Generative Modeling of Infinite Occluded Objects for Compositional Scene Representation

Jinyang Yuan [1] **Bin Li** [1] **Xiangyang Xue** [1]

Abstract

We present a deep generative model which explicitly models object occlusions for compositional scene representation. Latent representations of objects are disentangled into location, size, shape, and appearance, and the visual scene can be generated compositionally by integrating these representations and an infinite-dimensional binary vector indicating presences of objects in the scene. By training the model to learn spatial dependences of pixels in the unsupervised setting, the number of objects, pixel-level segregation of objects, and presences of objects in overlapping regions can be estimated through inference of latent variables. Extensive experiments conducted on a series of specially designed datasets demonstrate that the proposed method outperforms two state-of-the-art methods when object occlusions exist.

1. Introduction

Perceiving the seemingly chaotic visual scenes structurally and compositionally is essential for humans to understand and interact with the complex world (Lake et al., 2017). This presumably innate ability has been considerably studied for years in the fields of neuroscience and cognitive science. Finding the underlying mechanisms is usually termed as the *perceptual grouping* problem (Grossberg et al., 1997), which is a special type of the binding problem (Treisman, 1996) related to visual perception. The synchronization theory (Milner, 1974) and the feature integration theory (Treisman & Gelade, 1980) are two representative theories underpinning the binding problem based on extensive experimental findings. It is intriguing to build human-like AI systems that can automatically segregate the visual scene into conceptual entities (e.g. objects) through inferring latent compositional and disentangled representations (Bengio et al., 2013) that cause the scene, on the basis of these neurodynamical and psychological theories. Because visual scenes may comprise different number of objects, it is also desirable to adaptively determine the number of representations in the scene. Compared with learning a single complex representation for the entire visual scene, decomposing the visual scene into multiple conceptual entities and learn a relatively simple representation for each separately is more advantageous because better expressiveness and generalizability can be achieved. By decomposing the raw observations, organized and compact knowledge of visual scenes can be summarized (Bienenstock et al., 1997). The learned knowledge is applicable to an infinite number of novel scenes composed of objetcs similar to the ones that have been observed before (Biederman, 1987; van den Hengel et al., 2015).

Various methods which are more or less related to the synchronization theory (Milner, 1974) have been proposed for segregating visual scenes into objects since last century. (Wang & Terman, 1995; Rao et al., 2008; Reichert & Serre, 2014) use neuronal synchrony as the mechanism and segregate objects based on simulated phases of neurons. Although the achieved results are gratifying, the learned representations of objects are not compositional and disentangled, which limits the expressiveness and generalizability of the model. In recent years, a series of approaches (Greff et al., 2016b;a; 2017; Prémont-Schwarz et al., 2017) combining spatial mixture models with neural networks to segregate objects have been proposed to learn compositional representations. In these approaches, the visual scene is modeled as a pixel-wise weighted summation of the images generated by individual mixture components, where each object is expected to be modeled by a single component. Posterior probabilities of the latent component indicator variables at each pixel are used as the segregation criterion.

Attend-Infer-Repeat (AIR) (Eslami et al., 2016) is a variable-sized variational autoencoder (VAE) (Kingma & Welling, 2014) that structurally decompose images into objects. It can automatically determine the number of objects and choose the number of inference steps accordingly. AIR

[1] Shanghai Key Laboratory of Intelligent Information Processing; Fudan-Qiniu Joint Laboratory for Deep Learning; Shanghai Institute of Intelligent Electronics & Systems; School of Computer Science, Fudan University, China. Correspondence to: Bin Li <libin@fudan.edu.cn>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

attends to a cropped region of the visual scene and infers the latent representation of each object sequentially, which is similar to the main idea of the feature integration theory (Treisman & Gelade, 1980). Sequential Attend-Infer-Repeat (SQAIR) (Kosiorek et al., 2018) applies recurrent neural networks to model temporal relations between consecutive frames of videos, and is able to discover and track objects.

Existing deep generative models cannot determine the number and generate pixel-level segregation of objects simultaneously; in addition, they have not explicitly modeled object occlusions such that overlapping regions are not handled well without making use of temporal information (Greff et al., 2017; Kosiorek et al., 2018). In this paper, we focus on learning compositional representations for *static* visual scenes and aim to build a generative model that can determine the number of objects, generate pixel-level segregation of objects, and estimate presences of objects in overlapping regions without supervision. The proposed method takes orders of objects into consideration and explicitly model their occlusions. When generating a scene, only the forefront object at each pixel contributes to the result. During the inference, latent variables of all objects are estimated. By directly modeling shapes of objects which may be partially occluded in the visual scene, spatial dependencies of pixels can be better learned in overlapping regions.

The proposed method is evaluated on a series of datasets in which each image consists of a varying number of hollow shapes or handwritten digits. Compared with two state-of-the-art deep generative models, N-EM (Greff et al., 2017) and AIR (Eslami et al., 2016), which also learn compositional representations of visual scenes in an unsupervised manner, the proposed method achieves noticeable improvements in handling object occlusions in the absence of temporal information. Relative orders of objects are also estimated more accurately if objects can be well segregated.

2. Related Work

Several spatial mixture models have been proposed recently to learn compositional representations and generate pixel-level segregation of visual scenes. Reconstruction Clustering (RC) (Greff et al., 2016b) iteratively masks the input image with posterior probabilities and reconstructs the image in a compositional manner with a pretrained denoising autoencoder. Tagger (Greff et al., 2016a) applies a Ladder Network (Rasmus et al., 2015) to extract high-level features for tasks like classification and low-level features for tasks like perceptual grouping. RTagger (Prémont-Schwarz et al., 2017) extends Tagger to sequential data by substituting the Ladder Network with a Recurrent Ladder Network. Neural Expectation Maximization (N-EM) (Greff et al., 2017) uses neural networks to iteratively update model parameters, and the update rule is inspired by the Expectation-Maximization (EM) algorithm. Under the unsupervised setting, N-EM achieves competitive performance with much fewer parameters compared with Tagger. Relational Neural Expectation Maximization (van Steenkiste et al., 2018) combines N-EM with Message Passing Neural Network (Gilmer et al., 2017) and is able to perform physical reasoning.

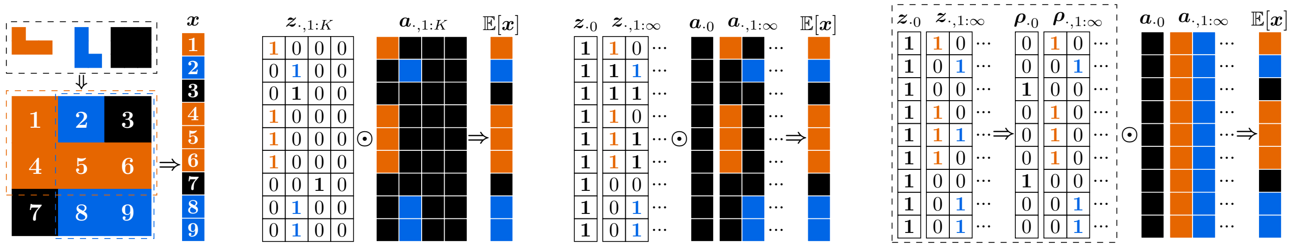
Attend-Infer-Repeat (AIR) (Eslami et al., 2016) is a variable-sized variational autoencoder (VAE) proposed for a task which also aims to describe scenes with compositional parts but does not require segregating images at pixel level. It can simultaneously estimate the number of objects as well as the representations of individual objects in the image. One limitation of AIR is that its performance degrades when occlusions of objects exist. Sequential Attend-Infer-Repeat (SQAIR) (Kosiorek et al., 2018) extends AIR to sequential data and is able to discover and track objects in the sequence of frames. It handles occlusions by utilizing temporal information for analyzing motions of objects.

Layered representations have been studied for years in the computer vision community, and several methods which handle occlusions have been proposed (Wang & Adelson, 1994; Williams & Titsias, 2004; Le Roux et al., 2011; Huang & Murphy, 2016; Moreno et al., 2016; Wu et al., 2017). We also consider occlusions, and propose a deep generative model that can extract latent representations which fully characterize objects and background in the visual scene.

The problem settings of N-EM (Greff et al., 2017) and AIR (Eslami et al., 2016) are closest to ours, and their main ideas are briefly described below.

2.1. Neural Expectation Maximization (N-EM)

N-EM (Greff et al., 2017) combines finite spatial mixture models with neural networks to estimate pixel-level segregations and extract compositional representations of objects in images. Each object is expected to be modeled by one mixture component, and prior probabilities that each pixel belongs to different objects are described by mixture weights. Let N, C, and K denote numbers of pixels, channels, and components in each image. The image $\boldsymbol{x} \in \mathbb{R}^{N \times C}$ is modeled by $p(\boldsymbol{x}) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(z_{n,k} = 1) p(\boldsymbol{x}_n | z_{n,k} = 1)$, where $\boldsymbol{z} \in \{0, 1\}^{N \times K}$ represents object assignments of pixels and each row of it is a one-hot vector. For real-valued images, $p(\boldsymbol{x}_n | z_{n,k} = 1)$ is chosen to be a normal distribution with mean $\boldsymbol{a}_{n,k} \in \mathbb{R}^C$, and the relation between $\boldsymbol{x}$, $\boldsymbol{z}$, and $\boldsymbol{a}$ is $\mathbb{E}[\boldsymbol{x}_n] = \sum_{k=1}^{K} z_{n,k} \boldsymbol{a}_{n,k}$. During inference, model parameters are estimated using neural networks based on the Expectation-Maximization (EM) (Dempster et al., 1977) framework. Because $\boldsymbol{a}_{n,k}$ is regularized to be close to a predefined value (e.g., expectation of the background pixels) if $P(z_{n,k} = 1 | \boldsymbol{x}_n)$ is small, shapes and appearances of objects are entangled in $\boldsymbol{a}$. Figure 1(b) illustrates possible values of $\boldsymbol{z}$ and $\boldsymbol{a}$ that generate the visual scene shown in Figure 1(a).



(a) Visual scene (b) N-EM (c) AIR (d) Proposed

Figure 1. Possible values of $\boldsymbol{z}$ and $\boldsymbol{a}$ that generate (a) the visual scene $\boldsymbol{x}$ in (b) N-EM, (c) AIR, and (d) the proposed method. For N-EM, all components (columns) of $\boldsymbol{a}$ contain information about the background, and the generated $\boldsymbol{x}$ is not affected by the values of $\boldsymbol{z}_3$ and $\boldsymbol{z}_7$. For AIR and the proposed method, entries not shown in $\boldsymbol{z}$ and $\boldsymbol{\rho}$ are all 0, and entries not shown in $\boldsymbol{a}$ have no influences on the generated $\boldsymbol{x}$. Symbol $\odot$ represents the operation that first performs element-wise multiplication and then sums the result along the row.

2.2. Attend-Infer-Repeat (AIR)

AIR (Eslami et al., 2016) is a type of variational autoencoder (VAE) that can automatically determine the number and locations, as well as obtaining compositional representations of objects in images. AIR assumes that the image $\boldsymbol{x}$ is drawn from the normal distribution $\mathcal{N}(\mathbb{E}[\boldsymbol{x}], \sigma^2\boldsymbol{I})$. It decomposes $\mathbb{E}[\boldsymbol{x}]$ into a binary component[1] $\boldsymbol{z}$ and a real-valued component $\boldsymbol{a}$, subject to $\mathbb{E}[\boldsymbol{x}_n]=\sum_{k=0}^{\infty} z_{n,k}\boldsymbol{a}_{n,k}$. Different from N-EM, AIR is an infinite latent feature model which does not restrict the number of 1's in each row of $\boldsymbol{z}$. During inference, $\boldsymbol{z}$ is jointly determined by outputs (bounding boxes) of a spatial transform network (STN) (Jaderberg et al., 2015) and a unary code vector representing the number of objects in the image, and $\boldsymbol{a}$ is generated by a VAE with crops of images in the bounding boxes as inputs. Figure 1(a) may be generated by AIR with values of $\boldsymbol{z}$ (note that the two dashed bounding boxes in Figure 1(a) correspond to $\boldsymbol{z}_{\cdot,1}$ and $\boldsymbol{z}_{\cdot,2}$) and $\boldsymbol{a}$ shown in Figure 1(c). Given the construction $\mathbb{E}[\boldsymbol{x}_5] = \sum_{k=0}^{\infty} z_{5,k}\boldsymbol{a}_{5,k} = \boldsymbol{a}_{5,1} + \boldsymbol{a}_{5,2}$, $\boldsymbol{a}_{5,1}$ and $\boldsymbol{a}_{5,2}$ cannot be orange and blue, respectively, in order to obtain the desired value (orange) of $\mathbb{E}[\boldsymbol{x}]$ which resembles the scene.

2.3. Limitations in Our Problem Setting

This paper considers learning compositional representations for *static* visual scenes and aim to build a generative model that can determine the number of objects, generate pixel-level segregation of objects, and estimate presences of objects in overlapping regions without supervision.

In N-EM, information of shape, appearance, location, and size of each object is entangled in one feature vector $\boldsymbol{s}_k$, and the complexity of $\boldsymbol{s}_k$ is relatively high in order to describe each object well. In addition, no prior distribution is defined to regularize $\boldsymbol{s}_k$. These two aspects limit the performance of N-EM when object occlusions exist, in the absence of additional information like relative motions of objects. Moreover, the number of objects in each image is not explicitly modeled and cannot be inferred directly.

AIR represents location, size, and attributes (shape and appearance) separately, and the complexity to describe each object is lower. Numbers of objects are modeled as random variables and estimated during the inference. However, pixel-level segregation of objects cannot be inferred directly like N-EM, because each pixel is not assumed to be generated by a single object. Moreover, overlapping regions cannot be handled well for the reason that the relation $\mathbb{E}[\boldsymbol{x}_n]=\sum_{k=0}^{\infty} z_{n,k}\boldsymbol{a}_{n,k}$ defined by AIR does not take object occlusions into consideration.

3. Proposed Approach

Inspired by advantages of N-EM and AIR, and realizing their shortcomings, we propose a deep generative model which explicitly models occlusions of objects for compositional scene representation. This method first constructs a binary matrix $\boldsymbol{z}$ with infinite columns describing shapes of objects, and then transforms $\boldsymbol{z}$ into another binary matrix $\boldsymbol{\rho}$ whose infinite columns indicate the perceived objects/background at each pixel. The representation for each object is disentangled into latent variables representing *location*, *size*, *shape*, and *appearance*. The background is modeled differently from the foreground objects because only the appearance of background may vary among images. The expectations of appearances of objects and background are represented by $\boldsymbol{a}$, in which entries of each column are identical[2]. Figure 1(d) shows one solution of $\boldsymbol{z}$, $\boldsymbol{\rho}$, and $\boldsymbol{a}$ that generates Figure 1(a).

[1] In (Eslami et al., 2016), $\boldsymbol{z}$ denotes latent representations of objects. To compare with N-EM in Figure 1(b) for a better illustration, here we use $\boldsymbol{z}$ to indicate bounding boxes of objects.

[2] For conciseness we only consider color as *appearance* here. It is straightforward to employ a neural network to generate textures as *appearance*; then entries of each column in $\boldsymbol{a}$ are not identical but composing coherent textures.

3.1. Generative Model

Each image $\boldsymbol{x}$ is assumed to be generated based on infinite-dimensional latent variables $\boldsymbol{s}$, $\boldsymbol{\nu}$, $\boldsymbol{z}^{\text{ind}}$, and $\boldsymbol{z}^{\text{dep}}$. The distributions that generate latent variables and the relations between observed and latent variables are described below.

$$
\begin{aligned}
\boldsymbol{s}_{\cdot k} &\sim \text{Normal}\left(\tilde{\boldsymbol{\mu}}, \text{diag}(\tilde{\boldsymbol{\sigma}}^2)\right), && k \geq 0 \\
\nu_k &\sim \text{Beta}(\alpha, 1), && k \geq 1 \\
z_k^{\text{ind}} &\sim \text{Bernoulli}\left(\textstyle\prod_{k'=1}^{k} \nu_{k'}\right), && k \geq 1 \\
z_{n,k}^{\text{dep}} &\sim \text{Bernoulli}\left(f_{\text{stn}}(f_{\text{shp}}(\boldsymbol{s}_{\cdot k}^{\text{shp}}), \boldsymbol{s}_{\cdot k}^{\text{stn}})_n\right), && k \geq 1 \\
z_{n,k} &= z_k^{\text{ind}} z_{n,k}^{\text{dep}}, && k \geq 1 \\
\rho_{n,k} &= \begin{cases} z_{n,k} \prod_{k'=1}^{k-1} \left(1 - z_{n,k'}\right), & k \geq 1 \\ 1 - \sum_{k'=1}^{\infty} \rho_{n,k'}, & k = 0 \end{cases} \\
\boldsymbol{a}_{n,k} &= \begin{cases} f_{\text{apc}}^{\text{obj}}(\boldsymbol{s}_{\cdot k}^{\text{apc}}), & k \geq 1 \\ f_{\text{apc}}^{\text{back}}(\boldsymbol{s}_{\cdot k}^{\text{apc}}), & k = 0 \end{cases} \\
\boldsymbol{x}_n &\sim \sum_{k=0}^{\infty} \rho_{n,k} \, \text{Normal}(\boldsymbol{a}_{n,k}, \hat{\sigma}^2 \boldsymbol{I})
\end{aligned}
$$

Variable $\boldsymbol{s}$ is a matrix of finite rows and infinite columns. Each column corresponds to one object/background, and the number of rows is the dimension of latent representation for each object/background. $\boldsymbol{s}_{\cdot k}$ can be further divided into $\boldsymbol{s}_{\cdot k}^{\text{apc}}$, $\boldsymbol{s}_{\cdot k}^{\text{shp}}$, and $\boldsymbol{s}_{\cdot k}^{\text{stn}}$, which describe *appearance*, *shape*, and relative *scale and translation* of the kth object/background. $\boldsymbol{\nu}$ contains parameters of the spatially independent Bernoulli distributions from which entries of $\boldsymbol{z}^{\text{ind}}$ are sampled, and $\boldsymbol{z}^{\text{ind}}$ is an infinite-dimensional binary vector of which the kth entry determines whether to include the object/background described by $\boldsymbol{s}_{\cdot k}$ in the image. $\boldsymbol{z}^{\text{ind}}$ can be seen as a binary matrix with one row and infinite columns which is obtained using the stick-breaking construction (Teh et al., 2007) for the Indian Buffet Process (IBP) (Ghahramani & Griffiths, 2006; Griffiths & Ghahramani, 2011). $\boldsymbol{z}^{\text{dep}}$ is a binary matrix with finite rows and infinite columns. The number of rows equals the number of pixels in each image. Entries of $\boldsymbol{z}^{\text{dep}}$ are sampled from spatially dependent Bernoulli distributions whose parameters are determined by $\boldsymbol{s}^{\text{shp}}$ and $\boldsymbol{s}^{\text{stn}}$. $\boldsymbol{z}_n^{\text{dep}}$ together with $\boldsymbol{z}^{\text{ind}}$ determines $\boldsymbol{z}_n$, which is in turn transformed into another binary vector $\boldsymbol{\rho}_n$. Because of the construction rule defined for $\rho_{n,k}$, it is guaranteed that $\boldsymbol{\rho}_n$ contains exactly one 1. $\boldsymbol{a}_{n,k}$ is the expectation of appearance of the kth object/background at the nth pixel. $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\sigma}}^2$, $\hat{\sigma}^2$, and α are hyperparameters of the distributions of latent variables. f_{stn}, f_{shp}, $f_{\text{apc}}^{\text{obj}}$, and $f_{\text{apc}}^{\text{back}}$ are neural networks that map latent representations $\boldsymbol{s}$ to other latent variables. The generative model is illustrated in Figure 2.

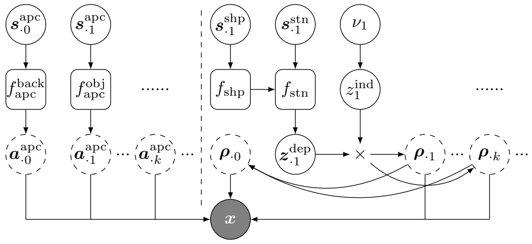

$\boldsymbol{s}_{\cdot 0}^{\text{apc}}$
$\boldsymbol{s}_{\cdot 1}^{\text{apc}}$
$\boldsymbol{s}_{\cdot 1}^{\text{shp}}$
$\boldsymbol{s}_{\cdot 1}^{\text{stn}}$
ν_1
$f_{\text{apc}}^{\text{back}}$
$f_{\text{apc}}^{\text{obj}}$
......
f_{shp}
f_{stn}
z_1^{ind}
......
$\boldsymbol{a}_{\cdot 0}^{\text{apc}}$
$\boldsymbol{a}_{\cdot 1}^{\text{apc}}$ ··· $\boldsymbol{a}_{\cdot k}^{\text{apc}}$ ···
$\boldsymbol{\rho}_{\cdot 0}$
$\boldsymbol{z}_{\cdot 1}^{\text{dep}}$
×
$\boldsymbol{\rho}_{\cdot 1}$ ··· $\boldsymbol{\rho}_{\cdot k}$ ···
x

Figure 2. The proposed generative model. Circles represent the observed and latent variables, and squares represent neural networks. Variables in dashed circles are deterministic given all the parents.

The background of the image spans all the pixels that are not occupied by any foreground objects, and is accordingly modeled by a special component with index $k = 0$. Latent variables $\boldsymbol{s}_{\cdot 0}^{\text{shp}}$, $\boldsymbol{s}_{\cdot 0}^{\text{stn}}$, ν_0, z_0^{ind}, and $\boldsymbol{z}_{\cdot 0}^{\text{dep}}$ are not sampled but fixed to constant values based on the prior knowledge of this property of the background. Let $\boldsymbol{h} = \{\boldsymbol{s}^{\text{apc}}, \boldsymbol{s}^{\text{shp}}, \boldsymbol{s}^{\text{stn}}, \boldsymbol{\nu}, \boldsymbol{z}^{\text{ind}}, \boldsymbol{z}^{\text{dep}}\}$ be the collection of latent variables, and $\boldsymbol{\theta}$ be the collection of all hyperparameters and neural networks' parameters. The joint probability of the image $\boldsymbol{x}$ and latent variables $\boldsymbol{h}$ is

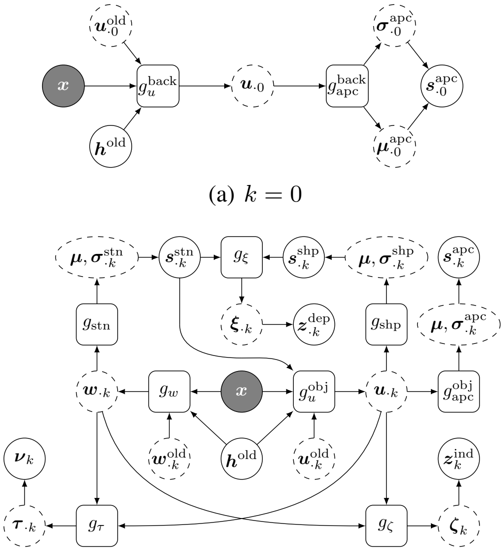
$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{h}) = &\prod_{k=0}^{\infty} p_{\boldsymbol{\theta}}(\boldsymbol{s}_{\cdot k}^{\text{apc}}) \prod_{k=1}^{\infty} p_{\boldsymbol{\theta}}(\boldsymbol{s}_{\cdot k}^{\text{shp}}) p_{\boldsymbol{\theta}}(\boldsymbol{s}_{\cdot k}^{\text{stn}}) p_{\boldsymbol{\theta}}(\nu_k) \\
&\prod_{k=1}^{\infty} \left(p_{\boldsymbol{\theta}}(z_k^{\text{ind}} | \boldsymbol{\nu}_{1:k}) \prod_{n=1}^{N} p_{\boldsymbol{\theta}}(z_{n,k}^{\text{dep}} | \boldsymbol{s}_{\cdot k}^{\text{shp}}, \boldsymbol{s}_{\cdot k}^{\text{stn}}) \right) \\
&\prod_{n=1}^{N} \sum_{k=0}^{\infty} \rho_{n,k} p_{\boldsymbol{\theta}}(\boldsymbol{x}_n | \boldsymbol{a}_{n,k})
\end{aligned} \tag{1}
$$

3.2. Variational Inference

The posterior probability $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{x})$ is computationally intractable. Latent variables $\boldsymbol{h}$ are thus inferred by approximating $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{x})$ with a variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{x})$ that minimizes the Kullback-Leibler (KL) divergence $D_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{x}) || p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{x}))$ under certain tractability constraints. Similar to the variational methods proposed by (Blei & Jordan, 2004) for the Dirichlet process, inference of the infinite-dimensional latent variables are handled by the truncated stick-breaking process bounded by K. For all $k > K$, $\prod_{k'=K+1}^{k} \nu_{k'}$ is assumed to be 0. As a result, $\rho_{n,k} = 0, \forall k > K$, and pixel $\boldsymbol{x}_n$ may only be drawn from $p_{\boldsymbol{\theta}}(\boldsymbol{x}_n|\boldsymbol{a}_{n,k})$ with indexes $0 \leq k \leq K$. The variational distribution is factorized as

$$
\begin{aligned}
q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{x}) = q_{\boldsymbol{\phi}}(\boldsymbol{s}_{\cdot 0}^{\text{apc}}) &\prod_{k=1}^{K} q_{\boldsymbol{\phi}}(\boldsymbol{s}_{\cdot k}^{\text{stn}}) q_{\boldsymbol{\phi}}(\boldsymbol{s}_{\cdot k}^{\text{shp}} | \boldsymbol{s}_{\cdot k}^{\text{stn}}) q_{\boldsymbol{\phi}}(\boldsymbol{s}_{\cdot k}^{\text{apc}} | \boldsymbol{s}_{\cdot k}^{\text{stn}}) \\
\prod_{k=1}^{K} &\left(q_{\boldsymbol{\phi}}(\nu_k | \boldsymbol{s}_{\cdot k}^{\text{stn}}) q_{\boldsymbol{\phi}}(z_k^{\text{ind}} | \boldsymbol{s}_{\cdot k}^{\text{stn}}) \prod_{n=1}^{N} q_{\boldsymbol{\phi}}(z_{n,k}^{\text{dep}} | \boldsymbol{s}_{\cdot k}^{\text{shp}}, \boldsymbol{s}_{\cdot k}^{\text{stn}}) \right)
\end{aligned} \tag{2}
$$

All the probability distributions on the right-hand side of (2) condition on $\boldsymbol{x}$, which is omitted for conciseness. The


$u_{\cdot 0}^{\text{old}}$
x
g_u^{back}
h^{old}
$u_{\cdot 0}$
$g_{\text{apc}}^{\text{back}}$
$\sigma_{\cdot 0}^{\text{apc}}$
$s_{\cdot 0}^{\text{apc}}$
$\mu_{\cdot 0}^{\text{apc}}$
$\mu, \sigma_{\cdot k}^{\text{stn}}$
$s_{\cdot k}^{\text{stn}}$
g_ξ
$s_{\cdot k}^{\text{shp}}$
$\mu, \sigma_{\cdot k}^{\text{shp}}$
$s_{\cdot k}^{\text{apc}}$
g_{stn}
$\xi_{\cdot k}$
$z_{\cdot k}^{\text{dep}}$
g_{shp}
$\mu, \sigma_{\cdot k}^{\text{apc}}$
$w_{\cdot k}$
g_w
x
g_u^{obj}
$u_{\cdot k}$
$g_{\text{apc}}^{\text{obj}}$
ν_k
$w_{\cdot k}^{\text{old}}$
h^{old}
$u_{\cdot k}^{\text{old}}$
z_k^{ind}
$\tau_{\cdot k}$
g_τ
g_ζ
ζ_k

(a) $k = 0$

(b) $1 \le k \le K$

Figure 3. The inference model. Solid circles denote the observed and latent variables, and squares denote neural networks. Dashed circles are parameters of the inference distribution as well as the intermediate nodes $\boldsymbol{w}$ and $\boldsymbol{u}$ which determine these parameters.

specific forms of these distributions are

$$\begin{aligned} q_{\boldsymbol{\phi}}(\boldsymbol{s}_{\cdot k}^{*}|\boldsymbol{s}_{\cdot k}^{\text{stn}}) &= \text{Normal}(\boldsymbol{s}_{\cdot k}^{*}; \boldsymbol{\mu}_{\cdot k}^{*}, \text{diag}({\boldsymbol{\sigma}_{\cdot k}^{*}}^{2})) \\ q_{\boldsymbol{\phi}}(\nu_k|\boldsymbol{s}_{\cdot k}^{\text{stn}}) &= \text{Beta}(\nu_k; \tau_{1,k}, \tau_{2,k}) \\ q_{\boldsymbol{\phi}}(z_k^{\text{ind}}|\boldsymbol{s}_{\cdot k}^{\text{stn}}) &= \text{Bernoulli}(z_k^{\text{ind}}; \zeta_k) \\ q_{\boldsymbol{\phi}}(z_{n,k}^{\text{dep}}|\boldsymbol{s}_{\cdot k}^{\text{shp}}, \boldsymbol{s}_{\cdot k}^{\text{stn}}) &= \text{Bernoulli}(z_{n,k}^{\text{dep}}; \xi_{n,k}) \end{aligned}$$

All parameters of the distributions above are generated by neural networks with image $\boldsymbol{x}$ and latent variables these distributions conditioned on as inputs. $\boldsymbol{\phi}$ in the subscripts of the distributions is the collection of all neural network parameters. $\tau_{1,k}$, $\tau_{2,k}$, ζ_k, as well as entries of $\boldsymbol{\mu}_{\cdot k}$ and $\boldsymbol{\sigma}_{\cdot k}^2$ that characterize $q_{\boldsymbol{\phi}}(\boldsymbol{s}_{\cdot k}^{\text{shp}}|\boldsymbol{s}_{\cdot k}^{\text{stn}})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{s}_{\cdot k}^{\text{apc}}|\boldsymbol{s}_{\cdot k}^{\text{stn}})$ depend on $\boldsymbol{s}_{\cdot k}^{\text{stn}}$. $\xi_{n,k}$ depends on $\boldsymbol{s}_{\cdot k}^{\text{shp}}$ and $\boldsymbol{s}_{\cdot k}^{\text{stn}}$. For notational simplicity, these dependences are omitted in the expressions.

Figure 3 describes the relations between observed/latent variables and parameters of the variational distribution. The neural networks modeling these relations for $k = 0$ and $1 \le k \le K$ are not identical, because the background ($k = 0$) differs from foreground objects ($1 \le k \le K$) in that $\boldsymbol{s}_{\cdot 0}^{\text{apc}}$ is the only latent variable required to be inferred. According to the feature integration theory (FIT) (Treisman & Gelade, 1980), humans attend to a particular location of the visual scene at a certain time when performing perceptual grouping. Latent variables of all objects are not inferred simultaneously in the proposed method in order to follow this theory. The coordinate ascent algorithm, which is commonly used in the mean-field variational inference, updates parameters of the variational distribution alternately and accords with FIT if parameters of each object are updated sequentially. Because of the non-linearities of neural networks, closed-form solutions of coordinate ascent cannot be derived for the proposed model. We utilize long short-term memories (LSTMs) g_w, g_u^{obj}, and g_u^{back} to imitate the procedure of coordinate ascent by alternately updating $\boldsymbol{w}_{\cdot k}$ and $\boldsymbol{u}_{\cdot k}$ which determine the variational distribution parameters of each object, conditioned on the image and the rest objects. $q_{\boldsymbol{\phi}}(\boldsymbol{s}_{\cdot k}^{\text{stn}})$ depends solely on $\boldsymbol{w}_{\cdot k}$, and $\boldsymbol{u}_{\cdot k}$ provides additional information for other distributions which are conditioned on $\boldsymbol{s}_{\cdot k}^{\text{stn}}$. In order to decrease the number of iterations and increase the inference speed, $\boldsymbol{w}_{\cdot k}$ and $\boldsymbol{u}_{\cdot k}$ are initialized with neural networks instead of random values. Approximating the behavior of coordinate descent using LSTMs is similar to learning to learn by gradient descent with LSTMs (Andrychowicz et al., 2016), whose idea is used by Neural Expectation Maximization (N-EM) (Greff et al., 2017) to concurrently update all randomly initialized mixture model parameters based on manually derived gradient expressions.

3.3. Structures of Neural Networks

In the generative model, f_{shp} is a convolutional neural network (CNN) that transforms latent representations $\boldsymbol{s}_{\cdot k}^{\text{shp}}$ to shapes of objects in object-level coordinates that are independent of sizes and locations of objects in the image. $f_{\text{apc}}^{\text{obj}}$ and $f_{\text{apc}}^{\text{back}}$ are multilayer perceptrons (MLPs) that compute the means of normal distributions of object/backgound appearances based on $\boldsymbol{s}_{\cdot k}^{\text{apc}}$. f_{stn} first uses *tanh* and *sigmoid* functions to compute the relative scales and translations of objects based on $\boldsymbol{s}_{\cdot k}^{\text{stn}}$, and then transforms $f_{\text{shp}}(\boldsymbol{s}_{\cdot k}^{\text{shp}})$ from object-level coordinates to the image-level coordinate.

In the inference model, g_w, g_u^{obj}, g_u^{back}, and the neural networks used for initializations are combinations of CNNs and LSTMs. g_ξ is a combination of CNN and affine transform. g_{stn}, g_{shp}, $g_{\text{apc}}^{\text{obj}}$, $g_{\text{apc}}^{\text{back}}$, g_τ, and g_ζ are MLPs.

3.4. Learning of Neural Networks

Parameters of all neural networks are jointly learned by minimizing the KL divergence $D_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{x}))$. An equivalent objective is to maximize the evidence lower bound (ELBO) defined by

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{h} \sim q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{h})] - D_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{h})) \quad (3)$$

By sampling latent variables $\boldsymbol{h}$ from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{x})$ and estimating the gradients of (3), neural network parameters can be optimized using gradient-based methods. Although z_k^{ind} and $z_{n,k}^{\text{dep}}$ are discrete latent variables, applying the generic but relatively high-variance black box methods such as BBVI (Ranganath et al., 2014) and NVIL (Mnih & Gregor, 2014) to compute their gradients is not necessary, because expectations of $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{h})$ with respect to z_k^{ind} and $z_{n,k}^{\text{dep}}$ can be computed analytically. Except for ν_k, which is sampled from a beta distribution, all the continuous latent variables are sampled from normal distributions and the reparameterization trick (Salimans et al., 2013; Kingma & Welling, 2014) is applicable to reduce the variance of the gradient estimator. ν_k appears only in the KL divergence terms. $D_{\text{KL}}(q(\nu_k|\boldsymbol{s}_{:k}^{\text{stn}})||p(\nu_k))$ can be computed analytically, and ν_k in this term is marginalized out. The only problem is $\boldsymbol{\nu}_{1:k}$ in $D_{\text{KL}}(q_{\boldsymbol{\phi}}(z_k^{\text{ind}}|\boldsymbol{s}_{:k}^{\text{stn}})||p(z_k^{\text{ind}}|\boldsymbol{\nu}_{1:k}))$, which may be handled in two ways. The first is to utilize the generalized reparameterization gradient (Ruiz et al., 2016) to obtain low-variance estimate of the gradient. The second is to apply the multinomial approximation to $\mathbb{E}_{\boldsymbol{\nu}}[\log(1-\prod_{k'=1}^{k}\nu_{k'})]$ as mentioned in (Doshi et al., 2009) and obtain an upper bound of $D_{\text{KL}}(q_{\boldsymbol{\phi}}(z_k^{\text{ind}}|\boldsymbol{s}_{:k}^{\text{stn}})||p(z_k^{\text{ind}}|\boldsymbol{\nu}_{1:k}))$ that has a closed-form solution. We adopt the latter and optimize parameters of neural networks with respect to a lower bound of (3). Details are provided in the supplementary material.

4. Experiments

Datasets: The perceptual grouping performance of the compared methods are evaluated on a series of datasets derived from the publicly released datasets provided by (Greff et al., 2016b;a; 2017). The size of images in all datasets is 48×48, and each image may contain $2\sim4$ binary hollow shapes (referred as *Shapes*) or real-valued handwritten digits (referred as *MNIST*). To evaluate the perceptual grouping performance from different perspectives, these datasets differ from one another in multiple aspects. Samples of images in different datasets are illustrated in the first row of Figure 4. In all datasets, 50,000, 10,000, and 10,000 images are used for training, validation, and test, respectively.

Compared Methods: *N-EM* (Greff et al., 2017) and *AIR* (Eslami et al., 2016) are chosen as the compared methods. The number of objects in each image cannot be estimated based on posterior inference in N-EM, and pixel-level segregations of objects are not generated in AIR. To evaluate the performance of N-EM and AIR in these aspects without modifying their core algorithms, we apply heuristic post-processing to the results of N-EM and AIR. For N-EM, the number of objects in each image is determined by classifying each mixture component as modeling object or background, based on the posterior probabilities. For AIR, objects are segregated in pixel level heuristically based on the similarity between reconstructions of individual objects and the original image at each pixel. Pixels are assumed to belong to objects with the highest similarities. For the proposed method, both the number of objects and pixel-level segregation results can be inferred directly, and no additional operations are required. All three methods are trained on images containing 2 or 3 objects with $K=4$.

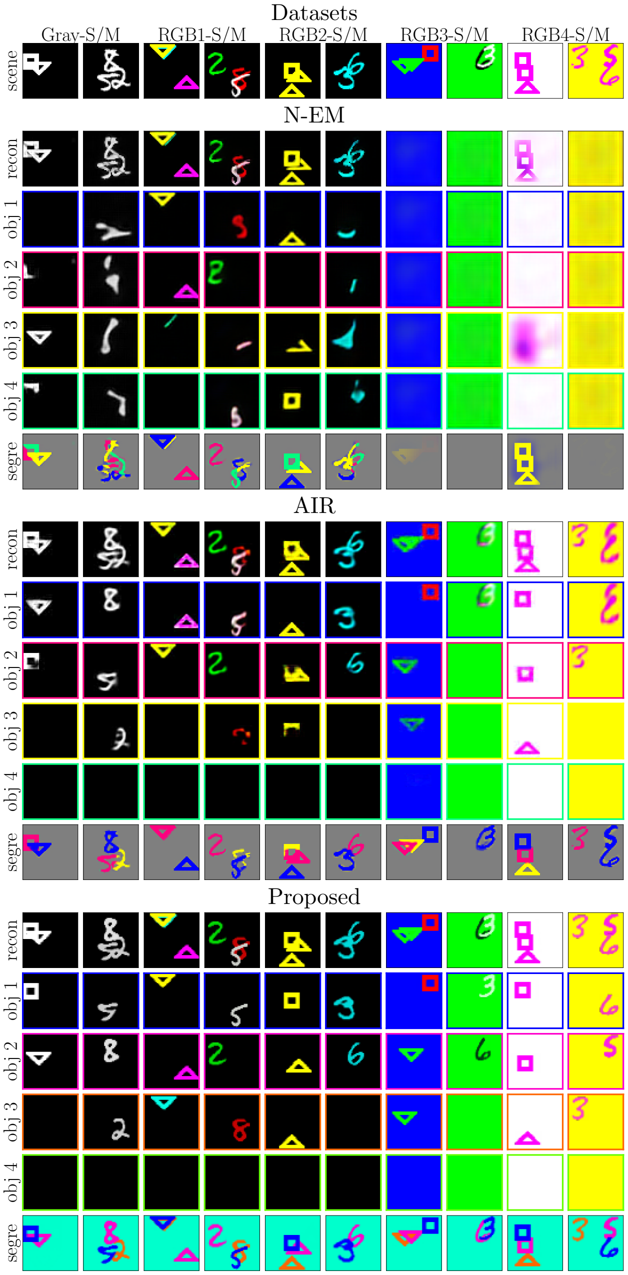


Figure 4. Examples of qualitative results evaluated on different datasets. "S/M" in the names of datasets stand for *Shapes* and *MNIST*, respectively. For each method, the 1st row presents reconstructions of scenes, the 2rd–5th rows illustrate reconstructions of individual objects, and the 6th row displays segregation results.

Evaluation Metrics: Four metrics are used to assess the performance: (1) *Adjusted Mutual Information* (AMI) is a normalized version of mutual information used to assess segregation results. (2) *Mean Squared Error* (MSE) measures similarities between reconstructions of individual objects and the ground truth. (3) *Object Counting Accuracy* (OCA) measures the quality of estimated number of objects in each

Table 1. Comparison of segregation and counting performance *with* existence of occlusion.

Data set	N-EM			AIR			Proposed		
	AMI (%)	MSE (e-3)	OCA (%)	AMI (%)	MSE (e-3)	OCA (%)	AMI (%)	MSE (e-3)	OCA (%)
Gray-S	77.3± 2e-1	10± 1e-1	56.2± 6e-1	85.4± 1e-1	6.5± 3e-2	80.9± 2e-1	**94.6**± 9e-2	**2.9**± 3e-2	**90.5**± 1e-1
Gray-M	30.5± 1e-1	22± 6e-2	13.5± 1.0	62.8± 4e-2	9.0± 9e-3	66.0± 2e-1	**71.1**± 1e-1	**7.5**± 4e-2	**77.6**± 3e-1
RGB1-S	81.8± 3e-1	5.6± 8e-2	74.2± 1.2	95.3± 6e-2	2.4± 1e-2	88.8± 2e-1	**98.3**± 5e-2	**1.1**± 1e-2	**95.1**± 2e-1
RGB1-M	57.0± 2e-1	9.4± 4e-2	16.3± 2.4	78.2± 8e-2	3.5± 8e-3	67.9± 5e-1	**82.0**± 5e-2	**3.1**± 1e-2	**74.8**± 4e-1
RGB2-S	66.2± 2e-1	9.0± 6e-2	60.8± 1.4	85.7± 4e-2	3.7± 7e-3	84.4± 1e-1	**92.3**± 1e-1	**2.2**± 3e-2	**86.3**± 3e-1
RGB2-M	34.9± 2e-1	13± 1e-2	12.5± 9e-1	64.1± 1e-1	4.8± 8e-3	69.8± 2e-1	**67.9**± 2e-1	**4.7**± 2e-2	**71.0**± 5e-1
RGB3-S	29.6± 1e-1	21± 8e-3	7.44± 6e-1	91.3± 8e-2	3.9± 9e-3	90.3± 2e-1	**97.4**± 6e-2	**1.4**± 2e-2	**92.5**± 2e-1
RGB3-M	15.4± 2e-1	22± 3e-1	2.30± 3e-1	67.5± 7e-2	5.4± 3e-3	60.5± 1e-1	**77.9**± 1e-1	**3.8**± 9e-3	**68.6**± 7e-1
RGB4-S	24.7± 3e-1	20± 4e-2	10.3± 2e-1	86.7± 2e-2	4.0± 3e-3	78.3± 1e-1	**90.7**± 8e-2	**2.5**± 2e-2	**83.3**± 9e-2
RGB4-M	3.82± 1e-1	32± 2e-1	2.35± 4e-1	56.9± 4e-2	6.3± 3e-3	58.2± 3e-1	**67.9**± 7e-2	**4.6**± 2e-2	**77.3**± 3e-1

image. (4) *Object Ordering Accuracy* (OOA) computes the weighted sum of pairwise order estimation accuracies. More details are provided in the supplementary material.

Table 2. Comparison of ordering performance.

OOA (%) with original order			
Data set	N-EM	AIR	Proposed
RGB1-S	49.3± 1.3	57.0± 0.2	**77.5**± 0.4
RGB1-M	49.6± 1.0	55.3± 0.4	**57.0**± 0.4
RGB3-S	49.6± 1.2	57.4± 0.2	**68.9**± 0.6
RGB3-M	50.1± 0.5	54.7± 0.2	**59.1**± 0.5

OOA (%) with adjusted order			
Data set	N-EM	AIR	Proposed
RGB1-S	51.2± 0.9	67.4± 0.7	**95.3**± 0.5
RGB1-M	45.4± 0.4	45.1± 0.8	**59.9**± 1.0
RGB3-S	47.0± 0.6	50.2± 0.2	**91.2**± 0.6
RGB3-M	47.1± 1.3	46.7± 0.5	**54.7**± 1.3

4.1. Qualitative Results

Some examples of qualitative results of the compared methods evaluated on different datasets are shown in Figure 4. Compared with N-EM and AIR, the proposed method performs noticeably better when occlusions exist; and its segregation results are sharper because representations of shapes and appearances are disentangled. By explicitly modeling object occlusions, the proposed method can infer the shapes of objects well in overlapping regions.

4.2. Performance of Segregation and Counting

In N-EM, information of background appearance is used to define a regularization term in the loss function in order to learn specialized representations. In AIR, no latent variable is assigned to the background. To compare with N-EM and AIR in a controlled setting, we fix $\boldsymbol{a}_{\cdot 0}$ to the background appearance and do not infer $\boldsymbol{s}^{\text{apc}}_{\cdot 0}$. Performance of the proposed method when the background appearance is unknown is provided in the supplementary material.

Without Existence of Occlusion: When no object occlusions exist, the average AMI/MSE/OCA scores of N-EM, AIR, and the proposed method are $53.3\%/13.8e\text{-}3/38.2\%$, $98.3\%/1.2e\text{-}3/96.4\%$, and $98.6\%/1.1e\text{-}3/97.5\%$, respectively. Detailed results are shown in the supplementary material. N-EM works well when images are composed of hollow shapes and backgrounds are black (Gray-S, RGB1-S, and RGB2-S), and its performance degrades when colors of backgrounds may vary among images (RGB3 and RGB4) or objects are relatively complex to represent (*MNIST*). It is probably because information of each object is entangled in a single feature vector, and no prior distribution is defined to regularize this vector. N-EM can handle moving handwritten digits even if occlusions exist (Greff et al., 2017). However, if the temporal information is not available, the high complexities of representations limit its performance on relatively complex scenes. Both AIR and the proposed method achieve gratifying and comparable performance on all datasets. When no object occlusion exists, they can accurately segregate objects, generate the visual scene containing a single object, and determine the number of objects.

With Existence of Occlusion: Performance evaluated when objects may be overlapped are shown in Table 1. Compared with the situation that no occlusion exists, all approaches achieve worse results. For AIR and the proposed method, performance on the datasets consist of hollow shapes degrades less than handwritten digits. The possible reason is that hollow shapes exhibit less variations and can be better reconstructed when partially occluded. Because the proposed method explicitly models object occlusions, it outperforms AIR and N-EM when objects may be overlapped.

Table 3. Comparison of generalizability *with* existence of occlusion (tested on images containing 4 objects with $K = 4$).

Data set	N-EM			AIR			Proposed		
	AMI (%)	MSE (E-3)	OCA (%)	AMI (%)	MSE (E-3)	OCA (%)	AMI (%)	MSE (E-3)	OCA (%)
Gray-S	75.0± 3E-1	16± 2E-1	69.5± 9E-1	81.5± 5E-2	9.6± 3E-2	52.8± 3E-1	**93.6**± 6E-2	**3.6**± 2E-2	**72.8**± 2E-1
Gray-M	38.6± 2E-1	31± 9E-2	1.35± 2E-1	56.4± 8E-2	14± 3E-2	18.5± 4E-1	**66.5**± 3E-2	**13**± 4E-2	**61.0**± 5E-1
RGB1-S	82.8± 2E-1	7.4± 7E-2	23.7± 4E-1	91.6± 6E-2	4.8± 1E-2	95.4± 2E-1	**97.0**± 5E-2	**1.9**± 1E-2	**95.9**± 2E-1
RGB1-M	65.1± 3E-1	13± 6E-2	0.79± 2E-1	73.1± 7E-2	5.9± 1E-2	**90.5**± 9E-2	**79.7**± 1E-1	**5.2**± 1E-2	88.6± 5E-1
RGB2-S	68.7± 1E-1	12± 7E-2	19.9± 1.9	81.1± 3E-2	5.9± 1E-2	58.6± 2E-1	**90.0**± 1E-1	**2.8**± 4E-2	**63.0**± 5E-1
RGB2-M	42.4± 6E-2	18± 3E-2	1.16± 2E-1	58.0± 4E-2	**7.5**± 1E-2	31.7± 5E-1	**64.1**± 5E-2	7.8± 2E-2	**55.6**± 2E-1
RGB3-S	28.0± 1E-1	22± 2E-2	7.92± 5E-1	77.0± 1E-2	8.0± 1E-2	83.6± 3E-1	**94.2**± 1E-1	**2.3**± 2E-2	**85.2**± 3E-1
RGB3-M	15.4± 1E-1	22± 3E-1	31.2[3]± 6E-1	60.9± 7E-2	8.0± 9E-3	68.9± 4E-1	**75.0**± 1E-1	**6.3**± 4E-2	**85.5**± 3E-1
RGB4-S	23.3± 2E-1	25± 4E-2	4.85± 3E-1	70.8± 7E-2	8.0± 3E-3	45.8± 2E-1	**87.6**± 7E-2	**3.3**± 2E-2	**54.8**± 6E-1
RGB4-M	3.64± 1E-1	37± 3E-1	**48.8**[3]± 5E-1	45.1± 5E-2	9.9± 8E-3	2.91± 1E-1	**60.8**± 3E-2	**7.4**± 1E-2	24.9± 2E-1

4.3. Performance of Ordering

The OOA scores of N-EM, AIR, and the proposed method evaluated under two settings are shown in Table 2.

Original Order: For N-EM, indices of mixture components are chosen as the estimated order. For AIR and the proposed method, the order determined by LSTMs are used. Because mixture components in N-EM are exchangeable, the OOA scores of N-EM are close to 50% (random guess). For AIR and the proposed method, the orders determined by LSTMs are better than random guess. The possible reason is that objects which are not occluded exhibit more integrities and are more likely to be noticed by the attention mechanism. Because the proposed method explicitly models object occlusions, it achieves higher OOA than AIR.

Adjusted Order: The similarities between the original image and reconstructions of individual objects can be used to define the estimated order. The main idea is that objects more similar to the original image in the regions occupied by them are less likely to be occluded. For N-EM and AIR, regions of objects are determined by the segregation results which contain shape information. The proposed method disentangles shapes and appearance, which allows the usage of inferred shapes as such regions. Using this scheme, the proposed method achieves significant improvements on datasets composed of hollow shapes. The performance on datasets containing handwritten digits are not changed much. It is probably because high variations prevent handwritten digits to be reconstructed well when occlusions exist.

4.4. Generalization to Novel Scenes

When object occlusions exist, the generalizabilities evaluated with $K = 4$ are presented in Table 3. All three methods generalize well in terms of segregation performance. The proposed method generalizes better than AIR on datasets in which background appearances may vary among images (RGB3 and RGB4). It is probably because occlusions are explicitly modeled, and representations of objects are independent of background. The counting performance of all methods generalizes less well compared with the segregation performance. The possible reason is that objects in test images are more likely to be heavily occluded, which makes the counting problem much harder. The results evaluated with $K = 10$ are included in the supplementary material.

Performances evaluated with $K = 4$ and 10 under the situation that no occlusions occur are provided in the supplementary material. The proposed method generalize well in terms of both segregation and counting performance.

5. Conclusion

In this paper, we have proposed a deep generative model which explicitly models object occlusions for compositional scene representation. By learning spatial dependencies of pixels in the unsupervised setting, this model can determine the number of objects, generate pixel-level segregation of objects, and estimate presences of objects in overlapping regions. We have demonstrated that the proposed method outperforms two state-of-the-art methods when object occlusions exist, and generalizes well to novel scenes.

The proposed method falls in the framework of Learnable Deep Priors (LDP) (Yuan et al., 2019), which facilitates integrating rich and structured prior knowledge. It has been observed that regularizing latent representations of objects with normal distributions alone is not sufficient when objects may be occluded and shapes of them exhibit high variations. Integrating stronger prior knowledge of spatial dependencies and learning compositional representations for more sophisticated scenes will be investigated in our future work.

[3] For N-EM, the OCA scores computed based on the heuristic post-processing are relatively high on these datasets when K equals the number of objects in each image.

Acknowledgments

This work was supported in part by National Key R&D Program of China (2017YFC0803700), NSFC Projects (61572138, U1611461), STCSM Project (16JC1420400, 18511103104), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01, 2018SHZDZX01), ZJLab, and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

References

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., and De Freitas, N. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3981–3989, 2016.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2013.

Biederman, I. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94 (2):115, 1987.

Bienenstock, E., Geman, S., and Potter, D. Compositionality, MDL priors, and object recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 838–844, 1997.

Blei, D. M. and Jordan, M. I. Variational methods for the Dirichlet process. In *International Conference on Machine Learning (ICML)*, pp. 12. ACM, 2004.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

Doshi, F., Miller, K., Van Gael, J., and Teh, Y. W. Variational inference for the Indian buffet process. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 137–144, 2009.

Eslami, S. A., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Hinton, G. E., et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3225–3233, 2016.

Ghahramani, Z. and Griffiths, T. L. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 475–482, 2006.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML)*, pp. 1263–1272, 2017.

Greff, K., Rasmus, A., Berglund, M., Hao, T., Valpola, H., and Schmidhuber, J. Tagger: Deep unsupervised perceptual grouping. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 4484–4492, 2016a.

Greff, K., Srivastava, R. K., and Schmidhuber, J. Binding via reconstruction clustering. In *International Conference on Learning Representations (ICLR) Workshop*, 2016b.

Greff, K., van Steenkiste, S., and Schmidhuber, J. Neural expectation maximization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 6691–6701, 2017.

Griffiths, T. L. and Ghahramani, Z. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(Apr):1185–1224, 2011.

Grossberg, S., Mingolla, E., and Ross, W. D. Visual brain and visual perception: How does the cortex do perceptual grouping? *Trends in Neurosciences*, 20(3):106–111, 1997.

Huang, J. and Murphy, K. Efficient inference in occlusion-aware generative models of images. In *International Conference on Learning Representations (ICLR) Workshop*, 2016.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2017–2025, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Kosiorek, A. R., Kim, H., Posner, I., and Teh, Y. W. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 8615–8625, 2018.

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

Le Roux, N., Heess, N., Shotton, J., and Winn, J. Learning a generative model of images by factoring appearance and shape. *Neural Computation*, 23(3):593–650, 2011.

Milner, P. M. A model for visual shape recognition. *Psychological Review*, 81(6):521, 1974.

Mnih, A. and Gregor, K. Neural variational inference and learning in belief networks. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 1791–1799, 2014.

Moreno, P., Williams, C. K., Nash, C., and Kohli, P. Overcoming occlusion with inverse graphics. In *European Conference on Computer Vision (ECCV) Workshop*, pp. 170–185. Springer, 2016.

Prémont-Schwarz, I., Ilin, A., Hao, T., Rasmus, A., Boney, R., and Valpola, H. Recurrent ladder networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 6009–6019, 2017.

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 814–822, 2014.

Rao, A. R., Cecchi, G. A., Peck, C. C., and Kozloski, J. R. Unsupervised segmentation with dynamical units. *IEEE Transactions on Neural Networks*, 19(1):168–182, 2008.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3546–3554, 2015.

Reichert, D. P. and Serre, T. Neuronal synchrony in complex-valued deep networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Ruiz, F. J., Titsias, M. K., and Blei, D. M. The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 460–468, 2016.

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

Teh, Y. W., Grür, D., and Ghahramani, Z. Stick-breaking construction for the Indian buffet process. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 556–563, 2007.

Treisman, A. The binding problem. *Current Opinion in Neurobiology*, 6(2):171–178, 1996.

Treisman, A. M. and Gelade, G. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

van den Hengel, A., Russell, C., Dick, A., Bastian, J., Pooley, D., Fleming, L., and Agapito, L. Part-based modelling of compound scenes from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 878–886, 2015.

van Steenkiste, S., Chang, M., Greff, K., and Schmidhuber, J. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *International Conference on Learning Representations (ICLR)*, 2018.

Wang, D. and Terman, D. Locally excitatory globally inhibitory oscillator networks. *IEEE Transactions on Neural Networks*, 6(1):283–286, 1995.

Wang, J. Y. and Adelson, E. H. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3 (5):625–638, 1994.

Williams, C. K. and Titsias, M. K. Greedy learning of multiple objects in images using robust statistics and factorial learning. *Neural Computation*, 16(5):1039–1062, 2004.

Wu, J., Tenenbaum, J. B., and Kohli, P. Neural scene derendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 699–707, 2017.

Yuan, J., Li, B., and Xue, X. Spatial mixture models with learnable deep priors for perceptual grouping. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.