Towards Unsupervised Learning of Generative Models for 3D Controllable Image Synthesis

Yiyi Liao[1,2,*] Katja Schwarz[1,2,*] Lars Mescheder[1,2,3,†] Andreas Geiger[1,2]

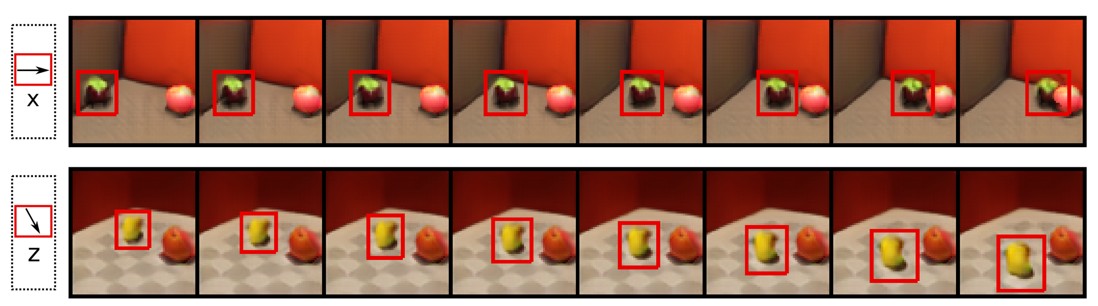
[1]Max Planck Institute for Intelligent Systems, Tübingen

[2]University of Tübingen [3]Amazon, Tübingen

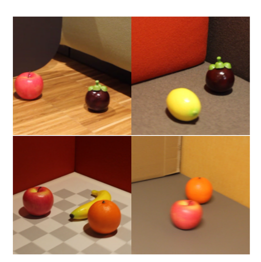
Motivation

Task: 3D Controllable Image Synthesis

- **3D controllability** is essential in many applications, e.g., gaming, simulation, virtual reality and data augmentation
- 3D controllable properties: 3D pose, shape, appearance of **multiple** objects and camera viewpoint
- Is it possible to learn the simulation pipeline including 3D content creation from **raw 2D image observations**?



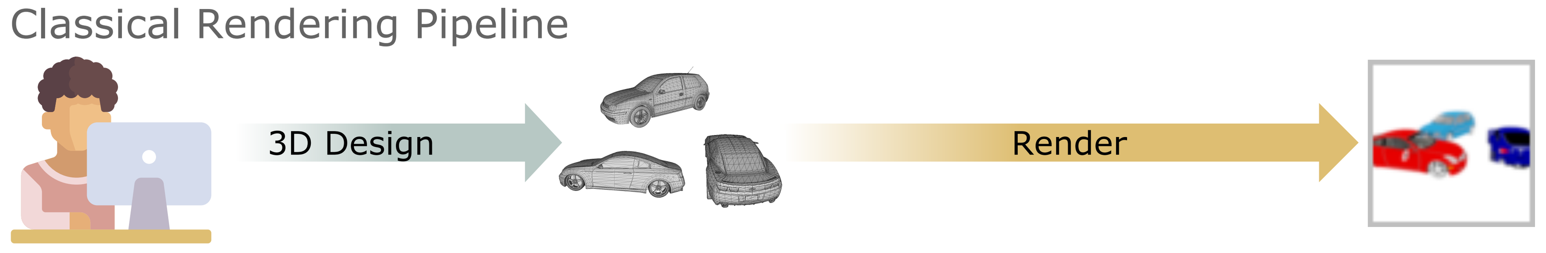
3D controllable image synthesis on Real-world Fruit Dataset



Training images

Generative Models

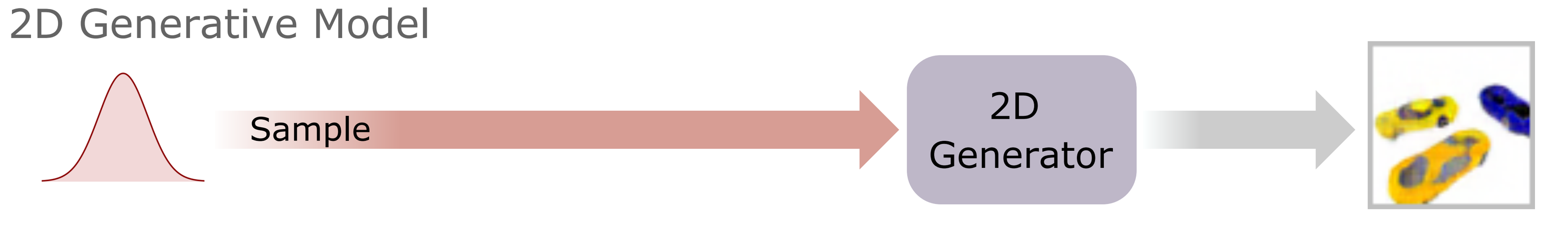
Classical Rendering Pipeline



+ 3D controllable

− Expensive and inefficient to design 3D models

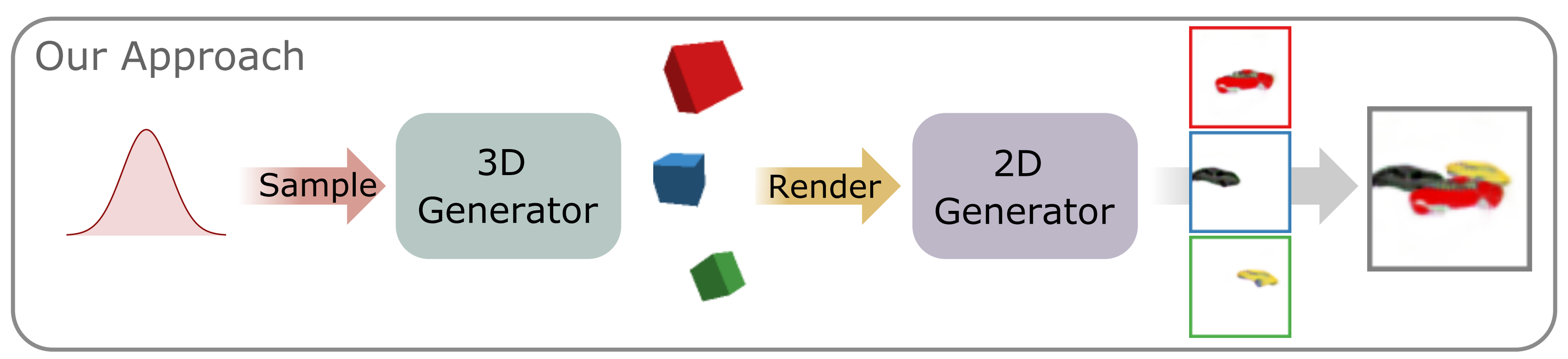
2D Generative Model



+ Efficient and can be learned from only 2D images

− Not 3D controllable

Our Approach

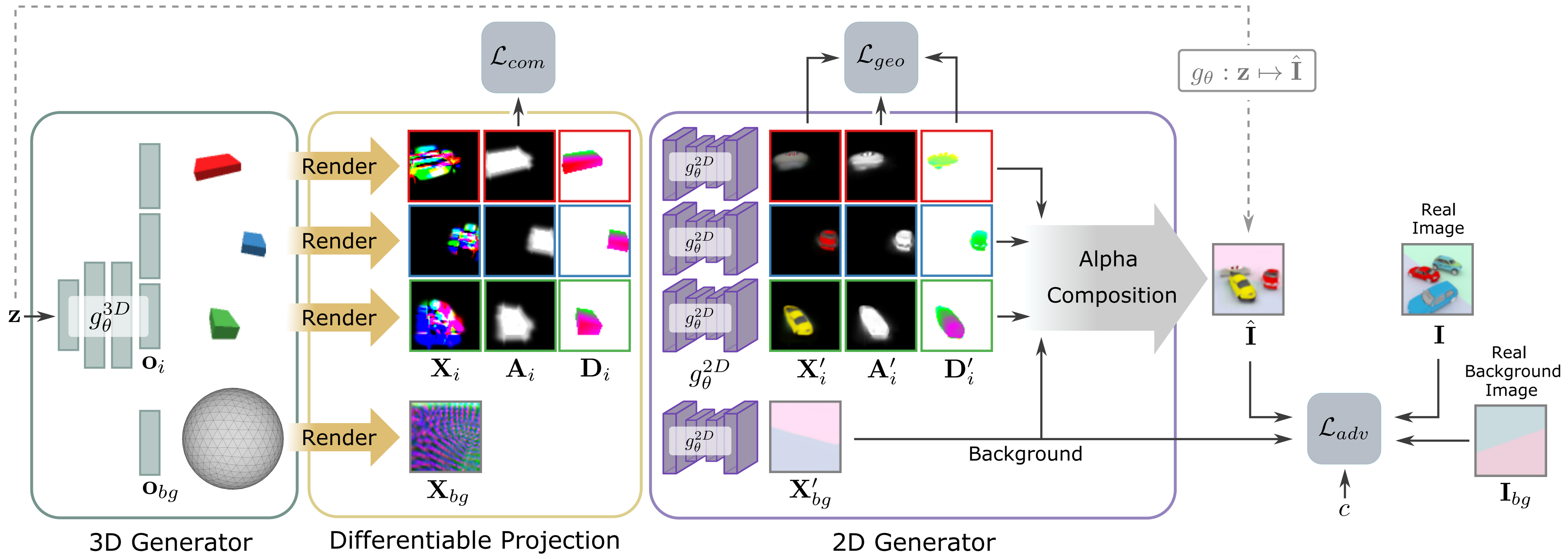


+ Efficient and can be learned from only 2D images

+ 3D controllable

Idea: Learning the image generation process jointly in 3D and 2D space

Method



3D Representations:

Foreground objects $\mathbf{o}_i$:

- $\mathbf{o}_i = (\mathbf{s}_i, \mathbf{R}_i, \mathbf{t}_i, \boldsymbol{\phi}_i)$
- $\boldsymbol{\phi}_i$: Appearance feature
- Primitive type: Point clouds, cuboids, spheres

Scene background $\mathbf{o}_{bg}$:

- Spherical environment map

Loss Functions:

- Adversarial Loss:

$$\mathcal{L}_{adv}(\theta, \psi, c) = \mathbb{E}_{p(\mathbf{z})}[f(d_\psi(g_\theta(\mathbf{z}, c), c))] + \mathbb{E}_{p_{\mathcal{D}}(\mathbf{I}|c)}[f(-d_\psi(\mathbf{I}, c))]$$

- Compactness Loss:

$$\mathcal{L}_{com}(\theta) = \mathbb{E}_{p(\mathbf{z})}\left[\sum_{i=1}^{N} \max\left(\tau, \frac{\|\mathbf{A}_i\|_1}{H \times W}\right)\right]$$

- Geometric Consistency Loss:

$$\mathcal{L}_{geo}(\theta) = \mathbb{E}_{p(\mathbf{z})}\left[\sum_{i=1}^{N} \|\mathbf{A}'_i \odot (\mathbf{X}'_i - \tilde{\mathbf{X}}'_i)\|_1\right] + \mathbb{E}_{p(\mathbf{z})}\left[\sum_{i=1}^{N} \|\mathbf{A}'_i \odot (\mathbf{D}'_i - \tilde{\mathbf{D}}'_i)\|_1\right]$$

Quantitative Results

Ablation Study on Different 3D Representations

	FID	$\text{FID}_\mathbf{t}$	$\text{FID}_\mathbf{R}$	FID_i	MVC[1]	$\mathbf{X}$	$\hat{\mathbf{A}}$	$\hat{\mathbf{I}}$
Vanilla GAN	50	–	–	41				
Point cloud	38	43	44	66	Good			
Cuboid	38	45	45	60	Good			
Sphere	33	45	45	53	Good			
Deformable primitive w/o g_θ^{2D}	69	71	74	69	Good			
Single primitive	30	38	44	–	Pool			

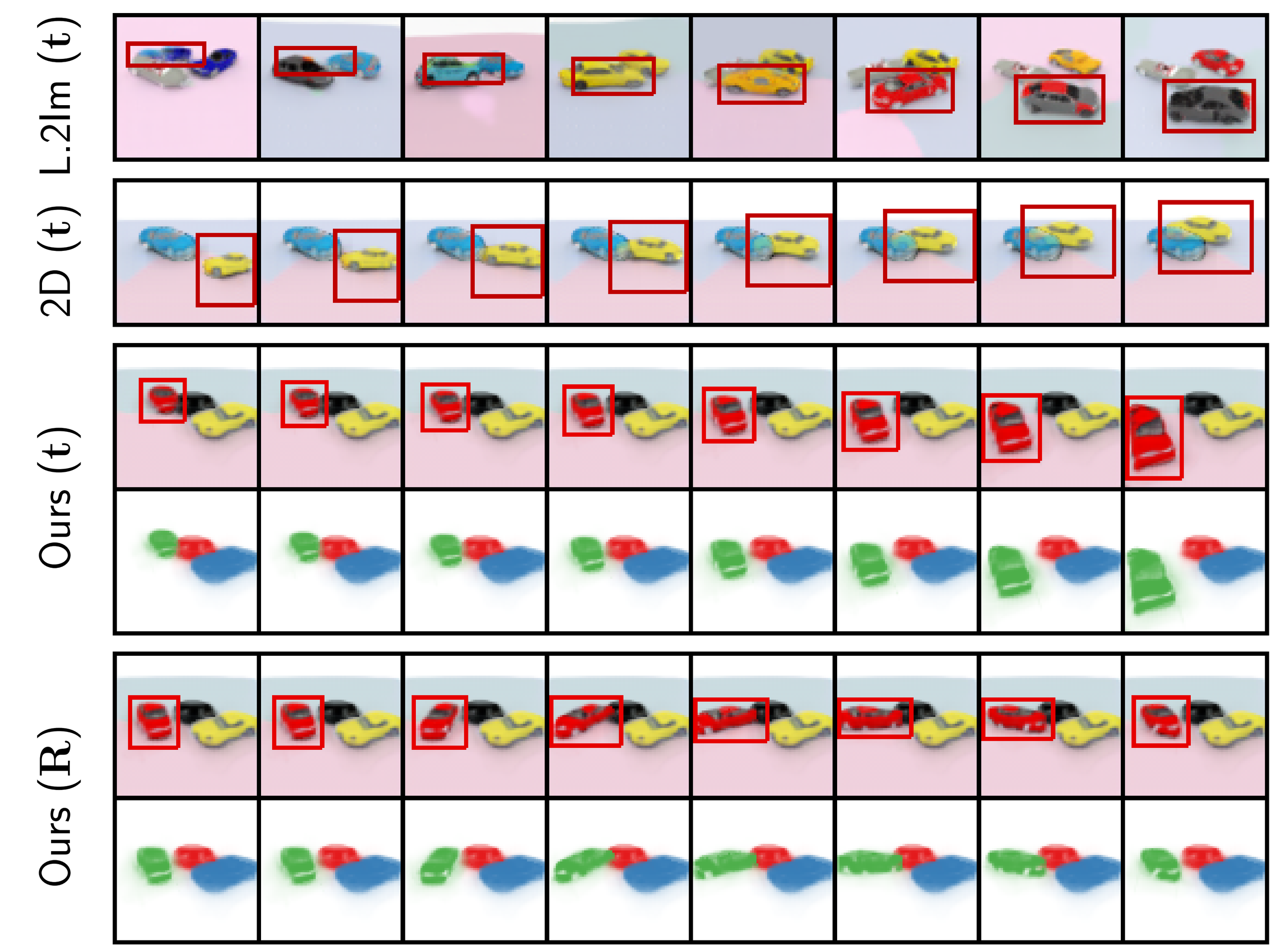
[1]Multi-view consistency

Comparison to Baselines

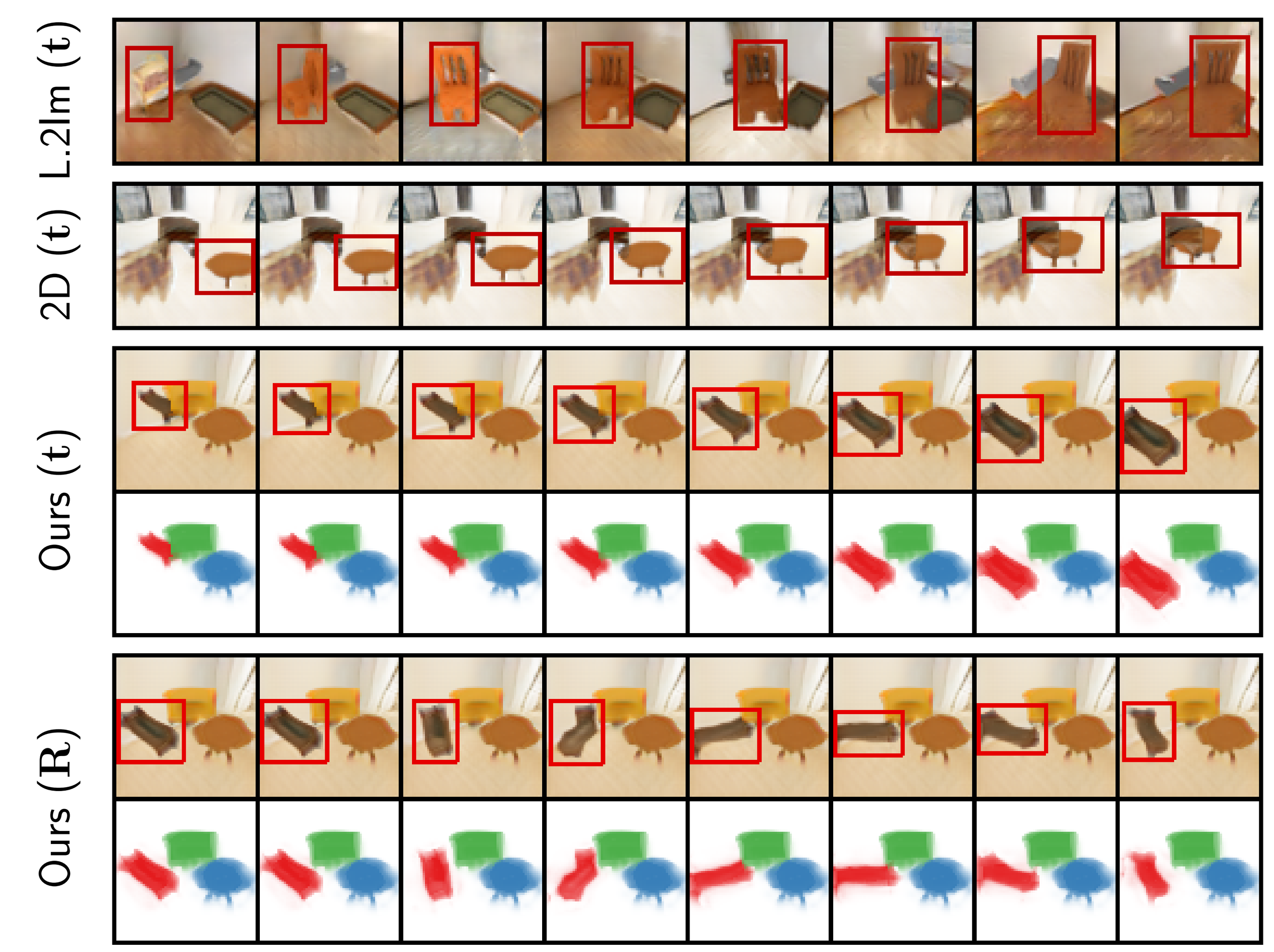
	Car			Indoor		
	FID	$\text{FID}_\mathbf{t}$	$\text{FID}_\mathbf{R}$	FID	$\text{FID}_\mathbf{t}$	$\text{FID}_\mathbf{R}$
Vanilla GAN	**43**	–	–	89	–	–
Layout2Im	**43**	56	–	**84**	93	–
2D Baseline	80	79	–	107	102	–
Ours (w/o c)	65	71	75	120	120	120
Ours	44	**54**	**66**	88	**90**	**100**

Qualitative Results

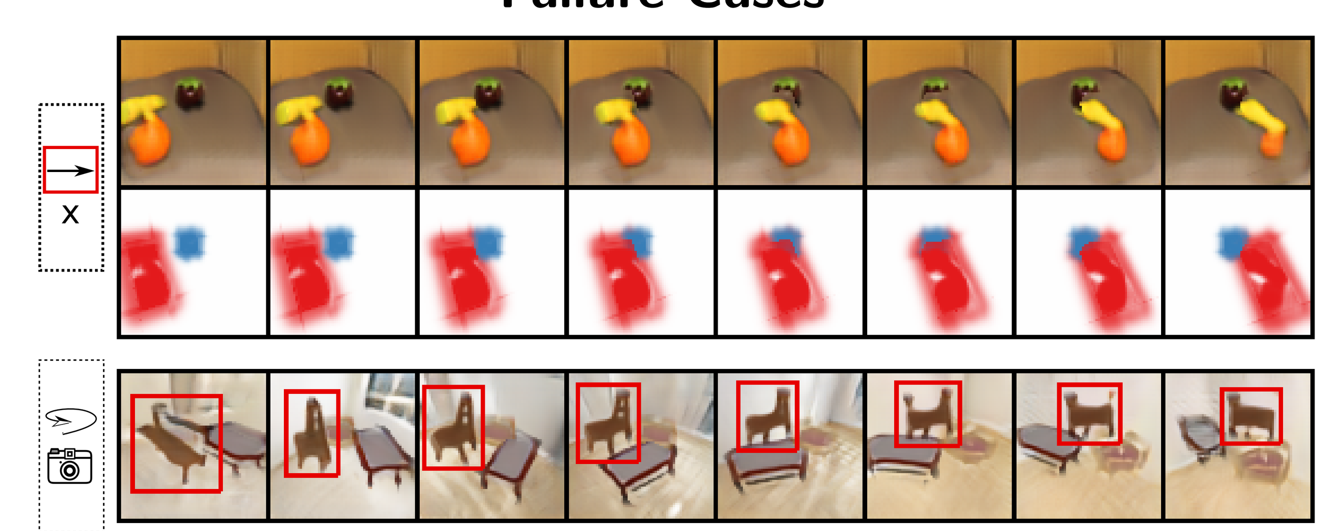
Car Dataset



Indoor Dataset



Failure Cases



https://github.com/autonomousvision/controllable_image_synthesis

{firstname.lastname}@tue.mpg.de