Photo-realistic 3D-aware Scene Generation

Lingjie Liu

Postdoc at Max Planck Institute for Informatics
Incoming Assistant Professor at the University of Pennsylvania

max planck institut informatik

Penn
UNIVERSITY of PENNSYLVANIA

Photo-realistic 3D-aware Scene Generation

Lingjie Liu

Postdoc at Max Planck Institute for Informatics
Incoming Assistant Professor at the University of Pennsylvania

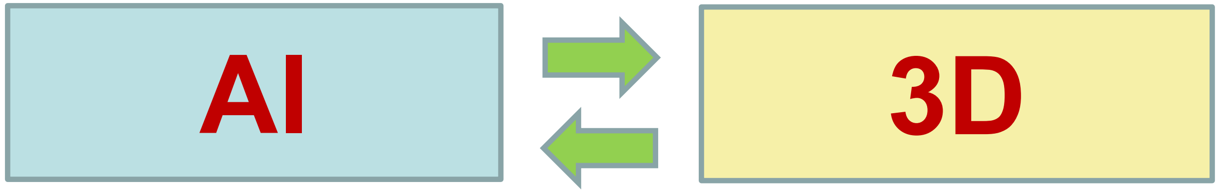
max planck institut
informatik

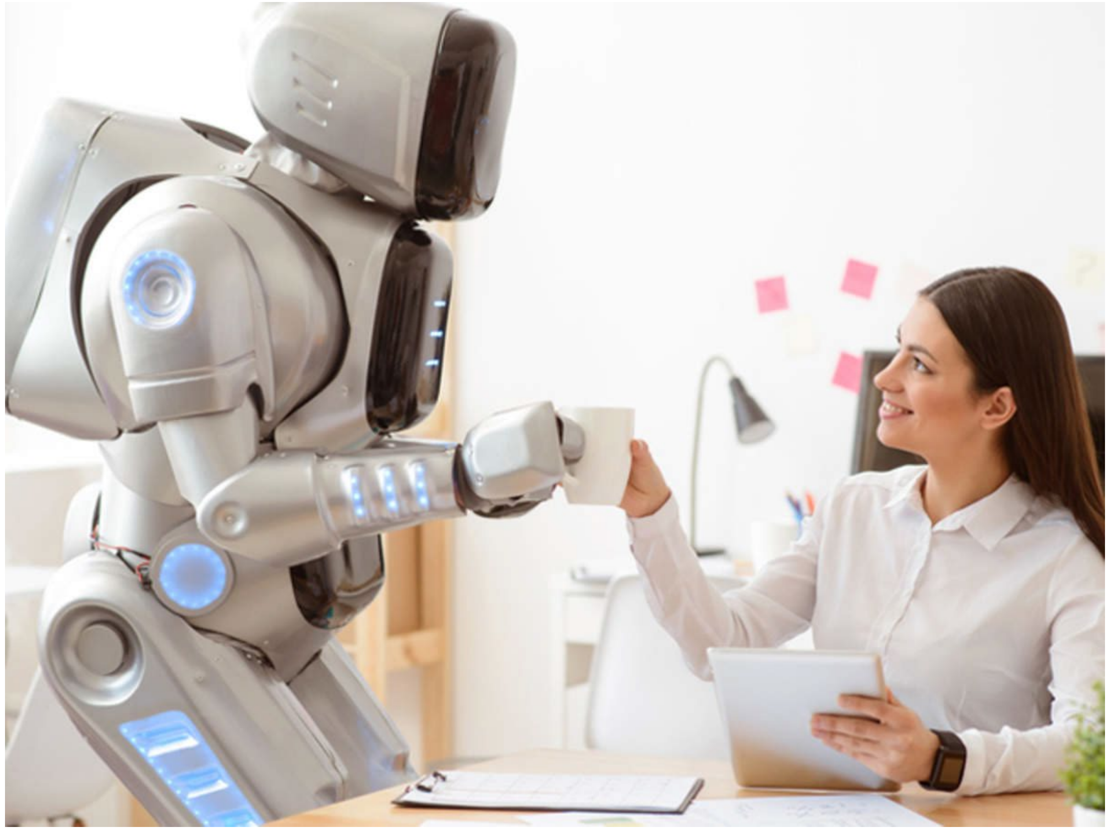
Penn
UNIVERSITY of PENNSYLVANIA

We Digitize Our World in 3D

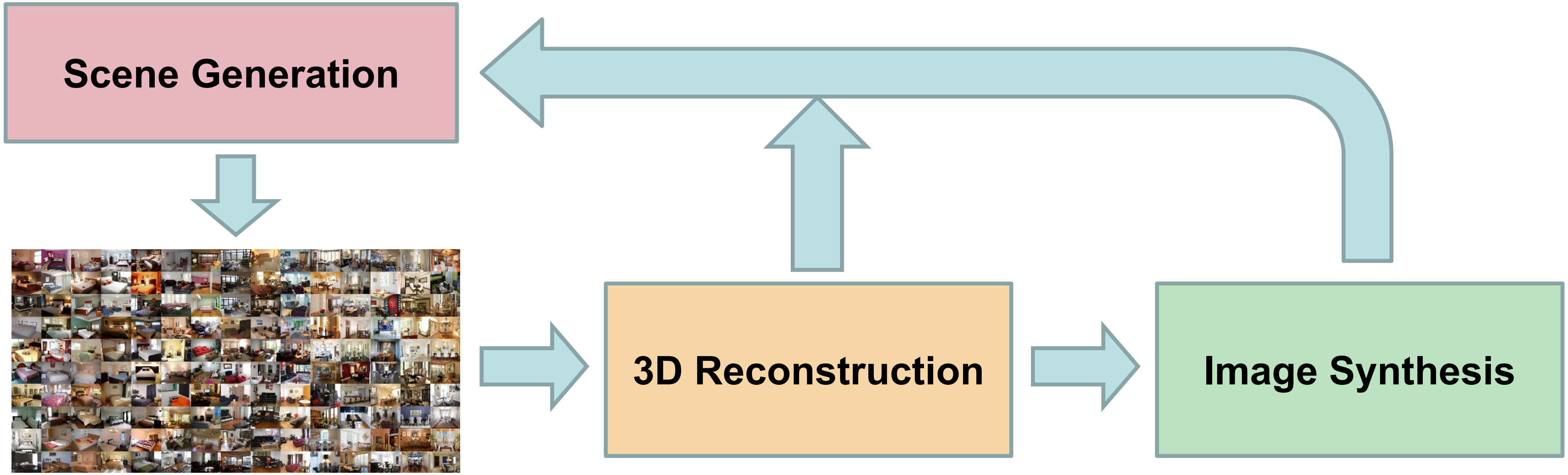
Future AI: Towards 3D Aware

Long-term Vision





Long-term Vision



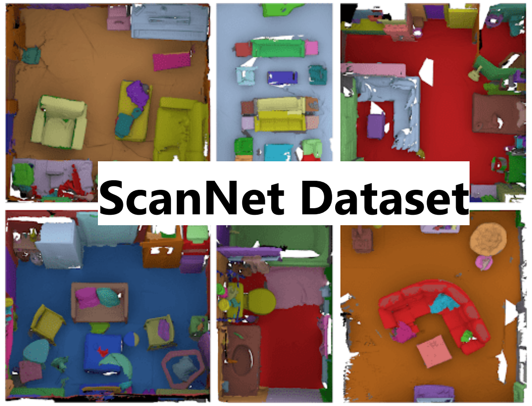
Bottleneck of Existing 3D Learning Models is the Lack of 3D Data

The size of 2D datasets can be as large as millions

The ImageNet dataset contains **millions** of images

Existing 3D data is far from sufficient

12k synthetic models



1k indoor scenes

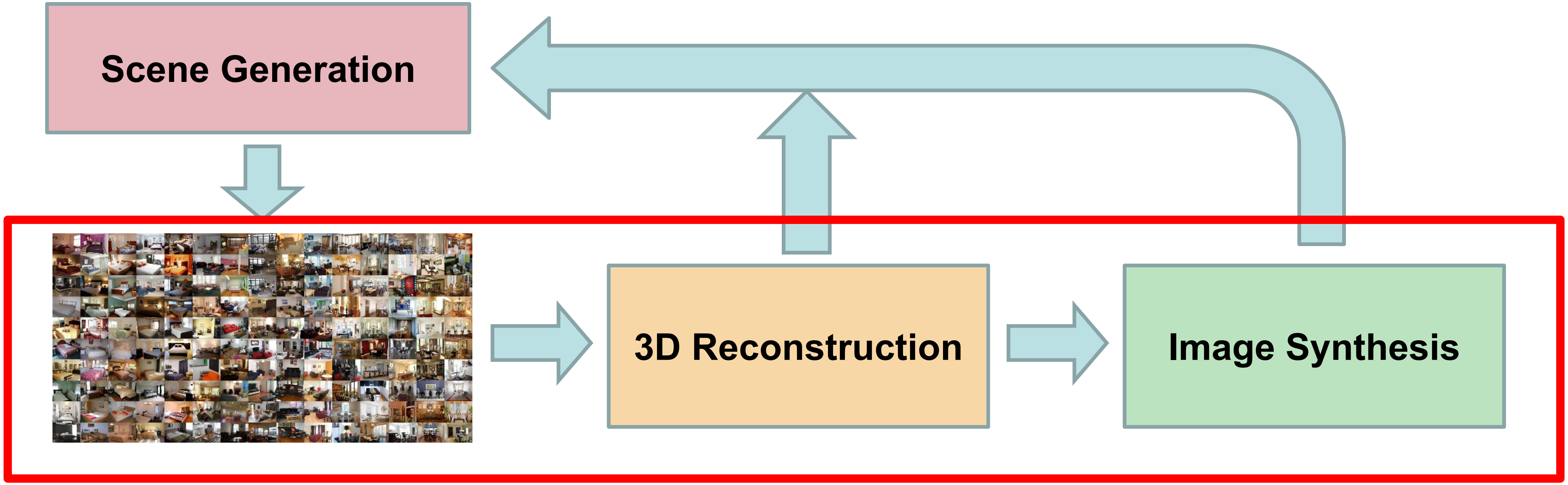
1k indoor scenes

7.2k demonstrations of a robot performing kitchen tasks

Lingjie Liu

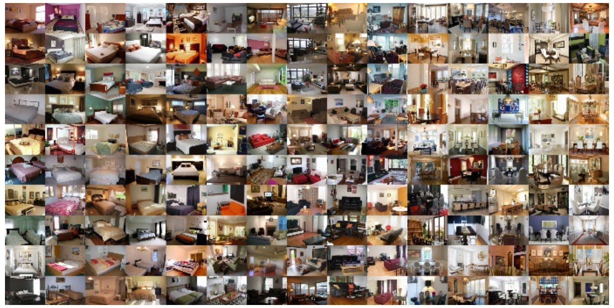
Why Challenging?

3D Reconstruction and Image Synthesis are Challenging



Lingjie Liu

Classical Computer Graphics Pipeline



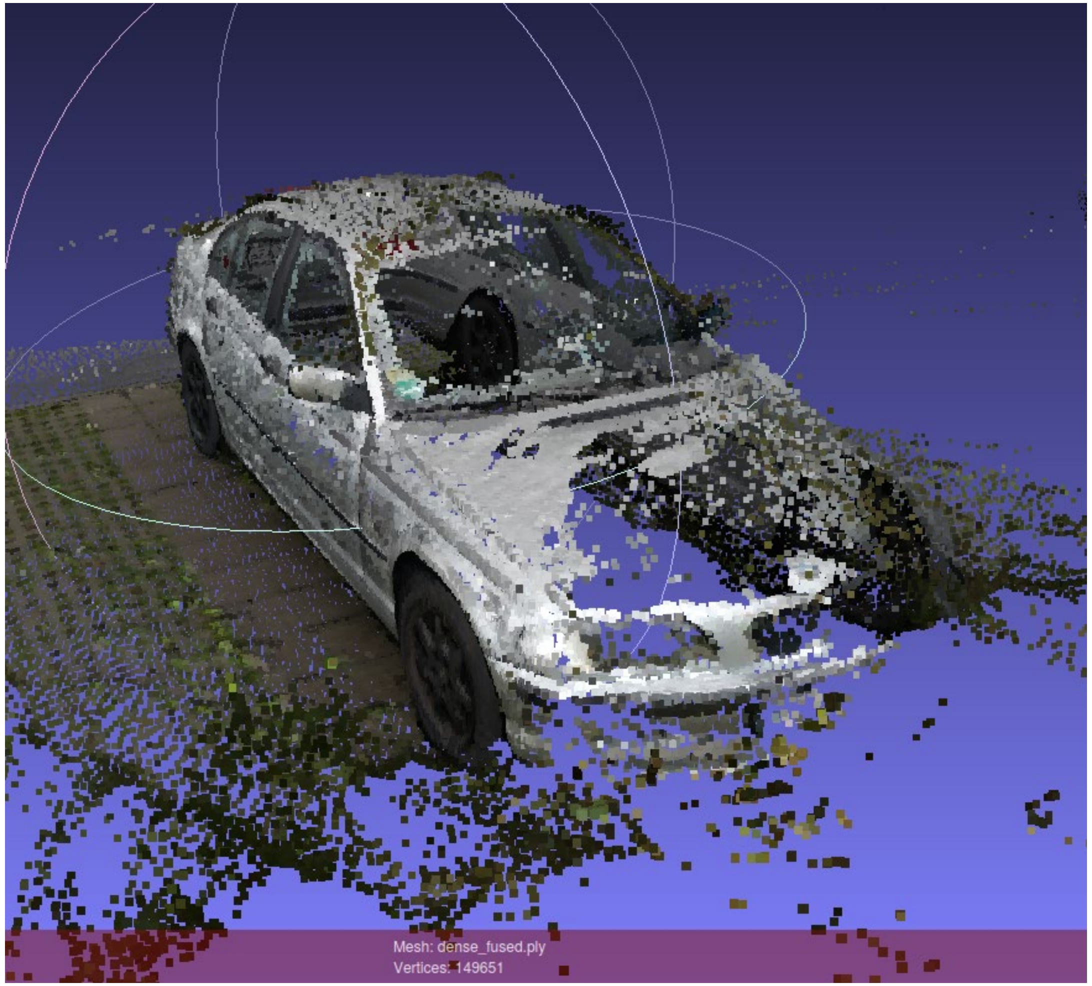
3D Reconstruction

Image-based 3D Reconstruction

Image Synthesis

Computer Graphics Rendering

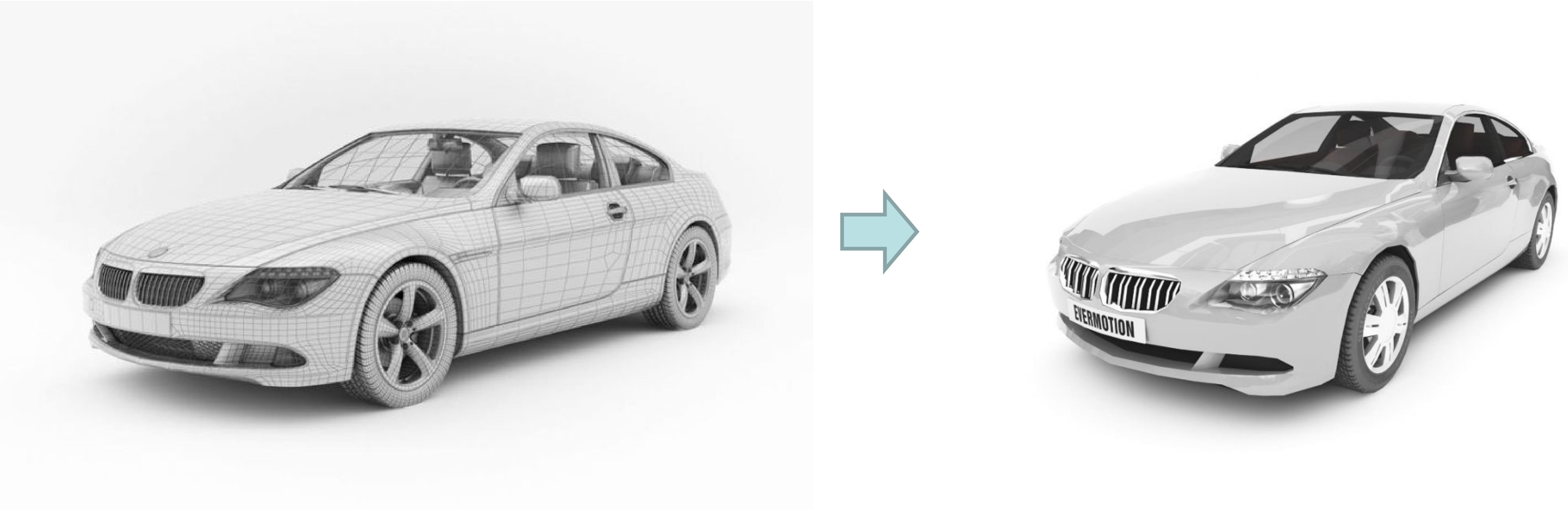
Image-based 3D Reconstruction



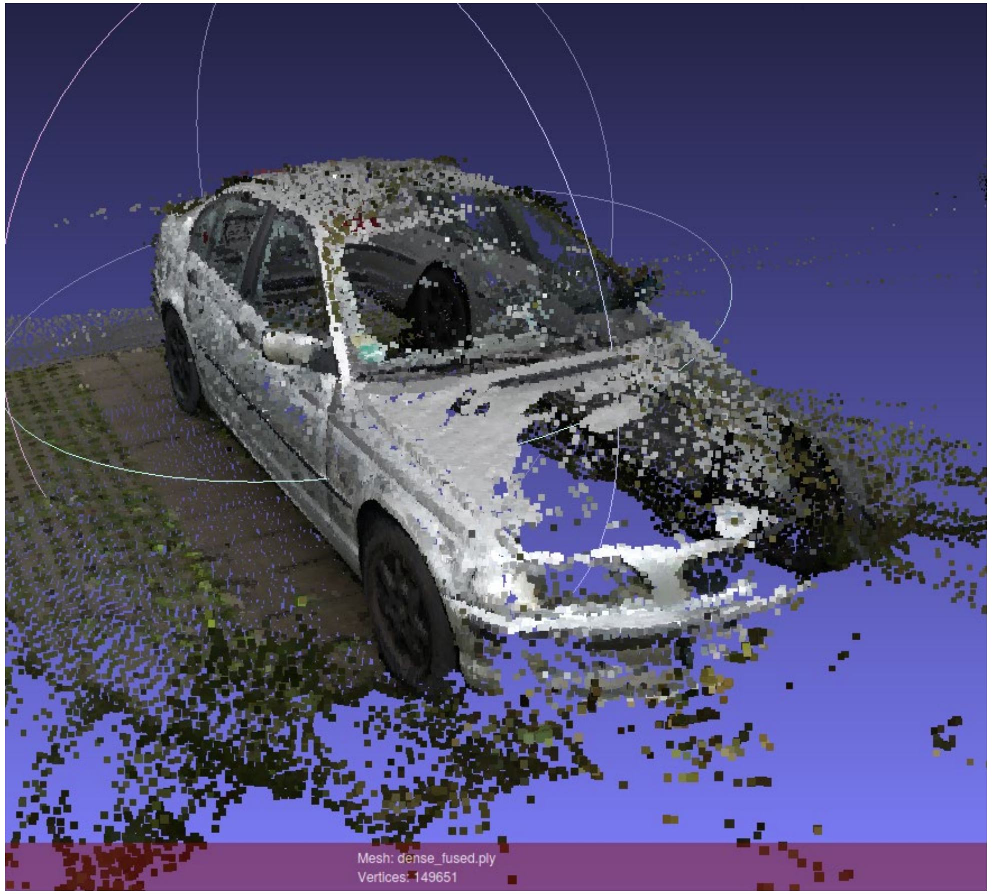
COLMAP [Johannes et al. 2016, Schoenberger et al. 2016]
(Input: 100 images)

Computer Graphics Rendering

Rendering requires very high-quality 3D models



Lingjie Liu



VS

Output of Image-based Reconstruction

Required Input for Photo-realistic Rendering

Photo-realistic Large-scale Scene Generation is Extremely Challenging

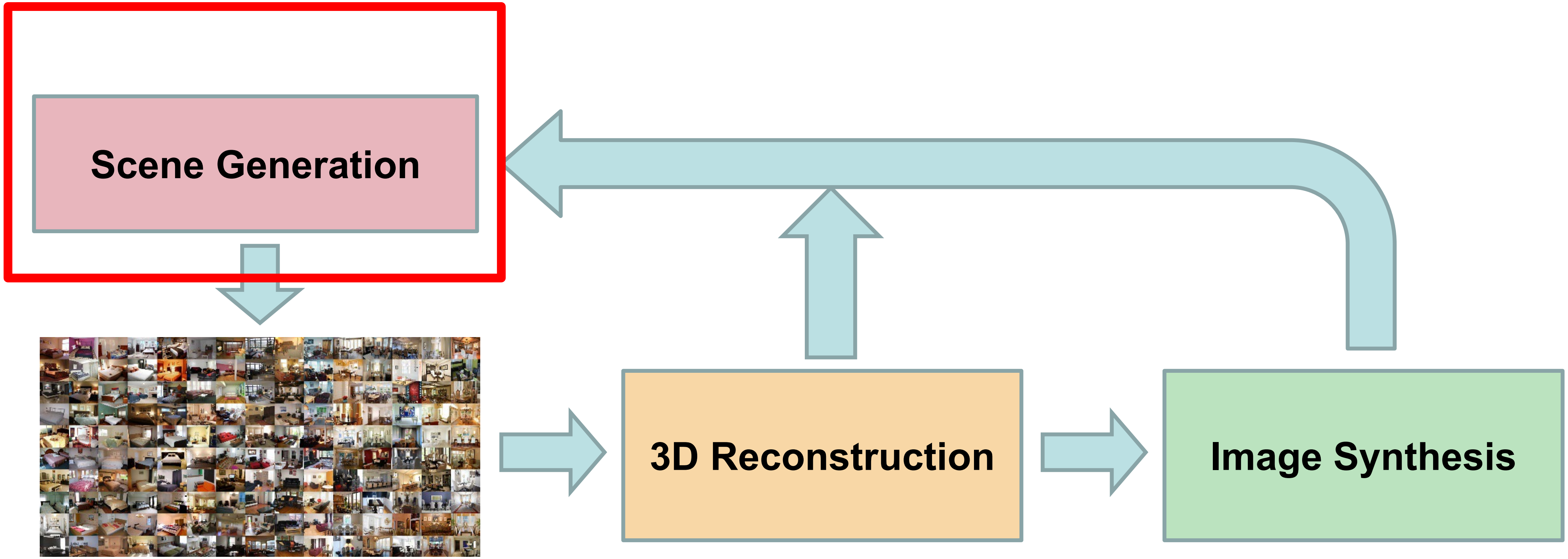
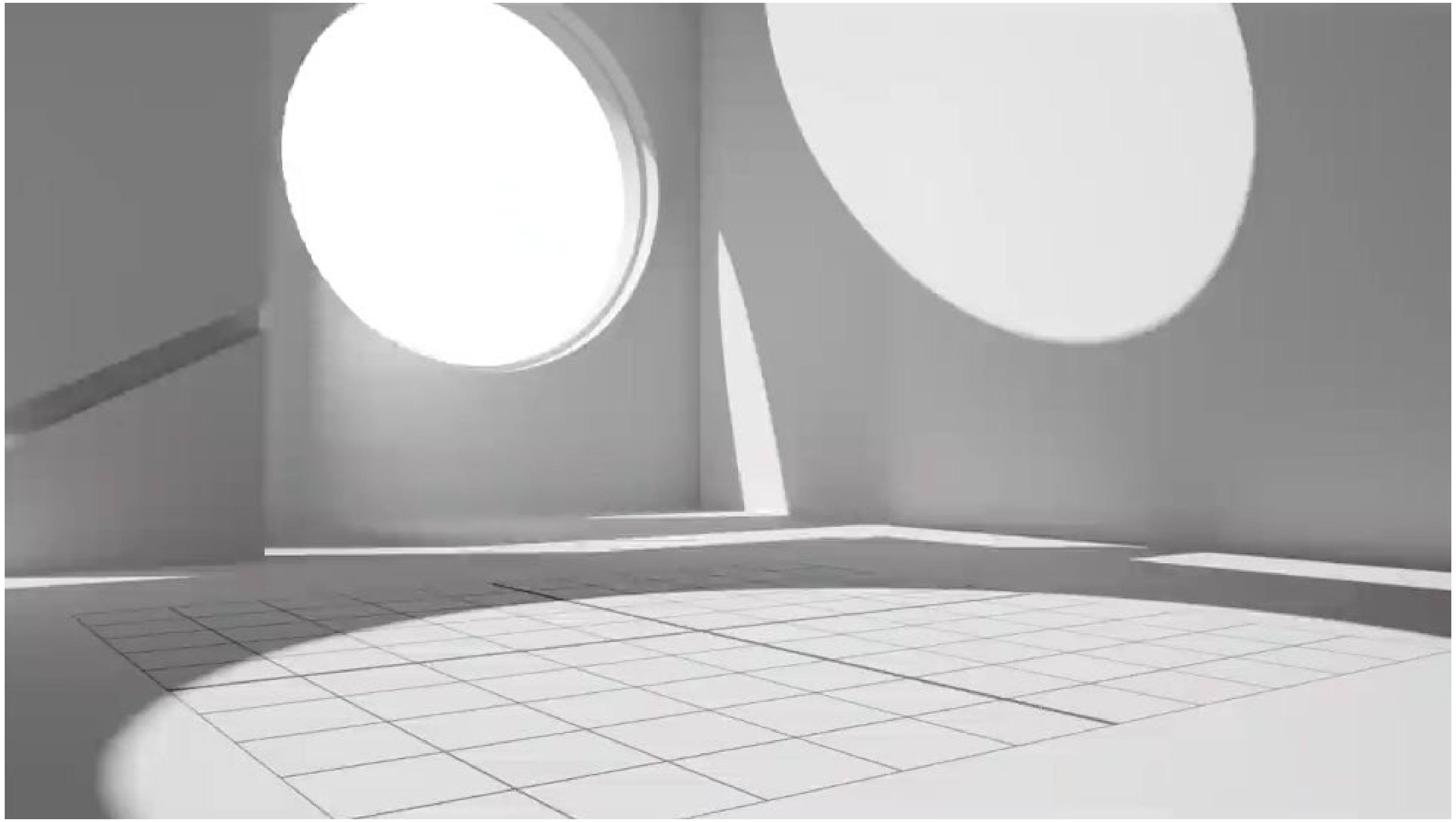


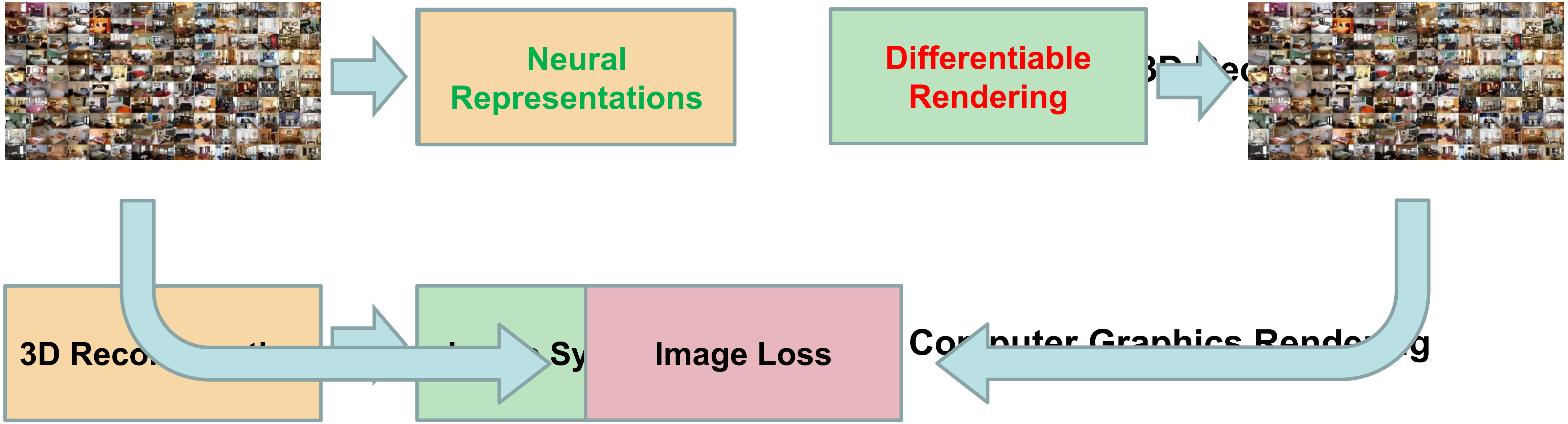
Photo-realistic Large-scale Scene Generation is Extremely Challenging

- Manually creating a scene is time-consuming



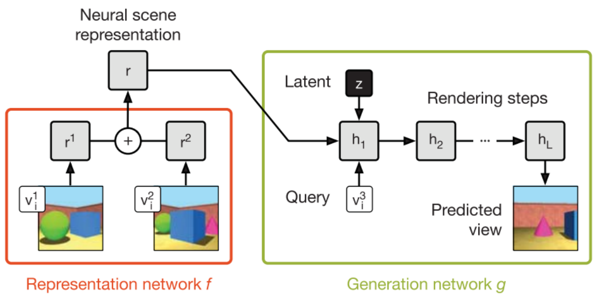
Self-supervised Learning of 3D Scenes

Allow the gradients of 3D objects to be calculated and propagated through images



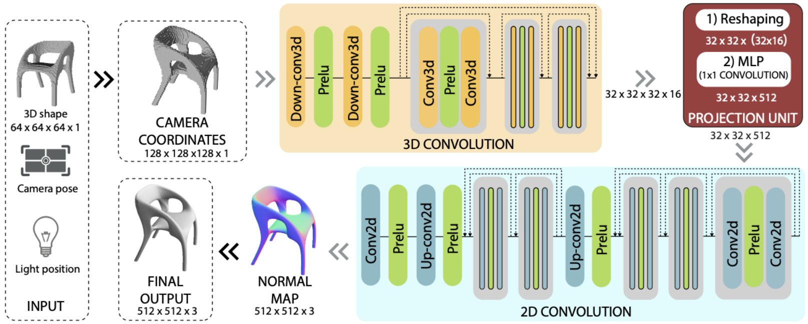
Lingjie Liu

Neural 3D Scene Representations

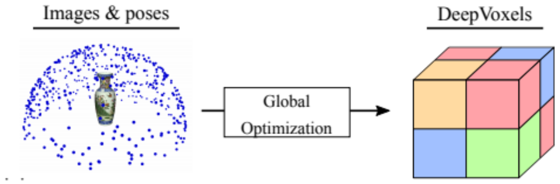


Generative Query Networks
[Eslami et al. 2018]

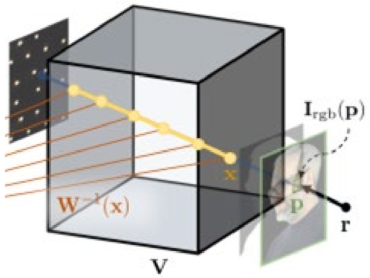
[Flynn et al., 2016; Zhou et al., 2018b; Mildenhall et al. 2019]
Multiplane Images (MPIs)



RenderNet [Nguyen-Phuoc et al. 2018]
Voxel Grids + CNN decoder

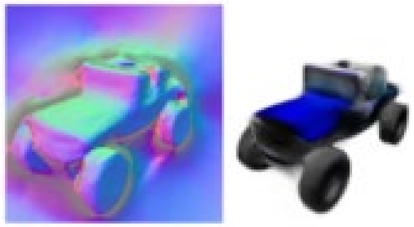


DeepVoxels
[Sitzmann et al. 2019]

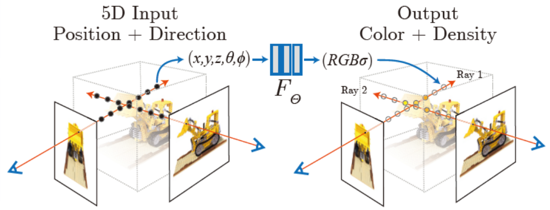


Neural Volumes
[Lombardi et al. 2019]

Voxel Grids + Ray Marching



SRN [Sitzmann et al. 2019b]

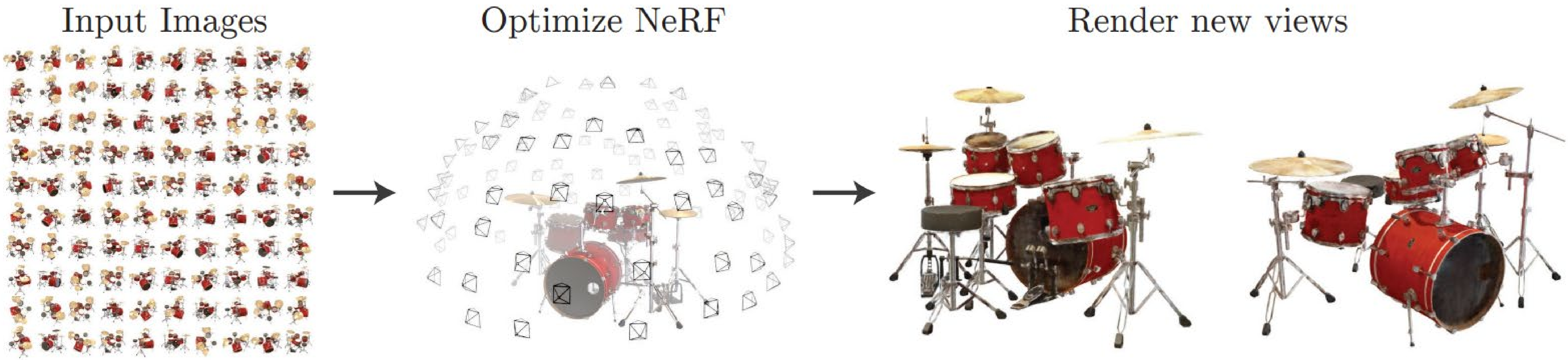


NeRF [Mildenhall et al. 2020]

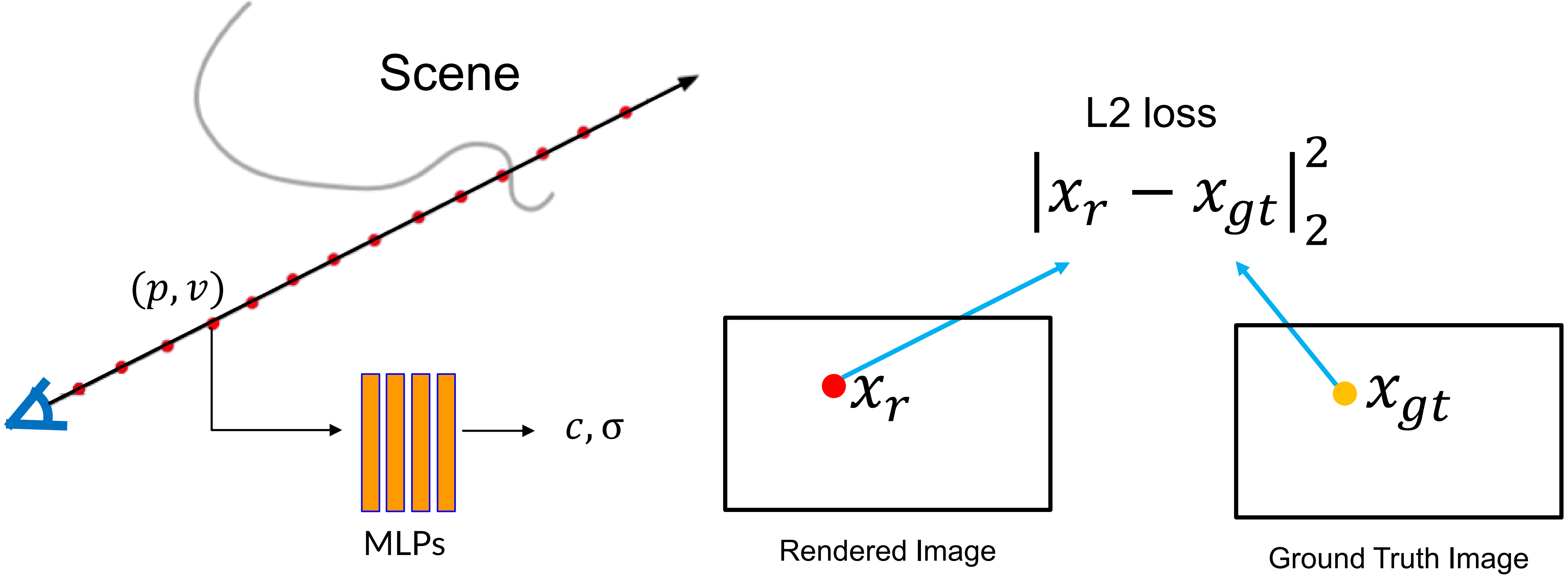
IDR [Yariv et al. 2020]

Implicit Fields

NeRF [Midenhall et al. 2020]



Neural Radiance Fields (NeRF)



[Mildenhall et al. 2020]

Hybrid Scene Representation for Fast Rendering

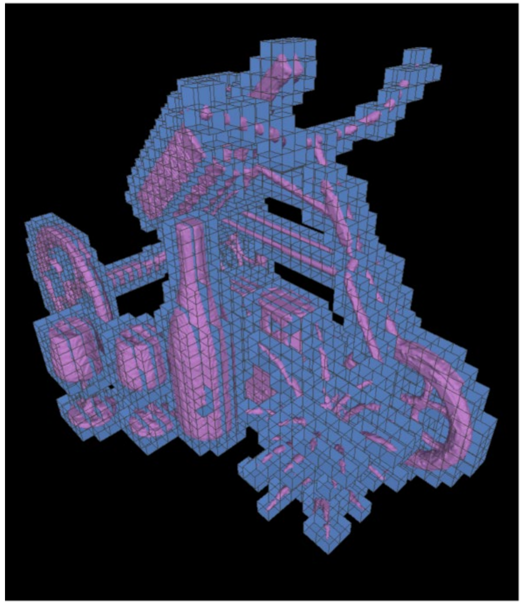
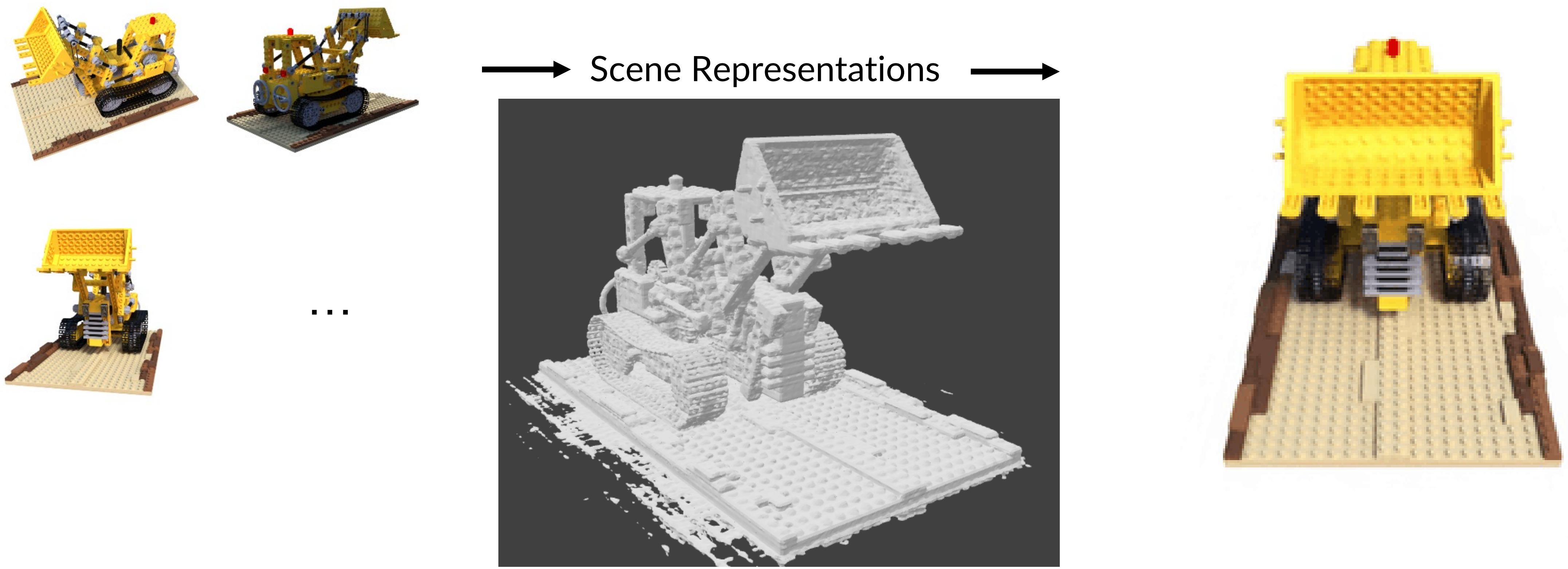


Illustration of
Neural Sparse Voxel Fields

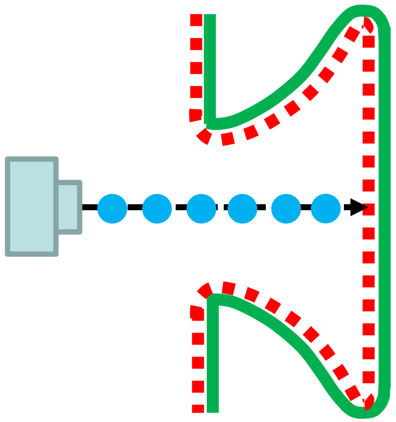
L. Liu, J. Gu, K.Z. Lin, T.S. Chua, C. Theobalt. Neural Sparse Voxel Fields, NeurIPS 2020 Spotlight

Surfaces Extracted from Learned Representation

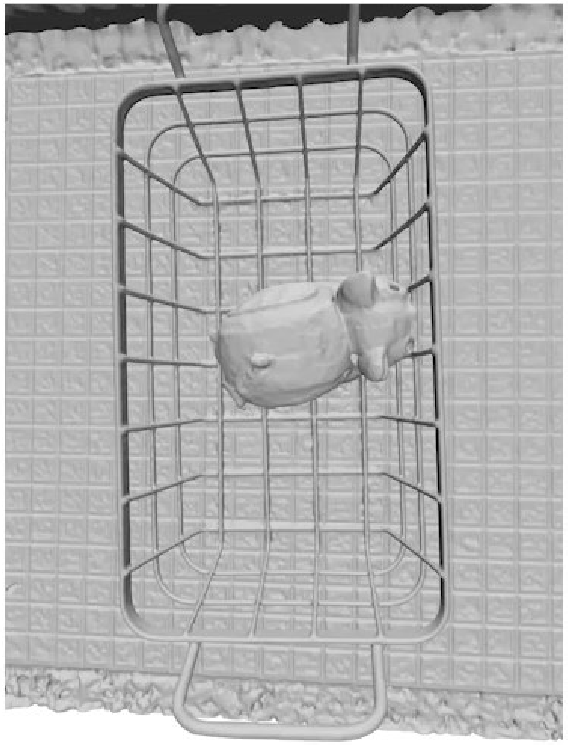


Volume density used as scene representation lacks surface constraints

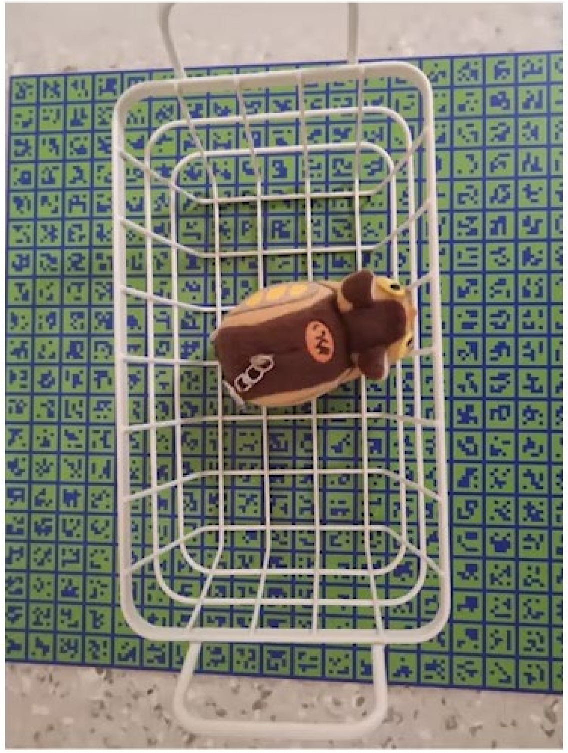
Neural Surface Representation for High-quality Reconstruction



Surface Representation
+ Volume Rendering



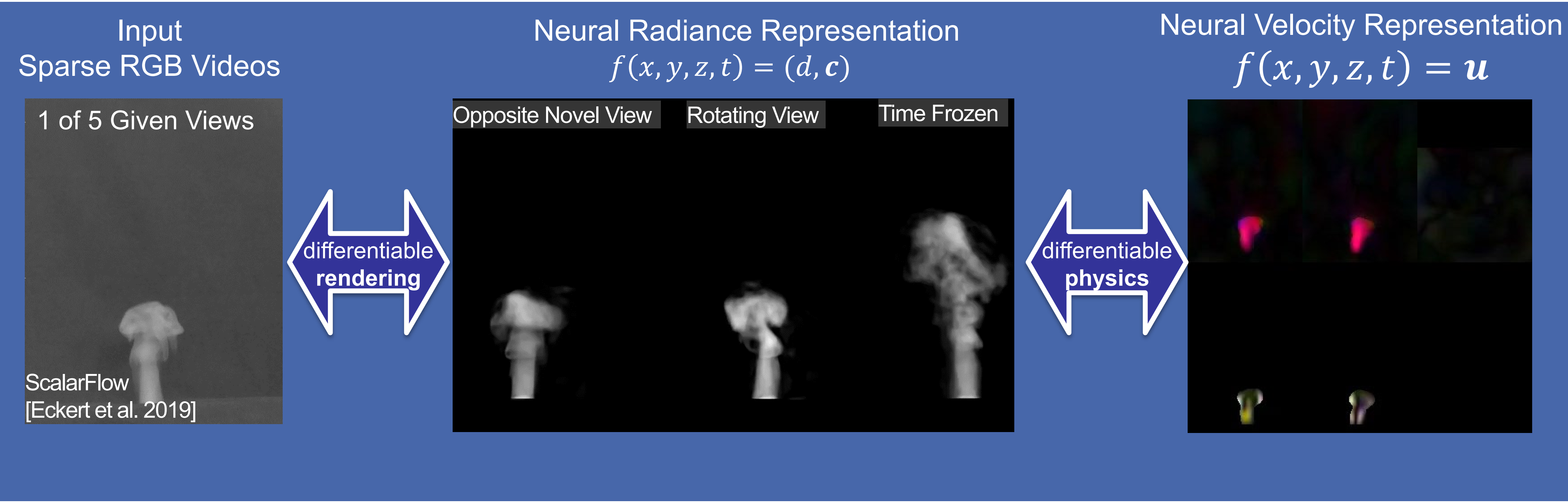
Our surface geometry
(w/o mask supervision)



Our rendering
(w/o mask supervision)

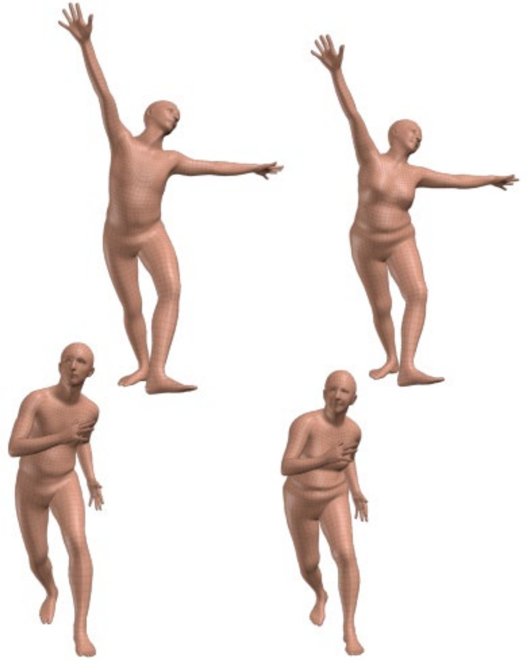
P. Wang, *L. Liu,* Y. Liu, C. Theobalt, T. Komura, W. Wang. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction, NeurIPS 2021 Spotlight

Physics Informed Scene Representation

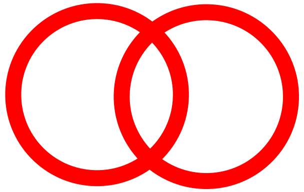


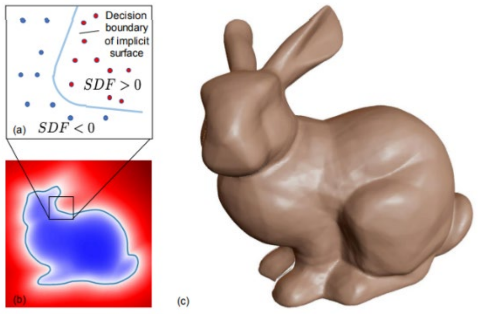
M. Chu, *L. Liu*, Q. Zheng, E. Franz, H.P. Seidel, C. Theobalt, R. Zayer.
Physics Informed Neural Fields for Smoke Reconstruction with Sparse Data, SIGGRAPH 2022 (Journal track)

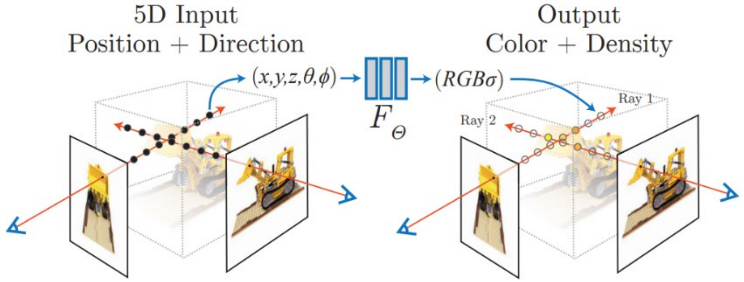
Neural Animatable Human Representation



Skinned Multi-person Linear Model (SMPL)







Neural Scene Representations

L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, C. Theobalt.
Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control, SIGGRAPH Asia 2021

Neural Animatable Human Representation

Input Driving Poses

Reference Video
of Driving Person

Our Result

L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, C. Theobalt.
Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control, SIGGRAPH Asia 2021

How to Generate New 3D Scenes?

Training Data? Multi-view Images?

How to Generate New 3D Scenes?

Training Data? Single-view Images? 😃

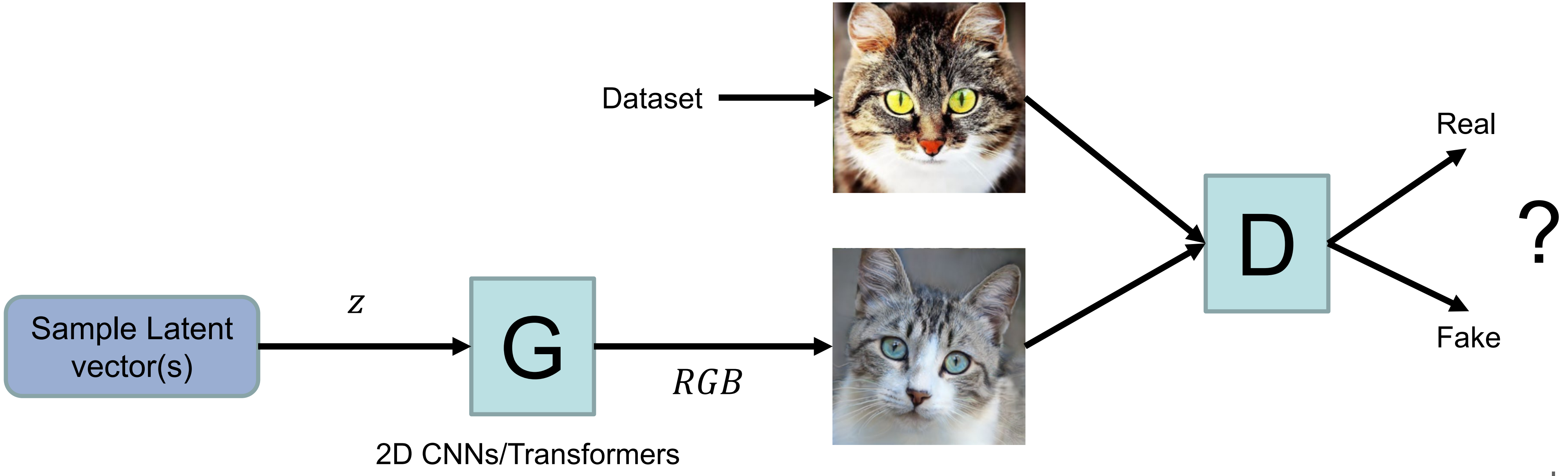
Lingjie Liu

How to Generate New 3D Scenes?

Model?

2D GANs? VAE?

- Generative Models
  - Likelihood-based (VAEs, Flow, DDPM, Autoregressive models, etc)
  - Likelihood-free (GANs)
- Generative Adversarial Networks (GANs)



Lingjie Liu

How to Generate New 3D Scenes?

Model?

2D GANs? VAE?

Results of the state-of-the-art GAN model (StyleGAN2)

Generate merely 2D images, without 3D information

HumanGAN: A Generative Model of Human Images

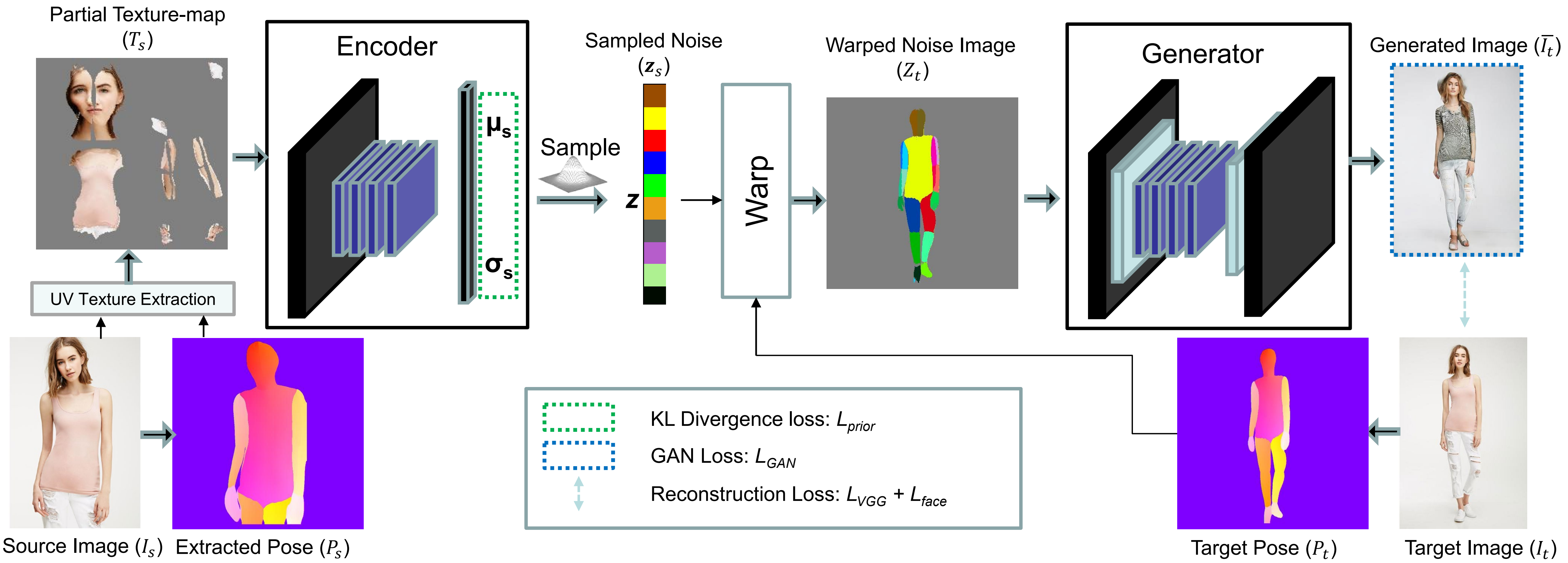
Appearance sampling on a given pose

Pose transfer on a given identity

Body parts sampling (HEAD)

K. Sarkar, *L. Liu*, V. Golyanik, C. Theobalt. HumanGAN: A Generative Model of Humans Images. 3DV 2021 (Oral)

HumanGAN: A Generative Model of Human Images



K. Sarkar, *L. Liu*, V. Golyanik, C. Theobalt. HumanGAN: A Generative Model of Humans Images. 3DV 2021 (Oral)

Lingjie Liu

Appearance Sampling

Input

K. Sarkar, *L. Liu*, V. Golyanik, C. Theobalt. HumanGAN: A Generative Model of Humans Images. 3DV 2021 (Oral)

Appearance Sampling

Part Sampling

- Head

Part Sampling

- Upper Body




Part Sampling

- Lower Body

K. Sarkar, *L. Liu*, V. Golyanik, C. Theobalt. HumanGAN: A Generative Model of Humans Images. 3DV 2021 (Oral)

Garment Transfer

Garments

Body

Garment Transfer

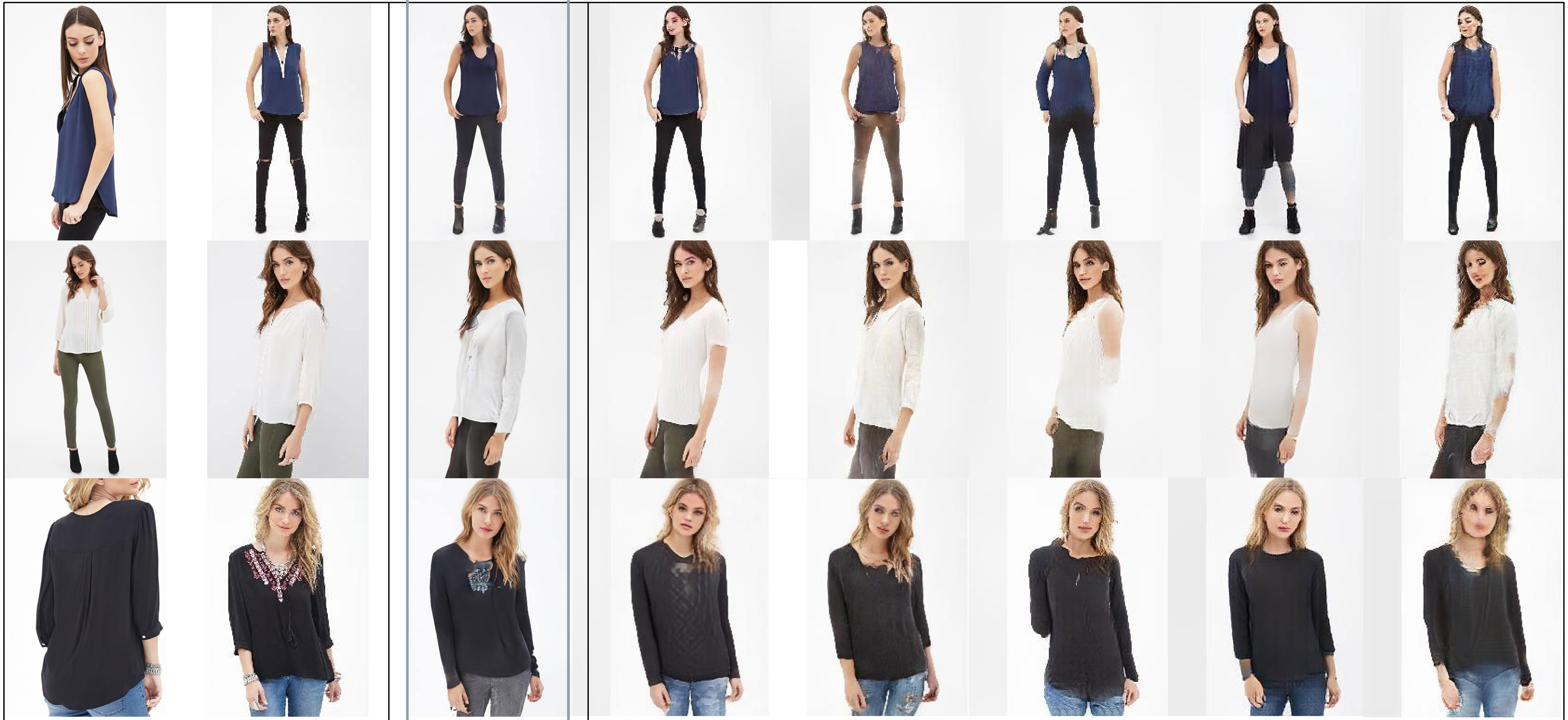
Garment Transfer

Garments | Body | Garment Transfer

Latent Space Interpolation

Latent Space Interpolation of the entire body
(Conditioning poses are not shown)

Pose Transfer



source target **ours** CBI NHRR DSC VUNet DPT

Lingjie Liu

Motion Transfer and Interpolation

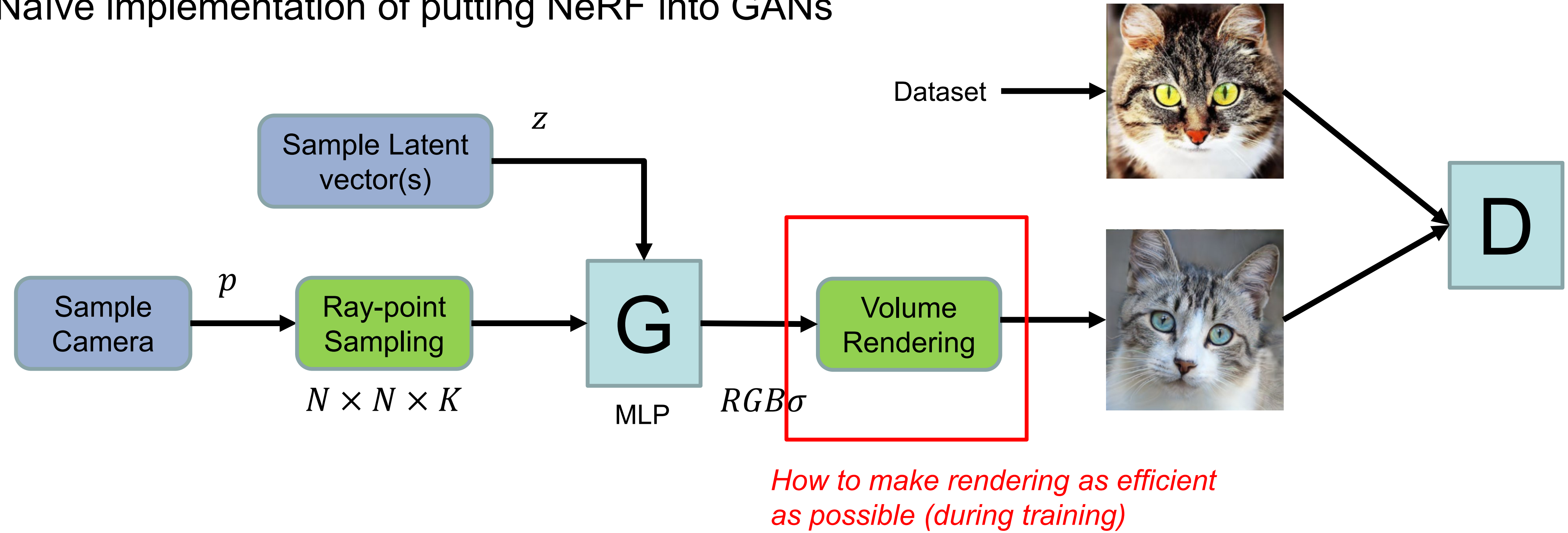
By changing both the pose and the latent vector,
we can perform ***motion transfer with varying appearances.***

3D GANs

J. Gu, *L. Liu,* P. Wang, C. Theobalt.
StyleNeRF: A Style-based 3D-Aware Generator for High-resolution Image Synthesis, ICLR 2022

3D GANs

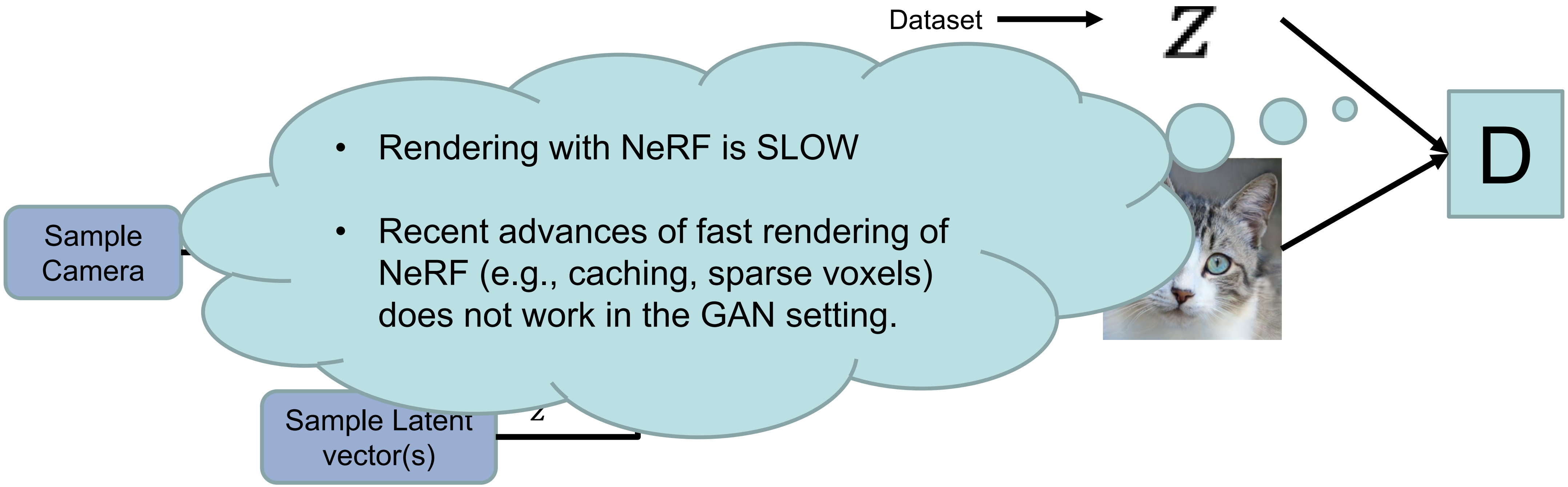
Naïve implementation of putting NeRF into GANs



Lingjie Liu

3D GANs

Naïve implementation of putting NeRF into GANs



- Rendering with NeRF is SLOW
- Recent advances of fast rendering of NeRF (e.g., caching, sparse voxels) does not work in the GAN setting.

Goal

- We propose to address the above issues simultaneously:
  - High-resolution
  - Efficient
  - Multi-view consistent

J. Gu, *L. Liu,* P. Wang, C. Theobalt.
StyleNeRF: A Style-based 3D-Aware Generator for High-resolution Image Synthesis, ICLR 2022

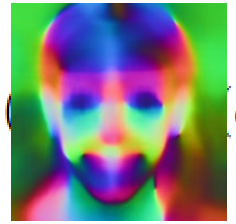
Method

- Approximated Volume Rendering

$$I_{\boldsymbol{w}}^{\mathrm{NeRF}}(\boldsymbol{r}) = \int_0^\infty p_{\boldsymbol{w}}(t)\boldsymbol{c}_{\boldsymbol{w}}(\boldsymbol{r}(t), \boldsymbol{d})dt, \quad \text{where } p_{\boldsymbol{w}}(t) = \text{ex}$$

$$\boldsymbol{r}(t))$$

$$I_{\boldsymbol{w}}^{\mathrm{Approx}}(\boldsymbol{r}) = \int_0^\infty p_{\boldsymbol{w}}(t) \cdot h_c \circ [\phi_{\boldsymbol{w}}^{n_c}(\boldsymbol{r}(t)), \zeta(\boldsymbol{d})]\, dt \approx h_c \circ [\phi_{\boldsymbol{w}}^{n_c, n_\sigma}($$



$$\boldsymbol{d})]$$

Early aggregation

$$\phi_{\boldsymbol{w}}^{n, n_\sigma}(\mathcal{A}(R_H)) \approx \texttt{Upsample}\left(\phi_{\boldsymbol{w}}^{n, n_\sigma}(\mathcal{A}(R_L))\right)$$

2D upsampling

Preserve 3D consistency

- Remove view direction input
  - We found that view direction will break the consistency and did not contribute to much quality (our dataset is single image)

- NeRF-path regularization

$$\mathcal{L}_{\text{NeRF-path}} = \frac{1}{|S|} \sum_{(i,j)\in S} \left(I_{\boldsymbol{w}}^{\text{Approx}}(R_{\text{in}})[i,j] - I_{\boldsymbol{w}}^{\text{NeRF}}(R_{\text{out}}[i,j]) \right)^2$$

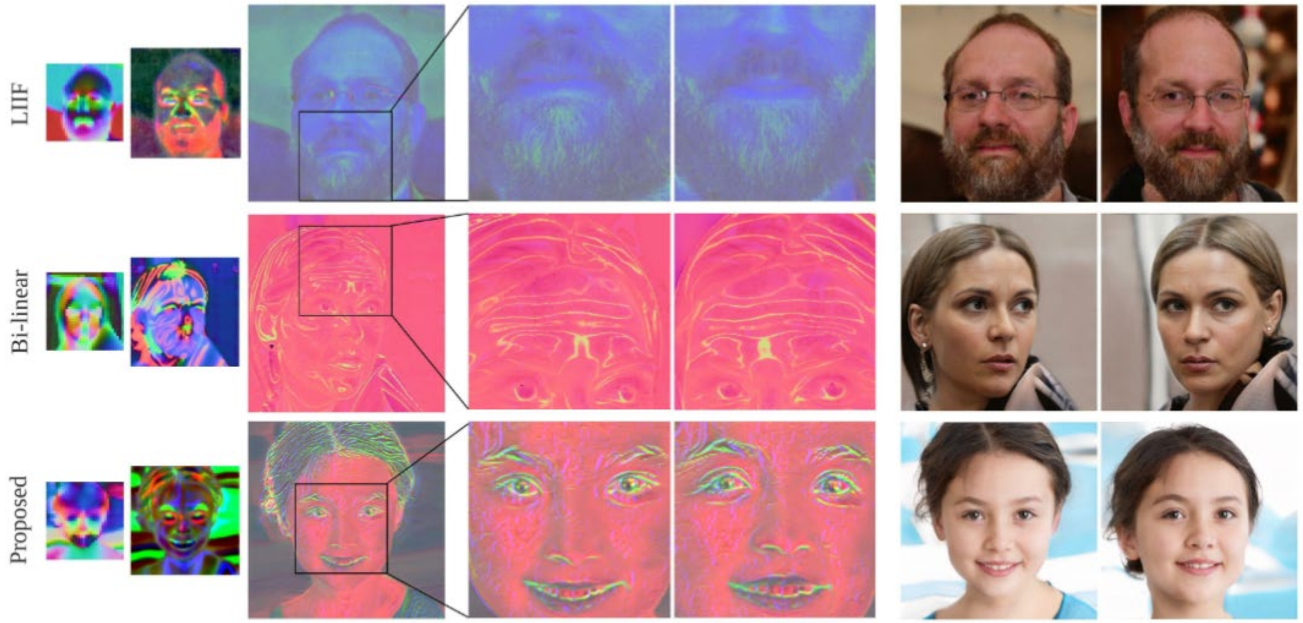
- Up-sampler design

$$\texttt{Upsample}(X) = \texttt{Conv2d}\left(\texttt{Pixelshuffle}\left(\texttt{Repeat}(X, 4) + \psi_\theta(X), 2\right), K\right)$$

Lingjie Liu

StyleNeRF

- Up-sampler: we have tested many ways
  - Filter-based (bilinear interpolation, FIR filters, etc) + MLP (1x1 Conv) will cause “bubble shape” artifacts
  - Learning-based (transposed conv, pixelshuffle, LIIF) will easily cause texture sticking artifacts
  - We combine these two methods



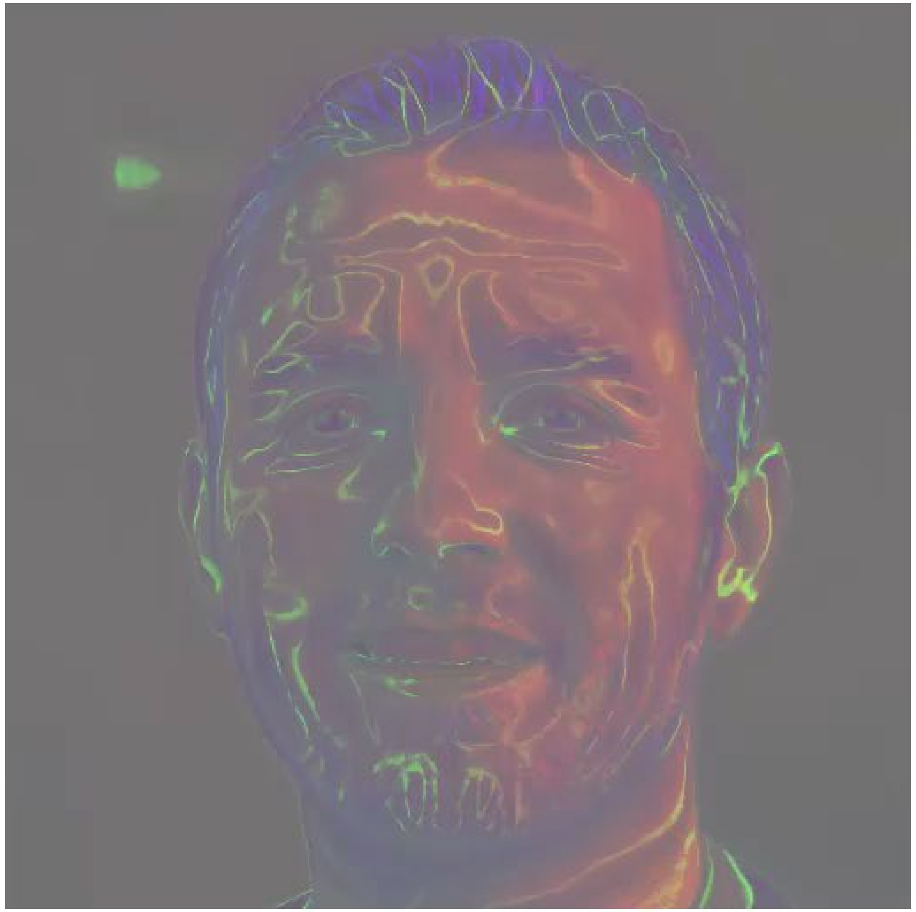
Ablation: Different Upsampling Operators

LIIF: Having the "texture sticking" artifacts

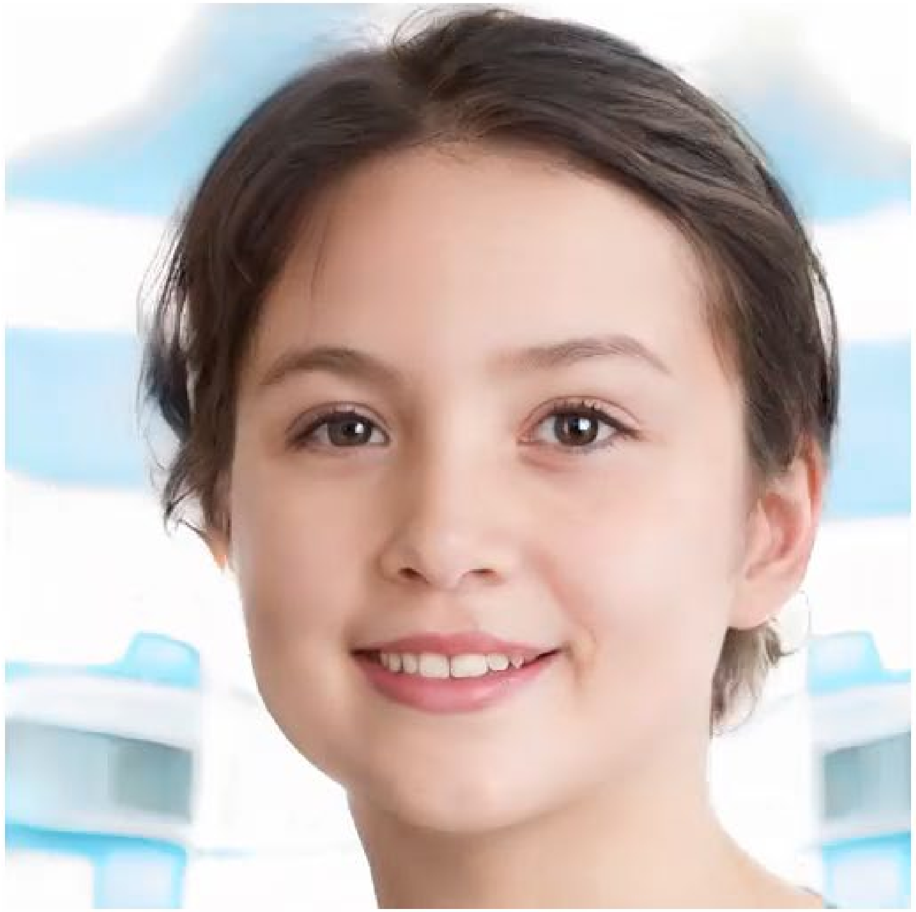
Lingjie Liu

Ablation: Different Upsampling Operators



Bilinear: Having the "bubble-shape" artifacts

Ablation: Different Upsampling Operators



Our proposed operator: Highly preserving 3D consistency
while getting rid of bubble-shape artifacts

Ablation: Importance of Progressive Training

Results of no progressive training

Our Results

This is the first time that a generative model can synthesize high-resolution images from novel views while preserving high 3D consistency

Results

- V.s. Existing works

Models	FFHQ 256^2 FID	FFHQ 256^2 KID	AFHQ 256^2 FID	AFHQ 256^2 KID	CompCars 256^2 FID	CompCars 256^2 KID	Rendering time (ms / image) 64	128	256	512	1024
2D GAN	4	1.1	9	2.3	3	1.6	-	-	46	51	53
HoloGAN	75	68.0	78	59.4	48	39.6	213	215	222	-	-
GRAF	71	57.2	121	83.8	101	86.7	61	246	990	3852	15475
π-GAN	85	90.0	47	29.3	295	328.9	58	198	766	3063	12310
GIRAFFE	35	23.7	31	13.9	32	23.8	8	-	9	-	-
Ours	8	3.7	14	3.5	8	4.3	-	-	65	74	98

- High resolution

Models	FFHQ 512^2 FID	FFHQ 512^2 KID	AFHQ 512^2 FID	AFHQ 512^2 KID	MetFace 512^2 FID	MetFace 512^2 KID	FFHQ 1024^2 FID	FFHQ 1024^2 KID
2D GAN	3.1	0.7	8.6	1.7	18.9	2.7	2.7	0.5
Ours	7.8	2.2	13.2	3.6	20.4	3.3	8.1	2.4

Results

- Consistency evaluation

Style Interpolation

Our synthesized results (512x512)

Applications: Style Mixing (Styles of Geometry)

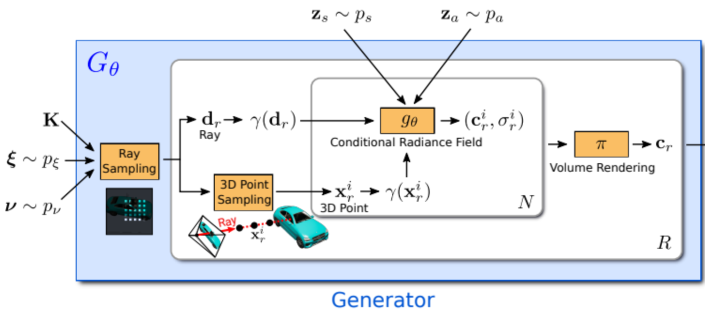
Our synthesized results (512x512)

Results

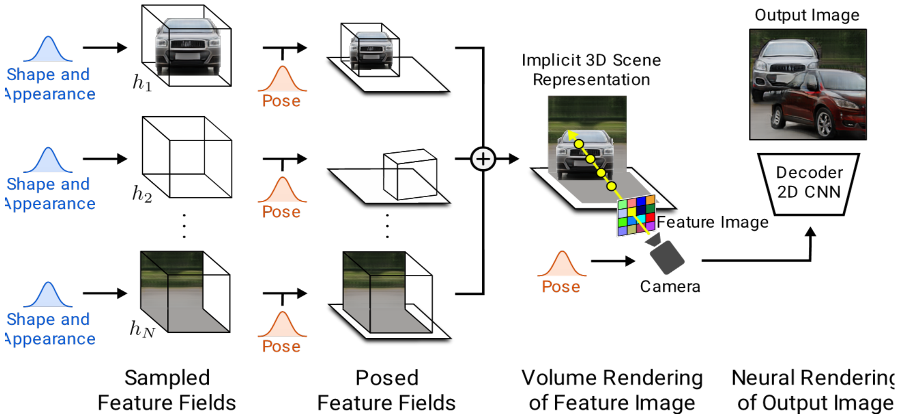
- Interactive Demo

  https://huggingface.co/spaces/facebook/StyleNeRF

Explosion of 3D GANs

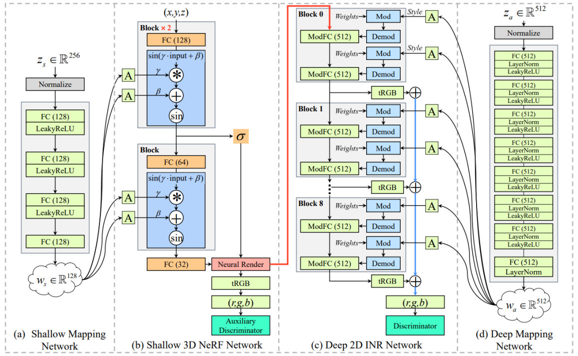


GRAF [Schwarz et al. 2020]

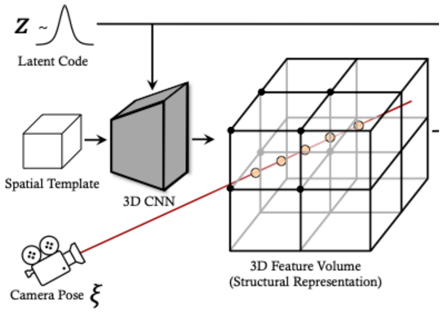


GIRAFFE [Niemeyer et al. 2020]

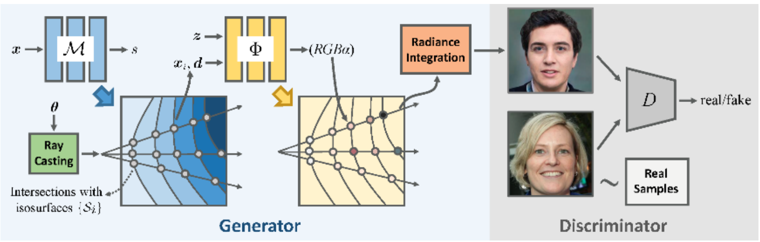
EG3D [Chan et al. 2022]



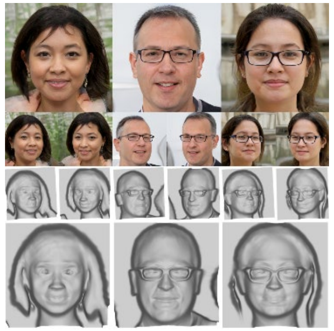
CIPS-3D [Zhou et al. 2021]



VolumeGAN [Xu et al. 2022]



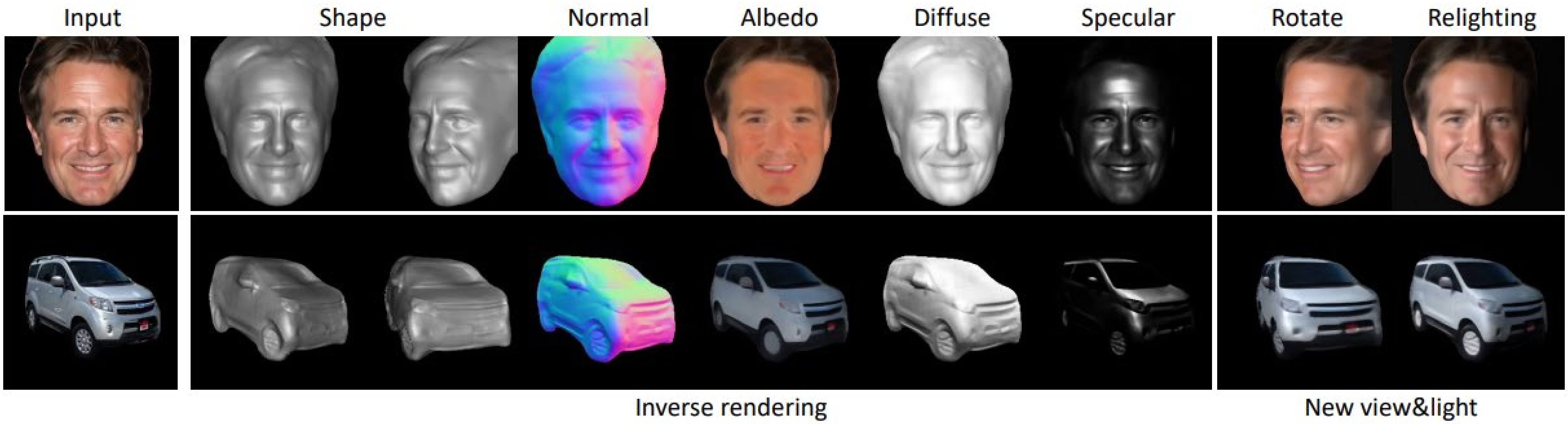
GRAM [Deng et al. 2022]



StyleSDF [Or-El et al. 2022]

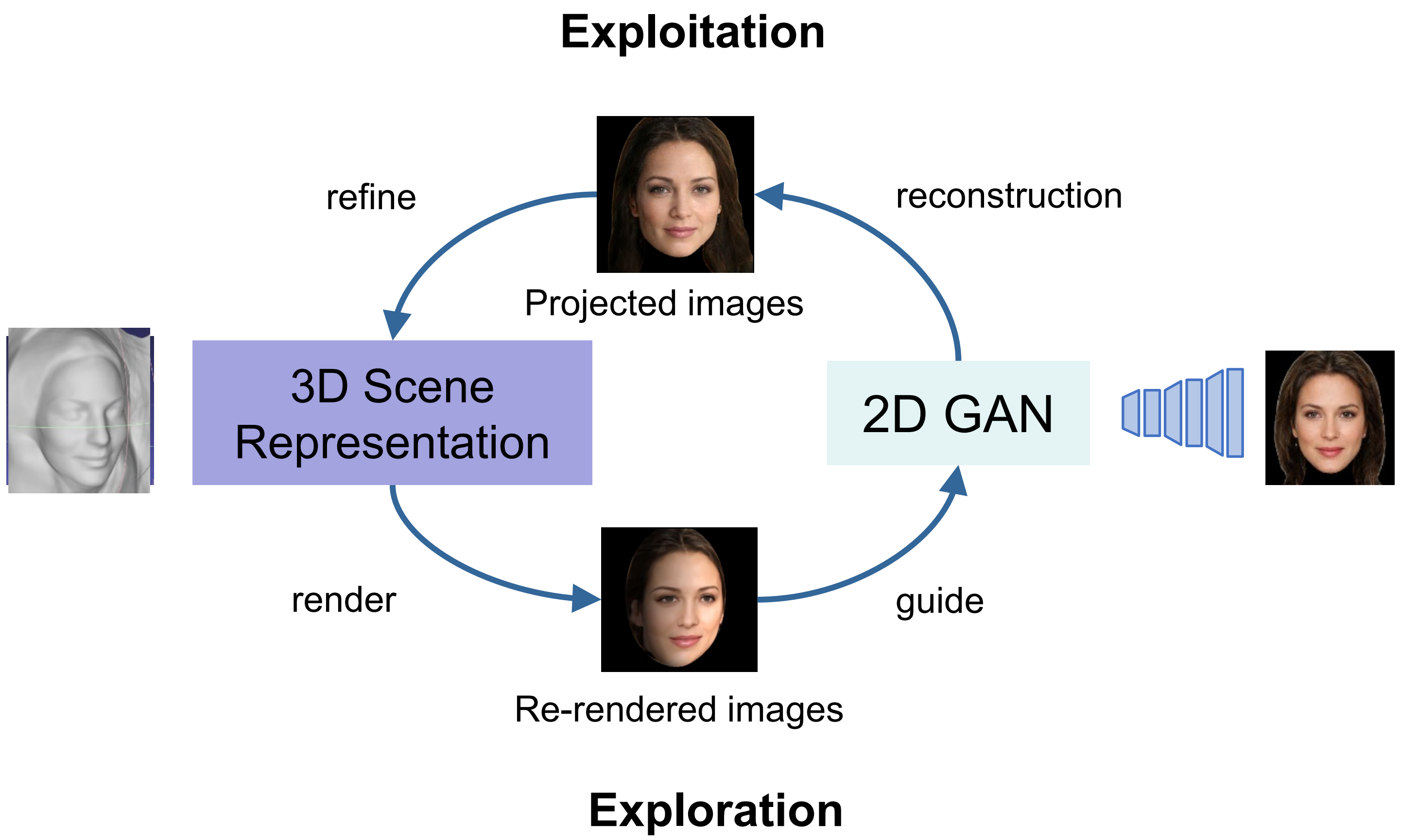
. . .

GAN2X: Non-Lambertian Inverse Rendering of Image GANs



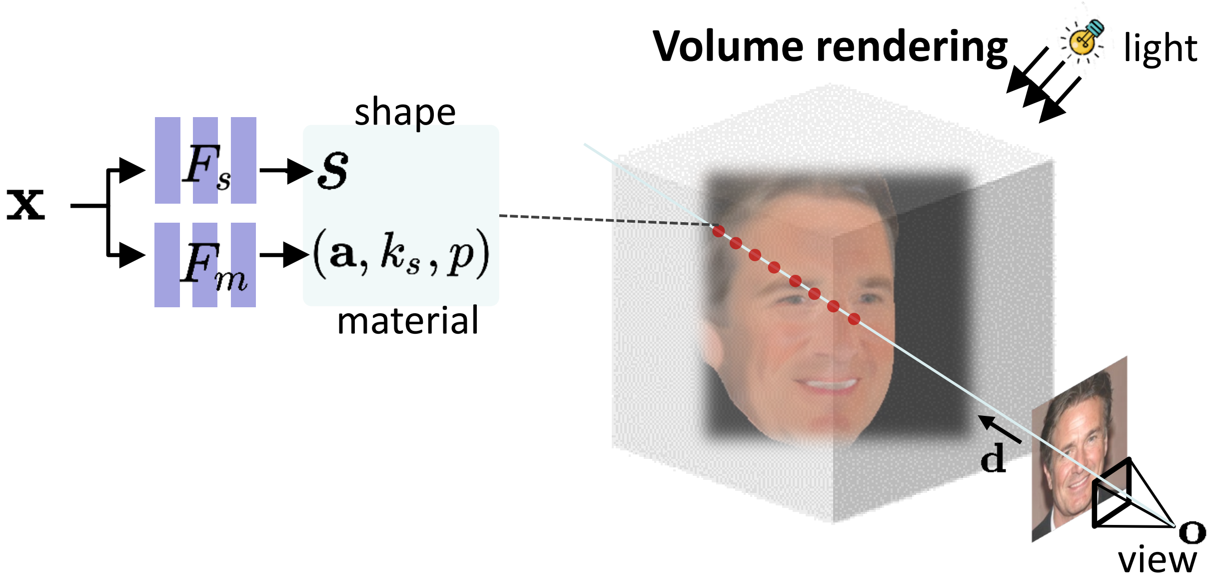
X. Pan, A. Tewari, *L. Liu,* C. Theobalt.
GAN2X: Non-Lambertian Inverse Rendering of Image GANs, 3DV 2022

Method



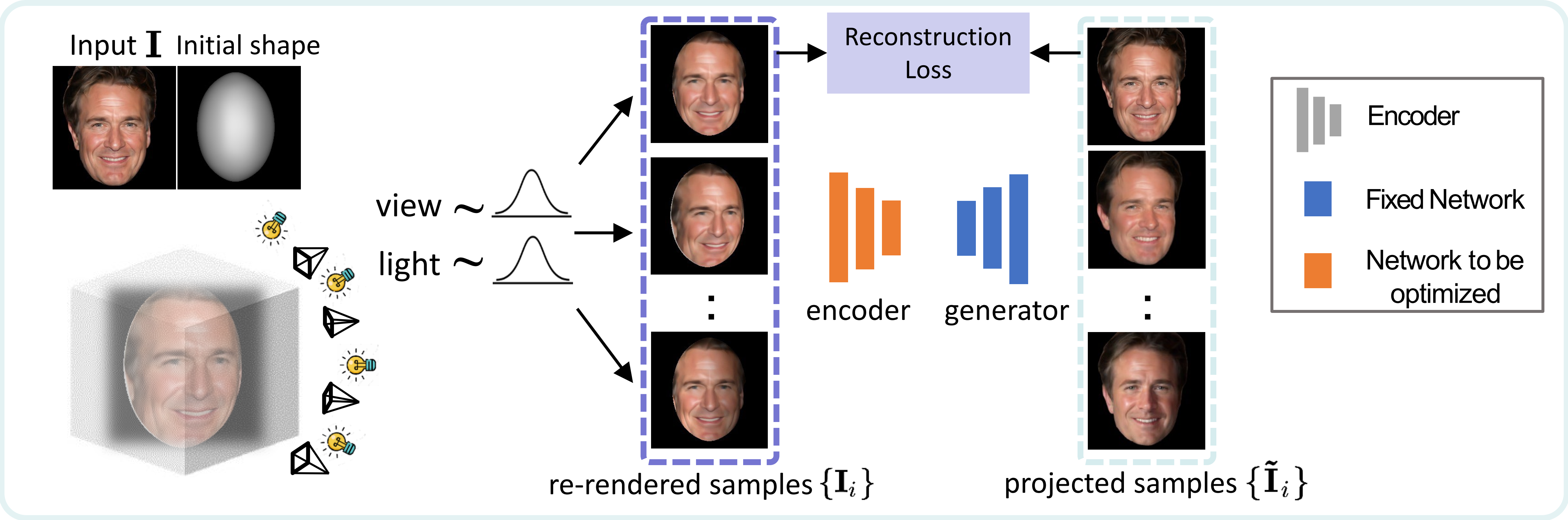
X. Pan, A. Tewari, *L. Liu,* C. Theobalt.
GAN2X: Non-Lambertian Inverse Rendering of Image GANs, 3DV 2022

Scene Representation

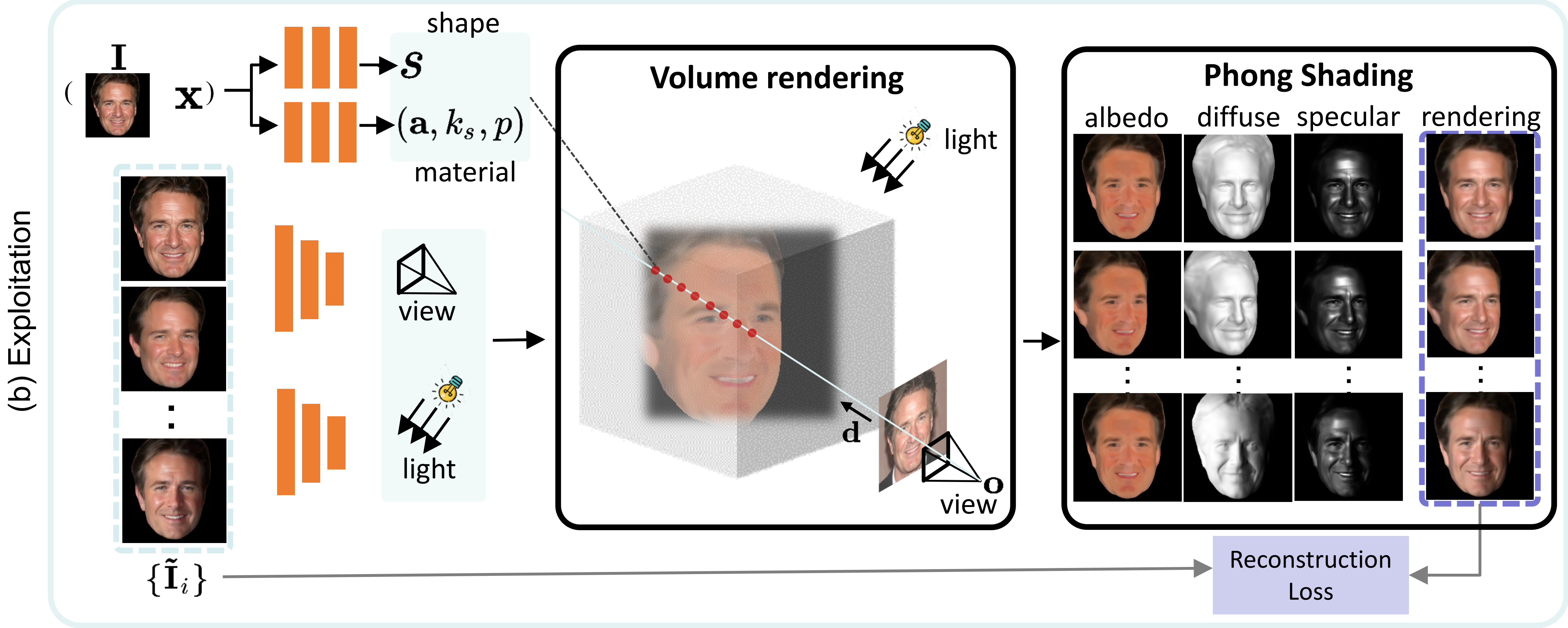


$\mathbf{x}$: 3D coordinate
s : signed distance
$\mathbf{A}/\mathbf{a}$: diffuse albedo
K_s/k_s : specular intensity
P/p : Shininess

Method: Exploration

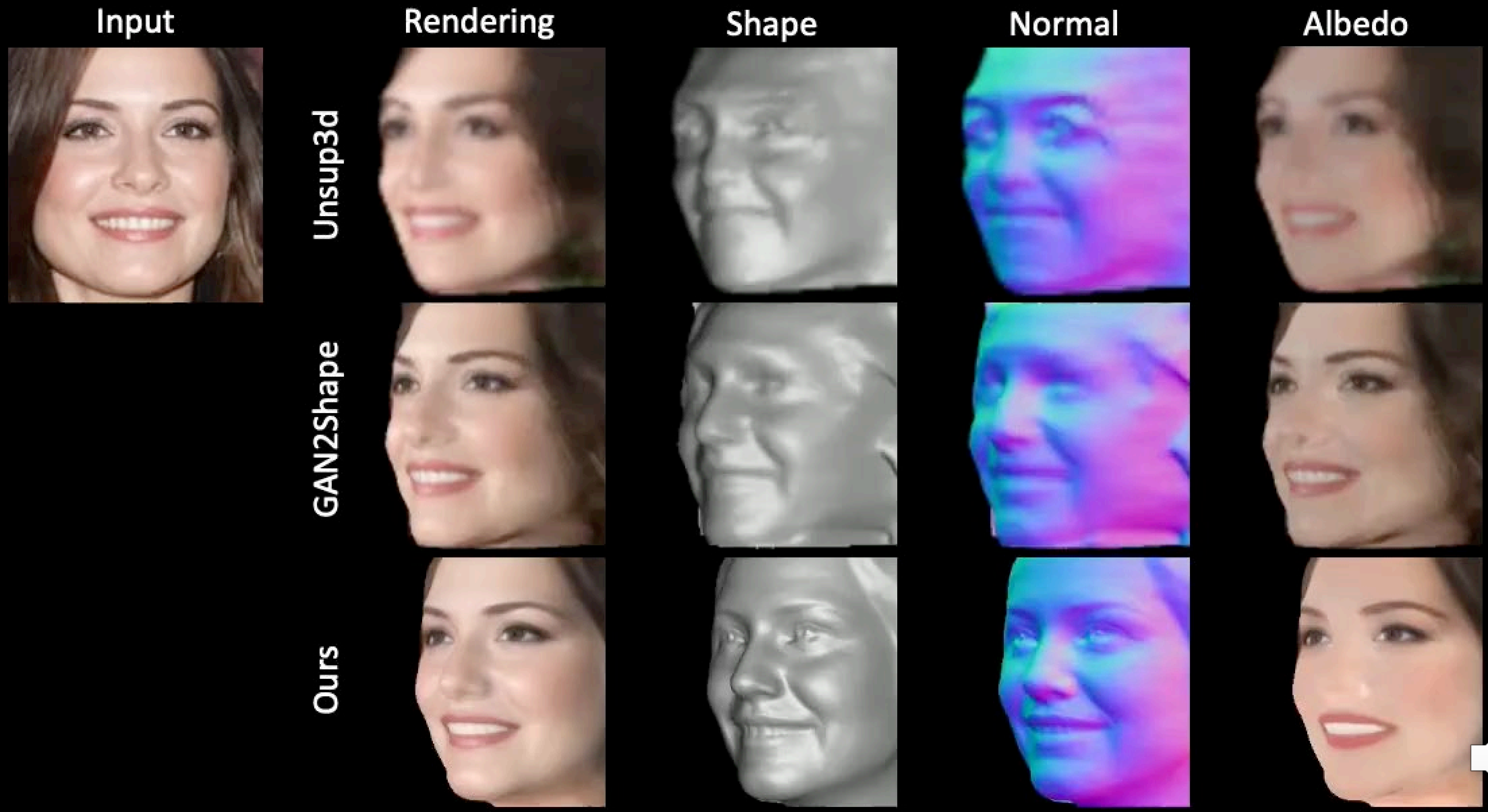


Method: Exploitation

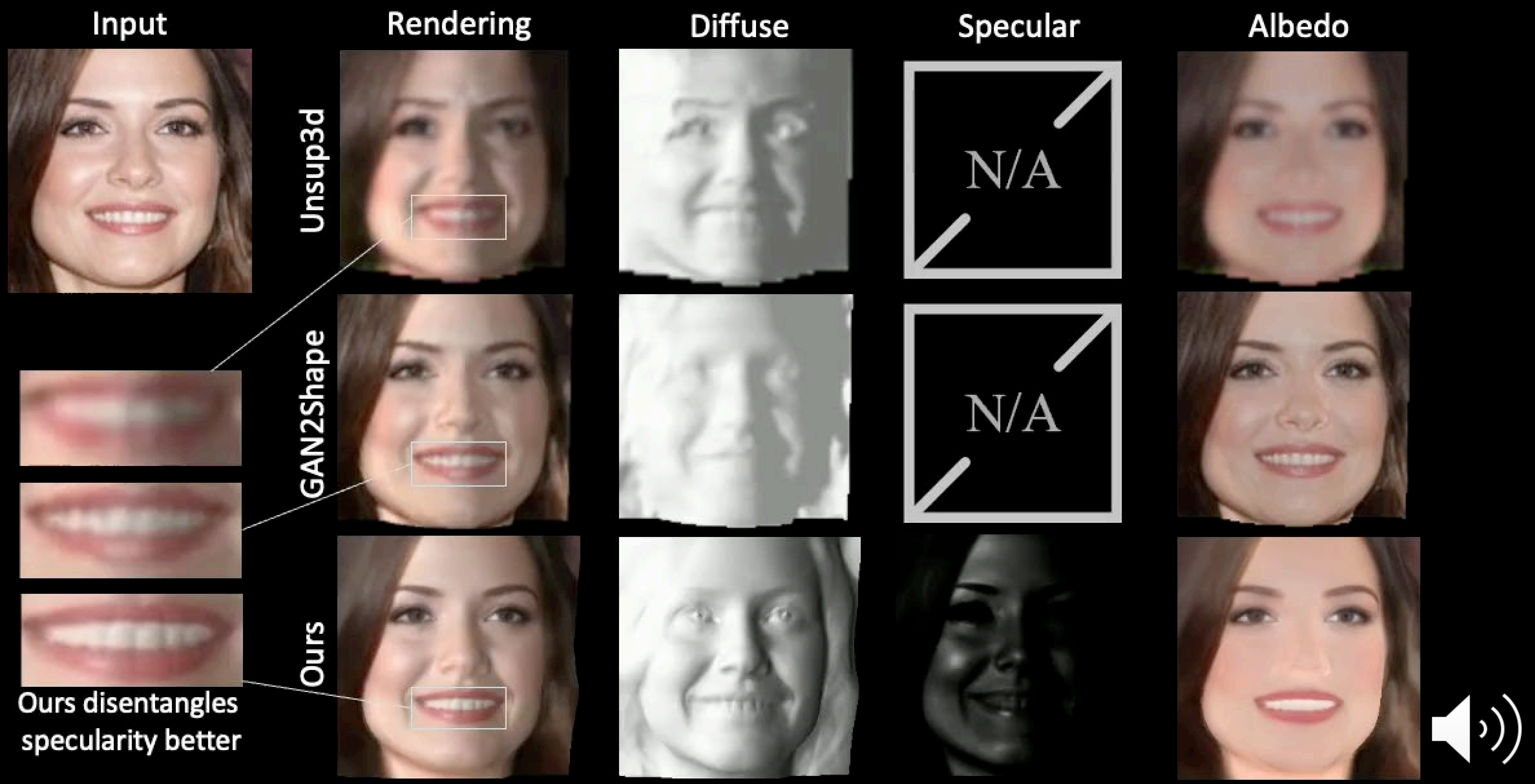


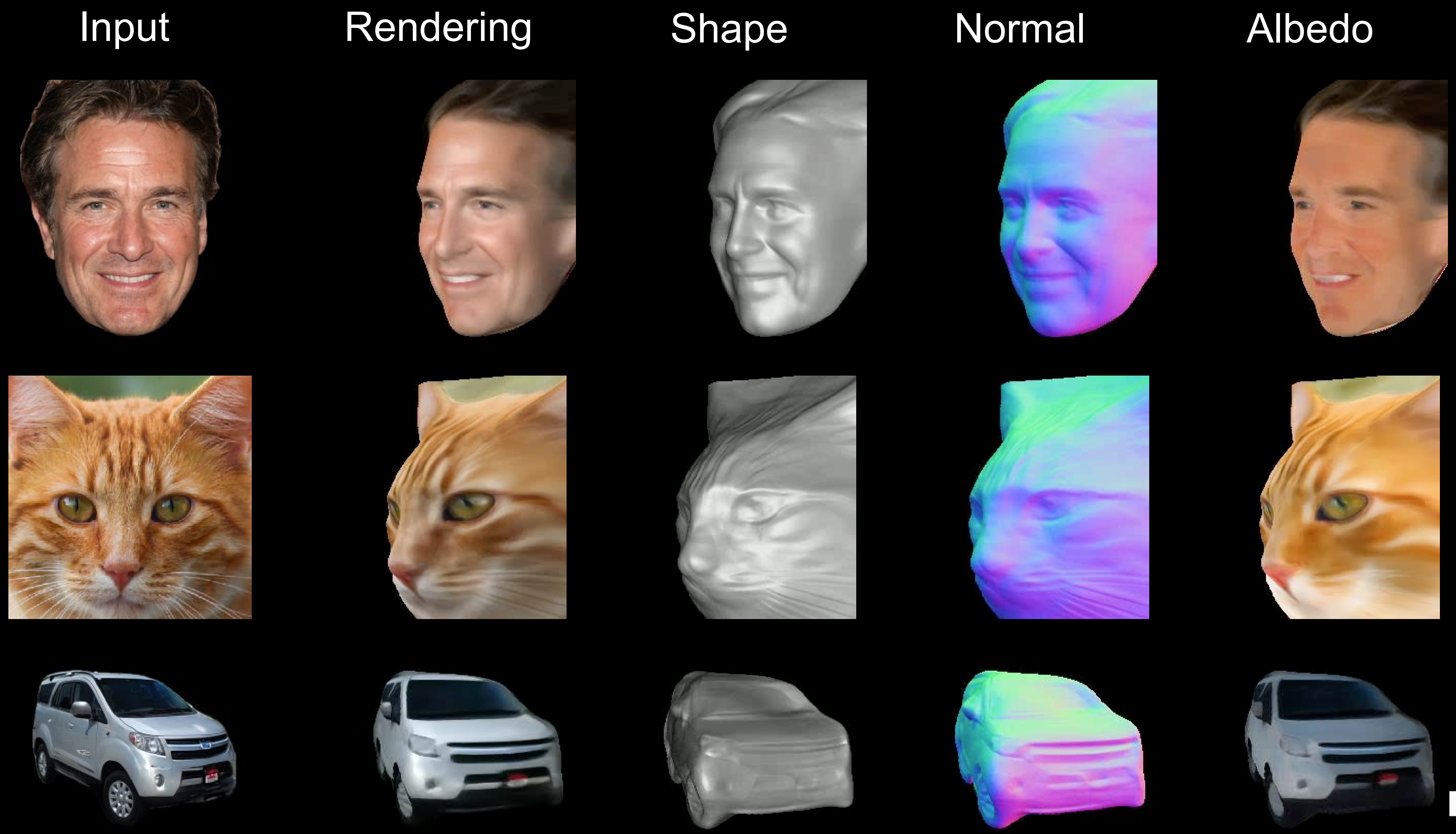
Lingjie Liu

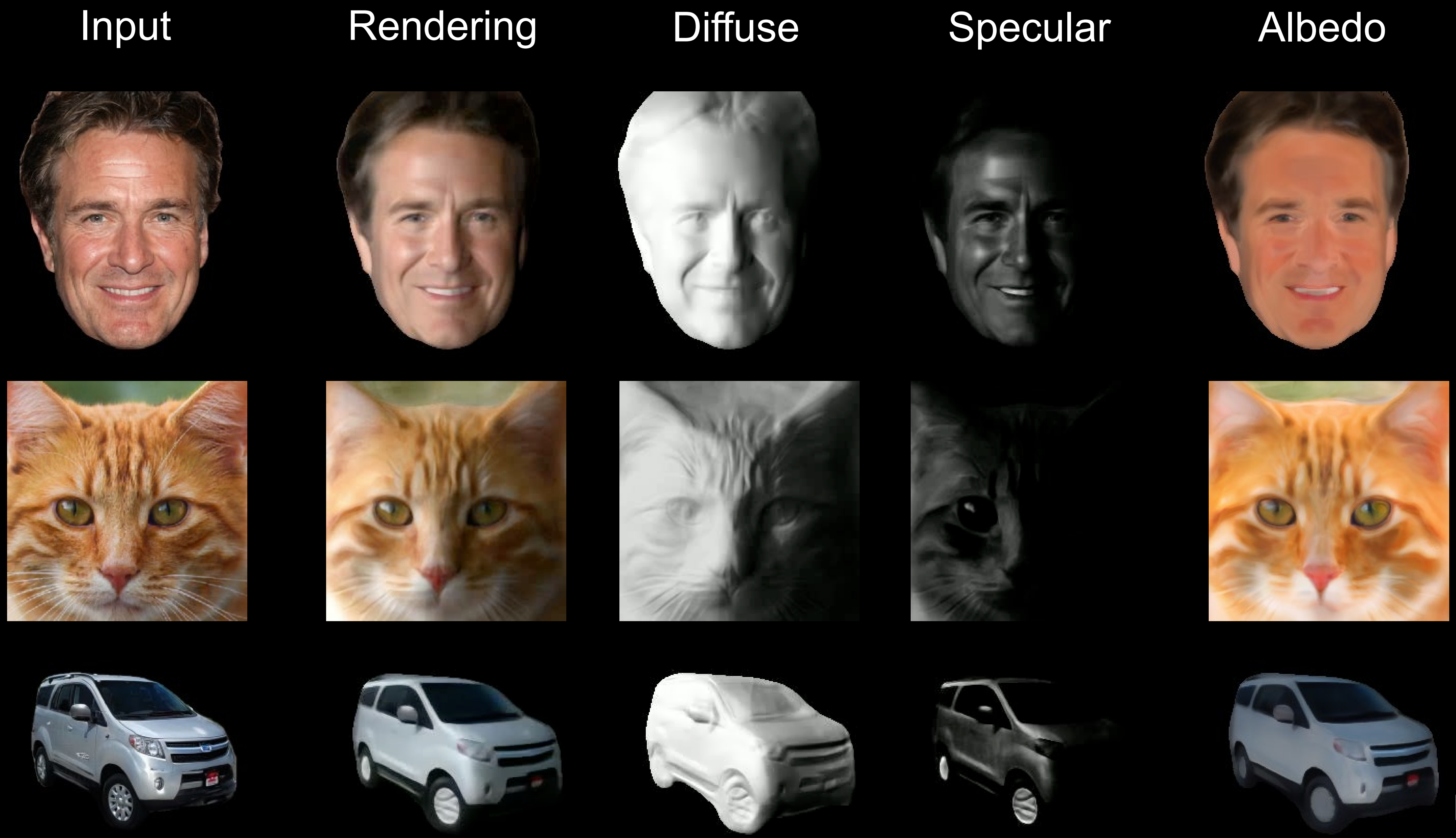
Qualitative Comparison on CelebA: Rotation



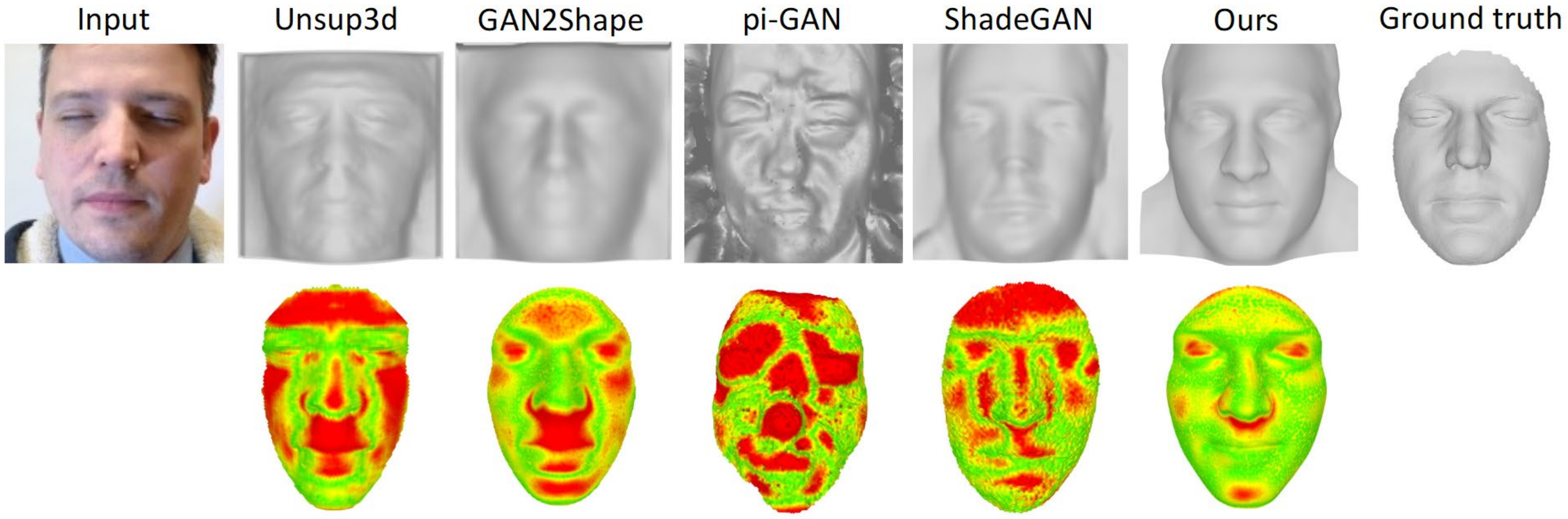
Qualitative Comparison on CelebA: Relighting




Input
Rendering
Shape
Normal
Albedo


Input
Rendering
Diffuse
Specular
Albedo

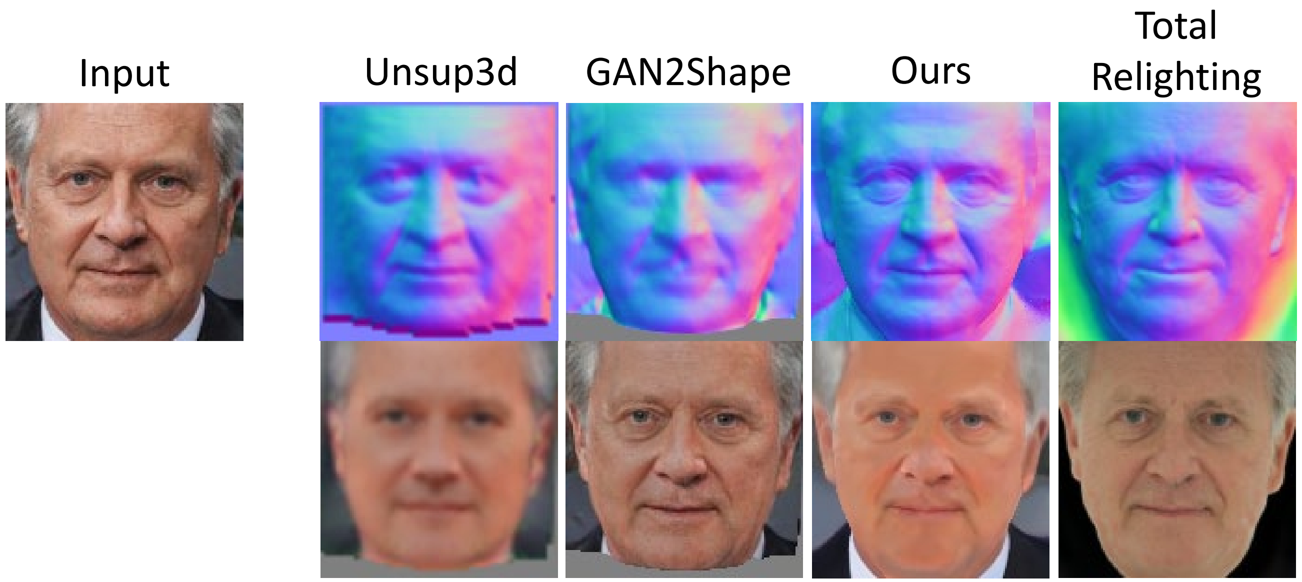
Quantitative Results



Single-view 3D reconstruction on the H3DS dataset.

Method	Unsup3d	GAN2Shape	pi-GAN	ShadeGAN	Ours(w/o SBR)	Ours
CD $\downarrow$	3.60	2.62	3.29	2.49	2.21	**2.08**

Quantitative Results



Quantitative comparison of albedo and surface normal on CelebA

	Unsup3d	GAN2Shape	Ours
SIE ($\times 10^{-2}$) $\downarrow$	3.21	3.05	**2.16**
MAD $\downarrow$	18.66	21.75	**12.67**

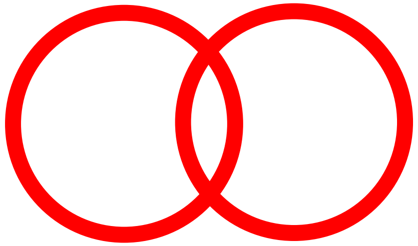
SIE: scale-invariant error
MAD: mean-angle deviation

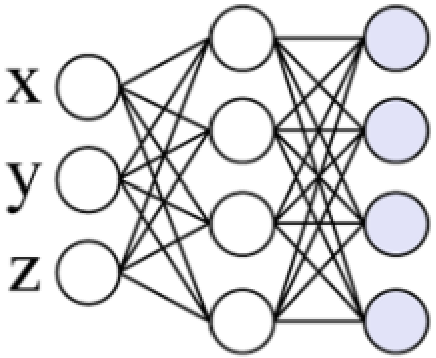
What’s Next?

3D-aware Generative Models Trained on More Diverse Datasets

Multi-modal Learning

Large-scale multimodal learning models

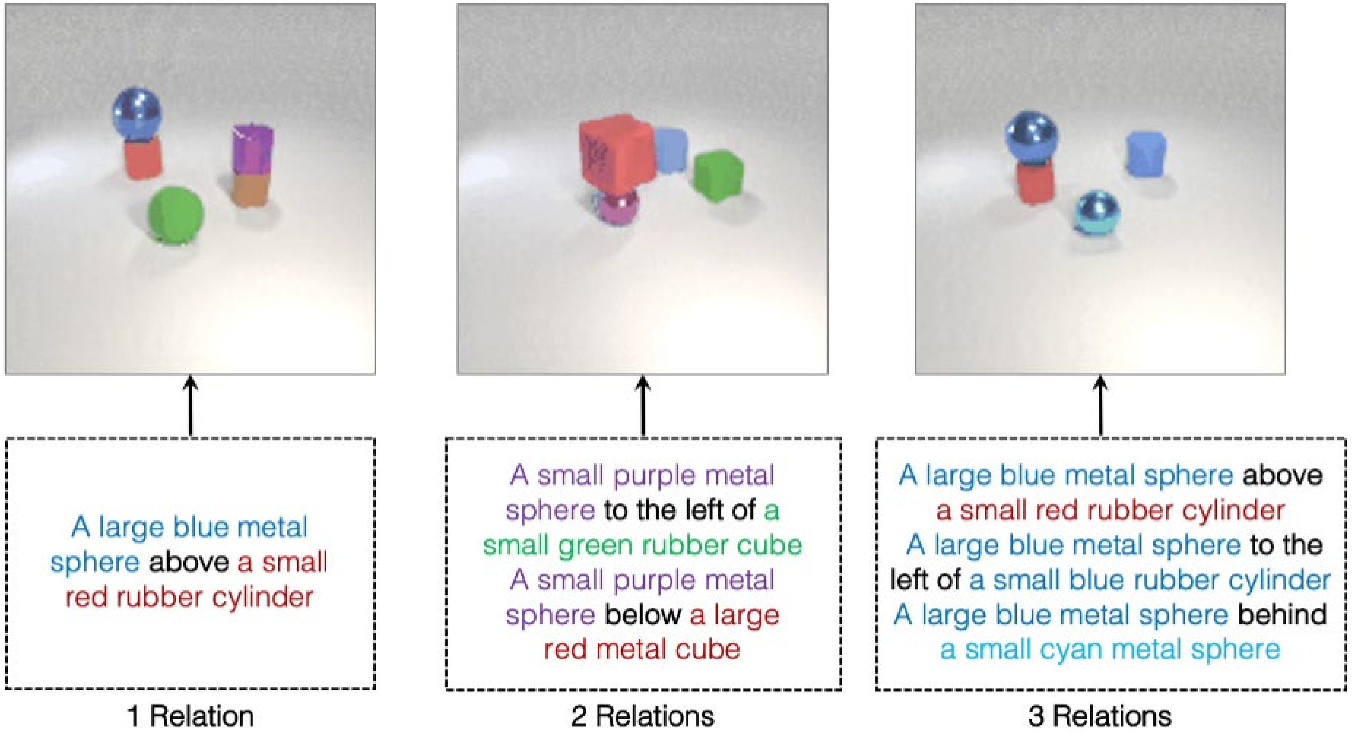




Neural scene representations

Multi-modal Learning

- Text-to-3D generation
- Language learning via 3D generation



Lingjie Liu

Thank you!