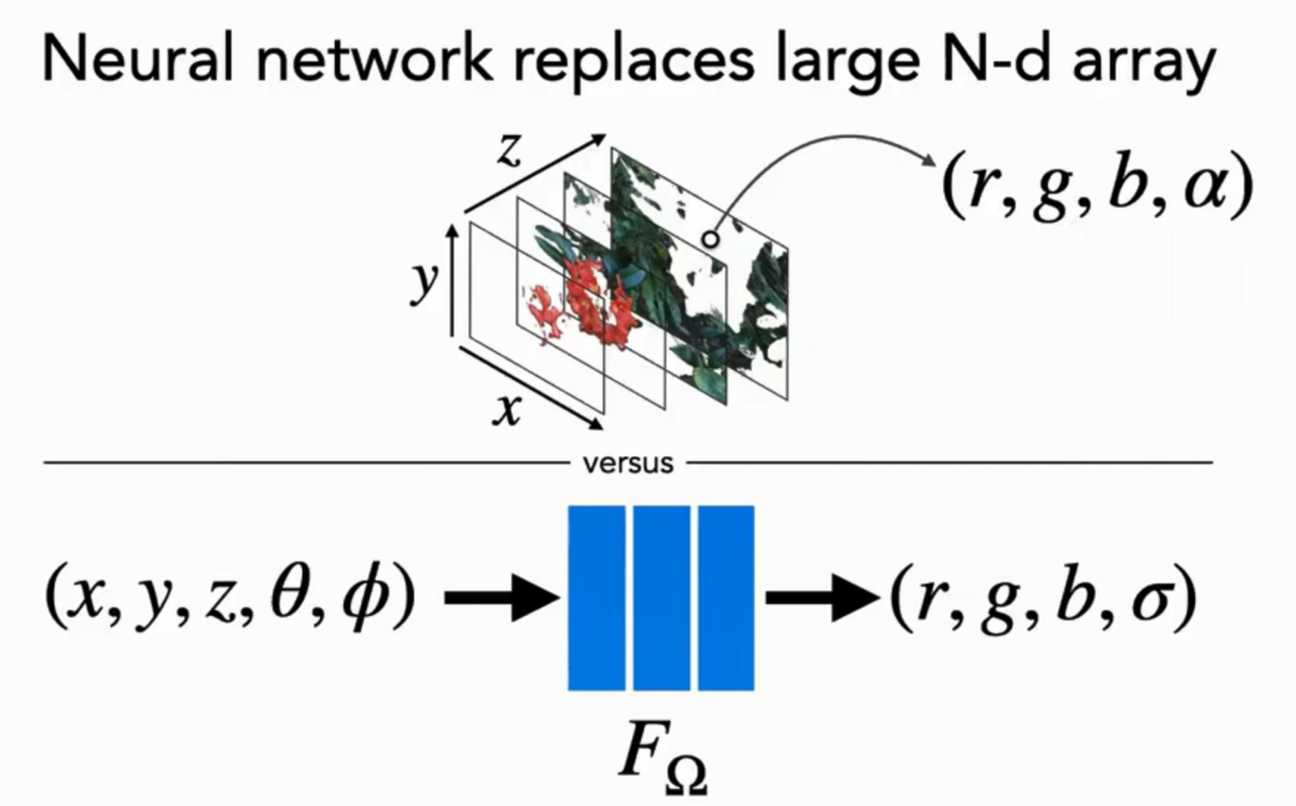
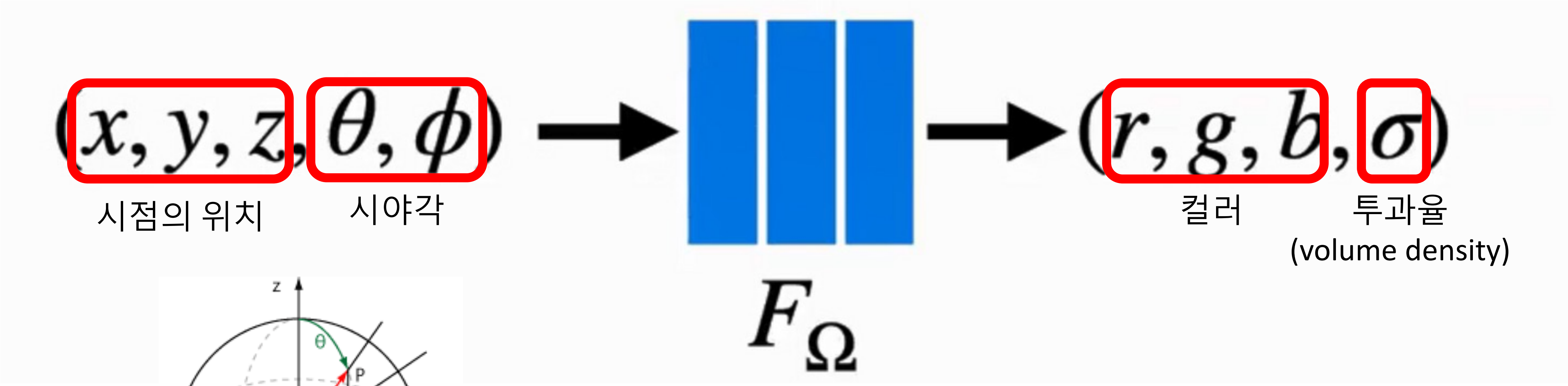
NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis

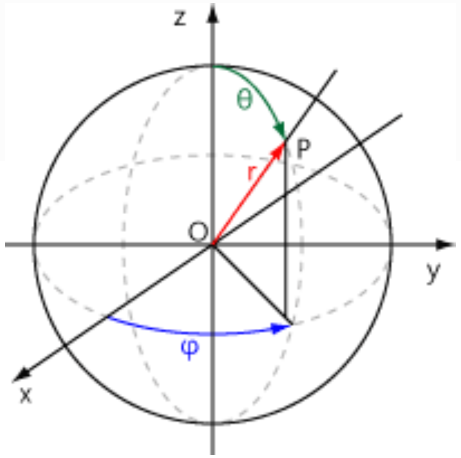
What is Implicit Representation?

- 어떤 정보를 인공 신경망을 통해서 저장하는 방법.



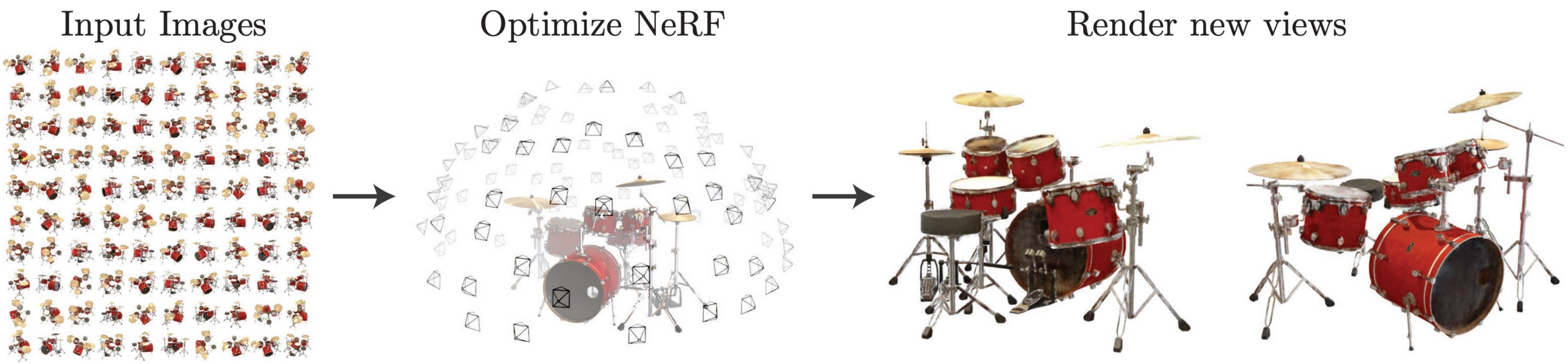
Overview





+ traditional rendering method

Overview



Contribution

- 복잡한 구조의 Continuous Scene을 표현하는 모델 학습.
- Hierarchical sampling을 사용하여 효과적 학습.
- Positional Encoding을 통해 high-frequency scene 학습.

Method - Overview

- Function structure
- Positional encoding
- Hierarchical volume sampling

Method – Function structure

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t),\mathbf{d})dt\,, \text{ where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right). \quad (1)$$

Sampling ↓

$$t_i \sim \mathcal{U}\left[t_n + \frac{i-1}{N}(t_f - t_n),\ t_n + \frac{i}{N}(t_f - t_n)\right]. \quad (2)$$

이산화 ↓

구적법 ↓

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i\,, \text{ where } T_i = \exp\left(-\sum_{j=1}^{i-1}\sigma_j\delta_j\right), \quad (3)$$

Continuous Ray

Discrete Ray

Method – Function structure

Rendering model for ray r(t) = o + td:

$$C \approx \sum_{i=1}^{N} T_i \alpha_i c_i$$

weights

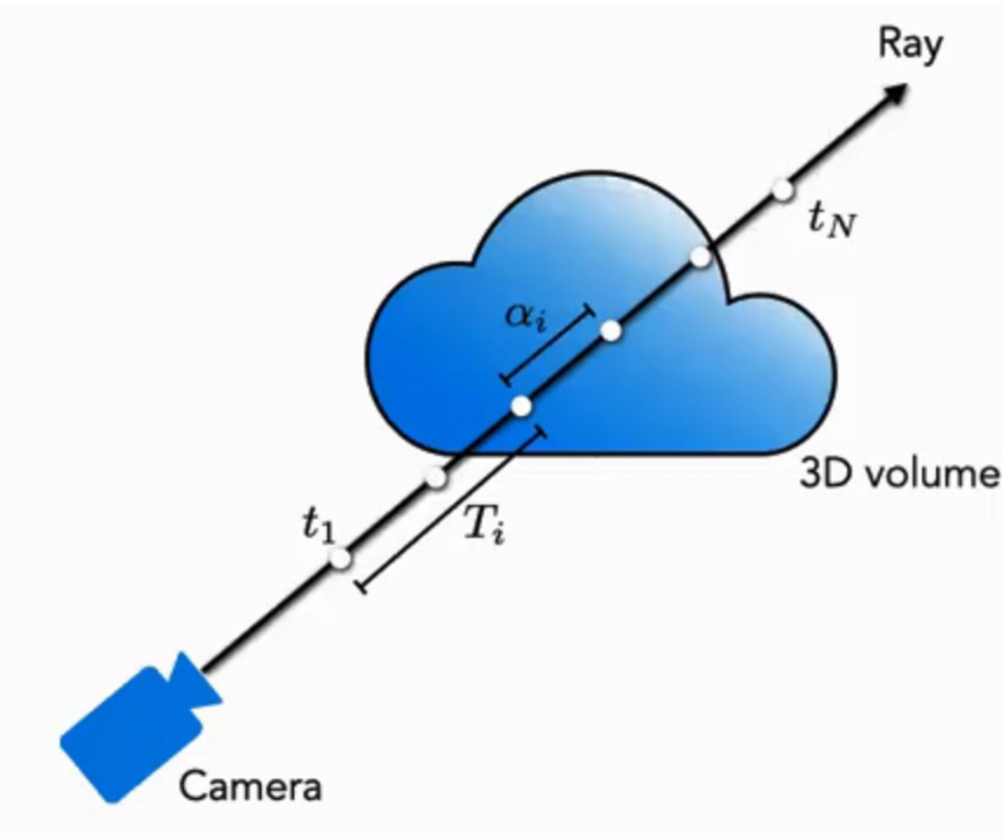
colors

How much light is blocked earlier along ray:

$$T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$$

How much light is contributed by ray segment *i*:

$$\alpha_i = 1 - e^{-\sigma_i \delta t_i}$$



Method – Positional encoding

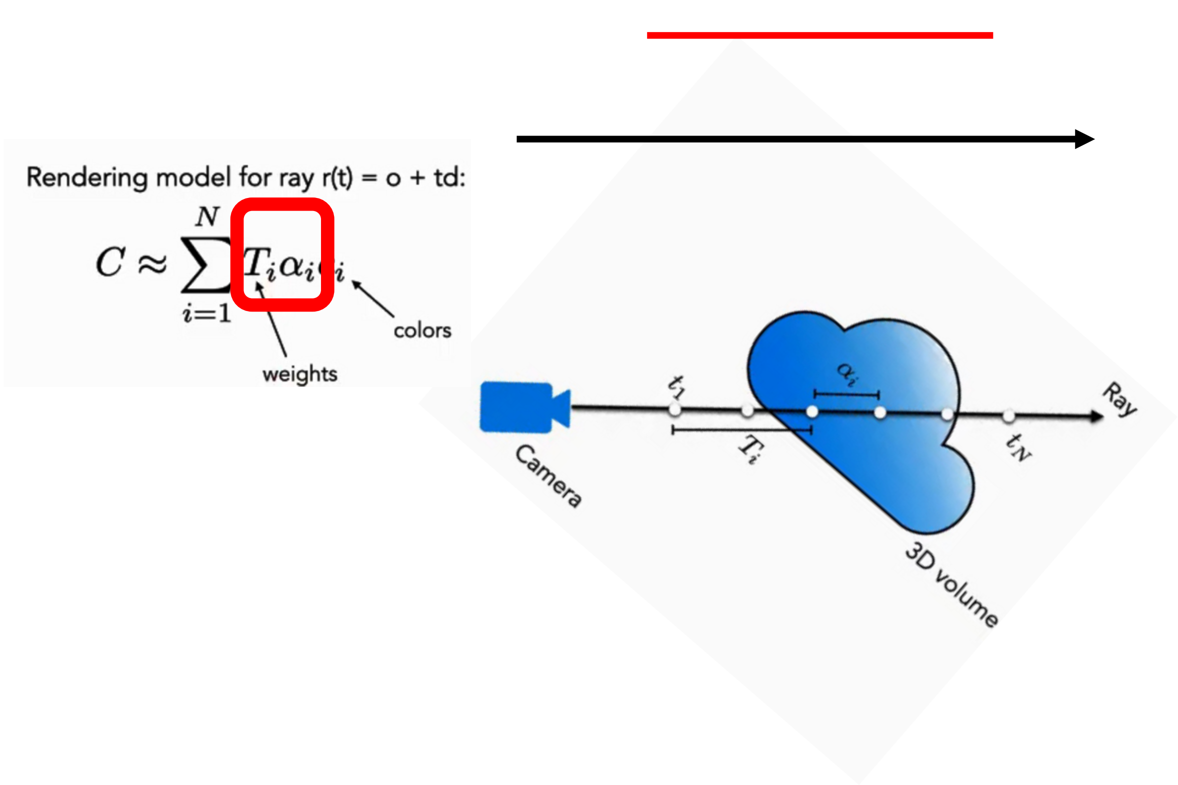
- 기존에는 복잡한 이미지 (high-frequency image)를 표현하지 못하는 이슈가 있었지만 “On the spectral bias of neural networks”에서 제시된 방법을 통해 이를 해결
  - high-frequency를 표현하기 위해서는 입력을 그대로 쓰기 보다 high-dimension에 mapping후 사용.

$$\gamma(p) = \left(\sin\left(2^0 \pi p\right), \cos\left(2^0 \pi p\right), \cdots, \sin\left(2^{L-1} \pi p\right), \cos\left(2^{L-1} \pi p\right) \right). \qquad (4)$$

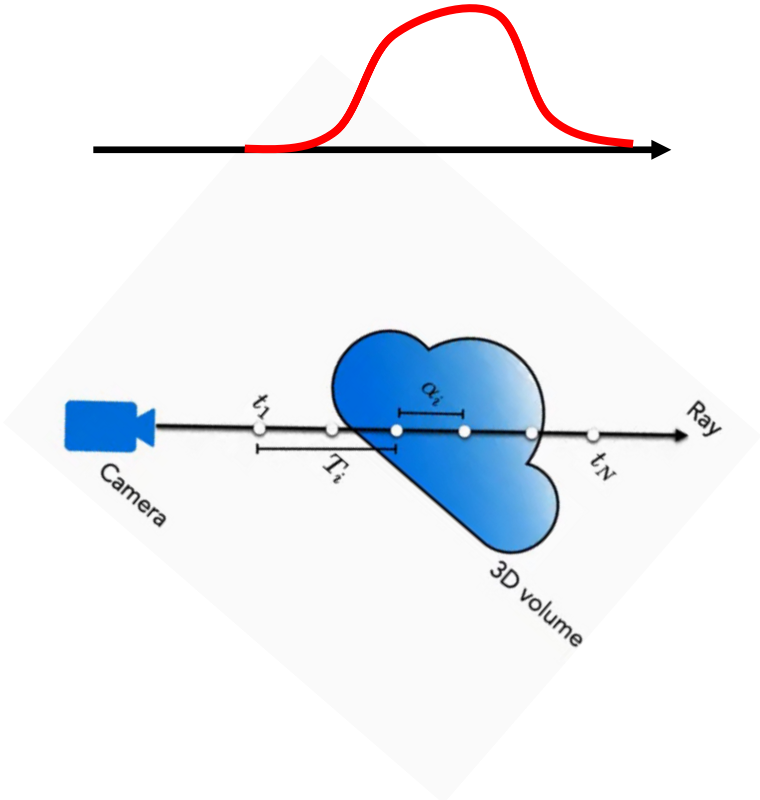
- 시점은 L을 20, 시야각은 4로 설정

Method – Hierarchical volume sampling

- 전형적인 Coarse – Refine구조 사용
- Uniform sampling -> Coarse func -> Weighted sampling -> Refine Func

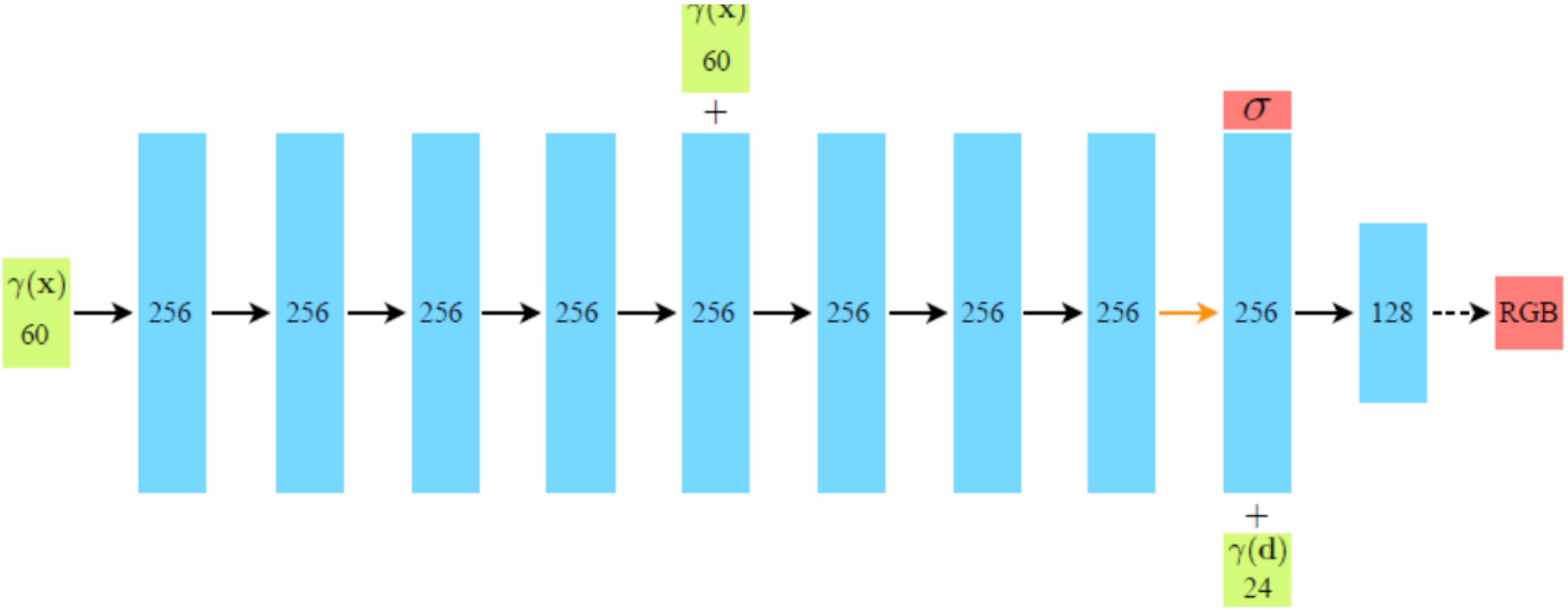


Inverse
Transform
Sampling



Training

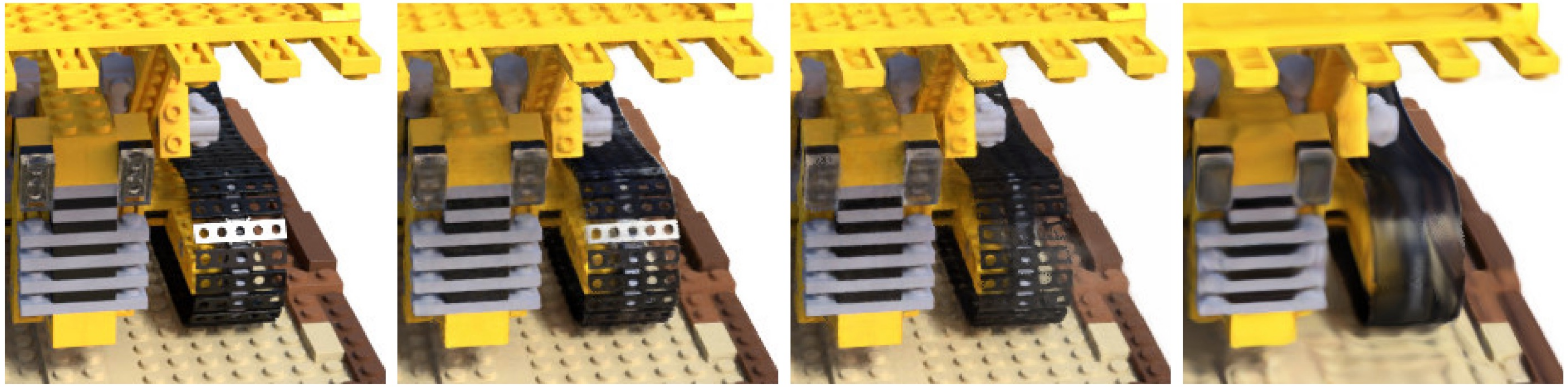
- 일정 시점 위치에서만 성능이 좋은 것을 방지하기 위해 다음과 같은 구조 채택



Result

- 이전의 SOTA보다 메모리 효율성 상승, 고해상, High-frequency 표현 가능

	Diffuse Synthetic 360° [41]			Realistic Synthetic 360°			Real Forward-Facing [28]		
Method	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



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

Materials

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

Ablation

	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**

Table 2: An ablation study of our model. Metrics are averaged over the 8 scenes from our realistic synthetic dataset. See Sec. 6.4 for detailed descriptions.