Object-Centric Slot Diffusion for Unsupervised Compositional Generation

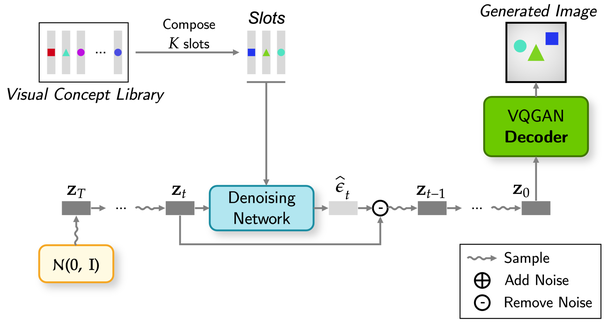
Jindong Jiang[1], Fei Deng[1], Gautam Singh[1], Sungjin Ahn[2]
[1]Rutgers University, [2]KAIST

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

- Unsupervised compositional generation in object-centric learning aims to synthesize novel images using visual concepts derived from existing images without supervised guidance. Existing methods are limited by constraints in image decoders, making them incompetent to handle complex realistic scenes.
- We propose ***Latent Slot Diffusion*** (LSD) which replaces traditional slot decoders with a slot-conditioned latent diffusion model, resulting in superior performance in **object segmentation, compositional generation,** and **component-based image editing** compared to state-of-the-art approaches.

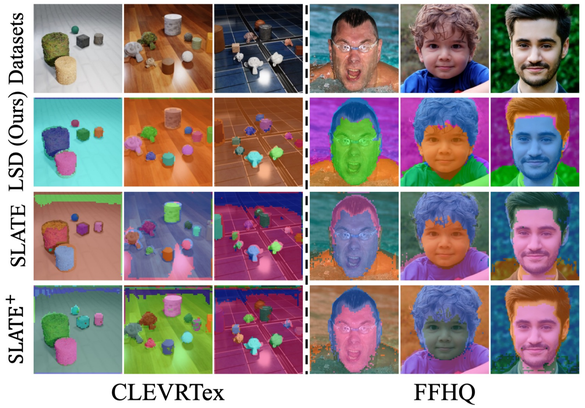
Latent Slot Diffusion



LSD consists of three main components: the **Object-Centric Encoder,** the **Visual Concept Library,** and the **Latent Slot Diffusion Decoder**.

- **Object-Centric Encoder** decomposes and represent an input image as a collection slots, where each slot represent a compositional entity in the image.
- **Visual Concept Library** utilizes clustering to generate a collection of visual concepts. Each concept is represented by a set of slots extracted from unlabeled images.
- **Latent Slot Diffusion Decoder** leverages recent advancements in generative modeling based on diffusion to enable high-quality slot-conditioned image generation.

Experiment: Unsupervised Object Segmentation

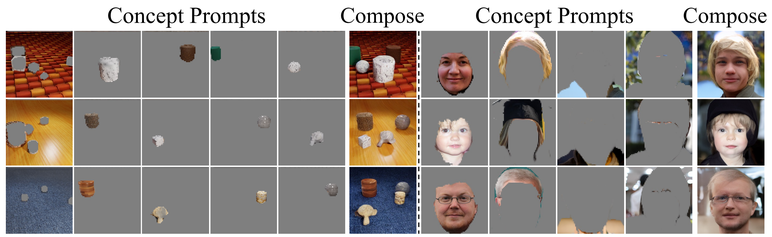


- By jointly training the object-centric encoder with the denoising objective, LSD achieves the ability of unsupervised object segmentation.
- Our experiments demonstrate that LSD outperforms state-of-the-art object-centric learning models on multi-object datasets.

Experiment: Compositional Image Generation

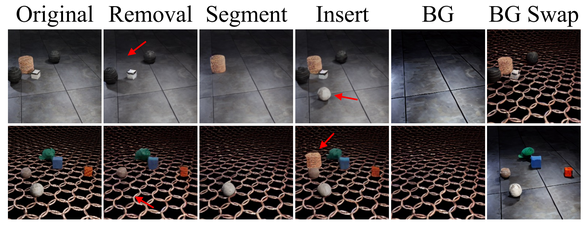
- Similar to text-to-image generative models, LSD has the capability to generate new images by taking unseen slot-based prompts at test time.
- To accomplish this, we create a slot-based prompt by sampling one slot representation from each visual concept library, and then utilize the diffusion decoder to generate the image.

- Followings are the visualization of concept prompts from the visual concept library and the corresponding generated image by LSD.



Experiment: Slot-Based Image Editing

- In addition to generating new images using visual concept library, LSD also allows images editing through slot manipulation.
- On the CLEVRTex dataset, we perform object removal, single-object extraction, object insertion, background extraction, and background swapping.



- On the FFHQ dataset, we demonstrate the face replacement task.