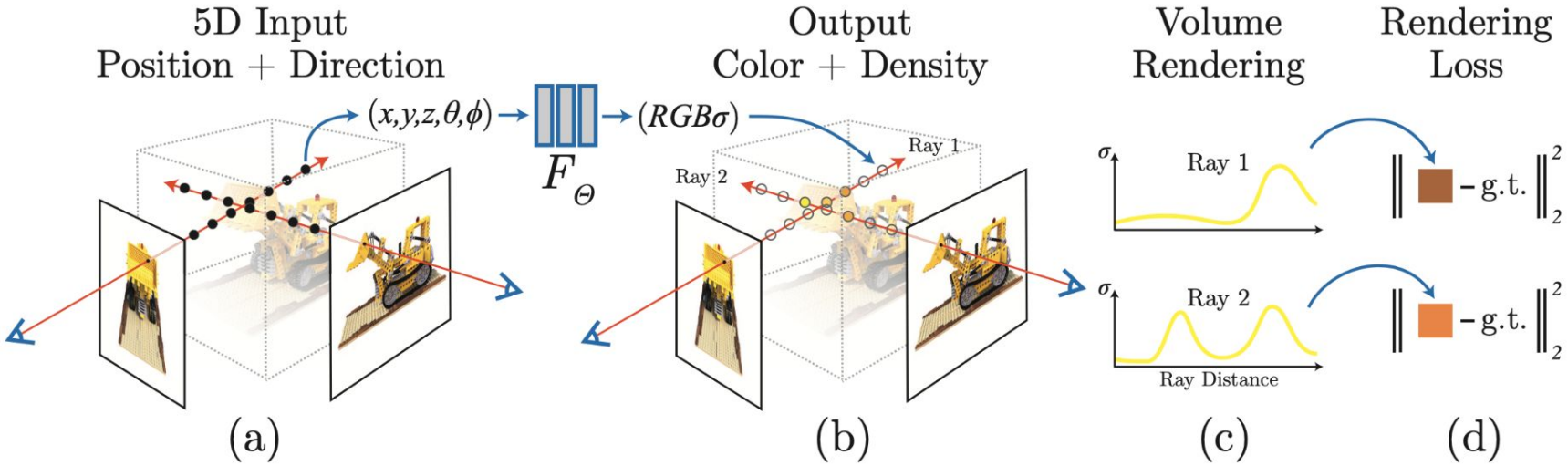
Полина Кадейшвили

Постановка задачи



Хотим на основе картинок одного объекта, сделанных с разных ракурсов, научиться представлять 3D сцену

Процесс обучения



- На вход подается точка в пространстве $\mathbf{x} = (x, y, z)$ и направление взгляда (θ, ϕ)
- На выходе мы получаем цвет точки $\mathbf{c} = (R, G, B)$ и плотность в этой точке (volum σ ensity)

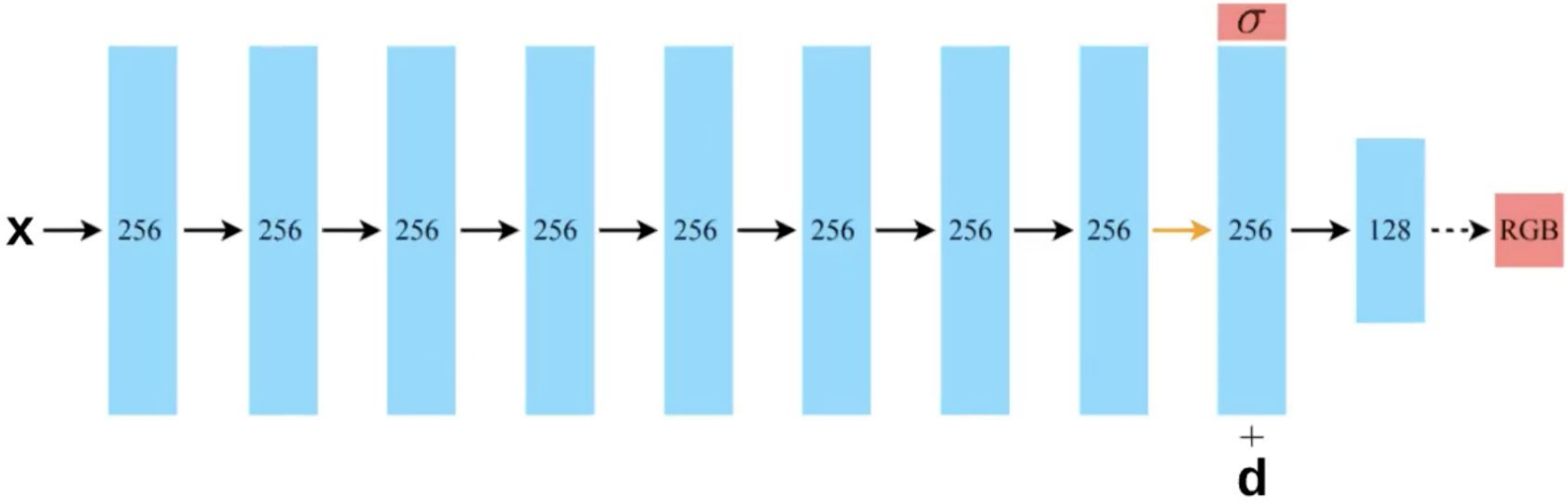
Визуализация

Направление взгляда

На практике вместо двух углов (θ, ϕ) используется трехмерный единичный вектор d в декартовой системе координат

$$\mathbf{d} = \begin{bmatrix} sin\theta \cos\phi \\ sin\theta sin\phi \\ cos\theta \end{bmatrix}$$

Архитектура модели



Черные стрелки - слой с активацией ReLU, оранжевые - без активации, пунктирная - сигмоидная активация

- NeRF - полносвязная сеть без сверток (MLP)
- σ - величина от 0 до бесконечности, которая не зависит от направления взгляда
- Цвет $\mathbf{c} = (R, G, B)$ зависит и от $\mathbf{x} = (x, y, z)$, и от $\mathbf{d}$

Предсказание цвета

Вероятность того, что луч пройдет из t_n в t, не встретив объект

Цвет в точке r(t) с позиции d

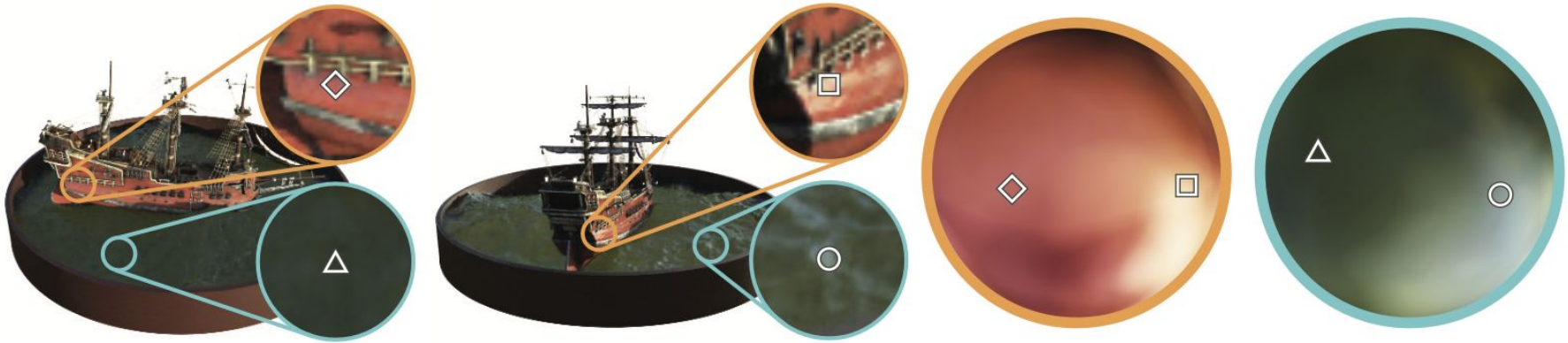
$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt \quad \text{, где} \quad T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right)$$

Предсказанный цвет

Плотность в точке

луч $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$

Иллюстрация предсказания цвета



(a) View 1 (b) View 2 (c) Radiance Distributions

Приближение интеграла

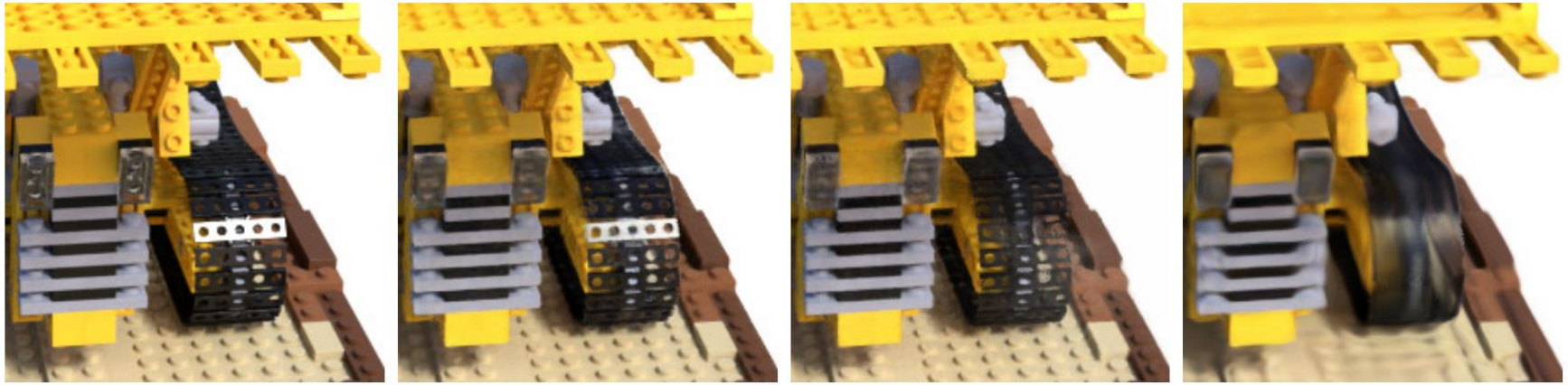
$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i \,, \text{ where } T_i = \exp\left(- \sum_{j=1}^{i-1} \sigma_j \delta_j \right)$$

$\delta_i = t_{i+1} - t_i$ -расстояние между соседними семплами на луче

$\mathbf{c}_i$ -значение $\mathbf{c}(\mathbf{r}(t), \mathbf{d})$ в точке t_i

$$t_i \sim \mathcal{U}\left[t_n + \frac{i-1}{N}(t_f - t_n),\ t_n + \frac{i}{N}(t_f - t_n) \right]$$

Визуализация результатов такой архитектуры



Ground Truth | Complete Model | No View Dependence | No Positional Encoding

Оптимизация модели

- Positional encoding
- Hierarchical volume sampling

Positional encoding

Проблема: сеть, которая работает только с входными $(\mathbf{x}, \mathbf{d})$ плохо визуализирует изменения цвета и геометрии.

Positional encoding

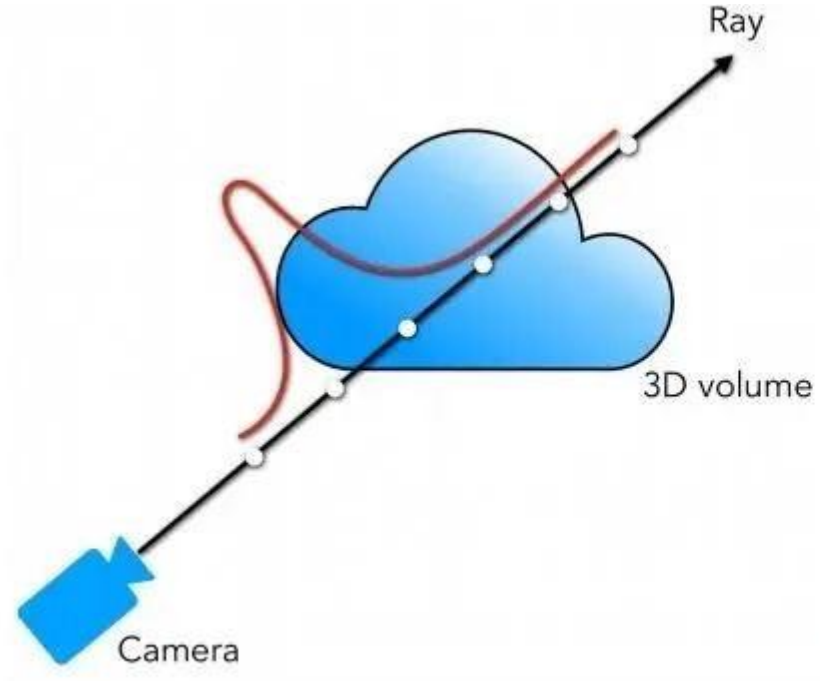
$$F_\Theta = F'_\Theta \circ \gamma$$

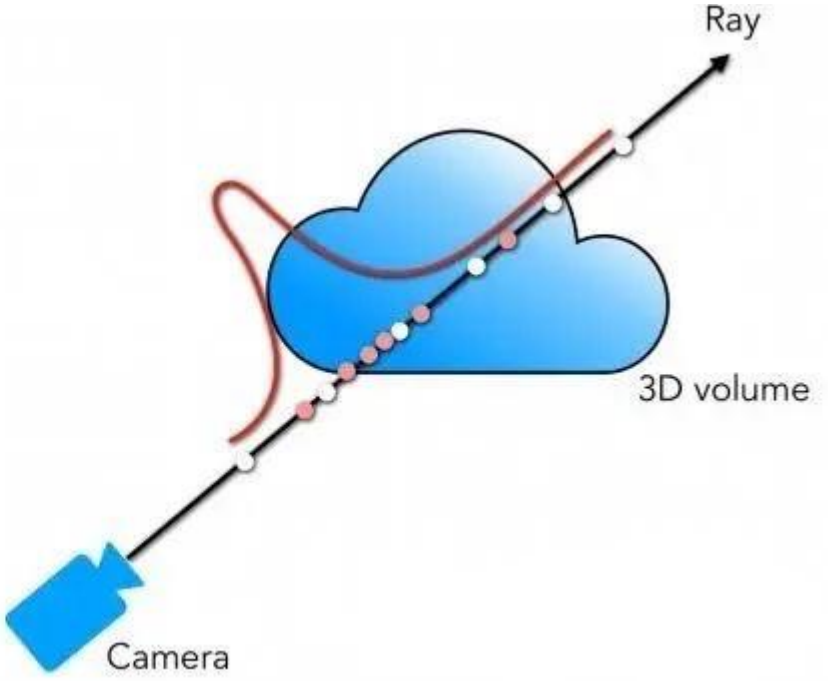
$$\gamma(p) = \left(\sin\left(2^0\pi p\right), \cos\left(2^0\pi p\right), \cdots, \sin\left(2^{L-1}\pi p\right), \cos\left(2^{L-1}\pi p\right)\right)$$

$$L = 10 \text{ for } \gamma(\mathbf{x}) \text{ and } L = 4 \text{ for } \gamma(\mathbf{d})$$

Hierarchical volume sampling

Проблема: в текущем подходе, когда мы берем точки на луче равномерно получается много неинформативных точек, которые не дают вклад в цвет пикселя.





Hierarchical volume sampling

- Возьмем 2 сети: “coarse” (грубая) и “fine” (точную)
- Сгенерируем точки N_c как раньше равномерно и обучим на них “coarse” модель
- Получим предсказания о цвете и плотности от “coarse” модели
- Сгенерируем более информативные точки N_f, основываясь на этих предсказаниях
- Оценим цвет с помощью “fine ” модели на N_f + N_c точках

Hierarchical volume sampling

Генерация новых значимых точек

Перепишем исходное равенство для N_c

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i\,, \text{ where } T_i = \exp\left(-\sum_{j=1}^{i-1}\sigma_j\delta_j\right)$$

$$\hat{C}_c(\mathbf{r}) = \sum_{i=1}^{N_c} w_i c_i\,, \qquad w_i = T_i(1 - \exp(-\sigma_i\delta_i))$$

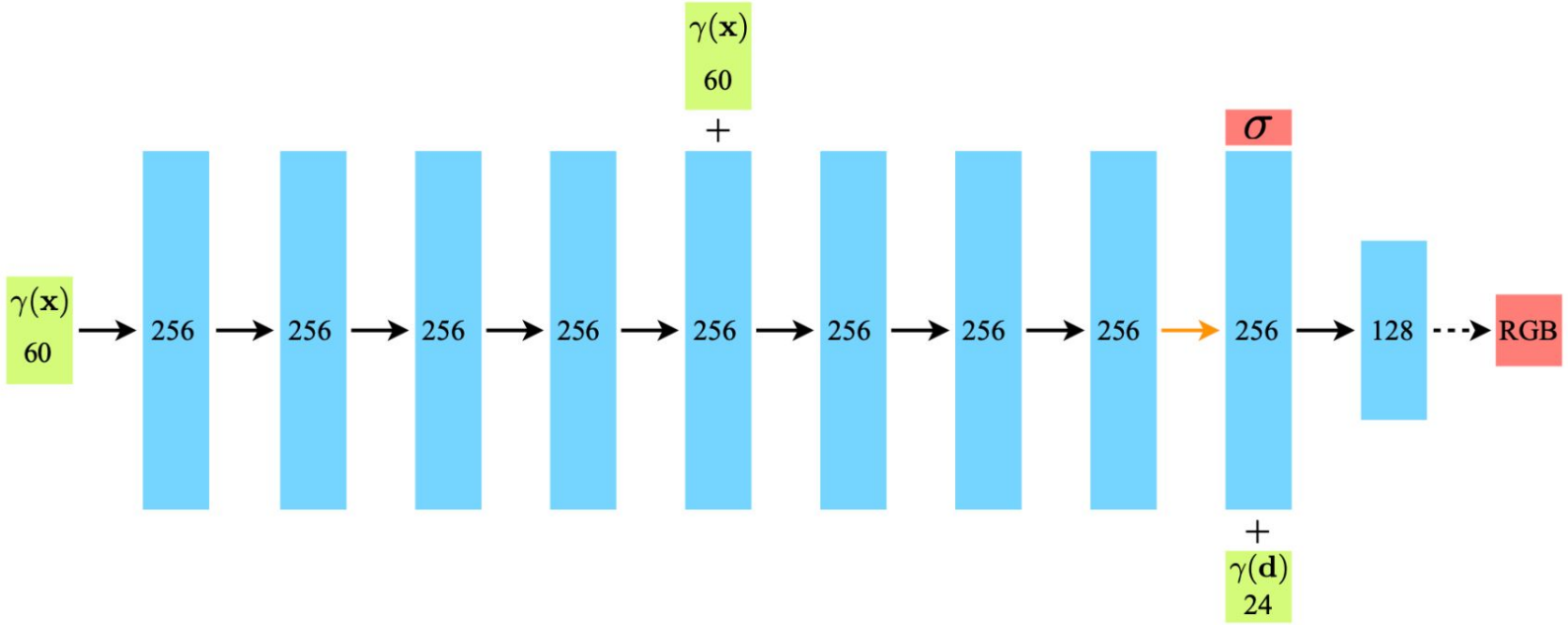
Отнормировав веса $\hat{w}_i = w_i / \sum_{j=1}^{N_c} w_j$, получим кусочно-заданную функцию плотности, из распределения которой будем генерировать точки N_f

Loss function

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}} \left[\left\| \hat{C}_c(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 + \left\| \hat{C}_f(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 \right]$$

- $\hat{C}_c(\mathbf{r})$ - предсказание цвета пикселя моделью “coarse”
- $\hat{C}_f(\mathbf{r})$ - предсказание цвета пикселя моделью “fine”
- $C(\mathbf{r})$ - настоящий цвет пикселя
- $\mathcal{R}$ - набор лучей в batch-е

Визуализация полной архитектуры



Черные стрелки - слой с активацией ReLU, оранжевые - без активации, пунктирная - сигмоидная активация

Эксперименты

	Input	#Im.	L	(N_c, N_f)	PSNR↑	SSIM↑	LPIPS↓
1) No PE, VD, H	xyz	100	-	(256, -)	26.67	0.906	0.136
2) No Pos. Encoding	$xyz\theta\phi$	100	-	(64, 128)	28.77	0.924	0.108
3) No View Dependence	xyz	100	10	(64, 128)	27.66	0.925	0.117
4) No Hierarchical	$xyz\theta\phi$	100	10	(256, -)	30.06	0.938	0.109
5) Far Fewer Images	$xyz\theta\phi$	25	10	(64, 128)	27.78	0.925	0.107
6) Fewer Images	$xyz\theta\phi$	50	10	(64, 128)	29.79	0.940	0.096
7) Fewer Frequencies	$xyz\theta\phi$	100	5	(64, 128)	30.59	0.944	0.088
8) More Frequencies	$xyz\theta\phi$	100	15	(64, 128)	30.81	0.946	0.096
9) Complete Model	$xyz\theta\phi$	100	10	(64, 128)	**31.01**	**0.947**	**0.081**

Метрики:

- PSNR - peak signal to noise ratio
- SSIM - structural similarity index
- LPIPS - learned perceptual image patch similarity

Сравнение с другими методами

Method	Diffuse Synthetic 360° [41]			Realistic Synthetic 360°			Real Forward-Facing [28]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SRN [42]	33.20	0.963	0.073	22.26	0.846	0.170	22.84	0.668	0.378
NV [24]	29.62	0.929	0.099	26.05	0.893	0.160	-	-	-
LLFF [28]	34.38	0.985	0.048	24.88	0.911	0.114	24.13	0.798	**0.212**
Ours	**40.15**	**0.991**	**0.023**	**31.01**	**0.947**	**0.081**	**26.50**	**0.811**	0.250

Нейронные сети:

- SRN (scene representation networks) - рекуррентная сеть
- NV (neural volumes) - сверточная сеть, работает только с ограниченными сценами
- LLFF (local light field fusion) - сверточная сеть, заточена под реальные сцены

Метрики:

- PSNR - peak signal to noise ratio
- SSIM - structural similarity index
- LPIPS - learned perceptual image patch similarity

Ship

Lego

Ground Truth NeRF (ours) LLFF [28] SRN [42] NV [24]

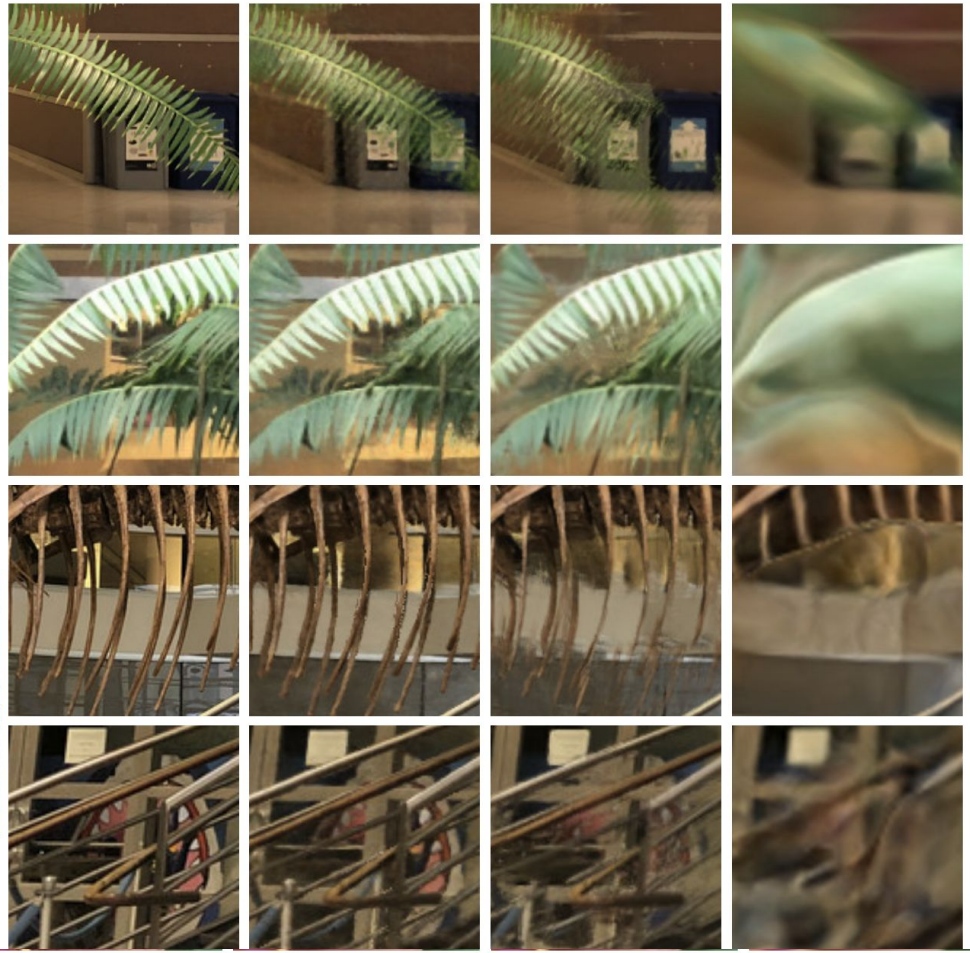
Microphone

Materials

Ground Truth NeRF (ours) LLFF [28] SRN [42] NV [24]

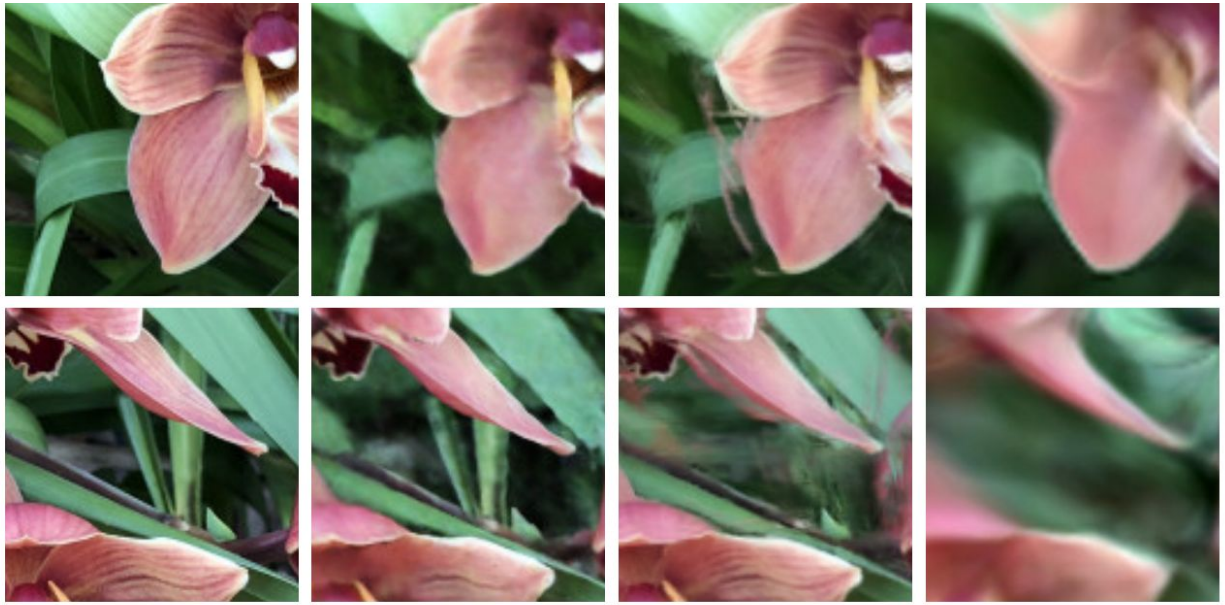
Fern

T-Rex



Ground Truth NeRF (ours) LLFF [28] SRN [42]

Orchid



Ground Truth NeRF (ours) LLFF [28] SRN [42]