NEURAL INFORMATION PROCESSING SYSTEMS

Visual Concept Tokenization

NeurIPS 2022

Tao Yang, Yuwang Wang, Yan Lu, Nanning Zheng

XIAN JIAOTONG UNIVERSITY 1896

Microsoft

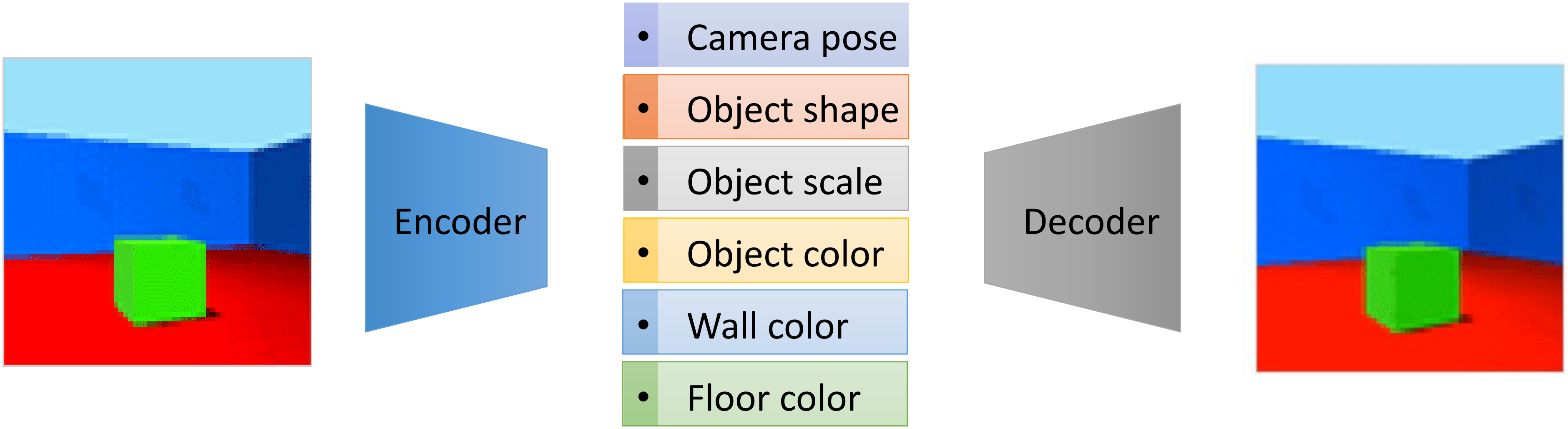
Motivation

Our Motivation is:

- Obtaining the human-like perception ability of abstracting visual concepts from concrete pixels has always been fundamental and important.
- We are particularly interested in finding a general way to learn visual concepts from pixels, which covers two branches, disentangled representation learning and scene decomposition.
- VCT extracts the visual concept inside a given image as a set of tokens, serving as a general solution for visual concept learning, similar to word embeddings in Natural Language Processing (NLP).

Background: Disentanglement

Disentangled representation should reflect the factors of variations behind the observed data of the world, and one latent unit is only sensitive to changes of an individual factor.



For **disentangled representation learning** , the goal is learning to extract such representations of the factors from the images.

Definition of visual concepts token (VCT)

- Concept Prototypes
  - Embeddings of different visual concepts prototypes (meta concepts).
  - A meta concept is represented by a single embedding.
  - Dataset-level concept (for the dataset).
  - A set of embeddings $\{c_1, c_2, \ldots, c_\mathrm{n}\}$ of a dataset.

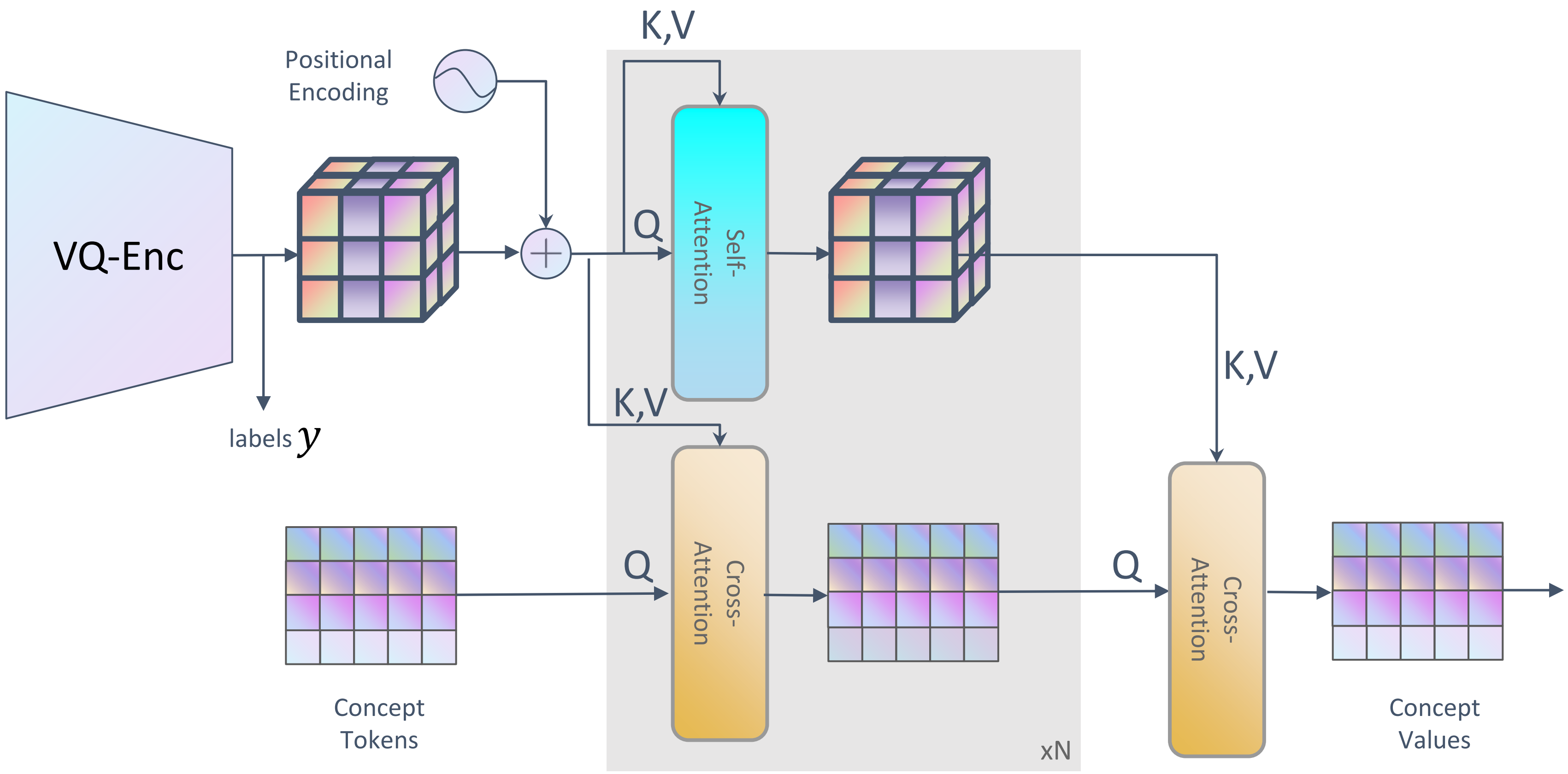
- Concept Tokens
  - The value embedding of different concept of a single image.
  - Image-level concept (for a single image)
  - A set of embeddings $\{v_{1i}, v_{2i}, \ldots v_{ni}\}$ of a single image.

Properties for VCT & disentanglement

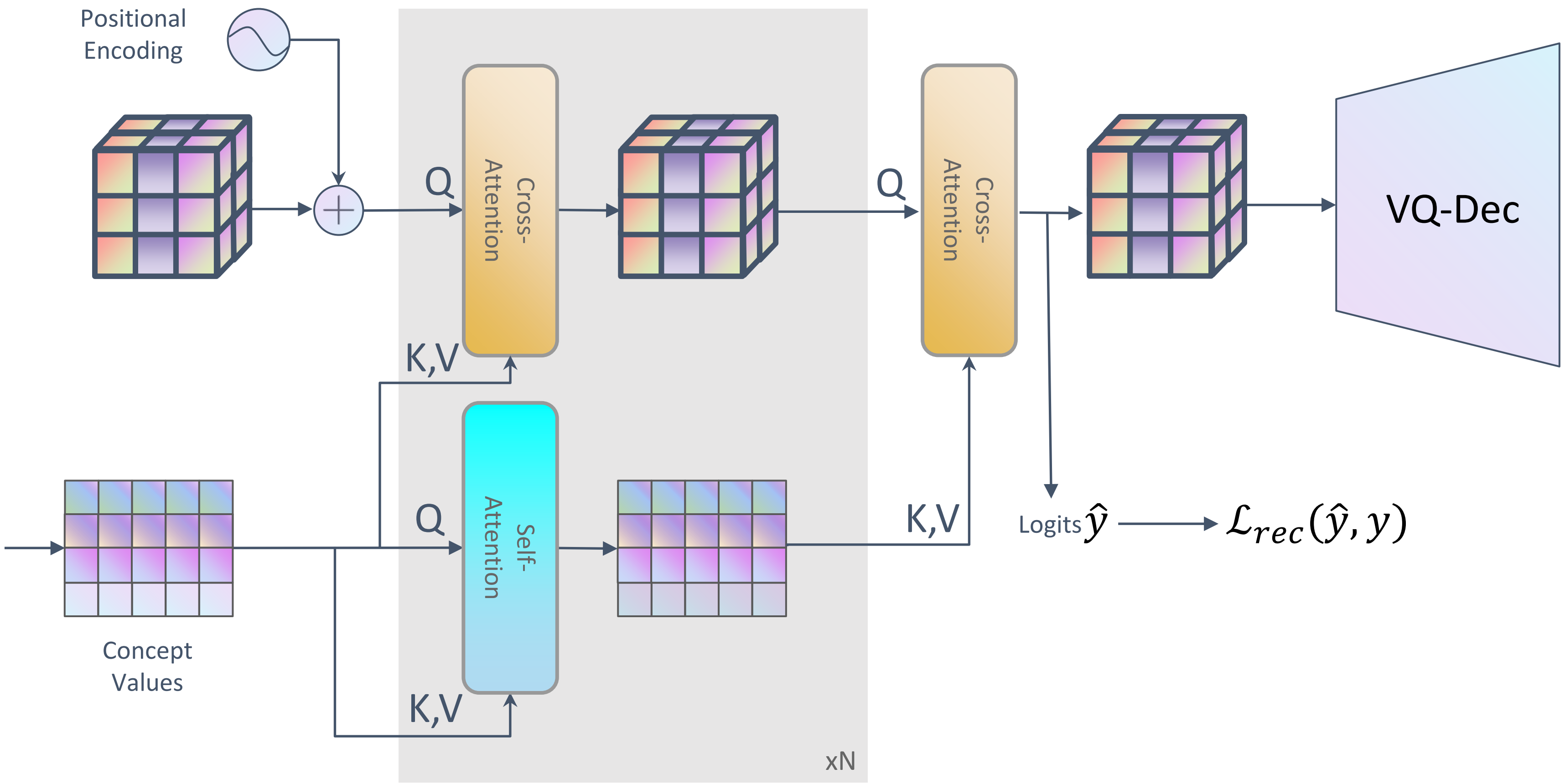
- What should be a good concept prototype?
  - A disordered set of dataset-level embedding
  - Query the corresponding concept from data
  - Different concept prototype are disentangled

- What should be a good concept tokens?
  - A disordered set of image-level embedding
  - The extraction of concept tokens should be independent
  - The concept tokens can be decoded back to data

Method (Encoder part)



Method (Decoder part)



Method

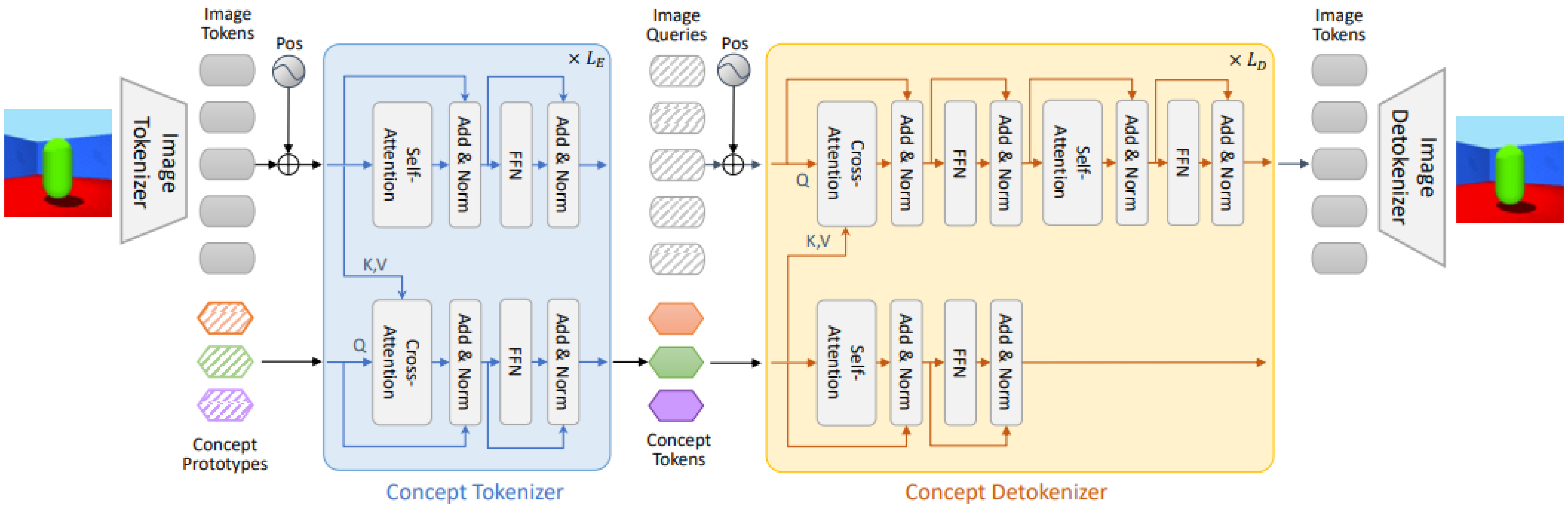


Figure 1: The framework of Visual Concept Tokenization (VCT). An image is represented as a set of concept tokens, and each token reflects a visual concept, such as green object color, blue background color. The concept prototypes and image queries are shared across different images.

Method

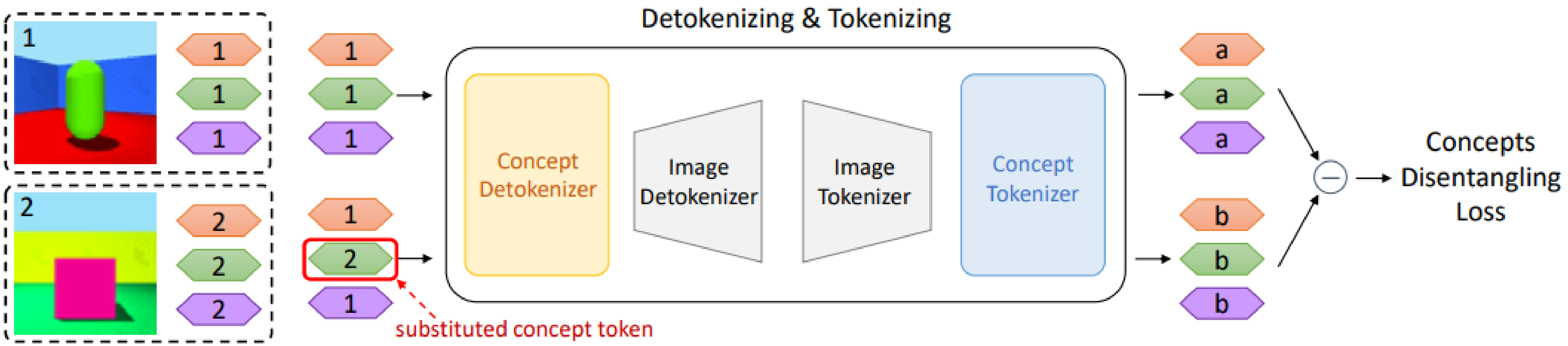


Figure 2: Illustration of Concept Disentangling Loss. The second concept token (labeled in green color) is substituted (from 1 to 2) to create the visual variation. Concept tokens {a,a,a} and {b,b,b} are the outputs for inputing {1,1,1} and {1,2,1} to detokenizing and tokenizing, respectively.

$$\Delta C = \mathcal{V}_T(\mathcal{I}_T(x_i')) - \mathcal{V}_T(\mathcal{I}_T(\hat{x}_i')).$$

$$\mathcal{L}_{dis} = CrossEntropy(norm(\Delta C), l),$$

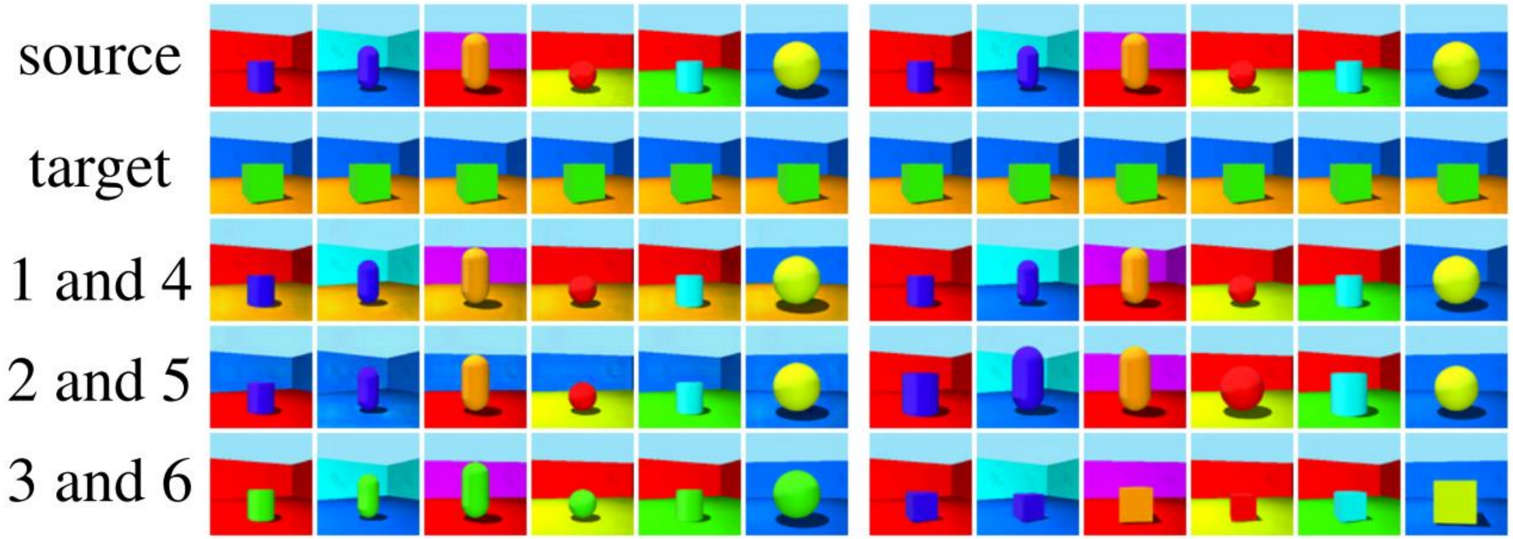
$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{dis}\mathcal{L}_{dis}$$

Experiments

- Disentanglement results

Method	Cars3D		Shapes3D		MPI3D	
	FactorVAE score	DCI	FactorVAE score	DCI	FactorVAE score	DCI
VAE-based:						
FactorVAE	0.906 ± 0.052	0.161 ± 0.019	0.840 ± 0.066	0.611 ± 0.082	0.152 ± 0.025	0.240 ± 0.051
β-TCVAE	0.855 ± 0.082	0.140 ± 0.019	0.873 ± 0.074	0.613 ± 0.114	0.179 ± 0.017	0.237 ± 0.056
GAN-based:						
InfoGAN-CR	0.411 ± 0.013	0.020 ± 0.011	0.587 ± 0.058	0.478 ± 0.055	0.439 ± 0.061	0.241 ± 0.075
Pre-trained GAN-based:						
LD	0.852 ± 0.039	0.216 ± 0.072	0.805 ± 0.064	0.380 ± 0.062	0.391 ± 0.039	0.196 ± 0.038
CF	0.873 ± 0.036	0.243 ± 0.048	0.951 ± 0.021	0.525 ± 0.078	0.523 ± 0.056	0.318 ± 0.014
GS	0.932 ± 0.018	0.209 ± 0.031	0.788 ± 0.091	0.284 ± 0.034	0.465 ± 0.036	0.229 ± 0.042
DS	0.871 ± 0.047	0.222 ± 0.044	0.929 ± 0.065	0.513 ± 0.075	0.502 ± 0.042	0.248 ± 0.038
DisCo	0.855 ± 0.074	0.271 ± 0.037	0.877 ± 0.031	0.708 ± 0.048	0.371 ± 0.030	0.292 ± 0.024
Concept-based:						
COMET	0.339 ± 0.008	0.024 ± 0.026	0.168 ± 0.005	0.002 ± 0.000	0.145 ± 0.024	0.005 ± 0.001
VCT (Ours)	$\mathbf{0.966 \pm 0.029}$	$\mathbf{0.382 \pm 0.080}$	$\mathbf{0.957 \pm 0.043}$	$\mathbf{0.884 \pm 0.013}$	$\mathbf{0.689 \pm 0.035}$	$\mathbf{0.475 \pm 0.005}$

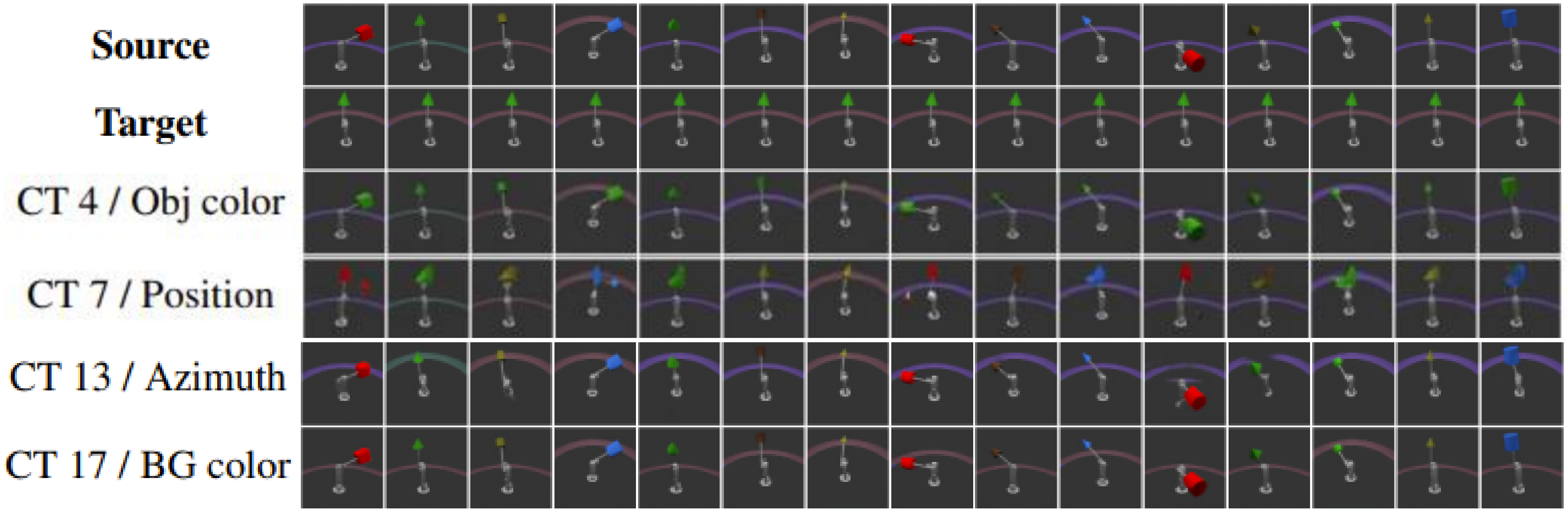
Qualitative Results (Shape3D) & Ablation study



Swapping latent concept on (a) Shapes3D

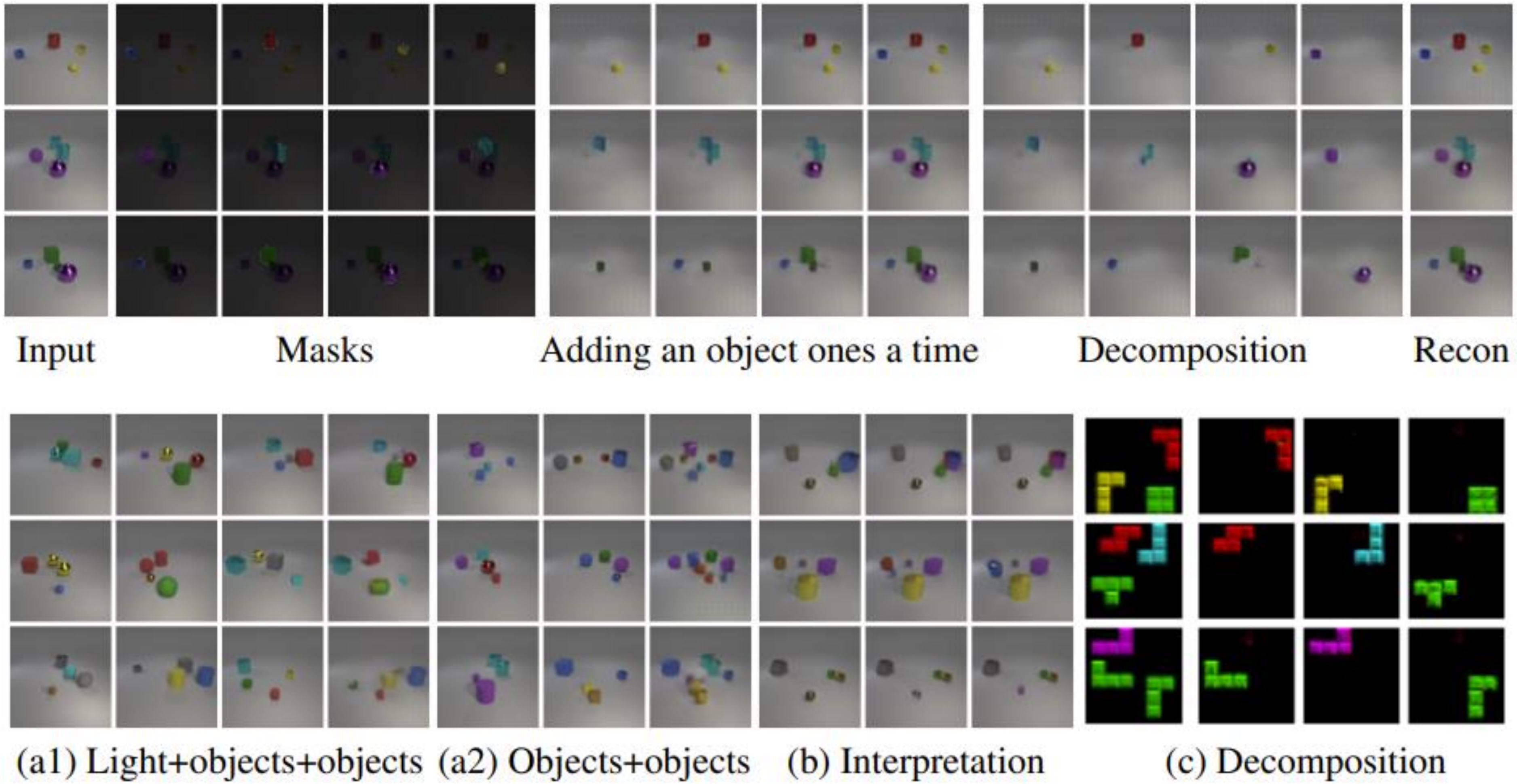
Method	MIG	DCI
Patch + VCT	0.361	0.668
AE + VCT	0.484	0.802
pretrained AE + VCT	**0.560**	0.849
pretrained VQ-VAE + VCT	0.525	**0.884**
AE + VCT wo $\mathcal{L}_{dis}$	0.165	0.692
pretrained VQVAE + VCT wo $\mathcal{L}_{dis}$	0.286	0.731
w/ self-attention	0.000	0.008
wo detach	0.392	0.871
w/ pos embedding	0.525	0.884
CNN DeTokenizer	0.157	0.847
Transformer DeTokenizer	0.467	0.821
Concept DeTokenizer	**0.525**	**0.884**
batchsize = 16	0.497	0.862
batchsize = 32	0.525	0.884
batchsize = 64	**0.535**	**0.900**
tokens number = 10	**0.533**	0.867
tokens number = 20	0.525	0.884
tokens number = 30	0.493	**0.885**

Qualitative Results (MPI-3D)



Experiments

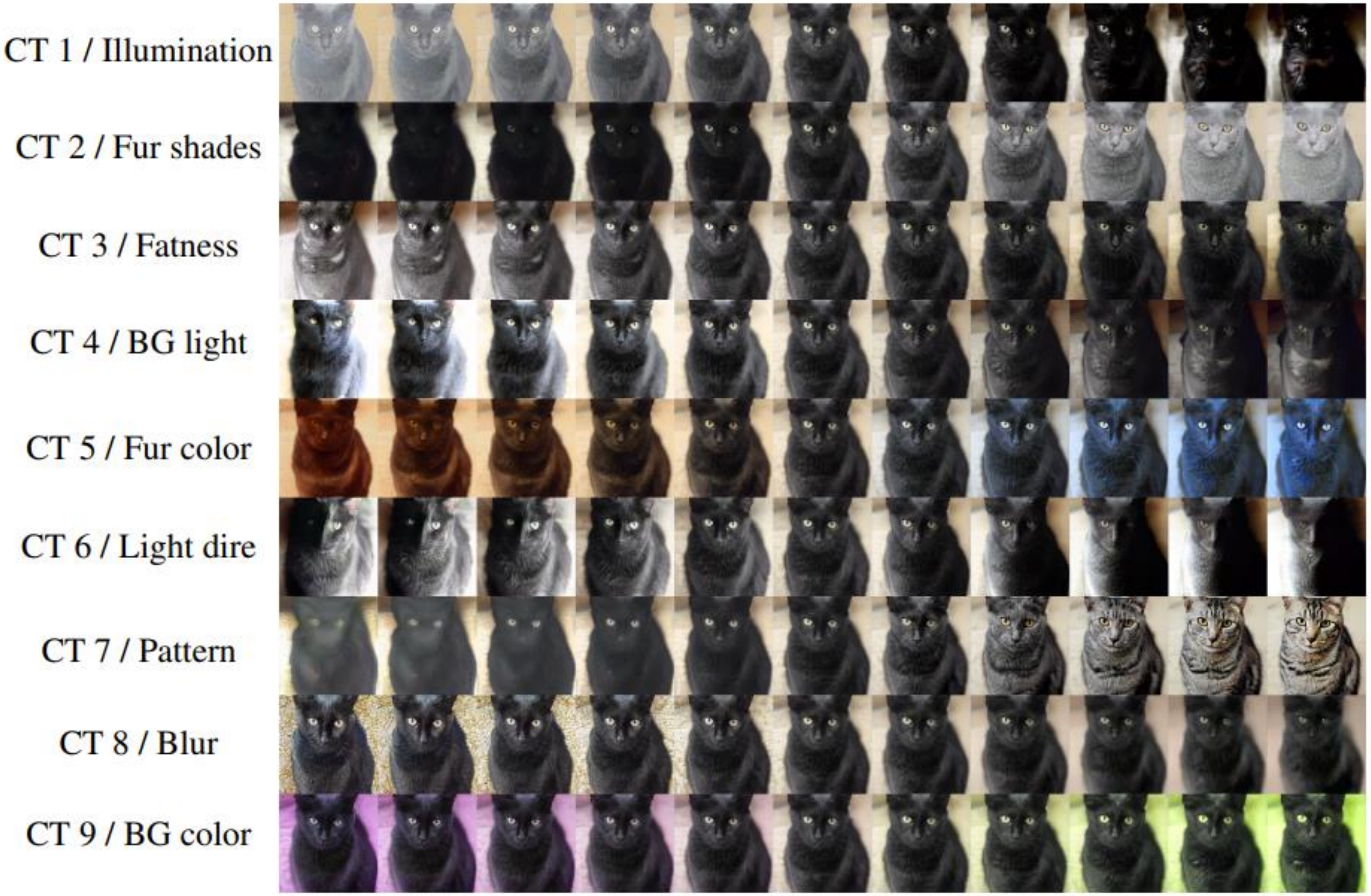
- Scene decomposition



Experiments

- On pretrained GAN

Shifted images on ImageNet dog



Shifted images on LSUN cat

Experiments

- CLIP results:

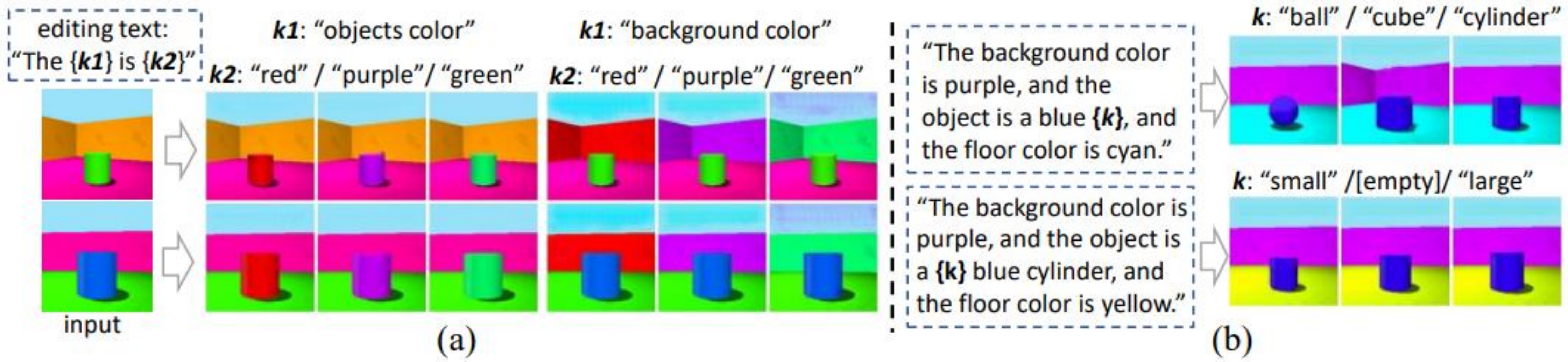


Figure 6: CLIP-based (a) text editing and (b) text decoding. The white arrow means decoding.

Summary

- We present a **general solution** to extract visual concepts from concrete pixels, which can achieve **disentangled representation learning** and **scene decomposition**.
- We build a transformer autoencoder, including Concept Tokenizer and Detokenizer, to represent an image into **a set of tokens**, and **each** token **reflects a visual concept**.
- We propose a Concept Disentangling Loss to facilitate the **mutual exclusivity** of the visual concept tokens.
- VCT can be deployed to the intermediate representations for learning visual concepts.

Code is available at: https://github.com/ThomasMrY/VCT