# Unsupervised Learning of Compositional Scene Representations from Multiple Unspecified Viewpoints

**Jinyang Yuan, Bin Li*, Xiangyang Xue***

Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University
{yuanjinyang, libin, xyxue}@fudan.edu.cn

## Abstract

Visual scenes are extremely rich in diversity, not only because there are infinite combinations of objects and background, but also because the observations of the same scene may vary greatly with the change of viewpoints. When observing a visual scene that contains multiple objects from multiple viewpoints, humans are able to perceive the scene in a compositional way from each viewpoint, while achieving the so-called "object constancy" across different viewpoints, even though the exact viewpoints are untold. This ability is essential for humans to identify the same object while moving and to learn from vision efficiently. It is intriguing to design models that have the similar ability. In this paper, we consider a novel problem of learning compositional scene representations from multiple unspecified viewpoints without using any supervision, and propose a deep generative model which separates latent representations into a viewpoint-independent part and a viewpoint-dependent part to solve this problem. To infer latent representations, the information contained in different viewpoints is iteratively integrated by neural networks. Experiments on several specifically designed synthetic datasets have shown that the proposed method is able to effectively learn from multiple unspecified viewpoints.

## Introduction

Vision is an important way for humans to acquire knowledge about the world. Due to the diverse combinations of objects and background that constitute visual scenes, it is hard to model the whole scene directly. In the process of learning from the world, humans are able to develop the concept of object (Johnson 2010), and is thus capable of perceiving visual scenes compositionally, which in turn leads to more efficient learning compared with perceiving the entire scene as a whole (Fodor and Pylyshyn 1988). Compositionality is one of the fundamental ingredients for building artificial intelligence systems that learn efficiently and effectively like humans (Lake et al. 2017). Therefore, instead of learning a single representation for the entire visual scene, it is desirable to build compositional scene representation models which learn *object-centric representations* (i.e., learn separate representations for different objects and background), so that the combinational property can be better captured.

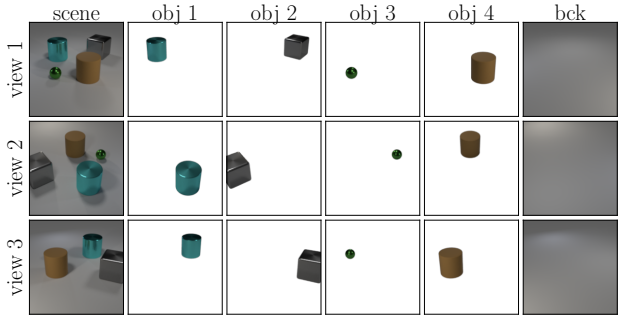



Figure 1: Humans are able to perceive visual scenes compositionally, while maintaining object constancy across different viewpoints (indexes of objects are arbitrarily chosen).

In addition, humans have the ability to achieve the so-called "object constancy" in visual perception, i.e., recognizing the same object from different viewpoints (Turnbull, Carey, and McCarthy 1997), possibly because of the mechanisms such as performing mental rotation (Shepard and Metzler 1971) or representing objects in a viewpoint-independent way (Marr 1982). When observing a multi-object scene from multiple viewpoints, humans are able to separate different objects from one another, and identify the same one from different viewpoints. As shown in Figure 1, given three images of the same visual scene observed from different viewpoints (column 1), humans are capable of decomposing each image into *complete* objects (columns 2-5) and background (column 6) that are *consistent* across viewpoints, even though the viewpoints are *unknown*, the poses of the same object may be significantly *different* across viewpoints, and some objects may be partially (object 2 in viewpoint 1) or even completely (object 3 in viewpoint 3) *occluded*. Observing visual scenes from multiple viewpoints gives humans a better understanding of the scenes, and it is intriguing to design compositional scene representation methods that are able to achieve object constancy and effectively learn from multiple viewpoints like humans.

In recent years, a variety of deep generative models have been proposed to learn compositional representations without object-level supervision. Most methods, such as AIR (Eslami et al. 2016), N-EM (Greff, van Steenkiste, and Schmidhuber 2017), MONet (Burgess et al. 2019), IODINE

*Corresponding author
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

(Greff et al. 2019), and Slot Attention (Locatello et al. 2020), however, are unsupervised methods that learn from only a *single* viewpoint. Only few methods, including MulMON (Li, Eastwood, and Fisher 2020) and ROOTS (Chen, Deng, and Ahn 2021), have considered the problem of learning from multiple viewpoints. These methods assume that the viewpoint annotations (under a certain global coordinate system) are given, and aim to learn viewpoint-independent object-centric representations *conditioned on* these annotations. Viewpoint annotations play fundamental roles in the initialization and updates of object-centric representations in MulMON, and in the computations of perspective projections in ROOTS. Therefore, without nontrivial modifications, existing methods *cannot* be applied to the novel problem of learning compositional scene representations from multiple unspecified viewpoints *without* any supervision.

The problem setting considered in this paper is very challenging, as the object-centric representations that are shared across viewpoints and the viewpoint representations that are shared across objects both need to be learned. More specifically, there are two major reasons. *Firstly*, the object constancy needs to be achieved *without the guidance* of viewpoint annotations, which are the only variable among images observed from different viewpoints and can be exploited to reduce the difficulty of learning the common factors. *Secondly*, the representations of images need to be disentangled into object-centric representations and viewpoint representations, even though there are *infinitely many* possible solutions, e.g., due to the change of global coordinate system.

In this paper, we propose a deep generative model called **O**bject-**C**entric **L**earning with **O**bject **C**onstancy (OCLOC) to learn object-centric representations from multiple viewpoints *without any supervision* (including viewpoint annotations), under the assumptions that 1) objects in the visual scenes are *static*, and 2) different visual scenes may be observed from *different* sets of *unordered* viewpoints. The proposed method models viewpoint-independent attributes of objects/background (e.g., 3D shapes and appearances in the global coordinate system) and viewpoints with separate latent variables, and adopts an amortized variational inference method that iteratively updates parameters of the approximated posteriors by integrating information of different viewpoints with inference neural networks.

To the best of the authors' knowledge, no existing object-centric learning method can learn from multiple unspecified viewpoints without viewpoint annotations. Thus, the proposed OCLOC cannot be directly compared with existing ones in the considered problem setting. Experiments on several specifically designed synthetic datasets have shown that OCLOC can effectively learn from multiple unspecific viewpoints without supervision, and *competes with* or *slightly outperforms* a state-of-the-art method that uses viewpoint annotations in the learning. Under an extreme condition that visual scenes are observed from one viewpoint, the proposed OCLOC is also comparable with the state-of-the-arts.

## Related Work

Object-centric representations are compositional scene representations that treat object or background as the basic entity of the visual scene and represent different objects or background separately. In recent years, various methods have been proposed to learn object-centric representations in an unsupervised manner, or using only scene-level annotations. Based on whether learning from multiple viewpoints and whether considering the movements of objects, these methods can be roughly divided into three categories.

**Single-Viewpoint Static Scenes:** CST-VAE (Huang and Murphy 2016), AIR (Eslami et al. 2016), and MONet (Burgess et al. 2019) extract the representation of each object sequentially based on the attention mechanism. GMIOO (Yuan, Li, and Xue 2019a) initializes the representation of each object sequentially and iteratively updates the representations, both with attentions on objects. SPAIR (Crawford and Pineau 2019) and SPACE (Lin et al. 2020) generate object proposals with convolutional neural networks and are applicable to large visual scenes containing a relatively large number of objects. N-EM (Greff, van Steenkiste, and Schmidhuber 2017), LDP (Yuan, Li, and Xue 2019b), IODINE (Greff et al. 2019), Slot Attention (Locatello et al. 2020), and EfficientMORL (Emami et al. 2021) first initialize representations of all the objects, and then apply some kind of competitions among objects to iteratively update the representations in parallel. GENESIS (Engelcke et al. 2020) and GNM (Jiang and Ahn 2020) consider the structure of visual scene in the generative models in order to generate more coherent samples. ADI (Yuan, Li, and Xue 2021) considers the acquisition and utilization of knowledge. These methods provide mechanisms to separate objects, and form the foundations of learning object-centric representations with the existences of object motions or from multiple viewpoints.

**Multi-Viewpoint Static Scenes:** MulMON (Li, Eastwood, and Fisher 2020) and ROOTS (Chen, Deng, and Ahn 2021) are two methods proposed to learn from static scenes from multiple viewpoints. MulMON extends the iterative amortized inference (Marino, Yue, and Mandt 2018) used in IODINE (Greff et al. 2019) to sequences of images observed from different viewpoints. Object-centric representations are first initialized based on the first pair of image and *viewpoint annotation*, and then iteratively refined by processing the rest pairs of data one by one. At each iteration, the previously estimated posteriors of latent variables are used as the current object-wise priors in order to guide the inference. ROOTS adopts the idea of using grid cells like SPAIR (Crawford and Pineau 2019) and SPACE (Lin et al. 2020), and generates object proposals in a bounded 3D region. The 3D center position of each object proposal is estimated and projected into different images with transformations that are computed based on the *annotated viewpoints*. After extracting crops of images corresponding to each object proposal, a type of GQN (Eslami et al. 2018) is applied to infer object-centric representations. As with our problem setting, different visual scenes are not assumed to be observed from the same set of viewpoints. However, because both methods heavily rely on the viewpoint annotations, they cannot be trivially applied to the fully-unsupervised scenario that the viewpoint annotations are unknown.

**Dynamic Scenes:** Inspired by the methods proposed for learning from single-viewpoint static scenes, several

methods, such as Relational N-EM (van Steenkiste et al. 2018), SQAIR (Kosiorek et al. 2018), R-SQAIR (Stanic and Schmidhuber 2019), TBA (He et al. 2019), SILOT (Crawford and Pineau 2020), SCALOR (Jiang et al. 2020), OP3 (Veerapaneni et al. 2020), and PROVIDE (Zablotskaia et al. 2021), have been proposed for learning from video sequences. The difficulties of this problem setting include modeling object motions and relationships, as well as maintaining the identities of objects even if objects disappear and reappear after full occlusion (Weis et al. 2021). Although these methods are able to identify the same object across adjacent frames, they cannot be directly applied to the problem setting considered in this paper for two major reasons: 1) images observed from different viewpoints are assumed to be unordered, and the positions of the same object may differ significantly in different images; and 2) viewpoints are shared among objects in the same visual scene, while object motions in videos do not have such a property.

## Generative Modeling

Visual scenes are assumed to be independent and identically distributed. For simplicity, the index of visual scene is omitted, and the procedure to generate images of a single visual scene is described. Let $M$ denote the number of images observed from different viewpoints (*may vary* in different visual scenes), $N$ and $C$ denote the respective numbers of pixels and channels in each image, and $K$ denote the maximum number of objects that may appear in the visual scene. The image of the $m$th viewpoint $\boldsymbol{x}_m \in \mathbb{R}^{N\times C}$ is assumed to be generated via a pixel-wise weighted summation of $K+1$ layers, with $K$ layers ($1 \le k \le K$) describing the objects and 1 layer ($k=0$) describing the background. The pixel-wise weights $\boldsymbol{s}_{m,0:K} \in [0,1]^{(K+1)\times N}$ as well as the images of layers $\boldsymbol{a}_{m,0:K} \in \mathbb{R}^{(K+1)\times N\times C}$ are computed based on latent variables. In the following, we first describe the latent variables and the likelihood function, and then express the generative model in the mathematical form.

### Viewpoint-Independent Latent Variables

Viewpoint-independent latent variables are the ones that are shared across different viewpoints, and are introduced in the generative model to achieve object constancy. These latent variables include $\boldsymbol{z}^{\text{attr}}$, $\boldsymbol{\rho}$, and $\boldsymbol{z}^{\text{prs}}$.

- $\boldsymbol{z}^{\text{attr}}_{0:K}$ characterize the viewpoint-independent attributes of objects ($1 \le k \le K$) and background ($k=0$). These attributes include the 3D shapes and appearances of objects and background in an automatically chosen global coordinate system. The dimensionalities of all the $\boldsymbol{z}^{\text{attr}}_k$ with $1 \le k \le K$ are identical, and are in general different from the dimensionality of $\boldsymbol{z}^{\text{attr}}_0$. For notational simplicity, this difference is not reflected in the expressions of the generative model. The priors of all the $\boldsymbol{z}^{\text{attr}}_k$ with $0 \le k \le K$ are standard normal distributions.
- $\boldsymbol{\rho}_{1:K}$ and $\boldsymbol{z}^{\text{prs}}_{1:K}$ are used to model the number of objects in the visual scene, considering that different visual scenes may contain different numbers of objects. The binary latent variable $\boldsymbol{z}^{\text{prs}}_k \in \{0,1\}$ indicates whether the $k$th object is included in the visual scene (i.e., the number of objects is $\sum_{k=1}^{K} \boldsymbol{z}^{\text{prs}}_k$), and is sampled from a Bernoulli distribution with the latent variable $\boldsymbol{\rho}_k$ as its parameter. The priors of all the $\boldsymbol{\rho}_k$ with $1 \le k \le K$ are beta distributions parameterized by hyperparameters $\alpha$ and $K$.

### Viewpoint-Dependent Latent Variables

Viewpoint-dependent latent variables may vary as the viewpoint changes. These latent variables include $\boldsymbol{z}^{\text{view}}$ and $\boldsymbol{z}^{\text{shp}}$.

- $\boldsymbol{z}^{\text{view}}_m$ determines the viewpoint (in an automatically chosen global coordinate system) of the $m$th image, and is drawn from a standard normal prior distribution.
- $\boldsymbol{z}^{\text{shp}}_{m,1:K,1:N} \in \{0,1\}^{K\times N}$ consist of binary latent variables that indicate the complete shapes of objects in the image coordinate system determined by the $m$th viewpoint. Each element of $\boldsymbol{z}^{\text{shp}}_{m,1:K,1:N}$ is sampled independently from a Bernoulli distribution, whose parameter is computed by transforming latent variables $\boldsymbol{z}^{\text{view}}_m$ and $\boldsymbol{z}^{\text{attr}}_k$ ($1 \le k \le K$) with a neural network $f_{\text{shp}}$ that captures the spatial dependencies among pixels. The sigmoid activation function in the last layer of $f_{\text{shp}}$ is explicitly expressed to clarify the output range of the neural network.

### Likelihood Function

All the pixels of the images $\boldsymbol{x}_{1:M,1:N}$ are assumed to be conditional independent of each other given all the latent variables $\boldsymbol{\Omega}$, and the likelihood function $p(\boldsymbol{x}|\boldsymbol{\Omega})$ is assumed to be factorized as the product of several normal distributions with varying mean vectors and constant covariance matrices, i.e., $\prod_{m=1}^{M}\prod_{n=1}^{N} \mathcal{N}(\sum_{k=0}^{K} s_{m,k,n}\,\boldsymbol{a}_{m,k,n},\, \sigma_{\text{x}}^2 \boldsymbol{I})$. To compute the mean vectors, intermediate variables $\boldsymbol{o}$, $\boldsymbol{s}$, and $\boldsymbol{a}$ need to be computed by transforming the sampled latent variables with deterministic functions.

- $\boldsymbol{o}_{m,1:K}$ characterize the depth ordering of objects in the image observed from the $m$th viewpoint. If multiple objects overlap, the object with the largest value of $o_{m,k}$ is assumed to occlude the others in a soft and differentiable way. To compute $o_{m,k}$, latent variables $\boldsymbol{z}^{\text{view}}_m$ and $\boldsymbol{z}^{\text{attr}}_k$ are first transformed by a neural network $f_{\text{ord}}$, and then the exponential function is applied to the output of $f_{\text{ord}}$ divided by $\lambda$. The exponential function ensures that the value of $o_{m,k}$ is greater than 0, and the hyperparameter $\lambda$ controls the softness of object occlusions.
- $\boldsymbol{s}_{m,0:K,1:N}$ indicate the perceived shapes of objects ($1 \le k \le K$) and background ($k=0$) in the $m$th image, and satisfy the constraints that $(\forall m,k,n)\, 0 \le s_{m,k,n} \le 1$ and $(\forall m,n) \sum_{k=0}^{K} s_{m,k,n} = 1$. These latent variables are computed based on $\boldsymbol{z}^{\text{prs}}_{1:K}$, $\boldsymbol{z}^{\text{shp}}_{m,1:K,1:N}$, and $\boldsymbol{o}_{m,1:K}$. Because $\boldsymbol{z}^{\text{prs}}$ and $\boldsymbol{z}^{\text{shp}}$ are binary variables, the perceived shape $\boldsymbol{s}_{m,0,1:N}$ of background is also binary, and equals 1 at the pixels that are not covered by any object. The computation of perceived shapes $\boldsymbol{s}_{m,1:K,n}$ of objects at each pixel can be interpreted as a masked softmax operation that only considers the objects covering that pixel. As the hyperparameter $\lambda$ in the computation of $\boldsymbol{o}$ approaches 0, the perceived shapes $\boldsymbol{s}_{m,0:K,n}$ of all the objects and background at each pixel approach a one-hot vector.

- $\boldsymbol{a}_{m,0:K,1:N}$ contain information about the complete appearances of objects ($1 \leq k \leq K$) and the background image ($k = 0$) in the $m$th image, and are computed by transforming latent variables $\boldsymbol{z}_m^{\text{view}}$ and $\boldsymbol{z}_k^{\text{attr}}$ with neural networks $f_{\text{bck}}$ (for $k = 0$) and $f_{\text{apc}}$ (for $1 \leq k \leq K$). Appearances of objects and the background image are computed differently because the dimensionality of $\boldsymbol{z}_0^{\text{attr}}$ is in general different from $\boldsymbol{z}_k^{\text{attr}}$ with $1 \leq k \leq K$.

### Generative Model

The mathematical expressions of the generative model are

$$
\begin{aligned}
\boldsymbol{z}_m^{\text{view}} &\sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right); & \boldsymbol{z}_k^{\text{attr}} &\sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right) \\
\rho_k &\sim \text{Beta}\left(\alpha/K, 1\right); & z_k^{\text{prs}} &\sim \text{Ber}\left(\rho_k\right) \\
z_{m,k,n}^{\text{shp}} &\sim \text{Ber}\left(\text{sigmoid}(f_{\text{shp}}(\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}})_n)\right) \\
o_{m,k} &= \exp\left(f_{\text{ord}}(\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}})/\lambda\right) \\
s_{m,k,n} &= \begin{cases} \prod_{k'=1}^{K}(1 - z_{k'}^{\text{prs}}\, z_{m,k',n}^{\text{shp}}), & k = 0 \\ \dfrac{(1 - s_{m,0,n})\, z_k^{\text{prs}}\, z_{m,k,n}^{\text{shp}}\, o_{m,k}}{\sum_{k'=1}^{K} z_{k'}^{\text{prs}}\, z_{m,k',n}^{\text{shp}}\, o_{m,k'}}, & 1 \leq k \leq K \end{cases} \\
\boldsymbol{a}_{m,k,n} &= \begin{cases} f_{\text{bck}}\left(\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}}\right)_n, & k = 0 \\ f_{\text{apc}}\left(\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}}\right)_n, & 1 \leq k \leq K \end{cases} \\
\boldsymbol{x}_{m,n} &\sim \mathcal{N}\Big(\sum_{k=0}^{K} s_{m,k,n}\, \boldsymbol{a}_{m,k,n},\, \sigma_{\text{x}}^2 \boldsymbol{I}\Big)
\end{aligned}
$$

In the above expressions, some of the ranges of indexes $m$ ($1 \leq m \leq M$), $n$ ($1 \leq n \leq N$), and $k$ ($0 \leq k \leq K$ for $\boldsymbol{z}^{\text{attr}}$, and $1 \leq k \leq K$ for $\boldsymbol{\rho}$, $\boldsymbol{z}^{\text{prs}}$, $\boldsymbol{z}^{\text{shp}}$, $\boldsymbol{o}$) are omitted for notational simplicity. $\alpha$, $\lambda$, and $\sigma_{\text{x}}$ are tunable hyperparameters. Let $\boldsymbol{\Omega} = \{\boldsymbol{z}^{\text{view}}, \boldsymbol{z}^{\text{attr}}, \boldsymbol{\rho}, \boldsymbol{z}^{\text{prs}}, \boldsymbol{z}^{\text{shp}}\}$ be the collection of all latent variables. The joint probability of $\boldsymbol{x}$ and $\boldsymbol{\Omega}$ is

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{\Omega}) = &\prod_{k=0}^{K} p(\boldsymbol{z}_k^{\text{attr}}) \prod_{k=1}^{K} p(\rho_k) p(z_k^{\text{prs}}|\rho_k) \\
&\prod_{m=1}^{M} p(\boldsymbol{z}_m^{\text{view}}) \prod_{k=1}^{K} \prod_{n=1}^{N} p(z_{m,k,n}^{\text{shp}}|\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}}) \\
&\prod_{m=1}^{M} \prod_{n=1}^{N} p(\boldsymbol{x}_{m,n}|\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_{0:K}^{\text{attr}}, \boldsymbol{z}_{1:K}^{\text{prs}}, \boldsymbol{z}_{m,1:K,n}^{\text{shp}})
\end{aligned} \tag{1}
$$

## Inference and Learning

The exact posterior distribution of latent variables $p(\boldsymbol{\Omega}|\boldsymbol{x})$ is intractable to compute. Therefore, we adopt amortized variational inference, which approximates the complex posterior distribution with a tractable variational distribution $q(\boldsymbol{\Omega}|\boldsymbol{x})$, and apply neural networks to transform the images $\boldsymbol{x}$ into parameters of the variational distribution. The neural networks $f_{\text{shp}}$, $f_{\text{ord}}$, $f_{\text{bck}}$, and $f_{\text{apc}}$ in the generative model, as well as the inference networks, are jointly optimized with the goal of maximizing the evidence lower bound (ELBO). Details of the inference and learning are described below.

### Inference of Latent Variables

The variational distribution $q(\boldsymbol{\Omega}|\boldsymbol{x})$ is factorized as

$$
\begin{aligned}
q(\boldsymbol{\Omega}|\boldsymbol{x}) = &\prod_{k} q(\boldsymbol{z}_k^{\text{attr}}|\boldsymbol{x}) \prod_{k} q(\rho_k|\boldsymbol{x}) q(z_k^{\text{prs}}|\boldsymbol{x}) \\
&\prod_{m} q(\boldsymbol{z}_m^{\text{view}}|\boldsymbol{x}) \prod_{k} \prod_{n} q(z_{m,k,n}^{\text{shp}}|\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}}, \boldsymbol{x})
\end{aligned} \tag{2}
$$

Algorithm 1: Inference of latent variables

**Input**: Images of $M$ viewpoints $\boldsymbol{x}_{1:M}$
**Output**: Parameters of $q(\boldsymbol{\Omega}|\boldsymbol{x})$

// Extract features and initialize intermediate variables
$\boldsymbol{y}_m^{\text{feat}} \leftarrow g_{\text{feat}}(\boldsymbol{x}_m), \quad \forall\, 1 \leq m \leq M$
$\boldsymbol{y}_m^{\text{view}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}^{\text{view}}, \text{diag}(\hat{\boldsymbol{\sigma}}^{\text{view}})), \quad \forall\, 1 \leq m \leq M$
$\boldsymbol{y}_k^{\text{attr}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}^{\text{attr}}, \text{diag}(\hat{\boldsymbol{\sigma}}^{\text{attr}})), \quad \forall\, 0 \leq k \leq K$
// Update intermediate variables $\boldsymbol{y}_{1:M}^{\text{view}}$ and $\boldsymbol{y}_{0:K}^{\text{attr}}$
**for** $t \leftarrow 1$ to $T$ **do** {$\forall\, 1 \leq m \leq M, 0 \leq k \leq K$ in the loop}
  $\boldsymbol{y}_{m,k}^{\text{full}} \leftarrow [\boldsymbol{y}_m^{\text{view}}, \boldsymbol{y}_k^{\text{attr}}]$
  $\boldsymbol{a}_{m,k} \leftarrow \text{softmax}_K\big(g_{\text{key}}(\boldsymbol{y}_m^{\text{feat}}) g_{\text{qry}}(\boldsymbol{y}_{m,0:K}^{\text{full}})/\sqrt{D_{\text{key}}}\big)$
  $\boldsymbol{u}_{m,k} \leftarrow \sum_N \text{softmax}_N(\log \boldsymbol{a}_{m,k})\, g_{\text{val}}(\boldsymbol{y}_m^{\text{feat}})$
  $[\boldsymbol{v}_{1:M,0:K}^{\text{view}}, \boldsymbol{v}_{1:M,0:K}^{\text{attr}}] \leftarrow g_{\text{upd}}(\boldsymbol{y}_{1:M,0:K}^{\text{full}}, \boldsymbol{u}_{1:M,0:K})$
  $\boldsymbol{y}_m^{\text{view}} \leftarrow \text{mean}_K(\boldsymbol{v}_{m,k}^{\text{view}})$
  $\boldsymbol{y}_k^{\text{attr}} \leftarrow \text{mean}_M(\boldsymbol{v}_{m,k}^{\text{attr}})$
**end for**
// Sample the background index and rearrange $\boldsymbol{y}_{0:K}^{\text{attr}}$
$\pi_k = \text{softmax}_K\big(g_{\text{sel}}(\boldsymbol{y}_{0:K}^{\text{attr}})\big), \quad \forall\, 0 \leq k \leq K$
$k^* \sim \text{Cat}(\pi_0, \ldots, \pi_K); \quad \boldsymbol{y}_{0:K}^{\text{attr}} \leftarrow [\boldsymbol{y}_{k^*}^{\text{attr}}, \boldsymbol{y}_{0:K\setminus k^*}^{\text{attr}}]$
// Convert $\boldsymbol{y}_{1:M}^{\text{view}}$ and $\boldsymbol{y}_{0:K}^{\text{attr}}$ to parameters of $q(\boldsymbol{\Omega}|\boldsymbol{x})$
$\boldsymbol{\mu}_0^{\text{attr}}, \boldsymbol{\sigma}_0^{\text{attr}} \leftarrow g_{\text{bck}}(\boldsymbol{y}_0^{\text{attr}})$
$\boldsymbol{\mu}_k^{\text{attr}}, \boldsymbol{\sigma}_k^{\text{attr}}, \boldsymbol{\tau}_k, \kappa_k \leftarrow g_{\text{obj}}(\boldsymbol{y}_k^{\text{attr}}), \quad \forall\, 1 \leq k \leq K$
$\boldsymbol{\mu}_m^{\text{view}}, \boldsymbol{\sigma}_m^{\text{view}} \leftarrow g_{\text{view}}(\boldsymbol{y}_m^{\text{view}}), \quad \forall\, 1 \leq m \leq M$
**return** $\boldsymbol{\mu}_{0:K}^{\text{attr}}, \boldsymbol{\sigma}_{0:K}^{\text{attr}}, \boldsymbol{\tau}_{1:K}, \boldsymbol{\kappa}_{1:K}, \boldsymbol{\mu}_{1:M}^{\text{view}}, \boldsymbol{\sigma}_{1:M}^{\text{view}}$

The ranges of indexes in Eq. (2) are identical to the ones in Eq. (1), and are omitted for simplicity. The choices of terms on the right-hand side of Eq. (2) are

$$
\begin{aligned}
q(\boldsymbol{z}_k^{\text{attr}}|\boldsymbol{x}) &= \mathcal{N}\left(\boldsymbol{z}_k^{\text{attr}}; \boldsymbol{\mu}_k^{\text{attr}}, \text{diag}(\boldsymbol{\sigma}_k^{\text{attr}})^2\right) \\
q(\rho_k|\boldsymbol{x}) &= \text{Beta}\left(\rho_k; \tau_{k,1}, \tau_{k,2}\right) \\
q(z_k^{\text{prs}}|\boldsymbol{x}) &= \text{Ber}\left(z_k^{\text{prs}}; \kappa_k\right) \\
q(\boldsymbol{z}_m^{\text{view}}|\boldsymbol{x}) &= \mathcal{N}\left(\boldsymbol{z}_m^{\text{view}}; \boldsymbol{\mu}_m^{\text{view}}, \text{diag}(\boldsymbol{\sigma}_k^{\text{attr}})^2\right) \\
q(z_{m,k,n}^{\text{shp}}|\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}}, \boldsymbol{x}) &= p(z_{m,k,n}^{\text{shp}}|\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}})
\end{aligned}
$$

In the variational distribution, $q(\boldsymbol{z}_k^{\text{attr}}|\boldsymbol{x})$ and $q(\boldsymbol{z}_m^{\text{view}}|\boldsymbol{x})$ are normal distributions with diagonal covariance matrices. $z_k^{\text{prs}}$ is assumed to be independent of $\rho_k$ given $\boldsymbol{x}$, and $q(\rho_k|\boldsymbol{x})$ and $q(z_k^{\text{prs}}|\boldsymbol{x})$ are chosen to be a beta distribution and a Bernoulli distribution, respectively. The advantage of this formulation is that the Kullback-Leibler (KL) divergence between $q(\rho_k|\boldsymbol{x})q(z_k^{\text{prs}}|\boldsymbol{x})$ and $p(\rho_k)p(z_k^{\text{prs}}|\rho_k)$ has a closed-form solution. For simplicity, $q(z_{m,k,n}^{\text{shp}}|\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}}, \boldsymbol{x})$ is assumed to be identical to $p(z_{m,k,n}^{\text{shp}}|\boldsymbol{z}_m^{\text{view}}, \boldsymbol{z}_k^{\text{attr}})$ in the generative model, so that no extra inference network is needed for $z_{m,k,n}^{\text{shp}}$. The procedure to compute the parameters $\boldsymbol{\mu}^{\text{attr}}$, $\boldsymbol{\sigma}^{\text{attr}}$, $\boldsymbol{\tau}$, $\boldsymbol{\kappa}$, $\boldsymbol{\mu}^{\text{view}}$, and $\boldsymbol{\sigma}^{\text{view}}$ of these distributions is presented in Algorithm 1, and the brief explanations are given below.

First, the feature maps $\boldsymbol{y}_m^{\text{feat}}$ of each image $\boldsymbol{x}_m$ are extracted by a neural network $g_{\text{feat}}$. Next, intermediate variables $\boldsymbol{y}^{\text{view}}$ and $\boldsymbol{y}^{\text{attr}}$ which fully characterize parameters of the viewpoint-dependent ($\boldsymbol{\mu}^{\text{view}}$ and $\boldsymbol{\sigma}^{\text{view}}$) and viewpoint-independent ($\boldsymbol{\mu}^{\text{attr}}$, $\boldsymbol{\sigma}^{\text{attr}}$, $\boldsymbol{\tau}$, and $\boldsymbol{\kappa}$) latent variables are not di-

rectly estimated, but instead randomly initialized from normal distributions with learnable parameters ($\hat{\boldsymbol{\mu}}^{\text{view}}$, $\hat{\boldsymbol{\sigma}}^{\text{view}}$, $\hat{\boldsymbol{\mu}}^{\text{attr}}$, and $\hat{\boldsymbol{\sigma}}^{\text{attr}}$) and then iteratively updated, considering that there are infinitely many possible solutions (e.g., due to the change of global coordinate system) to disentangle the image representations into a viewpoint-dependent part and a viewpoint-independent part. In each step of the iterative updates, information of images observed from different viewpoints are integrated using neural networks $g_{\text{key}}$, $g_{\text{qry}}$, $g_{\text{val}}$, and $g_{\text{upd}}$, based on attentions between feature maps $\boldsymbol{y}^{\text{feat}}$ and intermediate variables $\boldsymbol{y}^{\text{view}}$ and $\boldsymbol{y}^{\text{attr}}$. To achieve permutation equivariance, which has been considered as an important property in object-centric learning (Emami et al. 2021), objects and background are not distinguished in the initialization and updates of $\boldsymbol{y}^{\text{attr}}$, and the index $k^*$ that corresponds to background is determined after the iterative updates, by applying a neural network $g_{\text{sel}}$ to transform $\boldsymbol{y}^{\text{attr}}$ into parameters $\boldsymbol{\pi}$ of a categorical distribution and sampling from the distribution. After rearranging $\boldsymbol{y}^{\text{attr}}_{0:K}$ based on $k^*$, parameters of the variational distribution are computed by transforming $\boldsymbol{y}^{\text{attr}}_{0}$, $\boldsymbol{y}^{\text{attr}}_{1:K}$, and $\boldsymbol{y}^{\text{view}}_{1:M}$ with neural networks $g_{\text{bck}}$, $g_{\text{obj}}$, and $g_{\text{view}}$, respectively. For further details, please refer to the Supplementary Material.

## Learning of Neural Networks

The neural networks used in both the generative model and the amortized variational inference (including learnable parameters $\hat{\boldsymbol{\mu}}^{\text{view}}$, $\hat{\boldsymbol{\sigma}}^{\text{view}}$, $\hat{\boldsymbol{\mu}}^{\text{attr}}$, and $\hat{\boldsymbol{\sigma}}^{\text{attr}}$), are jointly optimized by minimizing the negative value of evidence lower bound (ELBO) that serves as the loss function $\mathcal{L}$. The expression of $\mathcal{L}$ is briefly given below, and a more detailed version is included in the Supplementary Material.

$$
\begin{aligned}
\mathcal{L} = & -\sum\nolimits_m \sum\nolimits_n \mathbb{E}_{q(\boldsymbol{\Omega}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}_{m,n}|\boldsymbol{z}^{\text{view}}, \boldsymbol{z}^{\text{attr}}, \boldsymbol{z}^{\text{prs}}, \boldsymbol{z}^{\text{shp}})\right] \\
& + \sum\nolimits_m D_{\text{KL}}\left(q(\boldsymbol{z}^{\text{view}}_m|\boldsymbol{x})||p(\boldsymbol{z}^{\text{view}}_m)\right) \\
& + \sum\nolimits_k D_{\text{KL}}\left(q(\boldsymbol{z}^{\text{attr}}_k|\boldsymbol{x})||p(\boldsymbol{z}^{\text{attr}}_k)\right) \\
& + \sum\nolimits_k D_{\text{KL}}\left(q(\rho_k|\boldsymbol{x})||p(\rho_k)\right) \\
& + \sum\nolimits_k \mathbb{E}_{q(\rho_k|\boldsymbol{x})}\left[D_{\text{KL}}\left(q(z^{\text{prs}}_k|\boldsymbol{x})||p(z^{\text{prs}}_k|\rho_k)\right)\right] \qquad (3)
\end{aligned}
$$

In Eq. (3), the first term is negative log-likelihood, and the rest four terms are Kullback-Leibler (KL) divergences that are computed by $D_{\text{KL}}(q||p) = \mathbb{E}_q[\log q - \log p]$. The loss function is optimized using the gradient-based method. All the KL divergences have closed-form solutions, and the gradients of these terms can be easily computed. The negative log-likelihood cannot be computed analytically, and the gradients of this term is approximated by sampling latent variables $\boldsymbol{z}^{\text{view}}$, $\boldsymbol{z}^{\text{attr}}$, $\boldsymbol{z}^{\text{prs}}$, and $\boldsymbol{z}^{\text{shp}}$ from the variational distribution $q(\boldsymbol{\Omega}|\boldsymbol{x})$. To reduce the variances of gradients, the continuous variables $\boldsymbol{z}^{\text{view}}$ and $\boldsymbol{z}^{\text{attr}}$ are sampled using the reparameterization trick (Salimans and Knowles 2013; Kingma and Welling 2014), and the discrete variables $\boldsymbol{z}^{\text{prs}}$ and $\boldsymbol{z}^{\text{shp}}$ are approximated using a continuous relaxation (Maddison, Mnih, and Teh 2017; Jang, Gu, and Poole 2017). To learn the neural network $g_{\text{sel}}$ that computes parameters of the categorical distribution from which the background index $k^*$ is sampled, NVIL (Mnih and Gregor 2014) is applied to obtain low-variance and unbiased estimates of gradients.

# Experiments

In this section, we aim to verify that the proposed method[1]:

- is able to learn from multiple viewpoints *without any supervision*, which *cannot* be solved by existing methods;
- competes with existing state-of-the-art methods that use *viewpoint annotations* in the learning;
- is comparable to the state-of-the-arts under an *extreme condition* that scenes are observed from one viewpoint.

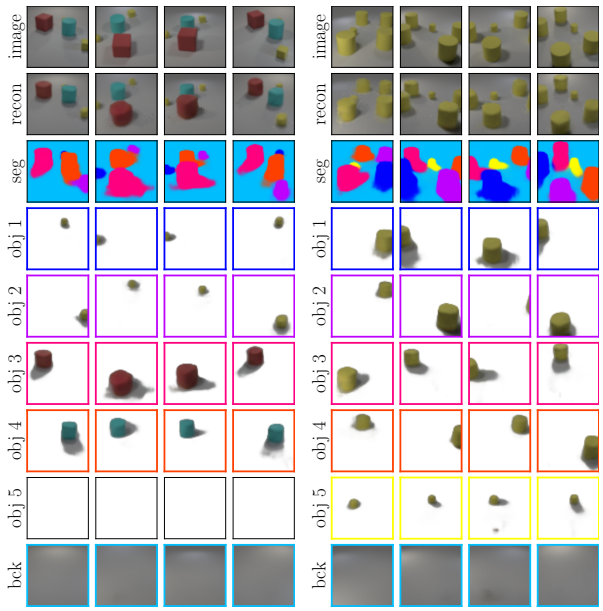
**Evaluation Metrics:** Several metrics are used to evaluate the performance from four aspects. 1) *Adjusted Rand Index* (ARI) (Hubert and Arabie 1985) and *Adjusted Mutual Information* (AMI) (Nguyen, Epps, and Bailey 2010) assess the quality of segmentation, i.e., how accurately images are partitioned into different objects and background. Previous work usually evaluates ARI and AMI only at pixels belong to objects, and how accurately background is separated from objects is unclear. We evaluate ARI and AMI under two conditions. ARI-A and AMI-A are computed considering both objects and background, while ARI-O and AMI-O are computed considering only objects. 2) *Intersection over Union* (IoU) and $F_1$ *score* (F1) assess the quality of amodal segmentation, i.e., how accurately complete shapes of objects are estimated. 3) *Object Counting Accuracy* (OCA) assesses the accuracy of the estimated number of objects. 4) *Object Ordering Accuracy* (OOA) as used in (Yuan, Li, and Xue 2019a) assesses the accuracy of the estimated pairwise ordering of objects. Formal definitions of these metrics are included in the Supplementary Material.

## Multi-Viewpoint Learning

**Datasets:** The experiments are performed on four multi-viewpoint variants (referred to as CLEVR-M1 to CLEVR-M4) of the commonly used CLEVR dataset that differ in the ranges to sample viewpoints and in the attributes of objects. CLEVR-M3/CLEVR-M4 is harder than CLEVR-M1/CLEVR-M2 in that the poses of objects are more dissimilar in different images of the same visual scene because viewpoints are sampled from a larger range. CLEVR-M2/CLEVR-M4 is harder than CLEVR-M1/CLEVR-M3 in that there are fewer visual cues to distinguish objects from one another because all the objects in the same visual scene share the same colors, shapes, and materials. Further details are described in the Supplementary Material.

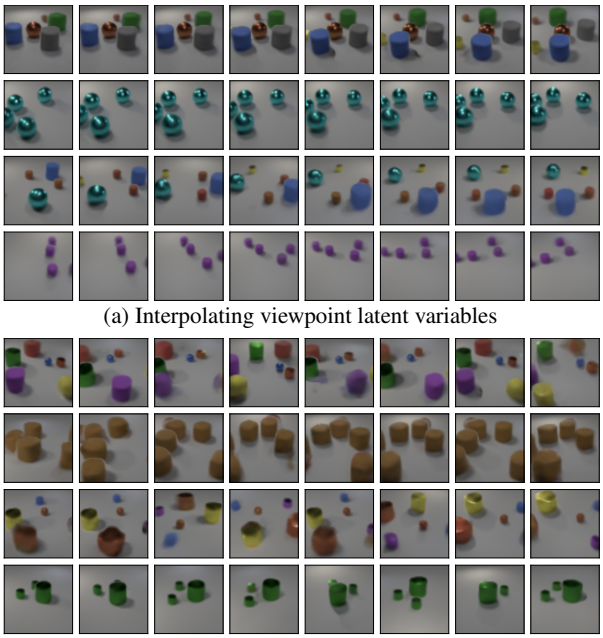
**Comparison Methods:** It is worth noting that the proposed method cannot be directly compared with existing methods in the novel problem setting considered in this paper. To verify that the proposed method can effectively achieve object constancy, a baseline method that does not maintain the identities of objects across viewpoints is compared with. This baseline method is derived from the proposed method by assigning each viewpoint a separate set of latent variables $\boldsymbol{z}^{\text{attr}}$, $\boldsymbol{\rho}$, and $\boldsymbol{z}^{\text{prs}}$ (all latent variables are viewpoint-dependent). To verify that the proposed method can effec-

[1]Code is available at https://git.io/JDnne.




(a) CLEVR-M1 (b) CLEVR-M2

Figure 2: Scene decomposition results of the proposed method in the multi-viewpoint learning setting. Objects are sorted based on the estimated $\boldsymbol{z}^{\text{prs}}$. Models are tested with $K=7$, and the last two objects with $z_k^{\text{prs}}=0$ are not shown.



(a) Interpolating viewpoint latent variables

(b) Sampling viewpoint latent variables

Figure 3: Results of interpolating and sampling viewpoints in latent space. The $i$th row of each sub-figure corresponds to the results evaluated on the CLEVR-M$\{i\}$ dataset.

tively learn without supervision, we compare it with MulMON (Li, Eastwood, and Fisher 2020), which solves a *simpler* problem by using viewpoint annotations in both learning and testing. Another representative partially supervised method ROOTS (Chen, Deng, and Ahn 2021) is not compared with because the official code is not publicly available.

**Scene Decomposition:** Qualitative results of the proposed method evaluated on the CLEVR-M1 and CLEVR-M2 datasets are shown in Figure 2. The proposed method is able to achieve *object constancy* even if objects are fully occluded (object 1 in columns 1 and 4 of sub-figure (a)). In addition, under the circumstances that objects are less identifiable and the poses of objects vary significantly across different viewpoints (objects $1\sim4$ in sub-figure (b)), the proposed method can also correctly identify the same objects across viewpoints. The proposed method tends to treat *shadows* as parts of objects instead of background, which is desirable because lighting effects are not explicitly modeled and the shadows will change accordingly as the objects move. More results can be found in the Supplementary Material.

Quantitative comparison of scene decomposition performance on all the datasets is presented in Table 1. The proposed method achieves high ARI-O, AMI-O, and OOA scores. As for ARI-A, AMI-A, IoU, and F1, the achieved performance is not so well. The major reason is that the proposed method tends to treat regions of shadows as objects, while they are considered as background in the ground truth annotations. MulMON also tends to incorrectly estimate shadows as objects, but slightly outperforms the proposed method in terms of ARI-A and AMI-A on all the datasets, possibly because MulMON does not explicitly model the complete shapes, the number, and the depth ordering of objects, but directly computes the perceived shapes using the softmax function, which makes it easier to learn the boundary regions of objects. For the similar reason, the IoU, F1, and OOA scores which require the estimations of complete shapes and depth ordering are not evaluated for MulMON. The OCA scores are computed based on the heuristically estimated number of objects (details in the Supplementary Material). The *unsupervised* proposed method achieves competitive or slightly better results compared to the *partially supervised* MulMON, which has validated the motivation of the proposed method.

**Generalizability:** Because visual scenes are modeled compositionally by the proposed method, the trained models are generalizable to novel scenes containing more numbers of objects than the ones used for training. Evaluations of generalizability are included in the Supplementary Material. Although the increased number of objects makes it more difficult to extract compositional scene representations, the proposed method performs reasonably well.

**Viewpoint Estimation:** The proposed method is able to estimate the viewpoints of images, under the condition that the viewpoint-independent attributes of objects and background are known. More specifically, given the approximate posteriors of object-centric representations, the proposed method is able to infer the corresponding viewpoint representations of different observations of the same visual scene. Please refer to the Supplementary Material for more details.

**Viewpoint Modification:** Multi-viewpoint images of the same visual scene can be generated by first inferring compositional scene representations and then modifying viewpoint

| Dataset | Method | ARI-A | AMI-A | ARI-O | AMI-O | IoU | F1 | OCA | OOA |
|---|---|---|---|---|---|---|---|---|---|
| CLEVR-M1 | Baseline | 0.512±9e-4 | 0.361±3e-3 | 0.269±1e-2 | 0.418±1e-2 | 0.171±3e-3 | 0.279±4e-3 | 0.004±5e-3 | 0.628±3e-2 |
| | MulMON | **0.615**±2e-3 | **0.560**±2e-3 | 0.927±5e-3 | 0.917±2e-3 | N/A | N/A | 0.446±5e-2 | N/A |
| | Proposed | 0.507±2e-3 | 0.486±2e-3 | **0.948**±3e-3 | **0.934**±2e-3 | **0.442**±3e-3 | **0.603**±3e-3 | **0.730**±5e-2 | **0.970**±1e-2 |
| CLEVR-M2 | Baseline | 0.505±1e-3 | 0.356±3e-3 | 0.274±1e-2 | 0.422±9e-3 | 0.167±4e-3 | 0.273±5e-3 | 0.004±5e-3 | 0.682±2e-2 |
| | MulMON | **0.602**±7e-4 | **0.550**±4e-4 | 0.939±3e-3 | 0.926±2e-3 | N/A | N/A | 0.570±5e-2 | N/A |
| | Proposed | 0.507±3e-3 | 0.479±2e-3 | **0.941**±3e-3 | **0.933**±2e-3 | **0.428**±3e-3 | **0.587**±4e-3 | **0.686**±3e-2 | **0.939**±2e-2 |
| CLEVR-M3 | Baseline | 0.531±1e-3 | 0.372±3e-3 | 0.278±1e-2 | 0.425±1e-2 | 0.173±5e-3 | 0.283±7e-3 | 0.000±0e-0 | 0.600±5e-2 |
| | MulMON | **0.591**±7e-3 | **0.552**±3e-3 | 0.938±2e-3 | 0.923±2e-3 | N/A | N/A | 0.424±5e-2 | N/A |
| | Proposed | 0.534±2e-3 | 0.498±2e-3 | **0.939**±5e-3 | **0.929**±3e-3 | **0.453**±3e-3 | **0.610**±4e-3 | **0.632**±3e-2 | **0.974**±8e-3 |
| CLEVR-M4 | Baseline | 0.519±1e-3 | 0.365±1e-3 | 0.280±6e-3 | 0.428±6e-3 | 0.170±4e-3 | 0.278±5e-3 | 0.000±0e-0 | 0.633±5e-2 |
| | MulMON | **0.640**±4e-4 | **0.578**±6e-4 | **0.936**±3e-3 | **0.927**±2e-3 | N/A | N/A | 0.490±3e-2 | N/A |
| | Proposed | 0.473±2e-3 | 0.452±2e-3 | 0.923±4e-3 | 0.922±2e-3 | **0.401**±1e-3 | **0.558**±1e-3 | **0.606**±8e-3 | **0.853**±2e-2 |

Table 1: Comparison of scene decomposition performance when learning from multiple viewpoints. All the methods are trained and tested with $M = 4$ and $K = 7$. The proposed *fully unsupervised* method achieves *competitive* or *slightly better* results compared with MulMON with *viewpoint supervision*.

| Dataset | Method | ARI-A | AMI-A | ARI-O | AMI-O | IoU | F1 | OCA | OOA |
|---|---|---|---|---|---|---|---|---|---|
| dSprites | Slot Attn | 0.142±3e-3 | 0.286±2e-3 | 0.935±1e-3 | 0.917±1e-3 | N/A | N/A | 0.000±0e-0 | N/A |
| | GMIOO | **0.969**±2e-4 | **0.922**±5e-4 | **0.956**±1e-3 | **0.950**±9e-4 | *0.860*±8e-4 | *0.911*±7e-4 | **0.874**±2e-3 | *0.891*±5e-3 |
| | SPACE | 0.946±7e-4 | 0.874±6e-4 | 0.858±1e-3 | 0.870±5e-4 | 0.729±7e-4 | 0.805±5e-4 | 0.587±4e-3 | 0.624±1e-2 |
| | Proposed | *0.959*±2e-4 | *0.907*±4e-4 | *0.939*±7e-4 | *0.922*±7e-4 | **0.861**±8e-4 | **0.912**±8e-4 | *0.813*±4e-3 | **0.899**±8e-3 |
| Abstract | Slot Attn | **0.940**±5e-4 | **0.877**±3e-4 | 0.935±8e-4 | 0.903±7e-4 | N/A | N/A | 0.888±5e-3 | N/A |
| | GMIOO | 0.832±2e-4 | 0.751±3e-4 | *0.941*±2e-3 | *0.927*±1e-3 | *0.750*±8e-4 | *0.848*±8e-4 | **0.955**±2e-3 | *0.940*±3e-3 |
| | SPACE | *0.888*±6e-4 | 0.797±6e-4 | 0.816±1e-3 | 0.817±2e-3 | 0.722±7e-4 | 0.798±8e-4 | 0.685±2e-3 | 0.799±5e-3 |
| | Proposed | 0.887±3e-4 | *0.812*±4e-4 | **0.947**±9e-4 | **0.933**±9e-4 | **0.801**±3e-4 | **0.883**±2e-4 | *0.940*±6e-3 | **0.962**±2e-3 |
| CLEVR | Slot Attn | 0.026±2e-4 | 0.240±3e-4 | **0.985**±6e-4 | **0.983**±3e-4 | N/A | N/A | 0.002±1e-3 | N/A |
| | GMIOO | *0.716*±5e-4 | *0.665*±4e-4 | 0.943±1e-3 | 0.955±8e-4 | *0.605*±2e-3 | 0.725±2e-3 | 0.683±2e-3 | 0.906±4e-3 |
| | SPACE | **0.860**±3e-4 | **0.796**±3e-4 | 0.976±3e-4 | 0.973±1e-4 | **0.776**±7e-4 | **0.863**±7e-4 | *0.711*±2e-3 | *0.936*±7e-3 |
| | Proposed | 0.649±1e-4 | 0.614±2e-4 | *0.982*±9e-4 | *0.978*±5e-4 | 0.591±6e-4 | *0.736*±7e-4 | **0.875**±8e-3 | **0.952**±5e-3 |

Table 2: Comparison of scene decomposition performance when learning from a single viewpoint. All the methods are trained and tested with $K=6$, $K=5$, and $K=7$ on the dSprites, Abstract, and CLEVR datasets, respectively. The proposed method is *comparable* with the state-of-the-arts under the extreme condition that visual scenes are observed from one viewpoint.

latent variables. Results of interpolating and sampling viewpoint latent variables are illustrated in Figure 3. The proposed method is able to appropriately modify viewpoints.

## Single-Viewpoint Learning

**Datasets:** Three datasets are constructed based on the dSprites (Matthey et al. 2017), Abstract Scene (Zitnick and Parikh 2013), and CLEVR (Johnson et al. 2017) datasets, in a way similar to the Multi-Objects Datasets (Kabra et al. 2019) but provides extra annotations (for evaluation only) of complete shapes of objects. These datasets are referred to as *dSprites*, *Abstract*, and *CLEVER* for simplicity. More details are provided in the Supplementary Material.

**Comparison Methods:** The proposed method is compared with three state-of-the-art compositional scene representation methods. Slot Attention (Locatello et al. 2020) is chosen because the proposed method adopts a similar attention mechanism in the inference. GMIOO (Yuan, Li, and Xue 2019a) and SPACE (Lin et al. 2020) are chosen because they are two representative methods that also explicitly model the varying number of objects, and can distinguish background from objects and determine the depth ordering of objects.

**Experimental Results:** Comparison of scene decomposition performance under the extreme condition that each visual scene is only observed from one viewpoint is shown in Table 2. The proposed method is competitive with the state-of-the-arts and achieves the best or the second-best scores in almost all the cases. The Supplementary Material includes discussions and further quantitative and qualitative results.

## Conclusions

In this paper, we have considered a novel problem of learning compositional scene representations from multiple unspecified viewpoints in a fully unsupervised way, and proposed a deep generative model called OCLOC to solve this problem. On several specifically designed synthesized datasets, the proposed fully unsupervised method achieves competitive or slightly better results compared with a state-of-the-art method with viewpoint supervision, which has validated the effectiveness of the proposed method.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62176060), STCSM project (No.20511100400), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), Shanghai Research and Innovation Functional Program (No.17DZ2260900), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

## References

Burgess, C. P.; Matthey, L.; Watters, N.; Kabra, R.; Higgins, I.; Botvinick, M.; and Lerchner, A. 2019. MONet: Unsupervised scene decomposition and representation. *ArXiv*, 1901.11390.

Chen, C.; Deng, F.; and Ahn, S. 2021. ROOTS: Object-centric representation and rendering of 3D scenes. *Journal of Machine Learning Research*, 22(259): 1–36.

Crawford, E.; and Pineau, J. 2019. Spatially invariant unsupervised object detection with convolutional neural networks. In *AAAI*, 3412–3420.

Crawford, E.; and Pineau, J. 2020. Exploiting spatial invariance for scalable unsupervised object tracking. In *AAAI*, 3684–3692.

Emami, P.; He, P.; Ranka, S.; and Rangarajan, A. 2021. Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations. In *ICML*, 2970–2981.

Engelcke, M.; Kosiorek, A. R.; Jones, O. P.; and Posner, I. 2020. GENESIS: Generative scene inference and sampling with object-centric latent representations. In *ICLR*.

Eslami, S.; Heess, N.; Weber, T.; Tassa, Y.; Szepesvari, D.; Kavukcuoglu, K.; and Hinton, G. E. 2016. Attend, infer, repeat: Fast scene understanding with generative models. In *NeurIPS*, 3225–3233.

Eslami, S.; Rezende, D. J.; Besse, F.; Viola, F.; Morcos, A. S.; Garnelo, M.; Ruderman, A.; Rusu, A. A.; Danihelka, I.; Gregor, K.; Reichert, D. P.; Buesing, L.; Weber, T.; Vinyals, O.; Rosenbaum, D.; Rabinowitz, N. C.; King, H.; Hillier, C.; Botvinick, M.; Wierstra, D.; Kavukcuoglu, K.; and Hassabis, D. 2018. Neural scene representation and rendering. *Science*, 360: 1204–1210.

Fodor, J.; and Pylyshyn, Z. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28: 3–71.

Greff, K.; Kaufman, R. L.; Kabra, R.; Watters, N.; Burgess, C. P.; Zoran, D.; Matthey, L.; Botvinick, M.; and Lerchner, A. 2019. Multi-object representation learning with iterative variational inference. In *ICML*, 2424–2433.

Greff, K.; van Steenkiste, S.; and Schmidhuber, J. 2017. Neural Expectation Maximization. In *NeurIPS*, 6694–6704.

He, Z.; Li, J.; Liu, D.; He, H.; and Barber, D. 2019. Tracking by animation: Unsupervised learning of multi-object attentive trackers. In *CVPR*, 1318–1327.

Huang, J.; and Murphy, K. 2016. Efficient inference in occlusion-aware generative models of images. In *ICLR Workshop*.

Hubert, L.; and Arabie, P. 1985. Comparing partitions. *Journal of Classification*, 2: 193–218.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical reparameterization with Gumbel-softmax. In *ICLR*.

Jiang, J.; and Ahn, S.-J. 2020. Generative neurosymbolic machines. In *NeurIPS*, 12572–12582.

Jiang, J.; Janghorbani, S.; de Melo, G.; and Ahn, S. 2020. SCALOR: Generative world models with scalable object representations. In *ICLR*.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. B. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 1988–1997.

Johnson, S. 2010. How infants learn about the visual world. *Cognitive Science*, 34 7: 1158–1184.

Kabra, R.; Burgess, C.; Matthey, L.; Kaufman, R. L.; Greff, K.; Reynolds, M.; and Lerchner, A. 2019. Multi-object datasets. https://github.com/deepmind/multi-object-datasets/. Accessed: 2021-08-29.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational Bayes. In *ICLR*.

Kosiorek, A. R.; Kim, H.; Posner, I.; and Teh, Y. 2018. Sequential attend, infer, repeat: Generative modelling of moving objects. In *NeurIPS*, 8615–8625.

Lake, B.; Ullman, T. D.; Tenenbaum, J.; and Gershman, S. 2017. Building machines that learn and think like people. *The Behavioral and Brain Sciences*, 40: e253.

Li, N.; Eastwood, C.; and Fisher, R. B. 2020. Learning object-centric representations of multi-object scenes from multiple views. In *NeurIPS*, 5656–5666.

Lin, Z.; Wu, Y.-F.; Peri, S.; Sun, W.; Singh, G.; Deng, F.; Jiang, J.; and Ahn, S. 2020. SPACE: Unsupervised object-oriented scene representation via spatial attention and decomposition. In *ICLR*.

Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; and Kipf, T. 2020. Object-centric learning with slot attention. In *NeurIPS*, 11515–11528.

Maddison, C. J.; Mnih, A.; and Teh, Y. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*.

Marino, J.; Yue, Y.; and Mandt, S. 2018. Iterative amortized inference. In *ICML*, 3403–3412.

Marr, D. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co., Inc.

Matthey, L.; Higgins, I.; Hassabis, D.; and Lerchner, A. 2017. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/. Accessed: 2021-08-29.

Mnih, A.; and Gregor, K. 2014. Neural variational inference and learning in belief networks. In *ICML*, 1791–1799.

Nguyen, X.; Epps, J.; and Bailey, J. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11: 2837–2854.

Salimans, T.; and Knowles, D. A. 2013. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4): 837–882.

Shepard, R.; and Metzler, J. 1971. Mental rotation of three-dimensional objects. *Science*, 171: 701–703.

Stanic, A.; and Schmidhuber, J. 2019. R-SQAIR: Relational sequential attend, infer, repeat. In *NeurIPS Workshop*.

Turnbull, O.; Carey, D.; and McCarthy, R. 1997. The neuropsychology of object constancy. *Journal of the International Neuropsychological Society*, 3 3: 288–98.

van Steenkiste, S.; Chang, M.; Greff, K.; and Schmidhuber, J. 2018. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *ICLR*.

Veerapaneni, R.; Co-Reyes, J. D.; Chang, M.; Janner, M.; Finn, C.; Wu, J.; Tenenbaum, J.; and Levine, S. 2020. Entity abstraction in visual model-based reinforcement learning. In *CoRL*, 1439–1456.

Weis, M. A.; Chitta, K.; Sharma, Y.; Brendel, W.; Bethge, M.; Geiger, A.; and Ecker, A. S. 2021. Benchmarking unsupervised object representations for video sequences. *Journal of Machine Learning Research*, 22(183): 1–61.

Yuan, J.; Li, B.; and Xue, X. 2019a. Generative modeling of infinite occluded objects for compositional scene representation. In *ICML*, 7222–7231.

Yuan, J.; Li, B.; and Xue, X. 2019b. Spatial Mixture Models with Learnable Deep Priors for Perceptual Grouping. In *AAAI*, 9135–9142.

Yuan, J.; Li, B.; and Xue, X. 2021. Knowledge-Guided Object Discovery with Acquired Deep Impressions. In *AAAI*, 10798–10806.

Zablotskaia, P.; Dominici, E. A.; Sigal, L.; and Lehrmann, A. M. 2021. PROVIDE: A probabilistic framework for unsupervised video decomposition. In *UAI*.

Zitnick, C. L.; and Parikh, D. 2013. Bringing semantics into focus using visual abstraction. In *CVPR*, 3009–3016.