- 2023 Fall EECS598-007 -

pixelNeRF paper review and description

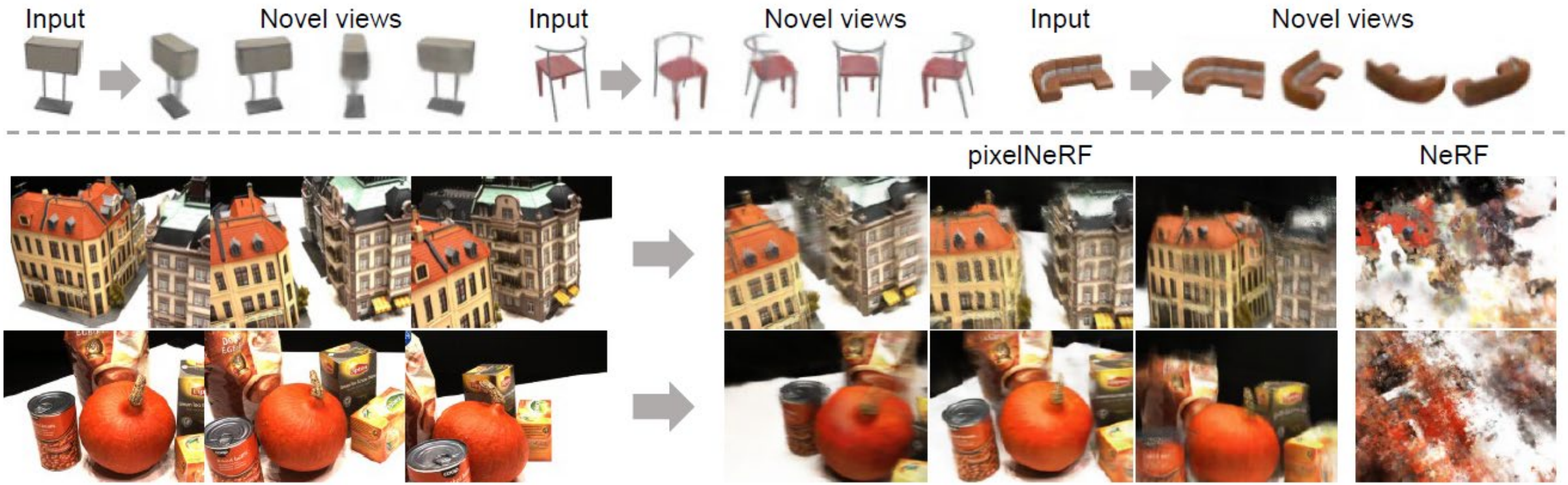
Wonseok Oh

Umich ECE master's student

Contents

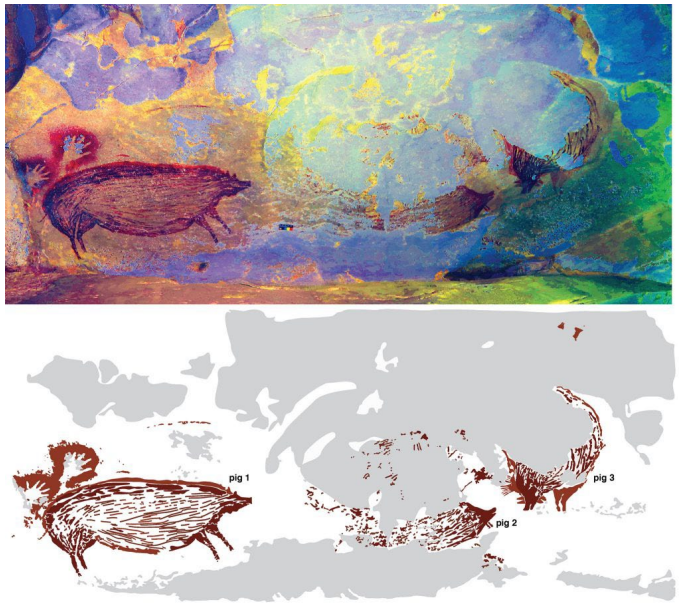
- Introduction
  - Problem definition (view synthesis)
- Motivation
  - Related work
  - Idea
- Methodology
  - Single-Image pixelNeRF
  - Multi-view pixelNeRF
- Experiments
- Conclusion
- Quizzes

Introduction



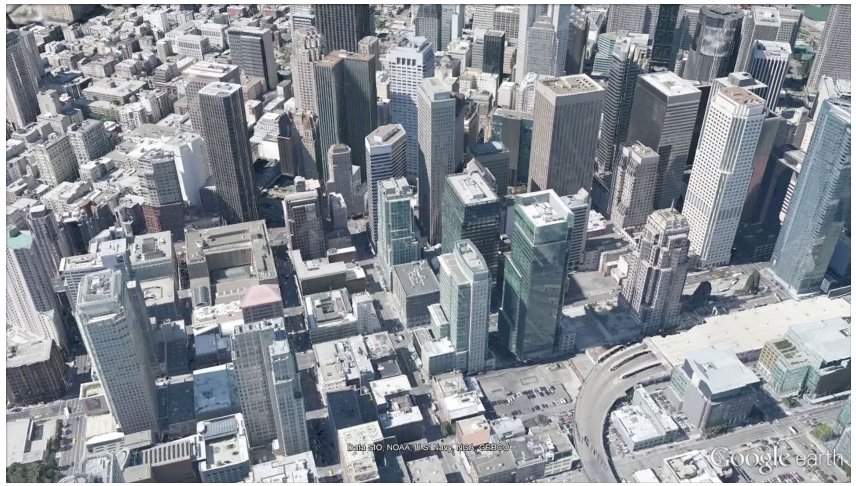
Introduction

- What is this? (Capturing the reality)



Earliest cave painting (45,500 years old) in Sulawesi, Indonesia

Monet's Cathedral series: study of light 1893-1894

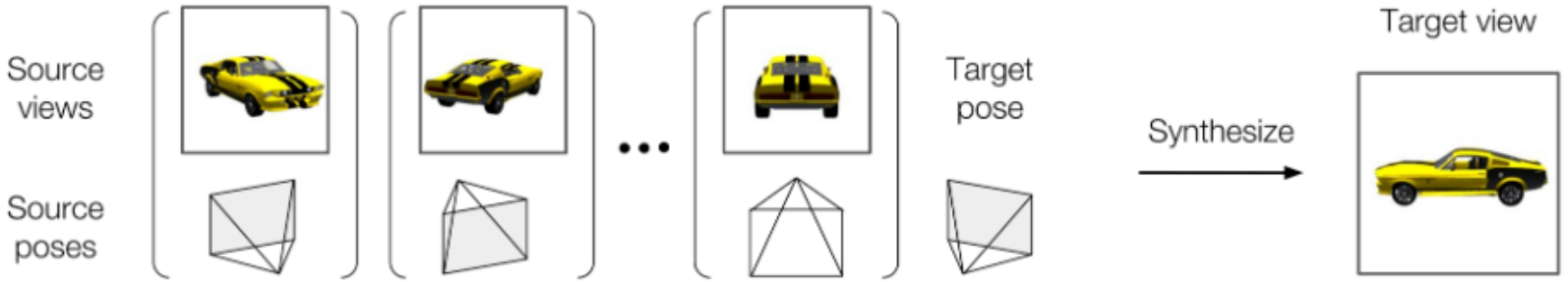


Google Earth 2016~

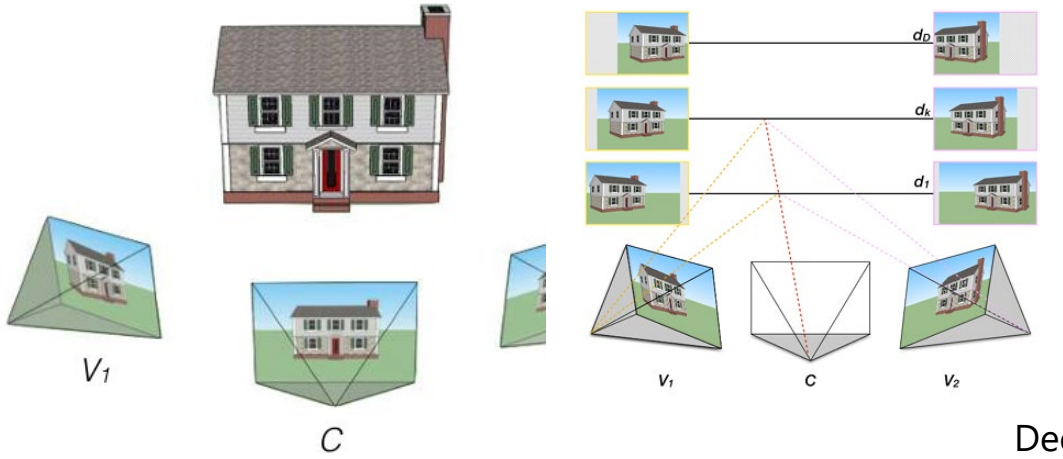
Images : A.Kanazawa

Problem definition

- View Synthesis



Multi-view to novel view [Sun et al., ECCV 2018]



DeepStereo [Fylnn et al., CVPR 2016]

Motivation

- Related Work (NeRF)

  Problem statement

Input: A set of calibrated Images

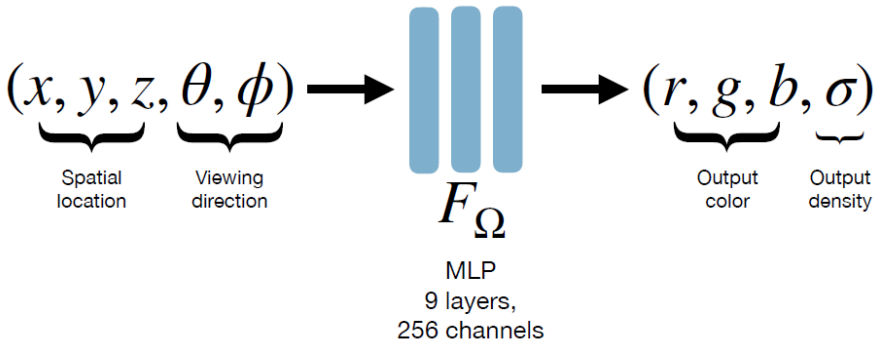
Output: A 3D scene representation that renders novel views

NeRF [B. Mildenhall et al., ACM 2021]

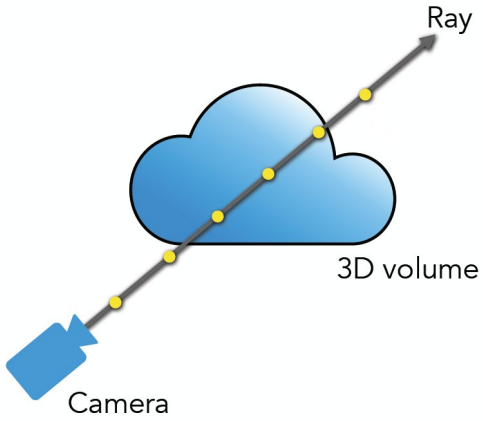
Motivation

- Related Work (NeRF)

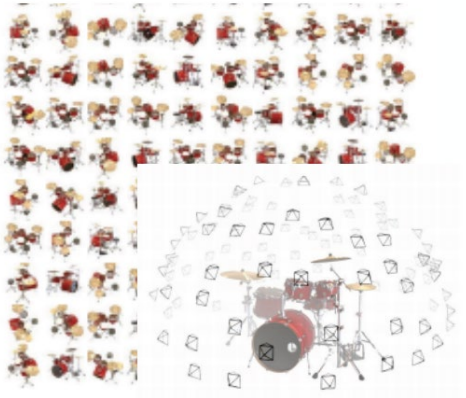
Three Key Components



Neural Volumetric
3D Scene Representation



Differentiable Volumetric
Rendering Function



Optimization via Analysis-by-Synthesis
Objective: Synthesize all training views

Images : A.Kanazawa

Motivation

- Representing a 3D scene as a continuous 5D function

The function here is defined as deep neural network

$$F_\Theta : (X, d) \rightarrow (c, \sigma)$$

- Input: position of pixel $X \in \mathbb{R}^3$ and the viewing direction unit vector $d \in \mathbb{R}^2$
- Output: color value c and density value σ

NeRF is a model for view synthesis

1. This is task of creating a 2D image of an object from a new angle by reconstructing light and perspective from N 2D images taken by a camera.

2. Creating a 2D image from a new angle means modeling the entire 3D object.

3. For this modeling, NeRF uses the neural radiance field (≈neural implicit representation), which is "a function that computes the RGB value of each pixel's coordinates given a value as input."

NeRF [B. Mildenhall et al., ACM 2021]

Motivation

- Volumetric Rendering

$$\hat{C}_r = \int_{t_n}^{t_f} T(t)\sigma(t)c(t)dt$$

How to volumetric Rendering using ***color*** and ***density?***

1. The color value computed by the function is the RGB value in three-dimensional coordinates.

2. To create a 2D image from different angles, we need to take into account things that are in front of us (when viewed from that direction) and things that are behind us.

3. Taking all of this into account, the formula for converting RGB values in three dimensions to RGB values in a 2D image is shown Left.

NeRF [B. Mildenhall et al., ACM 2021]

Motivation

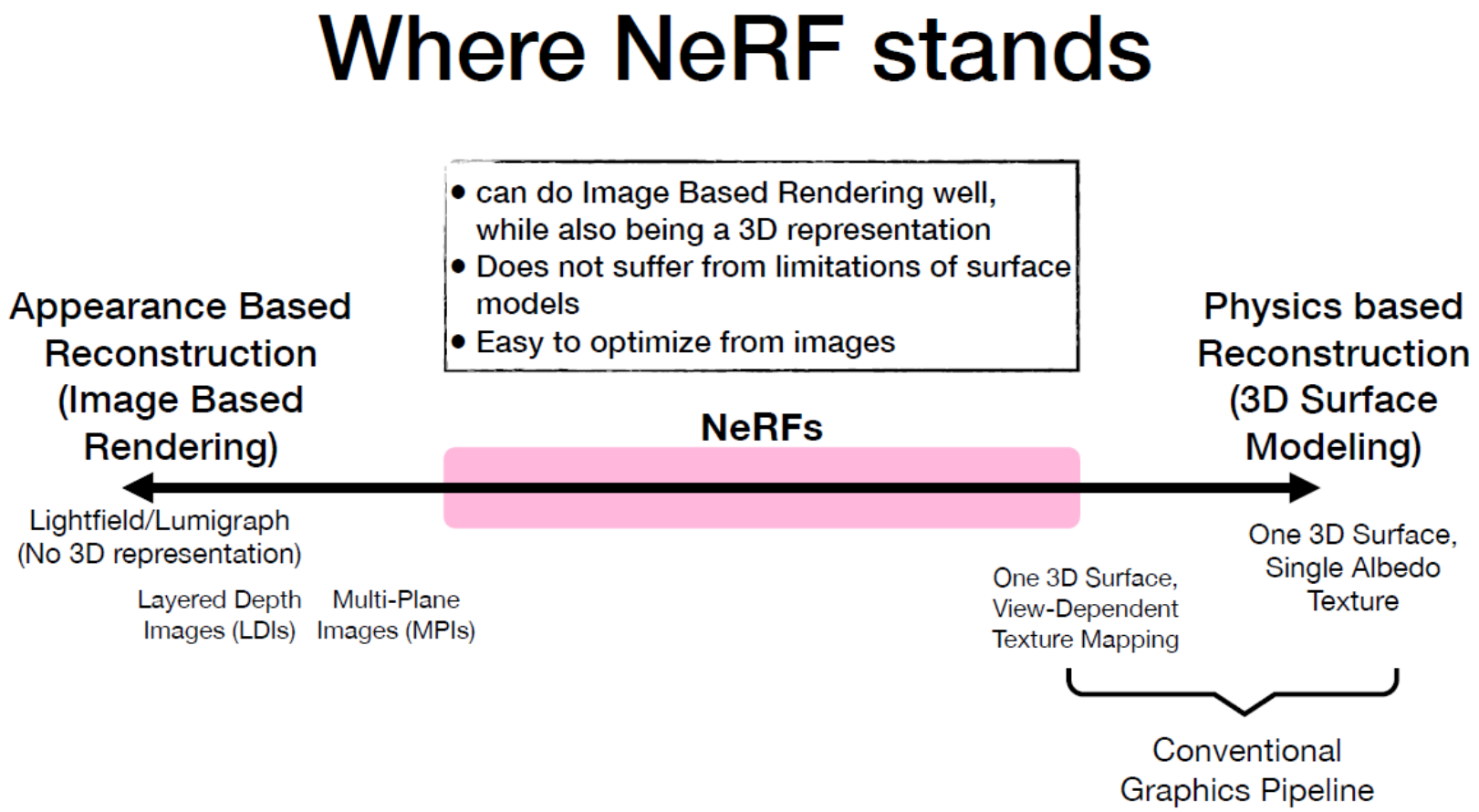
- Volumetric Rendering

$$\hat{C}_r = \int_{t_n}^{t_f} T(t)\sigma(t)c(t)dt$$

- camera ray $r(t) = o + td$
  - t: how far the object is (from the focus)
  - d: viewing direction unit vector
  - o: origin
- $T(t) = exp(-\int_{t_n}^{t} \sigma(s)ds)$
  : summation of the density values of the points blocking point t. ($\approx$ The probability that light rays will move from t_n to t without hitting other particles.)
- $\sigma(t)$: density value on point t
- $c(t)$: RGB value on point t

NeRF [B. Mildenhall et al., ACM 2021]

Motivation



Slide [A.Kanazawa et al., ECCV 2022]

Motivation - Idea

	NeRF	DISN	ONet	DVR	SRN	Ours
Learns scene prior?	✗	✓	✓	✓	✓	✓
Supervision	2D	3D	3D	2D	2D	2D
Image features	✗	Local	Global	Global	✗	Local
Allows multi-view?	✓	✓	✗	✗	✓	✓
View space?	-	✗	✗	✗	✗	✓

pixelNeRF [A. Yu et al., CVPR 2021]

Motivation - Idea

- Limitations?

1. Requires multiple angles of a single object to compose high quality image

2. Long time to optimize the model

- ➢ pixelNeRF addresses these limitations of NeRF and proposes a way to generate images from a new viewpoint in a much shorter amount to time with fewer images
- ➢ How?

pixelNeRF [A. Yu et al., CVPR 2021]

Methodology

- Two parts of Model

1. Fully-convolutional image encoder E

   The part that encodes the input image as a pixel-aligned feature

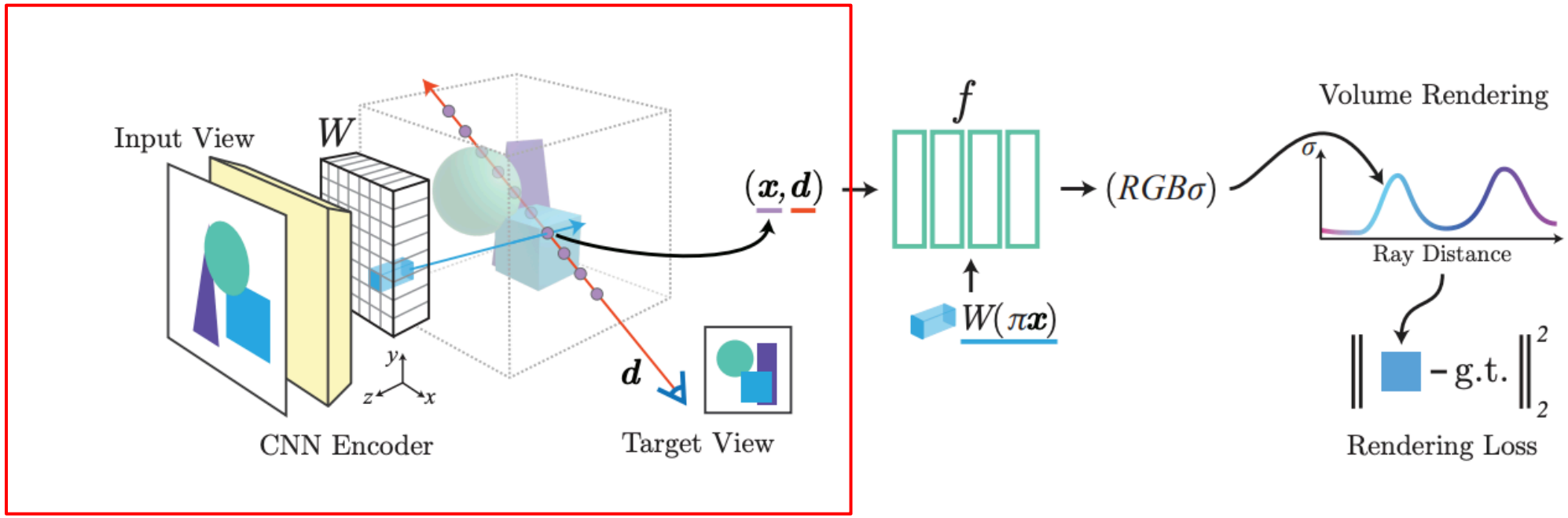
2. NeRF network

   The part that computes the color and density of an object

➢ The output of the encoder goes into the input of the nerf network

pixelNeRF [A. Yu et al., CVPR 2021]

Single-Image pixelNeRF (Single-shot learning method)



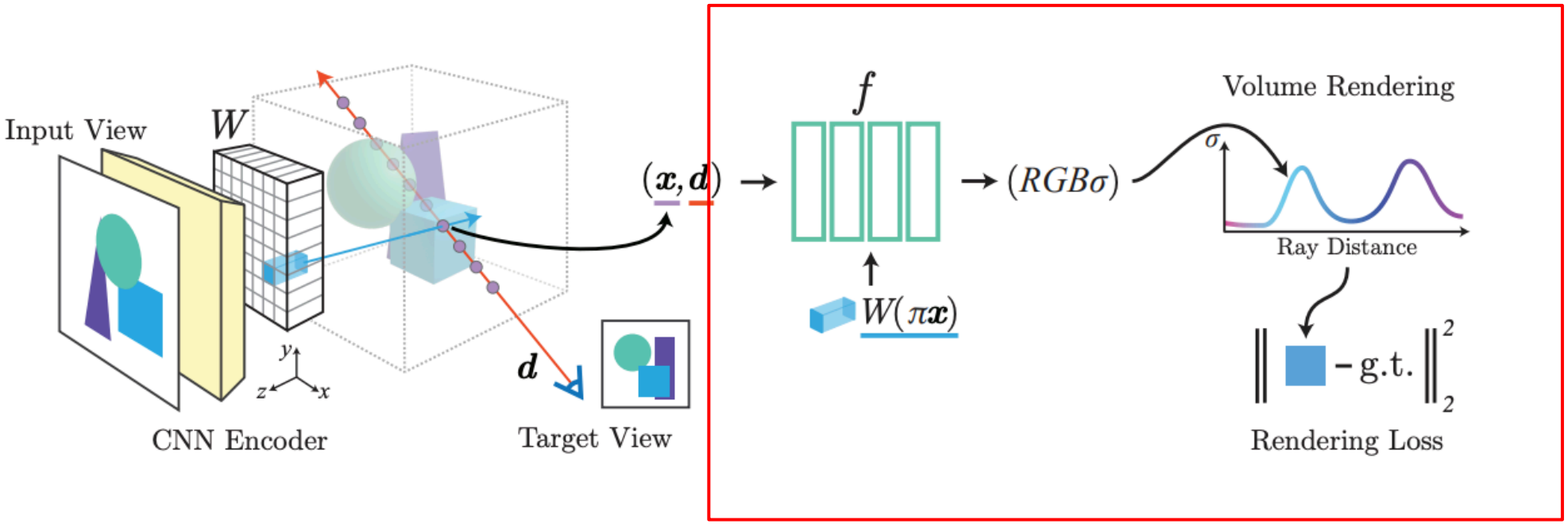
1. Put the Input Image *I* into the encoder E to extract the spatial feature vector W.

2. For the points on the camera ray *x*, we obtain the image feature corresponding to each.

- Project the camera ray *x* to the image plane and compute the corresponding coordinate $\pi(x)$.
- Use bilinear interpolation to obtain the spatial feature $W(\pi(x))$ corresponding to this coordinate

pixelNeRF [A. Yu et al., CVPR 2021]

Single-Image pixelNeRF (Single-shot learning method)



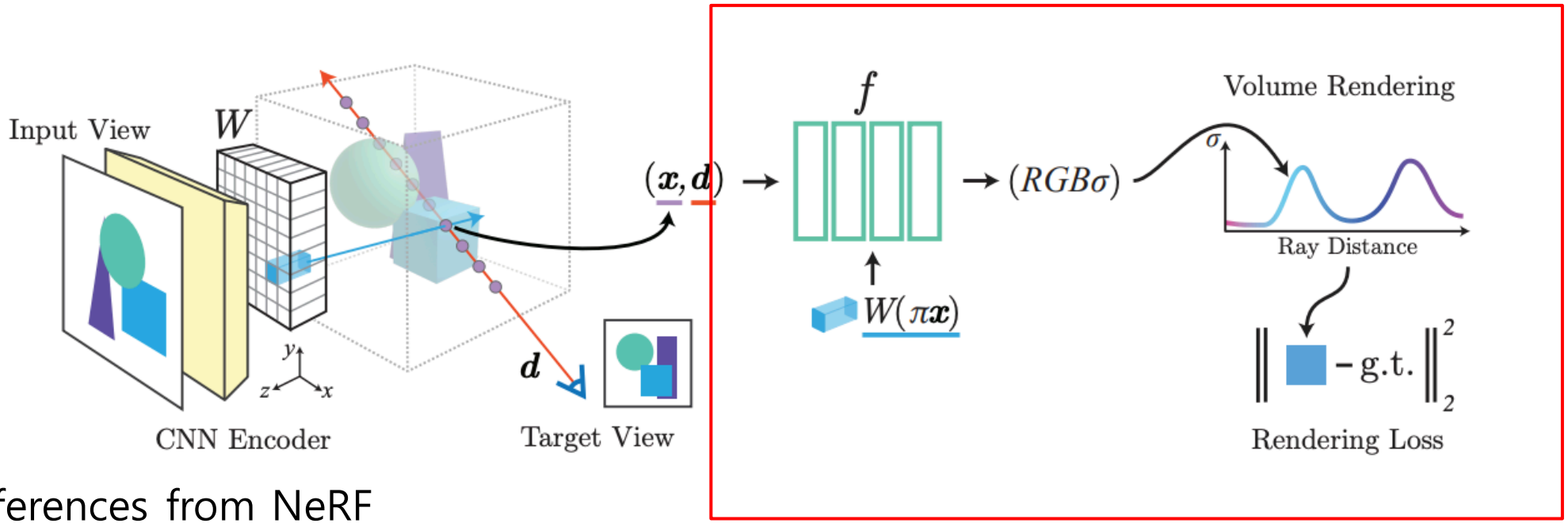
3. Put the $W(\pi(x))$, $\gamma(x)$, d into the NeRF network and obtain the color and density values as follows

$$f(\gamma(x), d; W(\pi(x))) = (\sigma, c)$$

4. Volumetric Rendering (Same way as NeRF)

pixelNeRF [A. Yu et al., CVPR 2021]

Single-Image pixelNeRF (Single-shot learning method)



Main differences from NeRF

Extracts the ***spatial features*** of the input image through pre-processing and adds them to the nerf network.

This inclusion of feature information allows the network to ***implicitly learn*** the organic relationship between ***individual pixel-level*** information, which enables it to make stable and accurate inferences with less data.

pixelNeRF [A. Yu et al., CVPR 2021]

Multi-view pixelNeRF (few-shot view synthesis)

We can view the importance of a particular image feature!

- If the input view and target direction are similar?
  The model can infer based on data learned by input

- Otherwise
  The existing learned prior will have to be utilized.

pixelNeRF [A. Yu et al., CVPR 2021]

Multi-view pixelNeRF (few-shot view synthesis)

Single-view vs Multi-view

1. To solve the multi-view task, the author first assumes that the relative camera positions of each image are known.

2. In each image $I(i)$, we convert the objects at the origin to coordinates at the target angle that we want to see.

$$P^{(i)} = [R^{(i)}\ t^{(i)}],\ x^{(i)} = P^{(i)}x,\ d^{(i)} = R^{(i)}d$$

pixelNeRF [A. Yu et al., CVPR 2021]

Multi-view pixelNeRF (few-shot view synthesis)

Single-view vs Multi-view

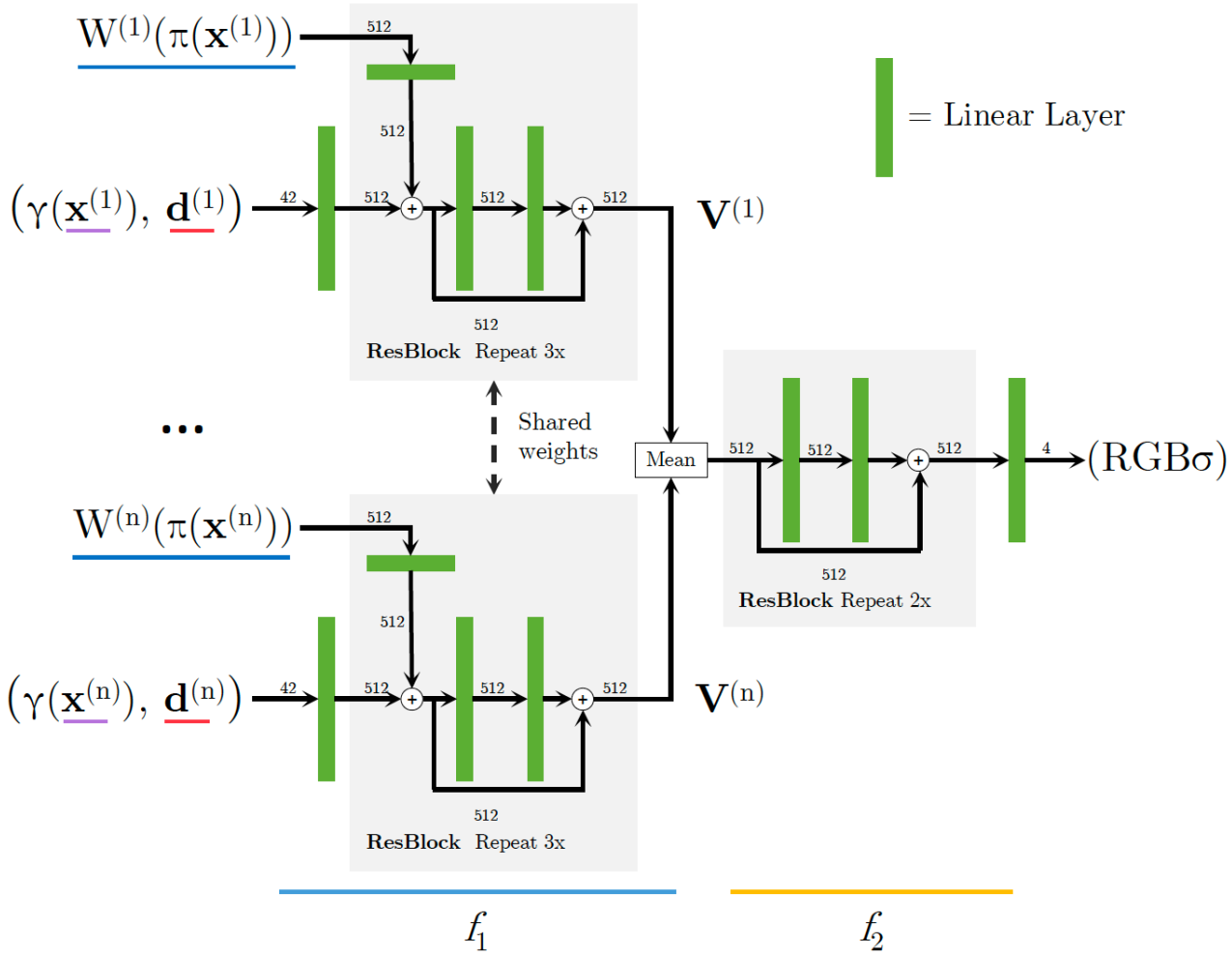
3. While selecting features through encoder, each view frame is pulled ***independently***, put into the NeRF network, and merge in the final layer of the NeRF network. This is to extract as many spatial features as possible from images from various angles.

$$\mathbf{V}^{(i)} = f_1\left(\gamma(\mathbf{x}^{(i)}), \mathbf{d}^{(i)}; \mathbf{W}^{(i)}(\pi(\mathbf{x}^{(i)}))\right) \qquad (\sigma, \mathbf{c}) = f_2\left(\psi\left(\mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(n)}\right)\right)$$

f_1 $V^{(i)}$ f_2 Note that these are initial, intermediate, final layer respectively

pixelNeRF [A. Yu et al., CVPR 2021]

Multi-view pixelNeRF (few-shot view synthesis)



pixelNeRF [A. Yu et al., CVPR 2021]

Experiments

- **Baseline & Datasets**

  Compared with SRN and DVR, which were SOTA models of the existing new-shot/single-shot view synthesis, and NeRF using a network of similar structures.

  Experiment on both ShapeNet, a benchmark dataset for 3D objects, and DTU datasets that resemble more real-life pictures, showing the performance of pixelNeRF.

- **Baseline & Datasets**

  For the performance indicator, widely used image qualifying metrics(PSNR, SSIM) are used.

  PSNR: $10log_{10}(\frac{R^2}{MSE})$

  SSIM: $\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$

pixelNeRF [A. Yu et al., CVPR 2021]

Experiments

- Evaluated pixelNeRF on category-specific and category-agnostic view synthesis task on ShapeNet

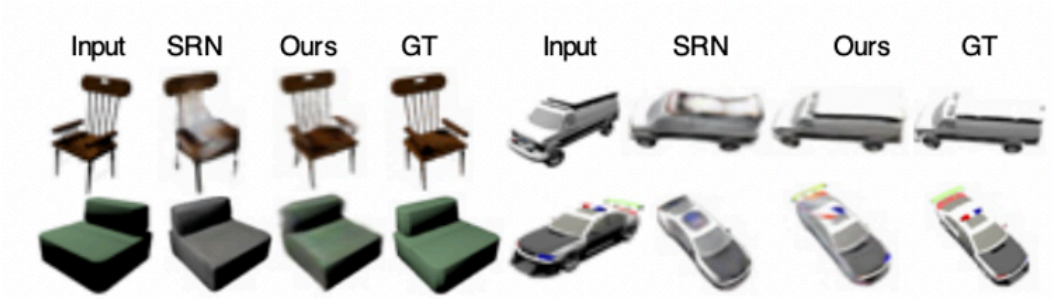


Figure 3: **Category-specific single-view reconstruction benchmark**. We train a separate model for cars and chairs and compare to SRN. The corresponding numbers may be found in Table 2.

		1-view		2-view	
		PSNR	SSIM	PSNR	SSIM
Chairs	GRF [43]	21.25	0.86	22.65	0.88
	TCO [40] *	21.27	0.88	21.33	0.88
	dGQN [9]	21.59	0.87	22.36	0.89
	ENR [8] *	22.83	-	-	-
	SRN [39]	22.89	0.89	24.48	0.92
	Ours *	**23.72**	**0.91**	**26.20**	**0.94**
Cars	SRN [39]	22.25	0.89	24.84	0.92
	ENR [8] *	22.26	-	-	-
	Ours *	**23.17**	**0.90**	**25.66**	**0.94**

Table 2: **Category-specific 1- and 2-view reconstruction**. Meth-

pixelNeRF [A. Yu et al., CVPR 2021]

Experiments

- Showed that view synthesis is also applicable to unseen categories or multi-object data in ShapeNet data, using pretrained parameters

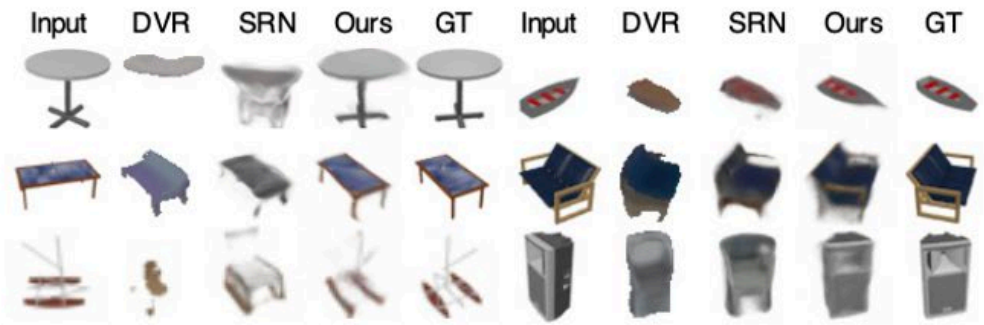


Figure 6: **Generalization to unseen categories.** We evaluate a

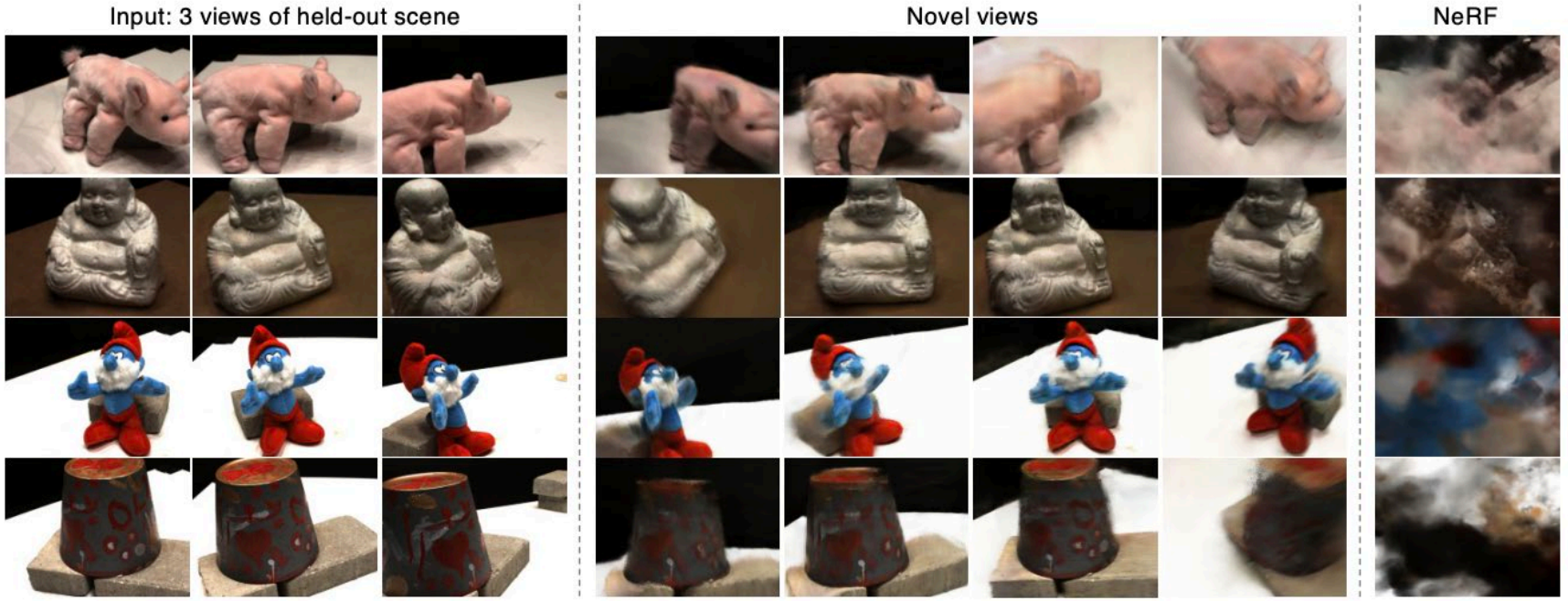
	Unseen category			Multiple chairs		
	↑ PSNR	↑ SSIM	↓ LPIPS	↑ PSNR	↑ SSIM	↓ LPIPS
DVR	17.72	0.716	0.240	-	-	-
SRN	18.71	0.684	0.280	14.67	0.664	0.431
Ours	**22.71**	**0.825**	**0.182**	**23.40**	**0.832**	**0.207**

Table 5: **Image quality metrics for challenging ShapeNet tasks.**

pixelNeRF [A. Yu et al., CVPR 2021]

Experiments

- View synthesis in Real view datasets (DTU, MVS)



pixelNeRF [A. Yu et al., CVPR 2021]

Conclusion

•pixelNeRF : Image conditioning enables learning powerful scene priors to reconstruct NeRFs from single(1)-fiew(3) views across objects, classes, and scenes.

•Practical Implications: Could enable photorealistic VR/AR from single view to few images and scalable content creation.

•Potential Impact: If extended to real-world datasets, could provide breakthroughs in 3D capture and graphics applications.

pixelNeRF [A. Yu et al., CVPR 2021]

Quizzes

1. What is the word inside the square below?

Main differences from NeRF

Extracts the  of the input image through pre-processing and adds them to the nerf network.

2. When using pixelNeRF, the 3D scene itself is formed compared to nerf, which does not form a 3D scene at all in the case of a single image. However, it is hard to say that it is a 3D scene in itself. What are your ideas for solving this? (hint : diffusion network)

pixelNeRF [A. Yu et al., CVPR 2021]

END

Thank you!