SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image

Dejia Xu*, Yifan Jiang*, Peihao Wang, Zhiwen Fan, Humphrey Shi, Zhangyang Wang
ECCV 2022

STRUCT Group Seminar
Presenter: Rundong Luo
2023.01.15

OUTLINE

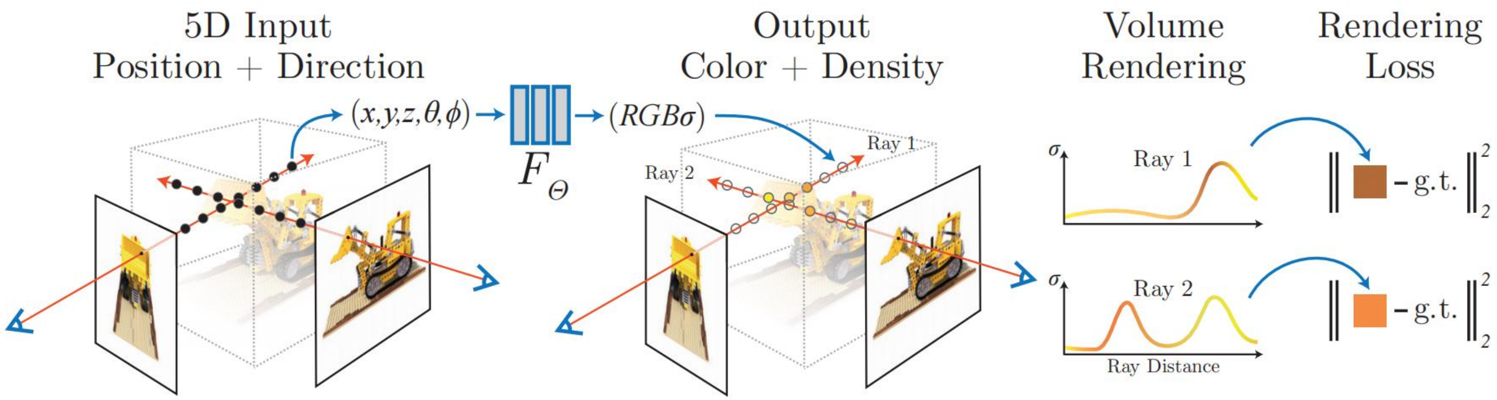
- Authorship
- Background
- Method
- Experiments
- Conclusion

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

BACKGROUND: NeRF

Overview



Volume rendering

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

BACKGROUND: NeRF

Critical points to consider when reading NeRF papers (personally)

- What's target task? Specifically, what's the input and output during inference?
- What's the training input? Multiview? Calibrated? etc.
- Is the model test-time optimization (one NeRF for each scene) or train-time optimization (one NeRF for multiple secens)

The original NeRF

- Training input: many (>10) calibrated images of the same scene
- Training output: a NeRF for this specific scene
- Inference input: camera pose
- Inference output: an image taken from this pose
- Test-time optimization

BACKGROUND: NeRF

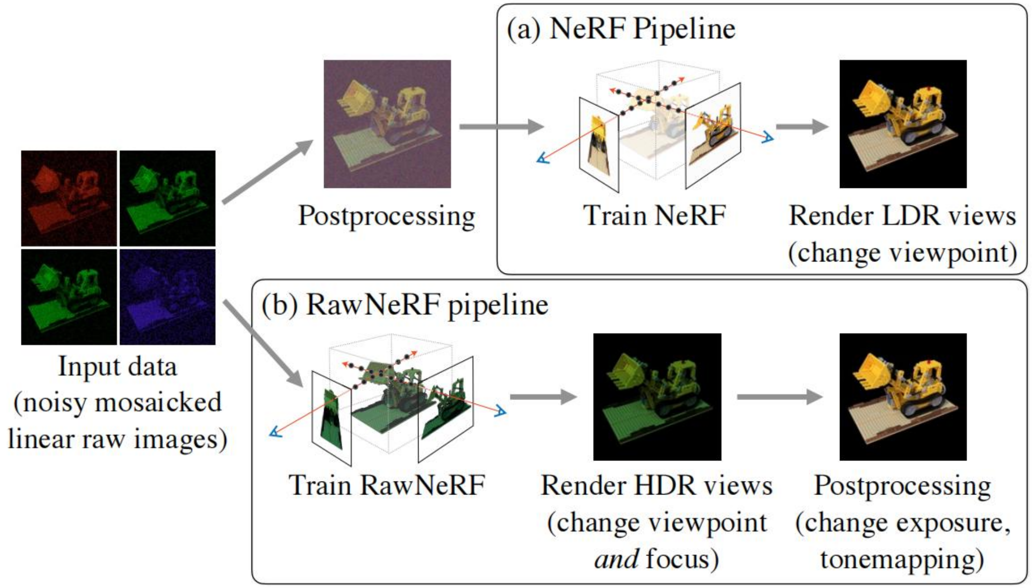
Fields of research on NeRF

- Faster training & inference
- Artistic effect, e.g. HDR NeRF
- Deformable
- Generalization, e.g. RawNeRF, GRAF, PixelNeRF, SinNeRF
- Compositionality, e.g. GIRAFFE, uORF

NeRF paper collection:
https://github.com/awesome-NeRF/awesome-NeRF

Previously on NeRF

RawNeRF (Mildenhall et al. CVPR 2022)

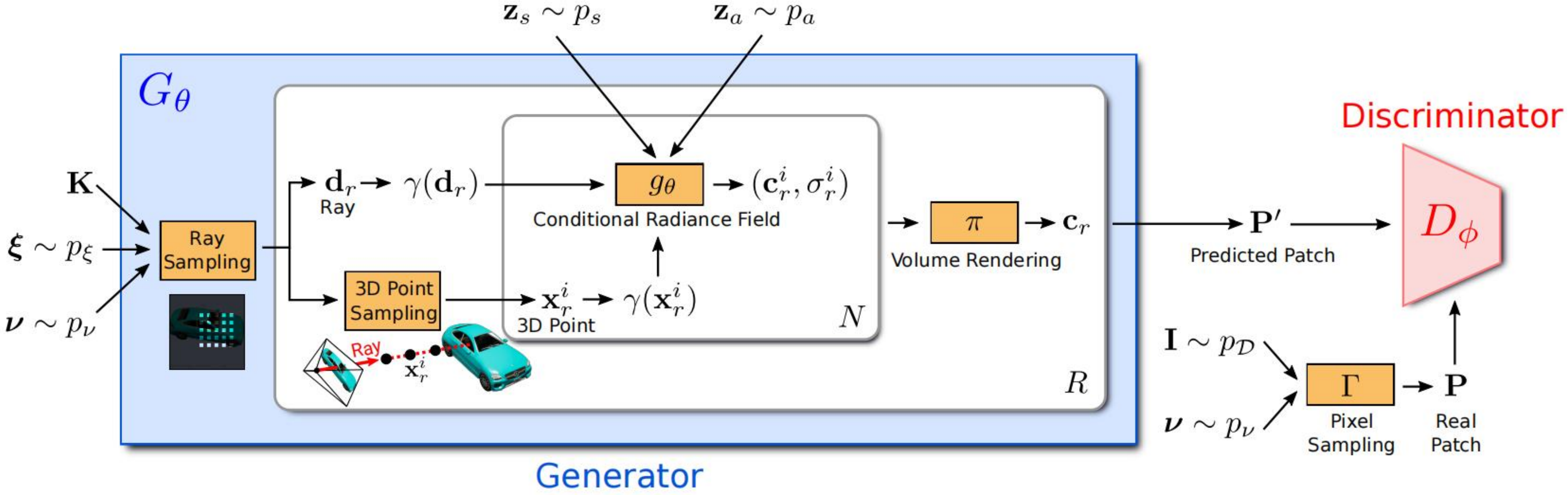


NeRF for denoising

- Training input identifcal to orignal NeRF
- Modified loss function & training inputs (multiple shutter speed)

Previously on NeRF

GRAF (Schwarz et al. NeurIPS 2020)

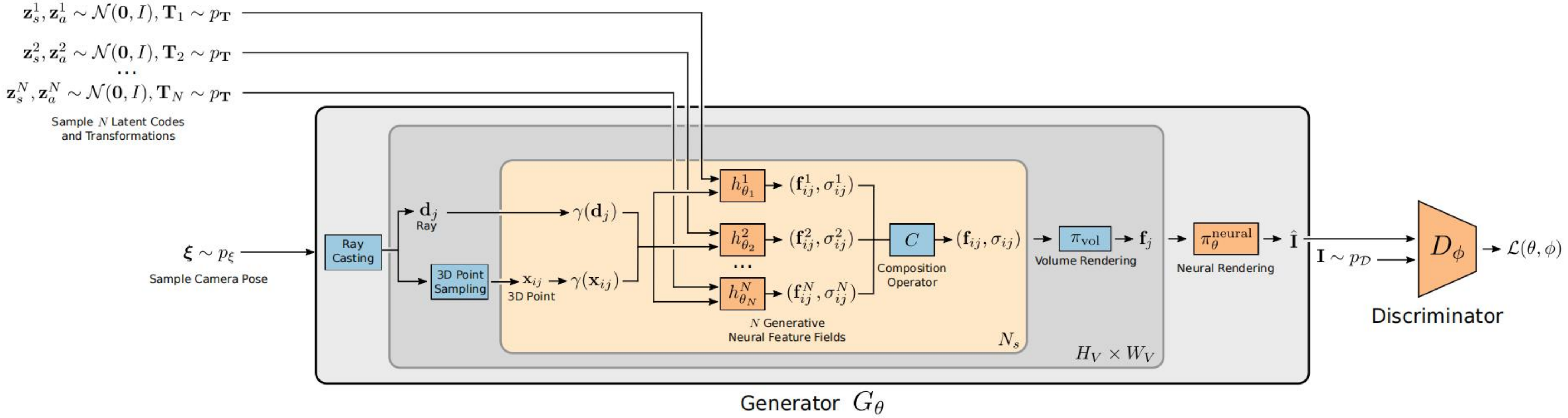


Generative Radiance Fields

- Training input: some real image patches of the same category (e.g. cars)
- Training output: NeRF for this type of input
- Trained on GAN loss (instead of pixel loss in original NeRF)
- Inference input: sampled from certain distributions, but no exact meaning
- Inference output: novel images of this type (conditioned on inputs)

Previously on NeRF

GIRAFFE (Niemeyer et al. CVPR 2021)

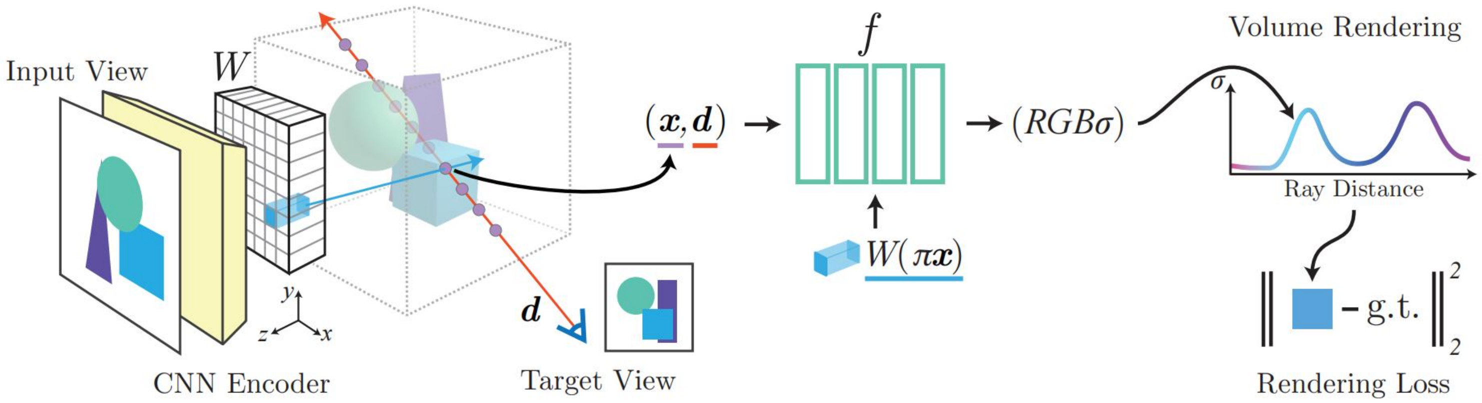


Generative Radiance Fields for Multiple Objects

- Training input: some real scenes with one or multiple objects
- Training output: encoder, NeRF in scene space
- Inference input: sampled from certain distributions. Inference latents, then stack together
- Inference output: novel images

Previously on NeRF

PixelNeRF (Yu et al. CVPR 2021)

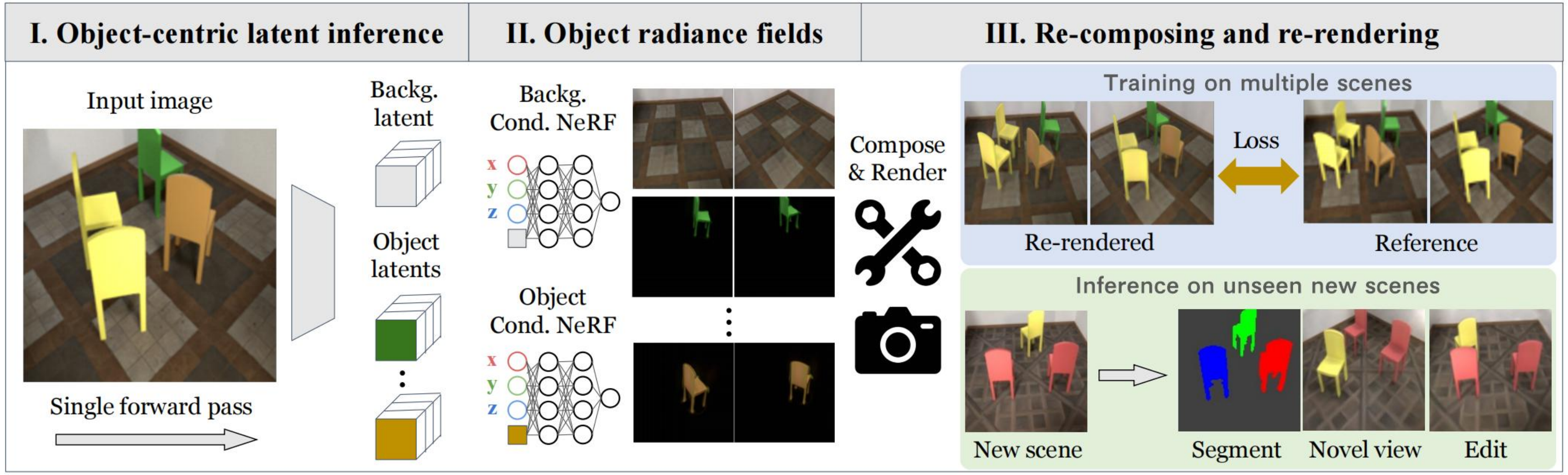


Single image NeRF for generating novel views

- Train-time optimization
- Feature encoder for synthesis
- Training input: multi-view (calibrated) scenes
- Training output: NeRF in view space, takes the reference view feature as an additional input
- Inference input: a single reference image and the desired camera pose (or relative pose to reference image)
- Inference output: novel images

Previously on NeRF

uORF (Yu et al. ICLR 2022)



Single image NeRF for scene editing & synthesis

- Train-time optimization
- Training input: multi-view (calibrated) scenes
- Training output: background NeRF in cononical space, object NeRF in view space
- Inference input: a single reference image
- Inference task: novel view synthesis, scene editing, scene segmentation

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

METHOD

Task description: generate novel views with only one image (test-time optimization)

Inputs

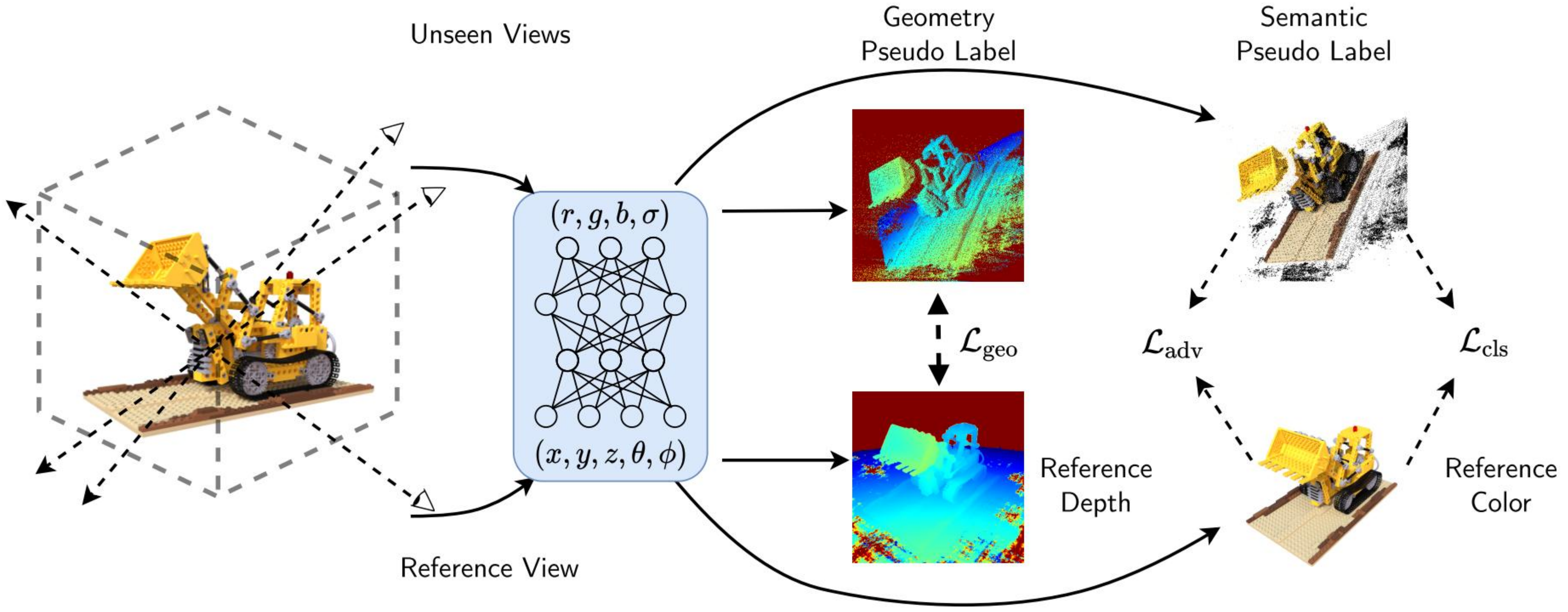
- Single RGB image
- Calibration (intrinsics & extrinsics)
- Depth

Supervision

- Rendering (pixel loss)
- Geometry
- Semantic

METHOD

Overview



$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pix}} + \lambda_1 \mathcal{L}_{\text{geo}} + \lambda_2 \mathcal{L}_{\text{adv}} + \lambda_3 \mathcal{L}_{\text{cls}}$$

METHOD

Geometry supervision

- Reconstruct 3D geometry through warping

Warp from reference view *i* to unseen view *j*

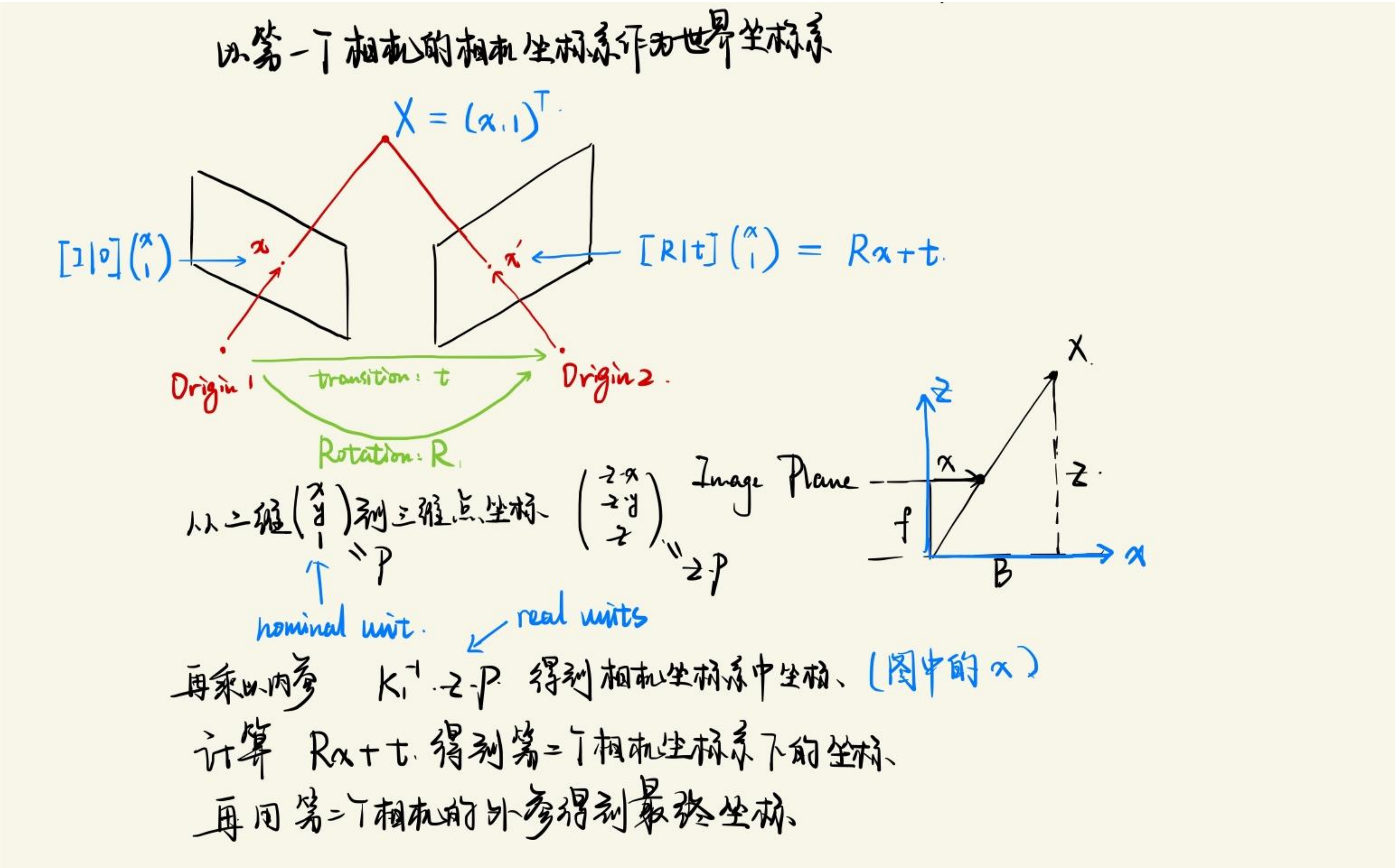
$$p_j = K_{\text{unseen}} T(K_{\text{ref}}^{-1} Z_i p_i)$$

Multi-view geometry revisited

成像原理：（以下都是 Homogeneous Coordinates）

$$\underbrace{\begin{pmatrix} \lambda x \\ \lambda y \\ \lambda \end{pmatrix}}_{\text{图像中像素点的位置}} = \underbrace{\begin{pmatrix} \alpha f & 0 & p_x \\ 0 & \alpha f & p_y \\ 0 & 0 & 1 \end{pmatrix}}_{\text{Intrinsics}} \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}_{\text{Projection}} \underbrace{\begin{pmatrix} R_{3\times 3} & t_{3\times 1} \\ 0 & 1 \end{pmatrix}}_{\substack{\text{Extrinsics} \\ \text{(世界坐标系到相机坐标系)}}} \underbrace{\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}}_{\text{三维坐标}}$$

Multi-view geometry revisited



Multi-view geometry revisited

文中的公式: $P_j = K_{unseen} T \cdot (K_{ref}^{-1} \cdot z_i \cdot p_i)$

实际版本:

① $T_{unseen} T_{ref}^{-1} \begin{pmatrix} K_{ref}^{-1} \cdot z_i \cdot \begin{pmatrix} p_i \\ 1 \end{pmatrix} \\ 1 \end{pmatrix}_{4\times3.}$

② 丢弃最后一维. 再左乘 K_{unseen} (3×1)

③ 从 Homogenous coordinates 转到 real coordinates

同时获得了 depth

METHOD

Geometry supervision: reconstruct 3D geometry through warping

- Warp from reference view *i* to unseen view *j*

$$p_j = K_{\text{unseen}} T(K_{\text{ref}}^{-1} Z_i p_i)$$

- Smoothness contraint

$$\mathcal{L}_{\text{smooth}}(d_i) = e^{-\nabla^2 \mathcal{I}(\mathbf{x}_i)} \left(|\partial_{xx} d_i| + |\partial_{xy} d_i| + |\partial_{yy} d_i| \right)$$

- Consistency contraint

$$\mathcal{L}_{\text{geo}} = \mathcal{L}_1(d_1, f(d_2)) + \mathcal{L}_1(f(d_1), d_2) + \lambda_4 \mathcal{L}_{\text{smooth}}$$

METHOD

Semantic supervision: local texture guidence & global structure guidance

- Discriminate between reference image patches and NeRF outputs
  Differentiable augmentation for collapse avoiding

$$
\begin{aligned}
\mathcal{L}_{\mathrm{D}} &= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}\left[f_D(-D(T(\boldsymbol{x})))\right] + \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[f_D(D(T(G(\boldsymbol{z}))))\right] \\
\mathcal{L}_{\mathrm{G}} &= \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[f_G(-D(T(G(\boldsymbol{z}))))\right],
\end{aligned}
$$

$$\mathcal{L}_{\text{adv}} = \mathcal{L}_{\mathrm{D}} + \mathcal{L}_{\mathrm{G}},$$

- Extract global structure information through ViT

$$\mathcal{L}_{\text{cls}} = ||f_{\text{vit}}(A) - f_{\text{vit}}(B)||^2$$

METHOD

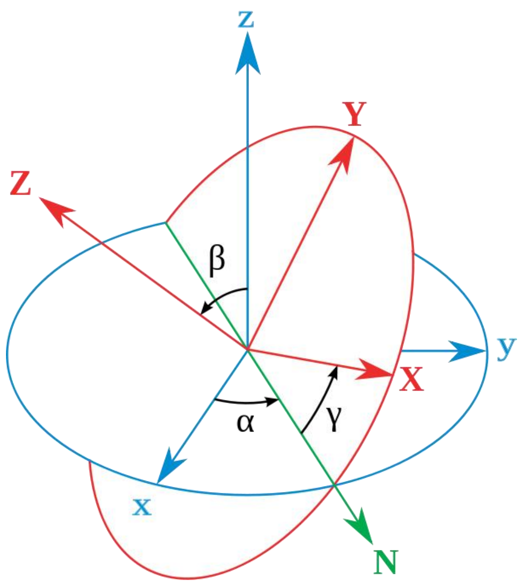
Training Techniques

- Progressive Strided Ray Sampling (from large stride to small stride)

$$\mathcal{P}(u, v, s) = \{(u + sx, v + sy) \mid x, y \in \{0, \dots, K\}\}$$

- Progressive Gaussian Pose Sampling (from small distortion to large distortion)

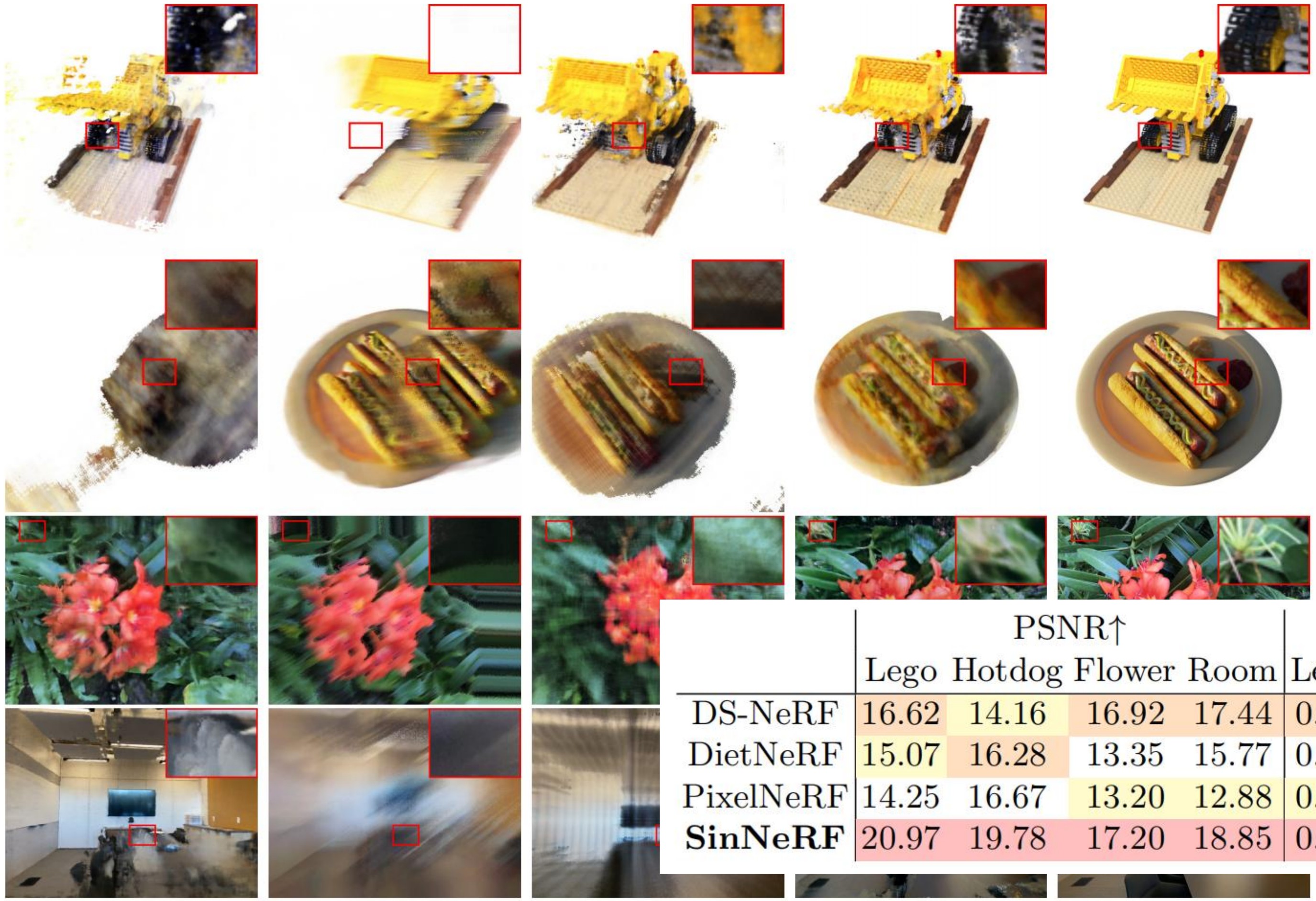
$$(\alpha, \beta, \gamma) \sim \mathcal{N}(0, \omega^2)$$



OUTLINE

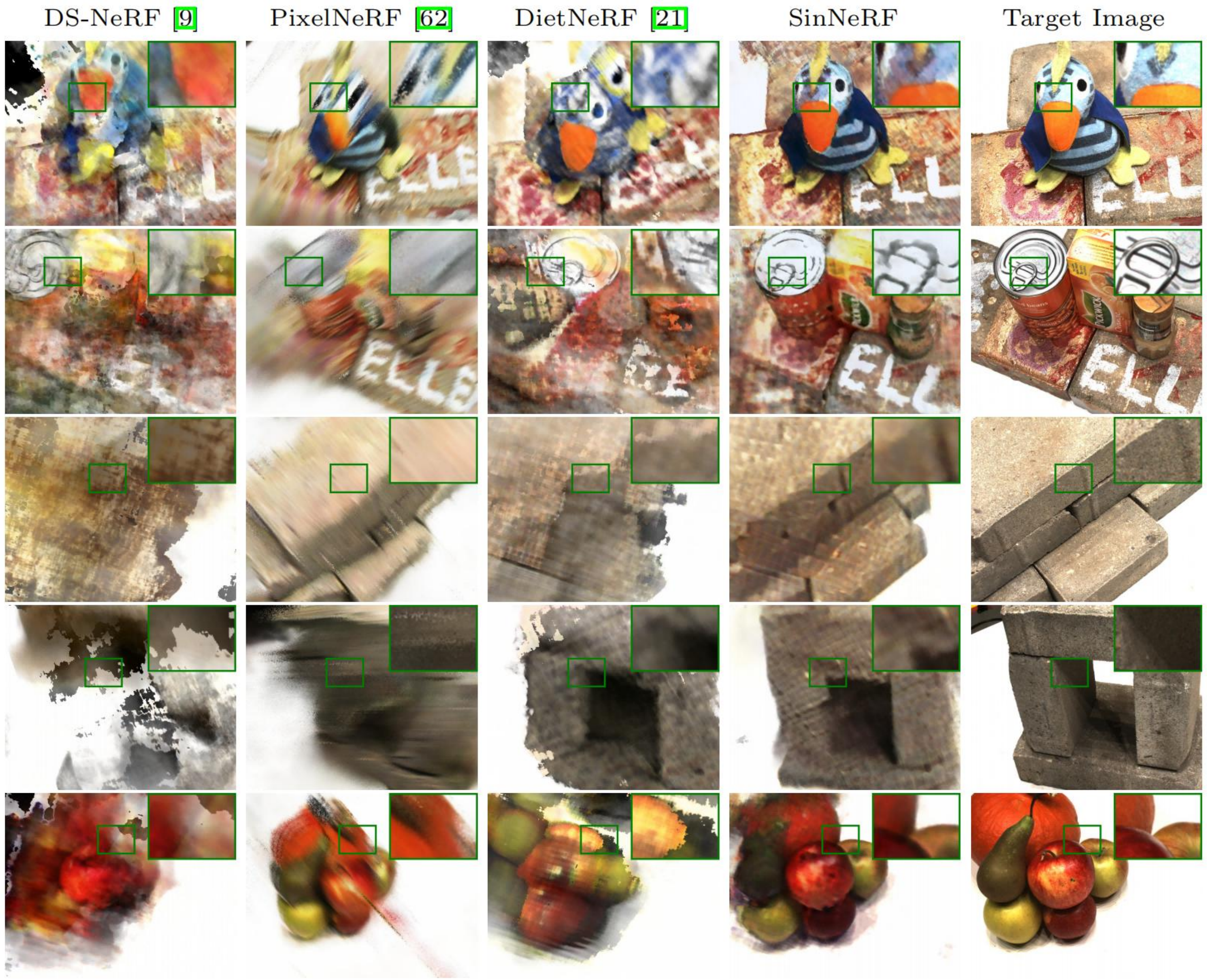
- Authorship
- Background
- Method
- Experiments
- Conclusion

EXPERIMENTS



	PSNR↑				SSIM↑				LPIPS↓			
	Lego	Hotdog	Flower	Room	Lego	Hotdog	Flower	Room	Lego	Hotdog	Flower	Room
DS-NeRF	16.62	14.16	16.92	17.44	0.77	0.67	0.41	0.65	0.1682	0.2956	0.3900	0.3986
DietNeRF	15.07	16.28	13.35	15.77	0.72	0.69	0.20	0.49	0.2063	0.2633	0.7526	0.7512
PixelNeRF	14.25	16.67	13.20	12.88	0.72	0.71	0.19	0.41	0.2171	0.2381	0.6378	0.7633
SinNeRF	20.97	19.78	17.20	18.85	0.82	0.77	0.41	0.67	0.0932	0.1700	0.3724	0.3796

EXPERIMENTS



EXPERIMENTS

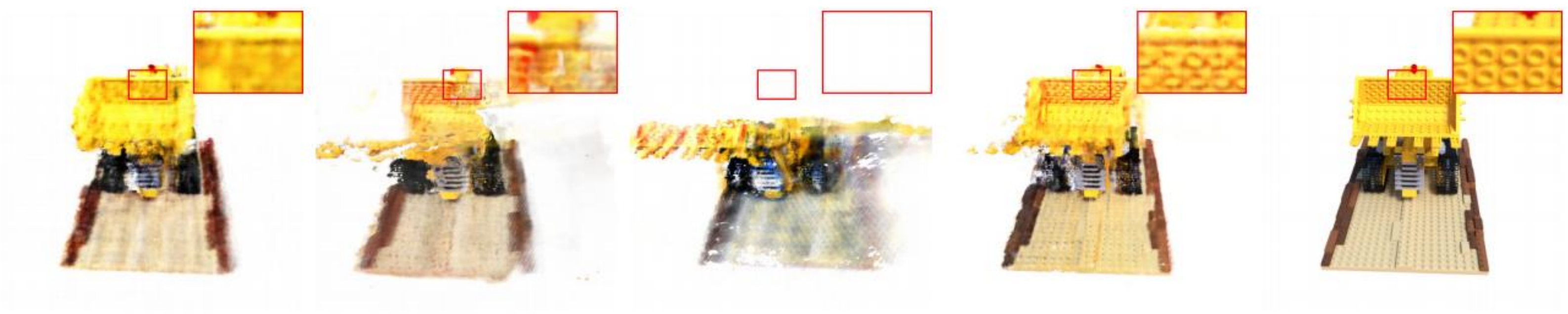


Fig. 5: Novel view synthesis from different variants of our proposed model.

Methods	PSNR↑	SSIM↑	LPIPS↓
w/o $\mathcal{L}_{geo}$	16.11 (-4.86)	0.74 (-0.08)	0.1919 (+0.0987)
w/o $\mathcal{L}_{cls}$	18.20 (-2.77)	0.76 (-0.06)	0.1348 (+0.0146)
w/o $\mathcal{L}_{adv}$	20.20 (-0.77)	0.79 (-0.03)	0.1306 (+0.0294)
Full Model	**20.97**	**0.82**	**0.0932**

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

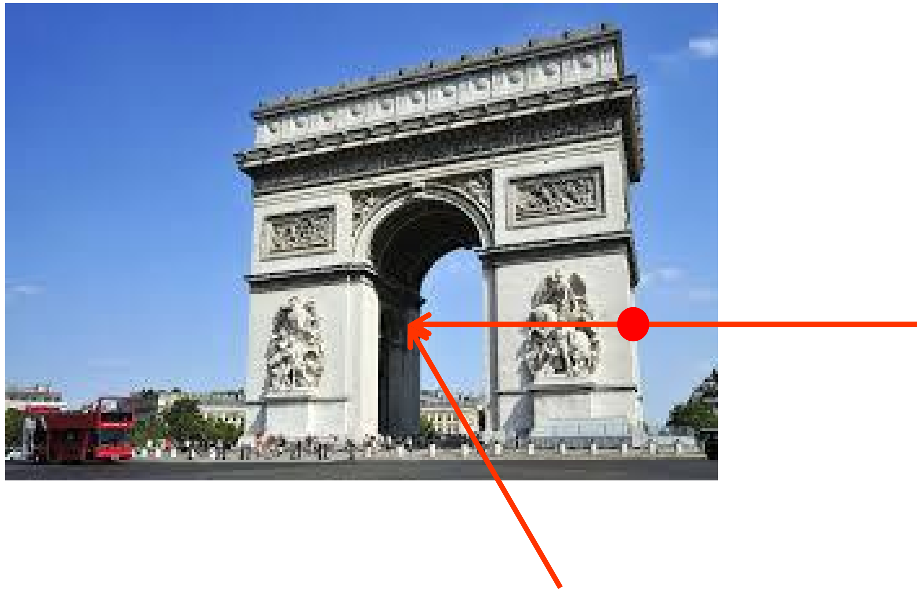
CONCLUSION

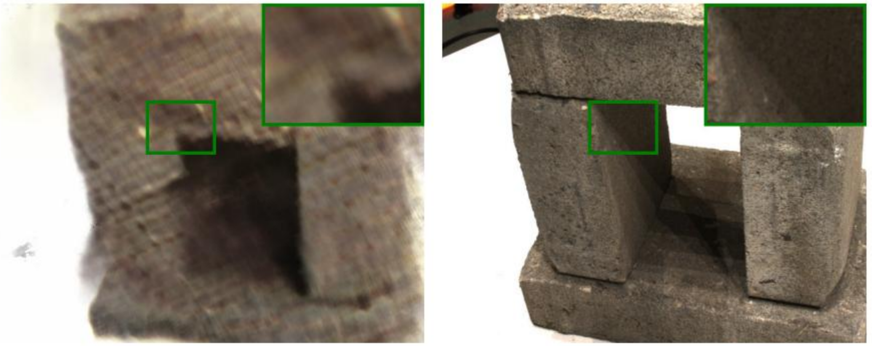
Pros

- The authors proposed SinNeRF, a novel view synthesis approach which only requires one calibrated view with depth
- SoTA on multiple benchmarks

Cons

- Test time optimization (same as original NeRF), i.e., one model per scene
- Simple scene
- Very complex design
- Cannot handle occlusion





Thanks for listening!