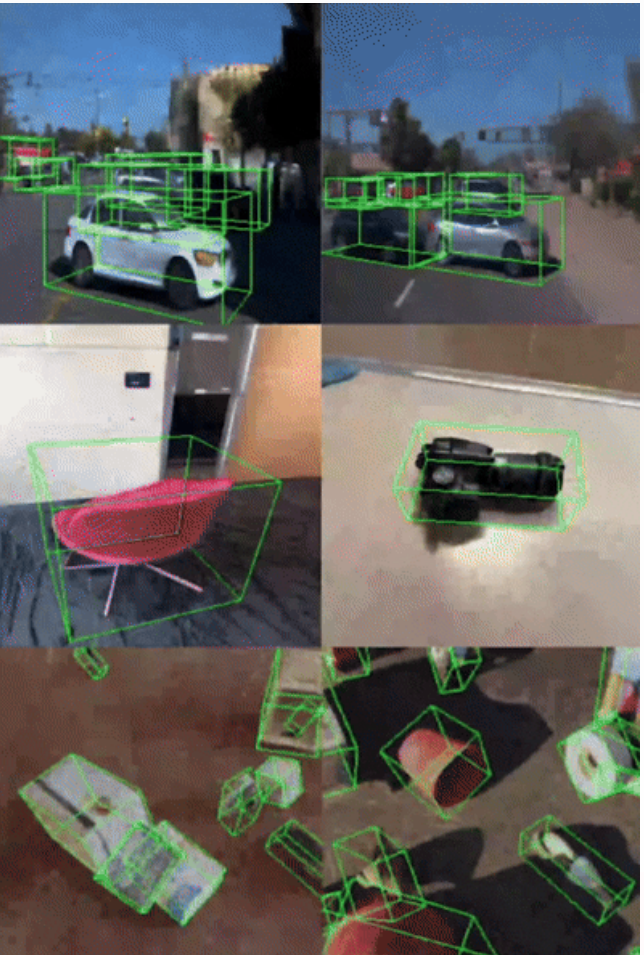


Google DeepMind

Neural Assets

Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew A. Hudson,
Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste,
Kelsey R. Allen, Thomas Kipf

Project page: https://neural-assets.github.io/

NeurIPS 2024

Spotlight

Presenter: Zhiguo Liu

Date: January 29

Georgia State University

Motivation

Traditional computer graphics workflows usually involve

1) the creation of 3D assets

2) the animation of them

Can recent generative models do so?

Deitke, Matt, et al. “Objaverse-xl: A universe of 10m+ 3d objects." NeurIPS. 2024.
Greff, Klaus, et al. "Kubric: A scalable dataset generator." CVPR. 2022.
Zheng, Yang, et al. "PointOdyssey: A large-scale synthetic dataset for long-term point tracking." ICCV. 2023.

Prior Works: 2D Spatial Control in DMs

Condition pre-trained DMs on **semantic** layouts

- 2D boxes, masks, points, etc.

Li, Yuheng, et al. "GLIGEN: Open-set grounded text-to-image generation." CVPR. 2023.
Wang, Xudong, et al. "InstanceDiffusion: Instance-level Control for Image Generation." CVPR. 2024.

Prior Works: 3D-Aware Image DMs

- 3DiM
- Zero-1-to-3

Condition DMs

Its core idea is to introduce additional conditional information during the diffusion denoising process, thereby controlling the generated results.

Condition Type	Description	Example Task
Text Condition (Text Prompt)	Describe the target content using natural language	Stable Diffusion generates "a cat sitting on a tree"
Image Condition (Image Condition)	Provide a reference image for the model to modify	Image inpainting, Super-resolution
Pose Condition (Pose Condition)	Provide 3D pose information to control the object's orientation	3DiM, Zero-1-to-3
Depth Map Condition (Depth Map Condition)	Use depth information to control scene structure	ControlNet Depth-to-Image
Sketch Condition (Sketch Condition)	Guide the model to generate a detailed image based on a sketch	ControlNet Sketch-to-Image
Segmentation Map Condition (Segmentation Map Condition)	Use semantic information to guide generation	Semantic-to-Image
Motion Condition (Motion Condition)	Control the dynamic behavior of objects	Video generation, Animation generation

Watson, Daniel, et al. "Novel view synthesis with diffusion models." ICLR. 2023. Liu, Ruoshi, et al. "Zero-1-to-3: Zero-shot one image to 3d object." ICCV. 2023.

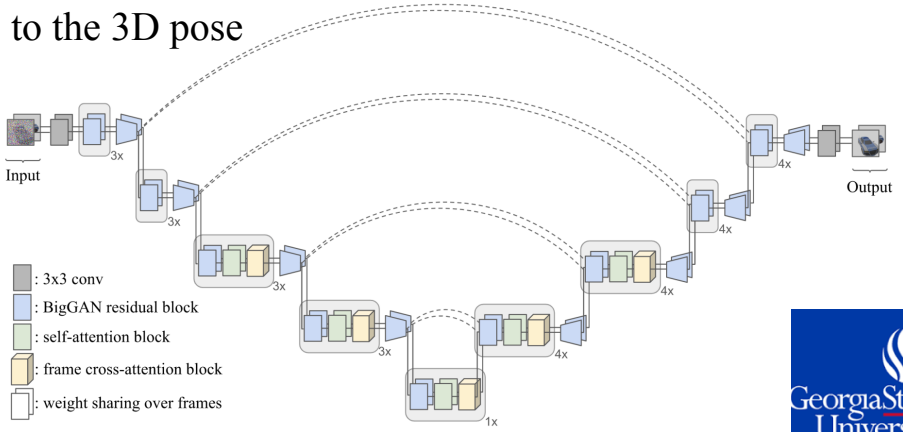
Prior Works: 3D-Aware Image DMs

Condition DMs on pose

- **Core idea:**
  - **Input:** 3D pose information of the object (such as displacement, rotation, depth) + 2D reference image
  - **Process:** In the U-Net network structure of the diffusion model, the 3D pose information is taken as a conditional input
  - **Output:** Generate a 2D image that conforms to the 3D pose

- 3DiM: Camera Pose Embedding
- Zero-1-to-3: Target Viewpoint

3DiM X-UNet Architecture

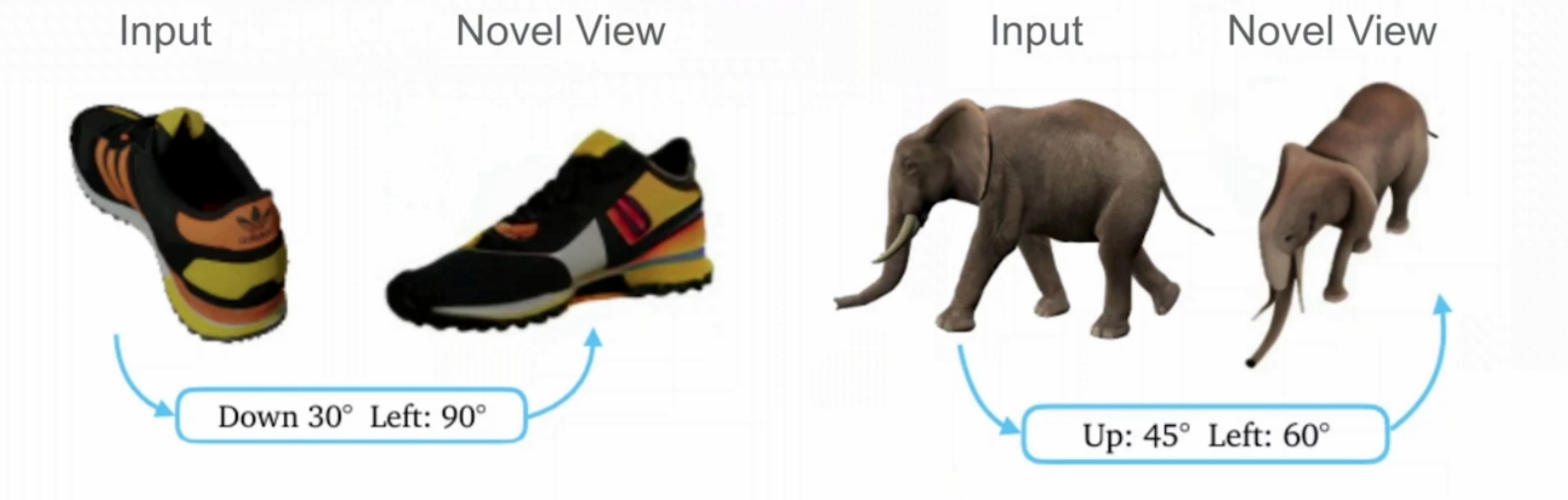


Watson, Daniel, et al. "Novel view synthesis with diffusion models." ICLR. 2023.
Liu, Ruoshi, et al. "Zero-1-to-3: Zero-shot one image to 3d object." ICCV. 2023.

Georgia State University

Prior Works: 3D-Aware Image DMs

- 3DiM & Zero-1-to-3
  - Condition DMs on pose
- **Limitations**: single-object, no background

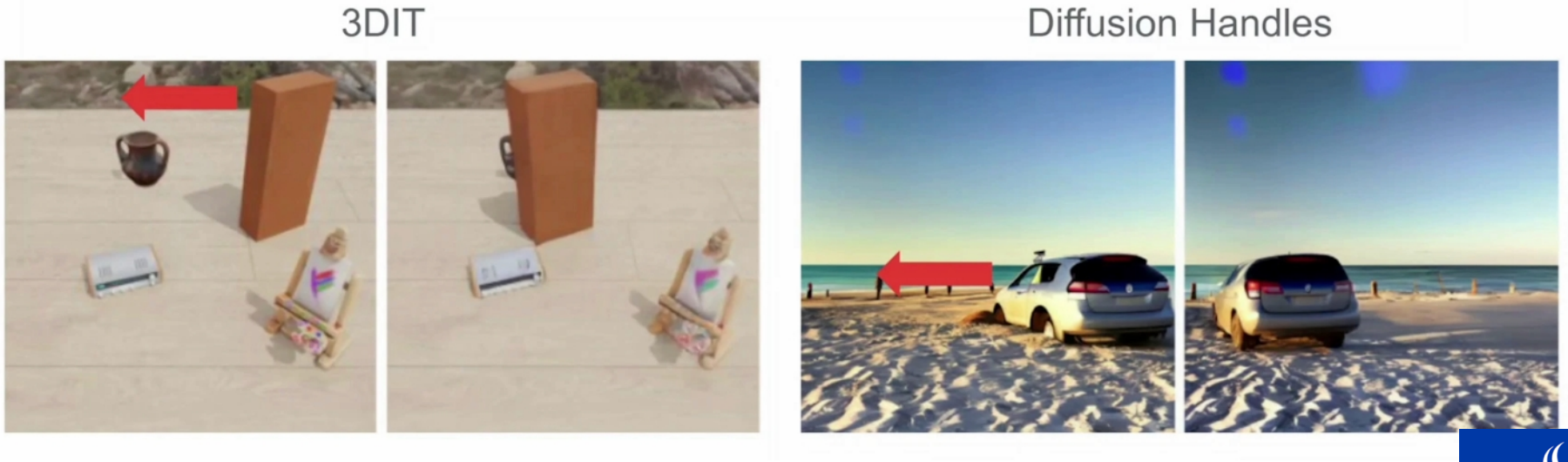


Watson, Daniel, et al. "Novel view synthesis with diffusion models." ICLR. 2023.
Liu, Ruoshi, et al. "Zero-1-to-3: Zero-shot one image to 3d object." ICCV. 2023.

Prior Works: 3D-Aware Image DMs

Recent works on images with backgrounds

- **3DIT**: trained on synthetic data
- **Diffusion Handles**: single-object, only small rotation

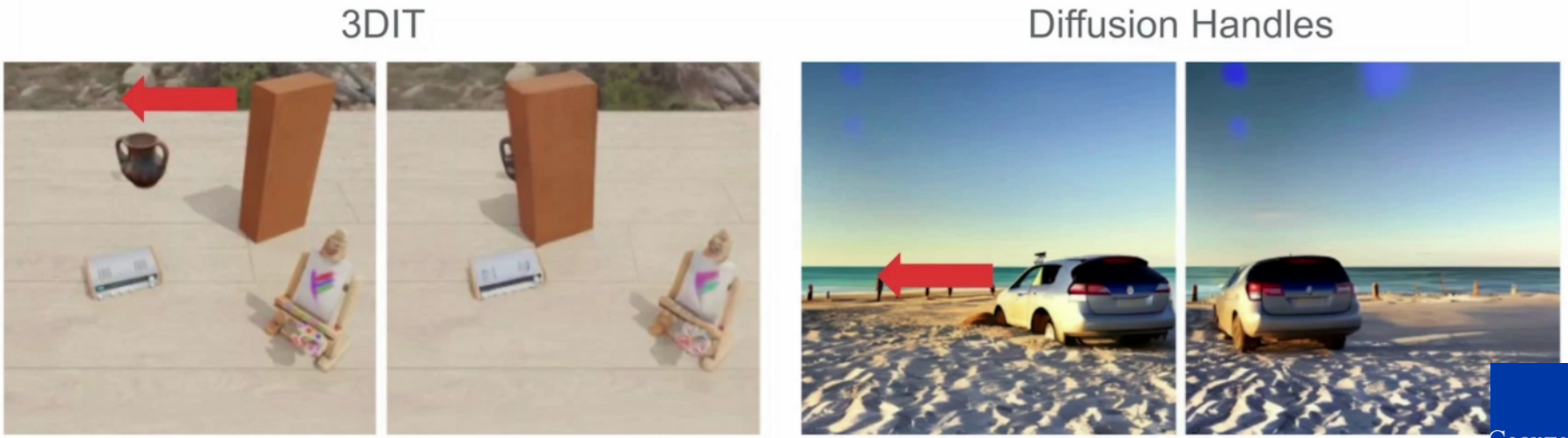


Michel, Oscar, et al. "Object 3DIT: Language-guided 3D-aware image editing." NeurIPS. 2023
Pandey, Karran, et al. "Diffusion Handles: Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D." CVPR. 2024.

Prior Works: 3D-Aware Image DMs

Recent works on images with backgrounds

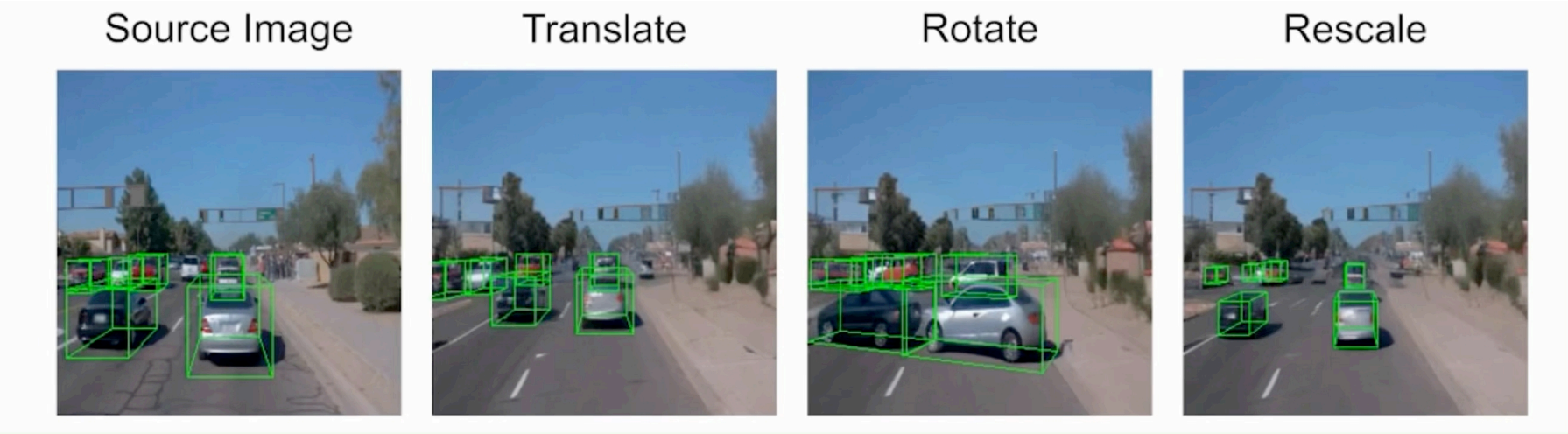
Method	Key Features	Drawbacks
3DiT	Generates 3D-aware scenes through **tokenization**	**Limited explicit control**; relies on token-based generation, making precise object manipulation difficult, can not individual modify the 3D object
Diffusion Handles	Allows users to mark control points in 2D images for localized editing	**Only provides 2D editing**; does not support full 3D-aware object manipulation



Michel, Oscar, et al. "Object 3DIT: Language-guided 3D-aware image editing." NeurIPS. 2023
Pandey, Karran, et al. "Diffusion Handles: Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D." CVPR. 2024.

Neural Assets

- Neural Assets – multi-object 3D control from a single source image
  - Keep correct **neural representation** of each object at diff scene.
  - Manipulate them in a **physically-plausible** way

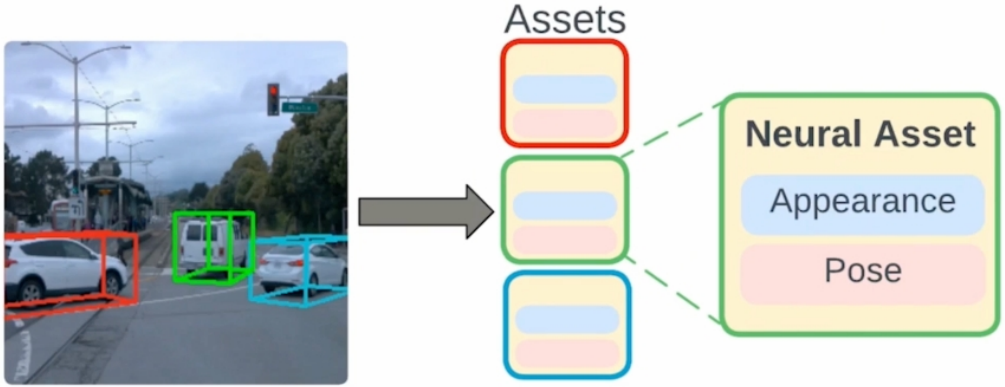


Waymo Open Dataset: Sun, Pei, et al. (CVPR. 2020)

Georgia State University

Method: Neural Assets

- Like Computer Graphics workflows
  - 3D assets are reusable, with a fixed appearance (e.g. canonical 3D shape & texture)
  - Creators only change their pose (e.g. rigid transformation & scaling)
- **Neural Asset:** per-object representation, factorized into **disentangled** appearance and pose tokens



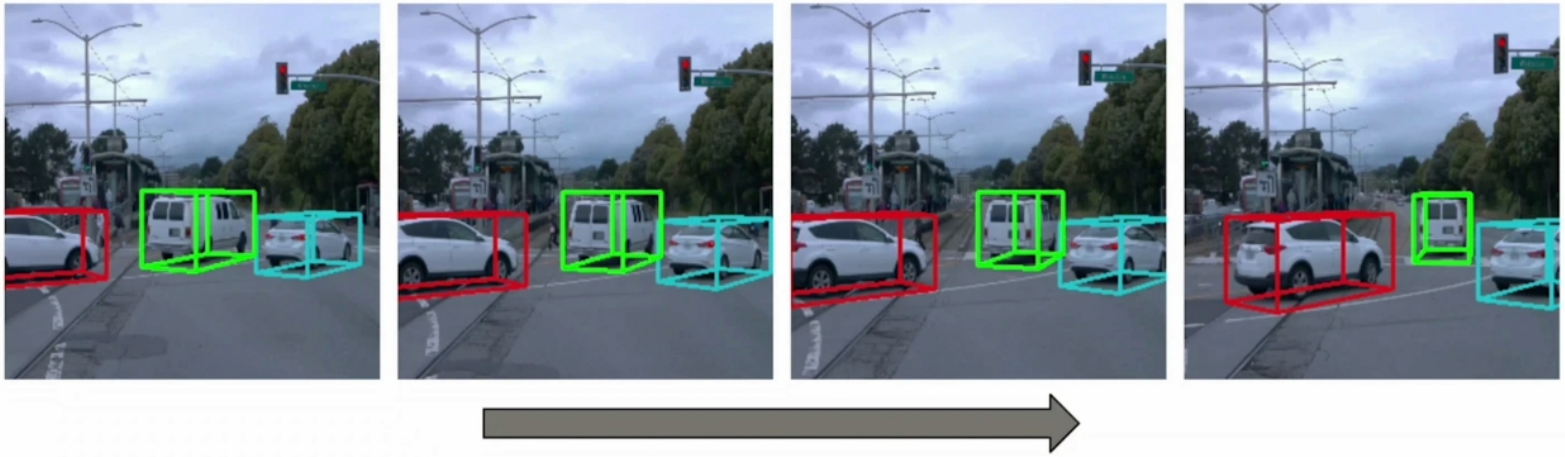
Georgia State University

Method: Neural Assets

- **Disentangle** appearance and pose.
  - **Need**: same object under different poses

Training data must ensure that the pose changes while the object's appearance remains unchanged.

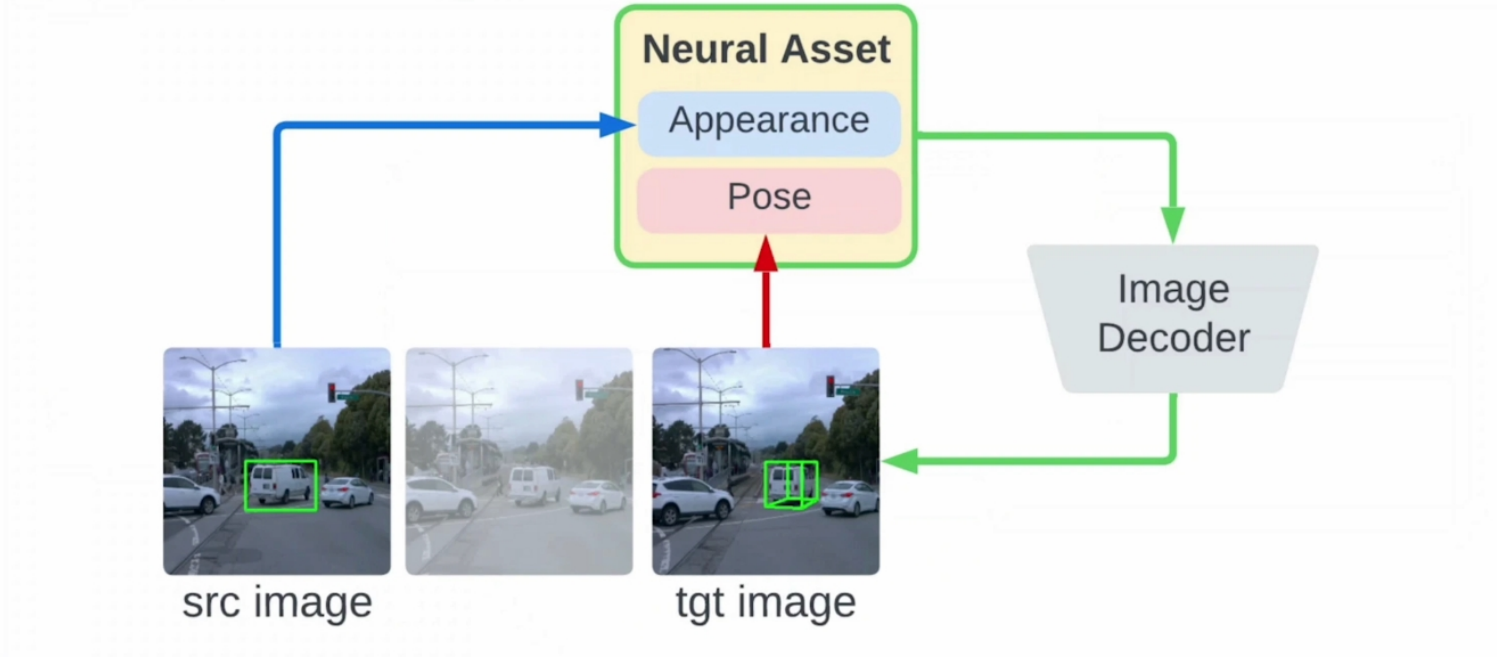
- Video as a scalable data source!



Learning Neural Assets

- **Disentangle** appearance and pose.
  - **High-level idea**: Encode them from different frames of video
  - Appearance is pose-invariant

It extracted two pictures from the video as a frame pair.



Neural Assets Architecture

- **RoIAlign:** Used for extracting the 2D visual features of the foreground objects

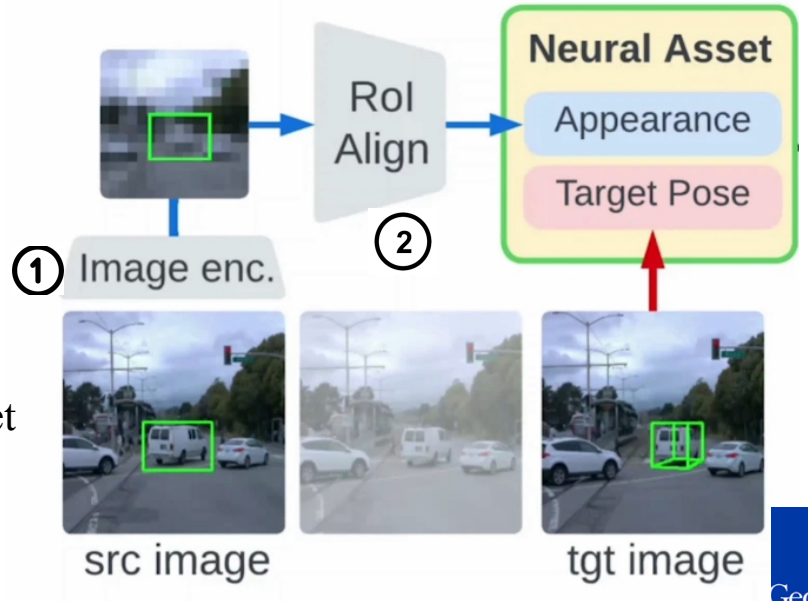
Appearance encoding: $A_i = \text{Flatten}(\text{RoIAlign}(H_i, b_i))\,, \quad H_i = \text{Enc}(x_{\text{src}})\,,$

A_i: Asset appearance tokens b_i: 2D bounding box

2. RoIAlign to crop the features of the target region from the Feature Map.

1. DINO as the visual encoder Enc Identify the target get Feature Map

DINO self-supervised pre-trained ViT-B/8

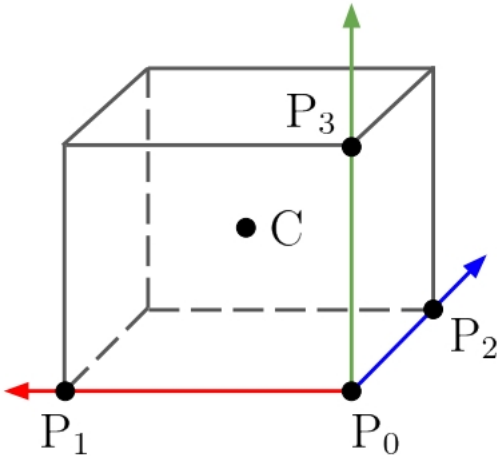


Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." IEEE/CVF. 2021.
RoIAlign: He, Kaiming, et al. "Mask R-CNN." ICCV. 2017.

Neural Assets Architecture

- **Pose**: MLP over 3D boxes

$$P_i = \text{MLP}(C_i)\,, \quad C_i = \text{Concat}[c_i^1, c_i^2, c_i^3, c_i^4]\,, \quad \{c_i^j = (h_i^j, w_i^j, d_i^j)\}_{j=1}^4$$



(a) Object pose as 3D bounding box

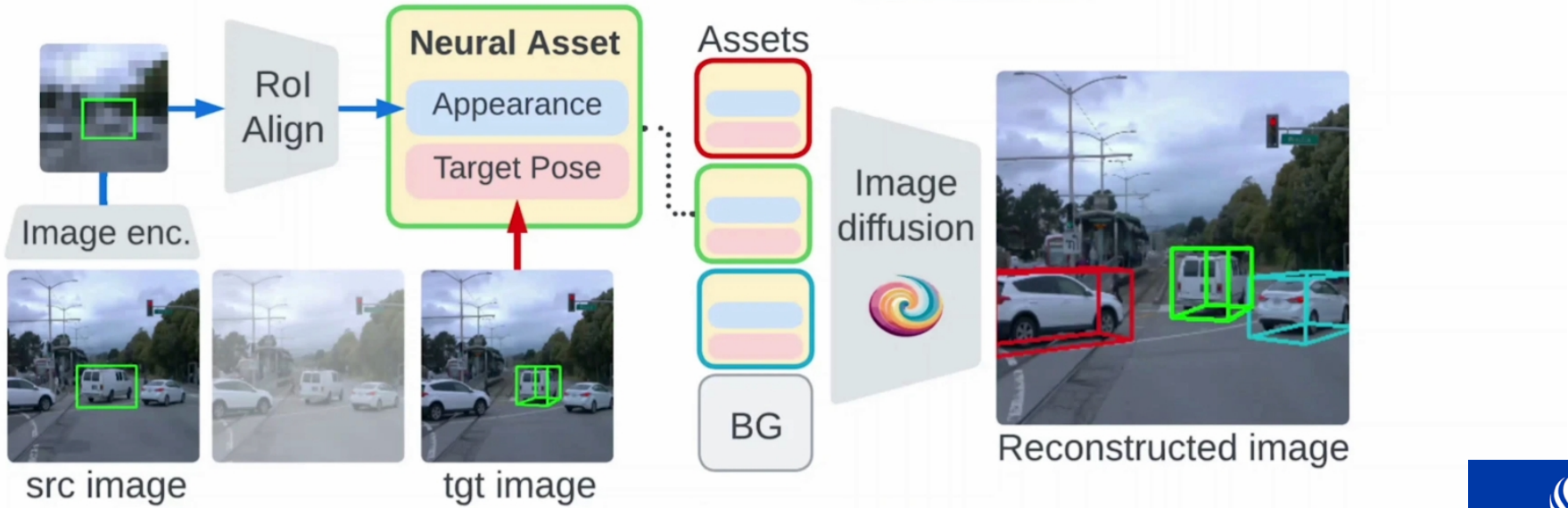
(b) Sample pose representation 1

(c) Sample pose representation 2

Neural Assets Architecture

- **Decoder**: an **fine-tuned** image diffusion model

The pre-trained **text-to-image** model **Stable Diffusion v2.1** is used as the image generator and **fine-tuned** to accept **Neural Assets tokens** as the conditioning signal.



Stable Diffusion v2.1

Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." IEEE/CVF . 2022.

Neural Assets Architecture

BG: independent background modeling

+ rel. camera pose

1. **Occlude the foreground objects:** Use the **2D bounding box** of Neural Assets to occlude the foreground objects in the image, leaving only the background part.

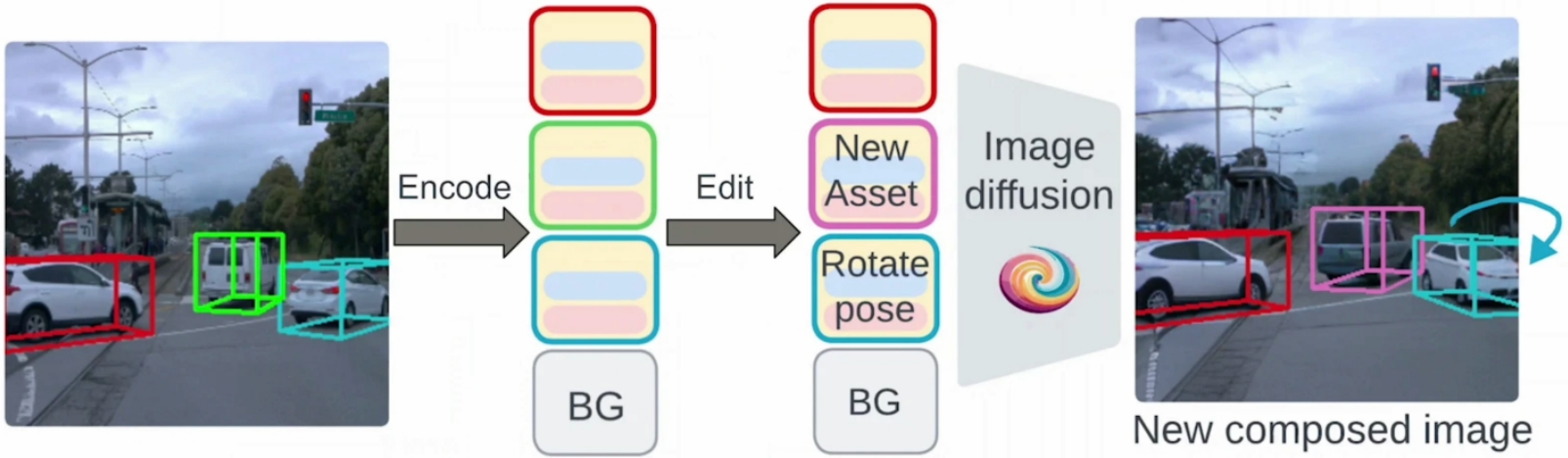
2. **Extract background features:** Input the occluded image into the image encoder Enc and extract the features of the background region using global **RoIAlign** to obtain the background appearance token Abg.

3. **Add pose token:** Concatenate Abg with the pose token Pbg to obtain the complete background token.

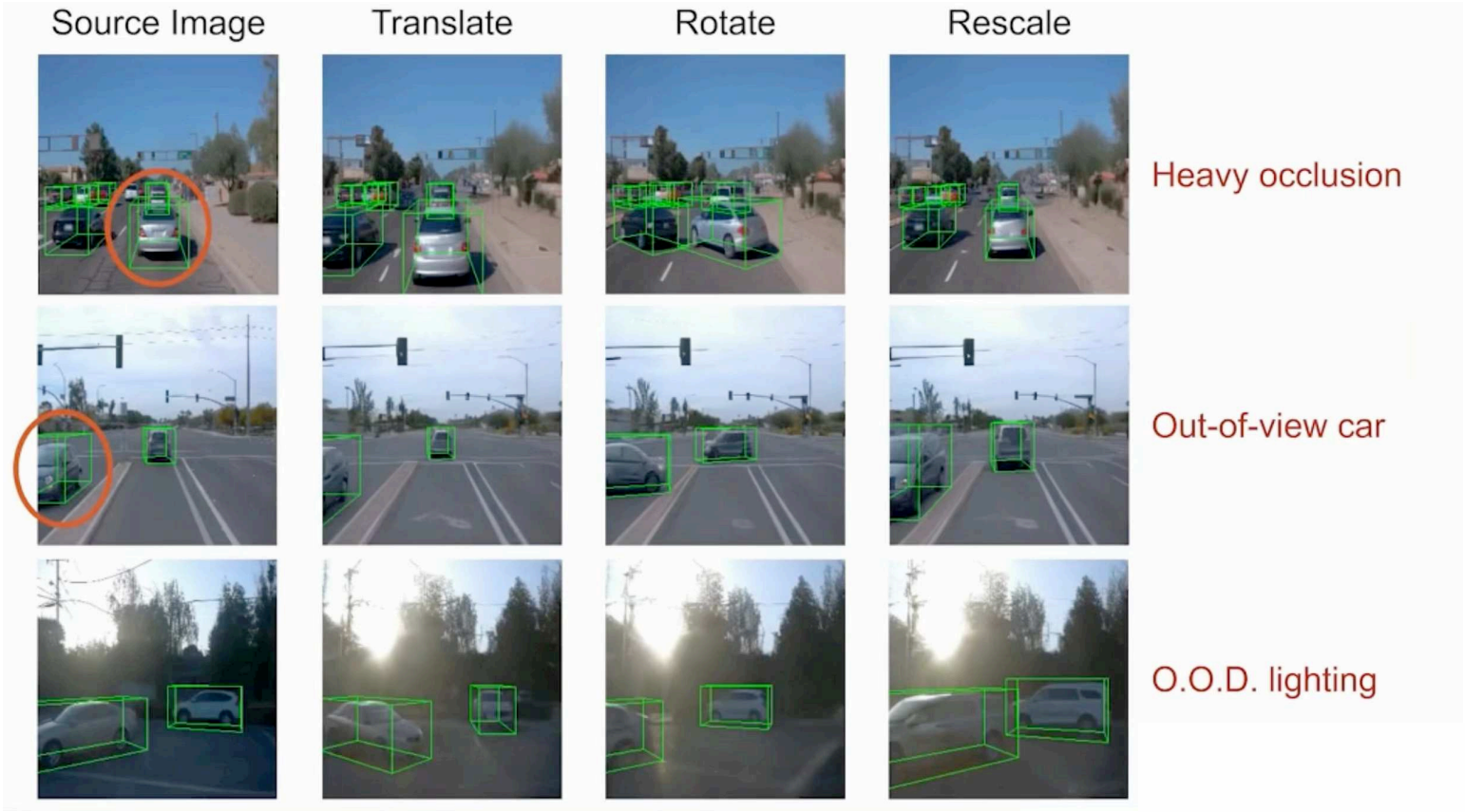
Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." IEEE/CVF. 2021.
RoIAlign: He, Kaiming, et al. "Mask R-CNN." ICCV. 2017.

Test-time Controllability

- Neural Assets as a unified interface
  - Pose control: transform 3D boxes
  - Compositional generation: replace assets

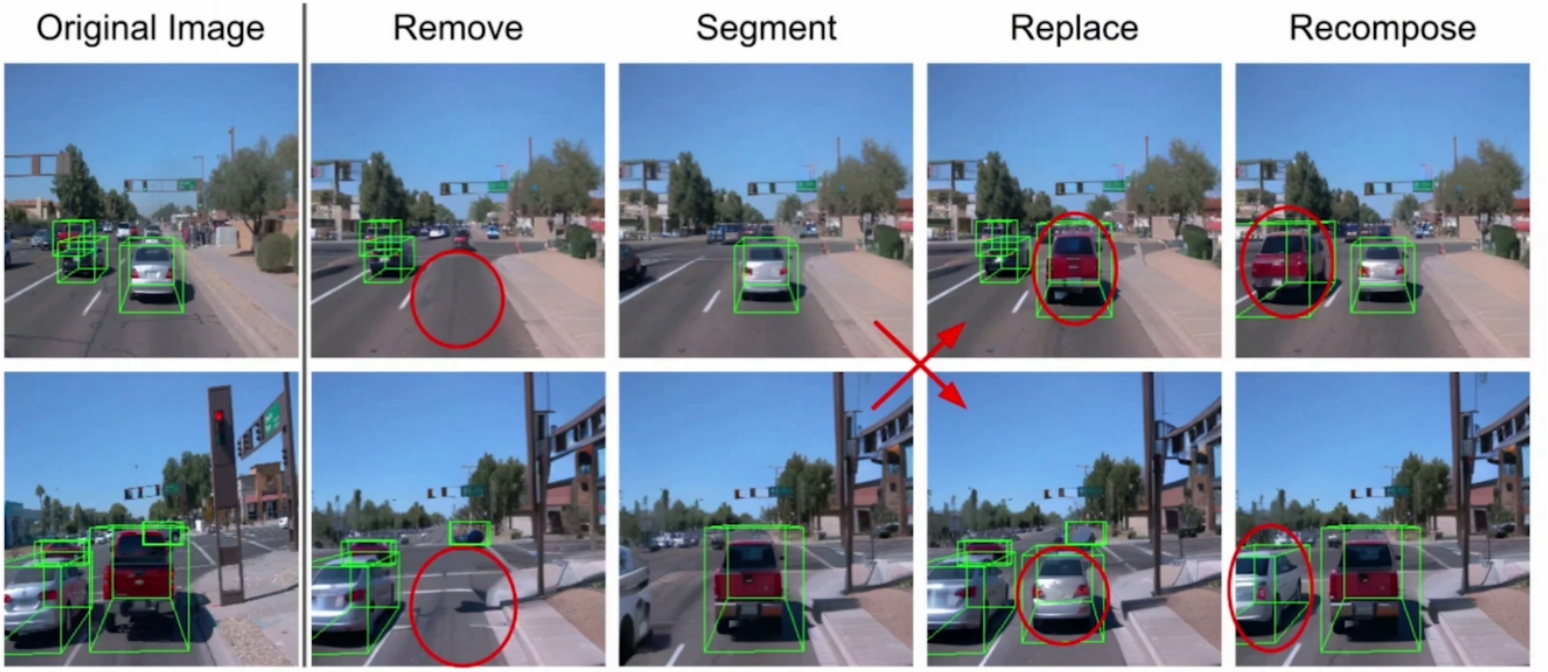


Controllable Scene Generation



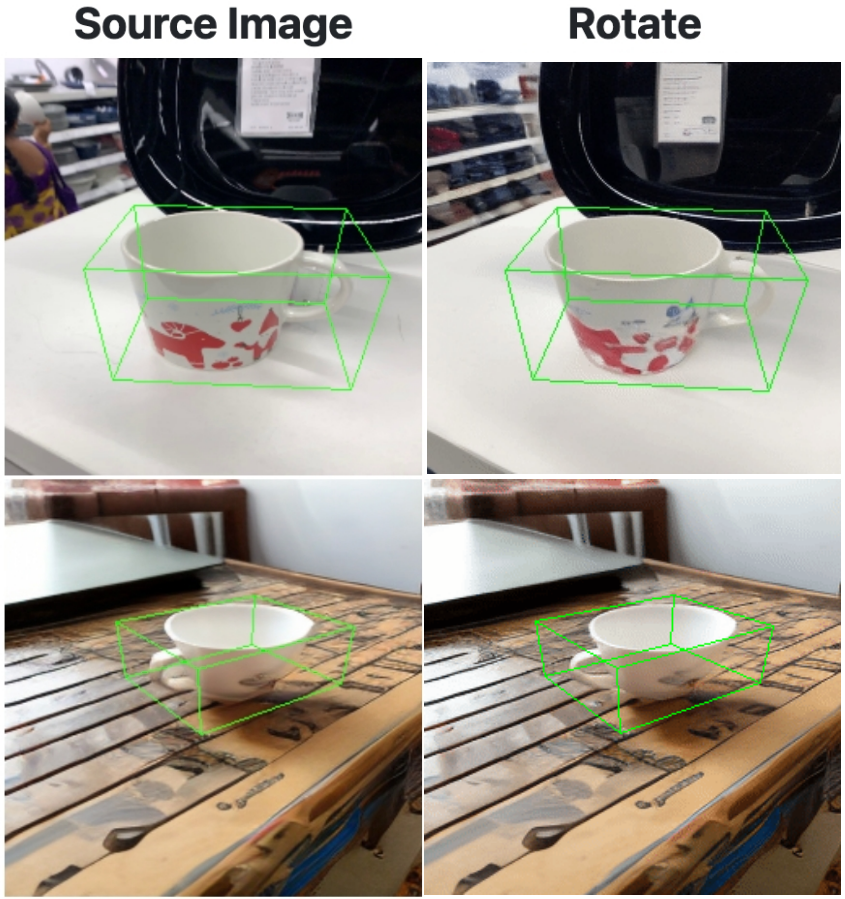
Controllable Scene Generation

Simply drop / replace / append Neural Assets across scenes



Global Background Replacement

Failure Case: symmetry ambiguity



One main failure case of this model is symmetry ambiguity.

As can be seen from the following rotation results, the handle of the cup gets **flipped** when it **rotates 180 degree**.

Q&A