Knowledge-Guided Object Discovery with Acquired Deep Impressions

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

We present a framework called Acquired Deep Impressions (ADI) which continuously learns knowledge of objects as "impressions" for compositional scene understanding. In this framework, the model first acquires knowledge from scene images containing a single object in a supervised manner, and then continues to learn from novel multi-object scene images which may contain objects that have not been seen before without any further supervision, under the guidance of the learned knowledge as humans do. By memorizing impressions of objects into parameters of neural networks and applying the generative replay strategy, the learned knowledge can be reused to generate images with pseudo-annotations and in turn assist the learning of novel scenes. The proposed ADI framework focuses on the acquisition and utilization of knowledge, and is complementary to existing deep generative models proposed for compositional scene representation. We adapt a base model to make it fall within the ADI framework and conduct experiments on two types of datasets. Empirical results suggest that the proposed framework is able to effectively utilize the acquired impressions and improve the scene decomposition performance.

Introduction

The world is complex not only in the great variations of objects that constitute it, but also in the diverse compositions of these objects. Figure 1 presents some toy examples of scenes constructed from various simple objects. In order to interact with the world efficiently and effectively, humans tend to understand the perceived visual scenes in a compositional and structured way by decomposing the complex scenes into relatively simple objects and organizing these objects structurally (Lake et al. 2017). Regularized by possibly inborn mental laws (Koffka 2013; Goldstein and Brockmole 2016) and based on the previously learned knowledge of objects, humans can decompose and understand novel visual scenes composed of familiar objects reliably, and build up new knowledge of novel objects efficiently with few examples of visual scenes containing these objects. The newly acquired knowledge is accumulated in the memory and assists the understanding of new scenes in the future.

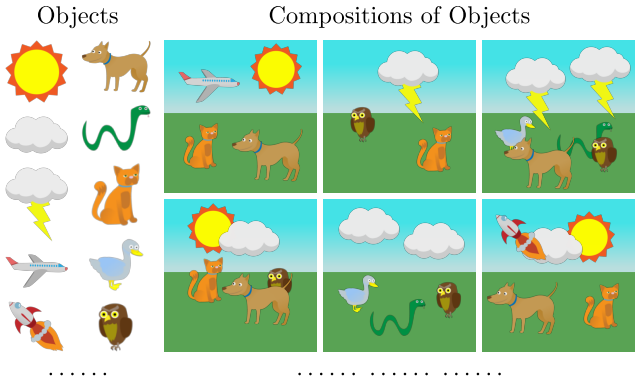


Figure 1: Scenes are diverse compositions of various objects.

In recent years, a large body of deep generative models such as variational autoencoders (VAE) (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014) and flow-based generative models (Dinh, Krueger, and Bengio 2015; Dinh, Sohl-Dickstein, and Bengio 2017; Kingma and Dhariwal 2018) have been proposed and provide mechanisms to represent images with latent variables by building nonlinear mappings between images and latent variables with neural networks. The prior distribution of the latent variable along with the neural network characterize the distribution of images and can thus be seen as the learned knowledge of images. By sampling latent variables from the prior distribution and transforming latent variables with the decoding neural network, novel images similar to those used for training can be generated. Most existing VAEs and flow-based models learn a single representation for the whole image, and cannot be used to decompose the scene image into individual objects directly.

It is intriguing to design human-like machines that are able to compositionally represent scenes with individual objects and continuously learn knowledge of objects which acts as priors to assist the decomposition of novel scenes. In contrast to the aforementioned approaches which treat the full scene as a whole and lack effective mechanisms of incorporating accumulated prior knowledge, such machines lower the complexity to represent visual scenes and facilitate understanding novel visual scenes. As shown in Figure 2, under the guidance of impressions of objects which are

*Corresponding author

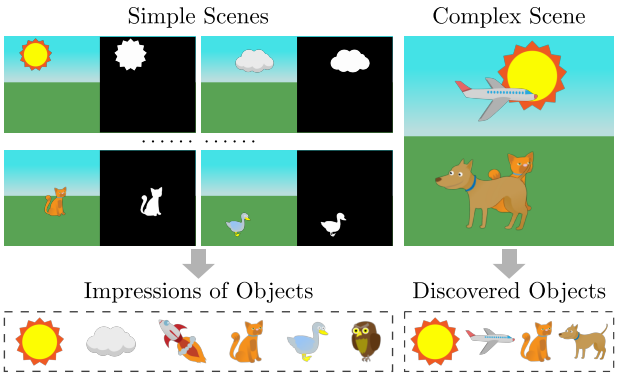


Figure 2: Complete objects in the complex scene can be discovered using impressions learned from simple scenes.

previously acquired in a supervised manner, it is possible to extract complete objects from the perceived visual scene even if the scene contains some novel objects that have not been seen before and occlusions exist.

To decompose visual scenes and achieve better compositionality, a number of deep generative models (Huang and Murphy 2016; Eslami et al. 2016; Greff, van Steenkiste, and Schmidhuber 2017; Crawford and Pineau 2019; Stelzner, Peharz, and Kersting 2019; Greff et al. 2019; Yuan, Li, and Xue 2019a; Burgess et al. 2019; Engelcke et al. 2020; Yang, Chen, and Soatto 2020) which learn a separate representation for each object in the scene in an unsupervised manner, have been proposed. Object-based representations extracted by these methods are fundamental to the acquisition of impressions of objects. However, these works focus mainly on the unsupervised decomposition of scenes instead of the acquisition and utilization of impressions.

In this paper, we propose a framework called Acquired Deep Impressions (ADI) which enables continuously learning knowledge of individual objects constituting the visual scenes as "impressions" in a methodical manner. In this framework, visual scenes are modeled as the composition of layers of objects and background using spatial mixture models. Each layer is associated with a latent variable, which can be transformed into appearance and shape of the layer by learnable mappings modeled by *deep* neural networks. The parameters of neural networks are *acquired* from data, and the prior distributions of latent variables together with the learned parameters of neural networks define the *impressions* of objects and background.

Complementary to existing deep generative models proposed for compositional scene representation, the proposed ADI framework provides a mechanism to effectively acquire and utilize knowledge. An existing base model is adapted to fall within the ADI framework, and experiments are conducted with this model on two types of datasets. Extensive empirical results suggest that the impressions previously acquired from relatively simple visual scenes play an important role as prior knowledge for the discovery of objects in complex scenes, and the proposed ADI framework is able to improve the scene decomposition performance by effectively acquiring and utilizing impressions.

Related Work

In recent years, various spatial mixture models have been proposed for the binding problem (Treisman 1996) and related higher-level tasks which require extracting compositional latent representations of objects. RC (Greff, Srivastava, and Schmidhuber 2016) iteratively reconstructs the visual scene compositionally with a pretrained denoising autoencoder. Tagger (Greff et al. 2016) utilizes a Ladder Network (Rasmus et al. 2015) to extract both high-level and low-level latent representations which are suitable for different tasks. RTagger (Prémont-Schwarz et al. 2017) is applicable to sequential data by replacing the Ladder Network used in (Greff et al. 2016) with the Recurrent Ladder Network. N-EM (Greff, van Steenkiste, and Schmidhuber 2017) infers latent representations of objects with a recurrent neural network (RNN) based on the Expectation-Maximization (EM) framework. Relational N-EM (van Steenkiste et al. 2018) tackles the problem of common-sense physical reasoning by modeling relations between objects. SMMLDP (Yuan, Li, and Xue 2019b) uses neural networks to model mixture weights of the spatial mixture model and infers latent representations based on the N-EM framework. In all these models, priors of latent representations are not defined, and inferences of latent representations are based on the maximum likelihood estimation (MLE). Knowledge of objects are not fully captured by the learned model because there is no natural way to sample latent representations and generate visual scenes similar to the ones used for training.

CST-VAE (Huang and Murphy 2016) models visual scenes by layered representations (Wang and Adelson 1994) and takes occlusions of objects into consideration. AIR (Eslami et al. 2016) is a type of variable-sized VAE that can determine the number of objects in the scene and extract object-based representations. SQAIR (Kosiorek et al. 2018) extends AIR to videos of moving objects by modeling relations between objects in consecutive frames. SPAIR (Crawford and Pineau 2019) combines ideas used in supervised object detection with AIR and is able to handle large and many-objects scenes. SuPAIR (Stelzner, Peharz, and Kersting 2019) substitutes VAEs in AIR with sum-product networks to increase the speed and robustness of learning. IODINE (Greff et al. 2019) jointly segments and represents objects based on the iterative amortized inference framework (Marino, Yue, and Mandt 2018). GMIOO (Yuan, Li, and Xue 2019a) models occlusions between objects and is able to determine the number of objects in the scene and segment them simultaneously. MONet (Burgess et al. 2019) proposes a novel recurrent attention network for inferring compositional latent representations. GENESIS (Engelcke et al. 2020) considers relationships between objects in the generative model and is able to generate more coherent scenes. (Yang, Chen, and Soatto 2020) integrates Contextual Information Separation and perceptual cycle-consistency into compositional deep generative model, and is able to perform unsupervised segmentation on manually generated scenes composed of objects with complex textures.

Object-based representations play a fundamental role in the proposed ADI framework. The aforementioned methods learn to extract object-based representations without

ground-truth annotations, and can be applied to discover objects in an unsupervised manner. Complementary to these methods, ADI focuses more on the utilization of previously learned knowledge and proposes a learning procedure which could empower these methods by exploiting the acquired impressions to assist the discovery of objects (including those that have not been seen before) in novel scenes. ADI aims to decompose a scene containing possibly multiple objects into object(s) and background, and thus differs from the methods proposed for single-object scenes (Singh, Ojha, and Lee 2019) or the methods that decompose scenes into hierarchical features (Zhao, Song, and Ermon 2017). Because ADI first acquires impressions from simple scenes in the supervised scenario and then continues to learn from more complex scenes without any further supervision, it is also different from the methods which incorporate pretrained models in the supervised learning pipeline (Ulutan, Iftekhar, and Manjunath 2020).

Acquired Deep Impressions

Representing visual scenes is complex due to the diverse combinations of objects in the scenes. In most existing deep generative models, the whole visual scene is encoded into a single latent representation. To obtain models which generalize well in novel scenes, a great number of data which cover all possible combinations need to be observed during training. Because the combinations of objects are in general extremely complex even if individual objects are simple to model, learning such a type of representation is not very data-efficient. If visual scenes can be decomposed into objects which are represented separately, improved sampling efficiency and generalizability may be achieved. Humans can decompose complex and novel scenes which are composed of familiar objects effectively even if occlusions exist, probably because impressions of complete objects are accumulated from simpler scenes which have been observed previously. Inspired by this phenomenon, we propose a framework called Acquired Deep Impressions (ADI) to facilitate the discovery of objects in novel scenes by learning *impressions* of objects in a methodical manner. The proposed ADI framework is complementary to existing compositional scene representation methods and is able to empower these methods to utilize previously acquired knowledge by modifying them to fall within the ADI framework.

Generative Model

In ADI, a visual scene $\boldsymbol{x} \in \mathbb{R}^{N\times C}$ is assumed to be generated by composing layers of objects and background using spatial mixture models. N and C are the respective numbers of pixels and channels in each image. Each layer k is associated with a latent variable $\boldsymbol{z}_k$ that is drawn independently from the prior distribution $p(\boldsymbol{z}_k;\boldsymbol{\theta}_{\text{bck}})$ or $p(\boldsymbol{z}_k;\boldsymbol{\theta}_{\text{obj}})$ for $k=0$ (background) or $k\geq 1$ (objects). The collection of all latent variables $\{\boldsymbol{z}_0,\boldsymbol{z}_1,\dots\}$ is denoted by $\boldsymbol{z}$. Each latent variable $\boldsymbol{z}_k$ is transformed to the appearance $\boldsymbol{a}_k \in \mathbb{R}^{N\times C}$ and the variable containing shape information $\boldsymbol{s}_k \in \mathbb{R}^N$ (e.g., mask of complete shape or logit of perceived shape) of the background ($k=0$) or object ($k\geq 1$), by a learnable mapping f_{bck} or f_{obj}. A compositing function f_{comp} (e.g. stick-breaking or softmax) which simultaneously takes all the variables containing shape information $\boldsymbol{s}=\{\boldsymbol{s}_0,\boldsymbol{s}_1,\dots\}$ as inputs is then applied to transform $\boldsymbol{s}$ into perceived shapes $\boldsymbol{m}=\{\boldsymbol{m}_0,\boldsymbol{m}_1,\dots\}$ ($\boldsymbol{m}_k \in \mathbb{R}^N, \forall k$). Each pixel $\boldsymbol{x}_n$ of the visual scene is assumed to be conditional independent of each other given all the latent variables $\boldsymbol{z}$. Let l_n denote the variable indicating which layer is observed at the nth pixel. The perceived shapes $\boldsymbol{m}$ and appearances $\boldsymbol{a}$ of layers are used as mixture weights $p(l_n=k|\boldsymbol{z})$ and parameters of mixture components $p(\boldsymbol{x}_n|\boldsymbol{z}_{l_n}, l_n=k)$, respectively. The joint probability of the visual scene $\boldsymbol{x}$ and latent variables $\boldsymbol{z}$ is factorized as $p(\boldsymbol{x},\boldsymbol{z})=p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$, where

$$p(\boldsymbol{z}) = \prod\nolimits_k p(\boldsymbol{z}_k) \tag{1}$$

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod\nolimits_n \sum\nolimits_k p(l_n=k|\boldsymbol{z}) p(\boldsymbol{x}_n|\boldsymbol{z}_{l_n}, l_n=k) \tag{2}$$

The generative process of the visual scene is illustrated in Figure 3. The detailed expressions are given by

$$\begin{aligned}
&\boldsymbol{z}_k \sim p(\boldsymbol{z}_k;\boldsymbol{\theta}_{\text{bck}}), \quad \boldsymbol{a}_k, \boldsymbol{s}_k = f_{\text{bck}}(\boldsymbol{z}_k); \quad k=0\\
&\boldsymbol{z}_k \sim p(\boldsymbol{z}_k;\boldsymbol{\theta}_{\text{obj}}), \quad \boldsymbol{a}_k, \boldsymbol{s}_k = f_{\text{obj}}(\boldsymbol{z}_k); \quad k\geq 1\\
&\boldsymbol{m}_0, \boldsymbol{m}_1, \cdots = f_{\text{comp}}(\boldsymbol{s}_0, \boldsymbol{s}_1, \dots)\\
&l_n \sim \text{Categorical}(m_{0,n}, m_{1,n}, \dots)\\
&\boldsymbol{x}_n \sim p(\boldsymbol{x}_n;\boldsymbol{a}_{l_n,n})
\end{aligned}$$

This framework is called Acquired Deep Impressions (ADI) because: 1) the learnable mappings from latent variables to objects and background are *acquired* from data; 2) these mappings are modeled by *deep* neural networks; and 3) the priors of latent variables together with the learned parameters of neural networks define the *impressions* of objects and background. The acquired impressions act as strong prior knowledge that assist the model to discover objects and extract latent variables of *complete* objects and background even if the perceived objects and background are incomplete due to occlusions.

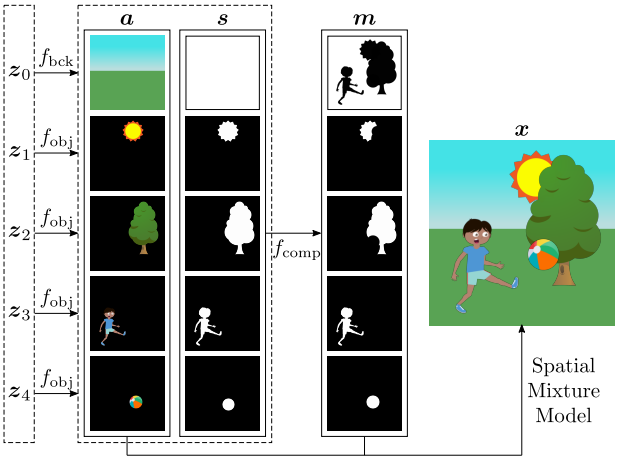


Figure 3: Illustration of the generative process. $\boldsymbol{z}_k$ is the latent variable of background ($k=0$) or object ($k\geq 1$). $\boldsymbol{a}$, $\boldsymbol{s}$, and $\boldsymbol{m}$ are the collections of appearances, masks of complete shapes (or logits of perceived shapes), and masks of perceived shapes, respectively. $\boldsymbol{x}$ is the generated visual scene.

Variational Inference

As shown in Figure 4, the layers of objects and background constituting the visual scene are fully characterized by latent variables $\boldsymbol{z}$, and the decomposition of visual scene is equivalent to the inference of latent variables of all the layers. Latent variables are inferred by a neural network g under the variational inference framework. The inference network g takes the visual scene $\boldsymbol{x}$ as inputs, and outputs parameters of the variational distribution $q(\boldsymbol{z}|\boldsymbol{x})$. By training the inference network to approximate the mapping from visual scene to latent variables, the model learns to perform scene decomposition in an amortized manner.

Learning Procedure

Learning to discover objects from scenes is in general difficult without supervision. ADI provides a way to lower the difficulty by dividing the learning into two stages which differ from each other mainly in the data used to train the model and the way the training is conducted:

- **Stage 1:** The model is first trained on scene images containing a single object, under the supervision of manually annotated or automatically generated ground truth of object shapes. Impressions of objects are saved as parameters of the decoding networks that map latent variables $\boldsymbol{z}$ to appearances and shapes of layers as well as the inference networks that output the approximated posterior distributions $q(\boldsymbol{z}|\boldsymbol{x})$. Because only one object may appear in each image, how to handle occlusions of object is not learned in this stage.
- **Stage 2:** The model then continues to learn from scene images comprising possibly multiple objects, without using any further supervision. Some images contain objects that have not been seen before. The model is expected to learn how to discover *complete* objects, including those not appearing in the first learning stage, even if the observed objects are incomplete due to occlusion. In order to improve the efficiency and effectiveness of learning, impressions acquired in the first learning stage are exploited using the generative replay strategy (Shin et al. 2017).

A commonly used loss function to train variational autoencoders is the negative evidence lower bound (ELBO) L_{elbo}, which can be decomposed into the negative log-likelihood (NLL) term L_{nll} and the Kullback–Leibler divergence (KLD) term L_{kld}, i.e., $L_{\text{elbo}} = L_{\text{nll}} + L_{\text{kld}}$. The detailed expressions of L_{nll} and L_{kld} are

$$L_{\text{nll}} = -\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\big[\log p(\boldsymbol{x}|\boldsymbol{z})\big] \tag{3}$$

$$L_{\text{kld}} = \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\big[\log q(\boldsymbol{z}|\boldsymbol{x}) - \log p(\boldsymbol{z})\big] \tag{4}$$

Beta-VAE (Higgins et al. 2017) adds an adjustable hyperparameter $\beta > 1$ as the coefficient of the KLD term in the ELBO to improve the quality of disentanglement, i.e. $L^{\beta}_{\text{elbo}} = L_{\text{nll}} + \beta L_{\text{kld}}$. As long as β is no less than 1, $-L^{\beta}_{\text{elbo}}$ is a valid lower bound of the log-evidence $\log p(\boldsymbol{x})$. We design the loss functions based on L^{β}_{elbo}, and add extra terms to incorporate supervision and perform generative replay. The illustration of the training procedure is shown in Figure 5.

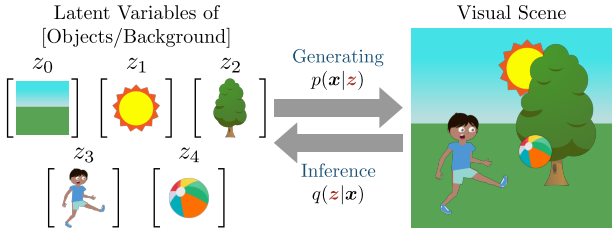


Figure 4: Objects and background are fully characterized by latent variables $\boldsymbol{z}$, and the decomposition of a visual scene is equivalent to the inference of $\boldsymbol{z}$.

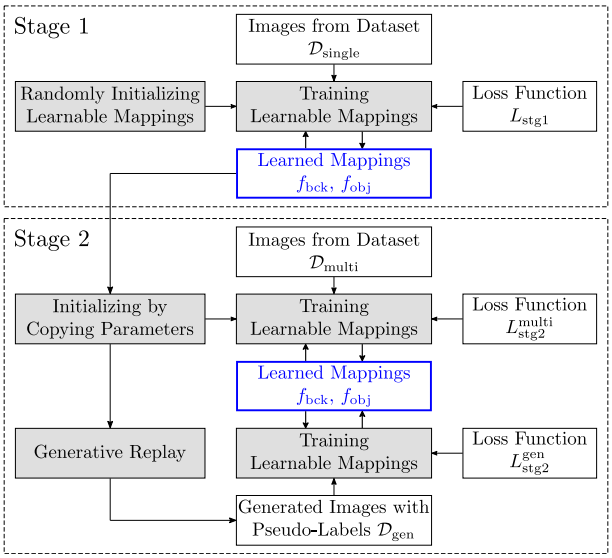


Figure 5: Illustration of the training procedure. In the first learning stage, learnable mappings f_{bck} and f_{obj} are iteratively refined by minimizing L_{stg1} on $\mathcal{D}_{\text{single}}$. In the second learning stage, f_{bck} and f_{obj} are iteratively refined by minimizing $L^{\text{multi}}_{\text{stg2}}$ and $L^{\text{gen}}_{\text{stg2}}$ on $\mathcal{D}_{\text{multi}}$ and $\mathcal{D}_{\text{gen}}$ simultaneously.

In Stage 1, Images $\mathcal{D}_{\text{single}}$ containing a *single* object are used to train the model in a *supervised* manner. The loss function of this learning stage consists of three parts:

$$L_{\text{stg1}} = L_{\text{nll}} + \beta_{\text{stg1}} L_{\text{kld}} + \alpha_{\text{stg1}} L_{\text{mask}} \tag{5}$$

In the above expression, $\beta_{\text{stg1}} \geq 1$ and $\alpha_{\text{stg1}} > 0$ are tunable hyperparameters. The first two parts are the NLL and KLD terms, and the third part is the supervision term. The supervision used during the training is the ground truth of mixture weights (perceived shapes) of all layers $\tilde{\boldsymbol{m}}$. The KL divergence between the ground truth $\tilde{\boldsymbol{m}}$ and the estimated mixture weights $\boldsymbol{m}$ are used as the supervision term L_{mask} in the loss function

$$L_{\text{mask}} = \sum\nolimits_{k,n} \tilde{m}_{k,n}(\log \tilde{m}_{k,n} - \log m_{k,n}) \tag{6}$$

In Stage 2, the model continues to learn from *unannotated* images $\mathcal{D}_{\text{multi}}$ containing *multiple* objects. When trained on these data $\mathcal{D}_{\text{multi}}$, the loss function is simply an upper bound of the negative log-evidence:

$$L^{\text{multi}}_{\text{stg2}} = L_{\text{nll}} + \beta_{\text{stg2}} L_{\text{kld}} \tag{7}$$

In order to guide the learning with the help of previously acquired impressions of objects, additional images $\mathcal{D}_{\text{gen}}$ are generated by following the generative process. f_{bck} and f_{obj} which are learned in the first learning stage are used to transform the sampled latent variables $\tilde{\boldsymbol{z}}$ into appearances $\tilde{\boldsymbol{a}}$ and variables containing shape information $\tilde{\boldsymbol{s}}$ of all the layers. $\tilde{\boldsymbol{m}} = f_{\text{comp}}(\tilde{\boldsymbol{s}})$ is used as the supervision, and the model is trained on the generated images $\mathcal{D}_{\text{gen}}$ using a loss function differing from Eq. (5) only in the chosen hyperparameters:

$$L_{\text{stg2}}^{\text{gen}} = L_{\text{nll}} + \beta_{\text{stg2}} L_{\text{kld}} + \alpha_{\text{stg2}} L_{\text{mask}} \tag{8}$$

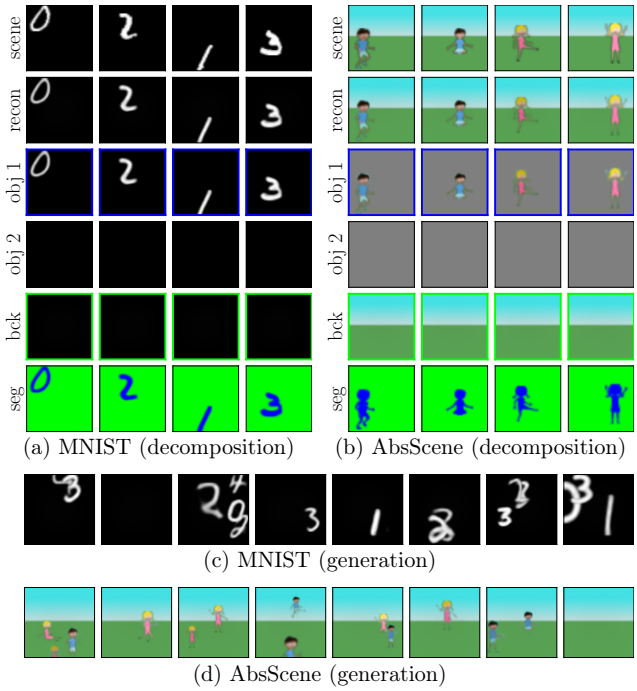
Experiments

Comprehensive experiments are conducted to validate the effectiveness of the acquired impressions. We adapt an existing method GMIOO (Yuan, Li, and Xue 2019a) proposed for compositional scene representation to make it fall within the ADI framework, and evaluate the adapted model on two types of datasets. [1] Details of the adaptation and choices of hyperparameters are provided in the supplementary material. Experimental results under different configurations demonstrate that the acquired impressions can greatly improve the discovery of objects in novel scenes. [2]

Datasets: The effectiveness of the proposed ADI framework is evaluated on two types of datasets. In the first type of datasets, images are composed of 70,000 variants of handwritten digits $0 \sim 9$ in the MNIST dataset (LeCun et al. 1998). In the second type of datasets, images are composed of 70 variants of boys and girls as well as 56 other types of abstract objects provided by the Abstract Scene Dataset (Zitnick and Parikh 2013; Zitnick, Parikh, and Vanderwende 2013). These two types of datasets are referred to as *MNIST* and *AbsScene*, respectively. In both types of datasets, the sizes of images are 64×64. The datasets $\mathcal{D}_{\text{single}}$ used in the first learning stage consist of 10,000 images containing a single object. Only some of the objects may appear in $\mathcal{D}_{\text{single}}$ (35,735 variants of digits $0 \sim 4$ for the MNIST dataset, and 70 variants of boys and girls for the AbsScene dataset). The datasets $\mathcal{D}_{\text{multi}}$ used in the second learning stage consist of 50,000 images containing $2 \sim 4$ objects, and all the objects may appear in $\mathcal{D}_{\text{multi}}$. To investigate the influence of object occlusion to the effectiveness of the proposed ADI framework, each dataset is divided into 2 subsets, which differ from each other in the average degree of occlusion ($0\% \sim 50\%$ and $50\% \sim 100\%$). The average degree of occlusion of each image is measured by first computing the ratio of overlapped area to the total area of bounding box for each object, and then averaging the ratios over all the objects in the image. 10,000 images containing $2 \sim 4$ or $5 \sim 6$ objects are used to evaluate the scene decomposition performance and generalizability of the trained models. Some examples of the two types of datasets can be found in the rows and columns labeled with "scene" in Figures 6 and 7.

[1] Code is available at https://github.com/jinyangyuan/acquired-deep-impressions.

[2] We have also tried to adapt another base model AIR (Eslami et al. 2016). Experimental results provided in the supplementary material verify that the ADI framework is also effective on AIR.



(a) MNIST (decomposition) (b) AbsScene (decomposition)

(c) MNIST (generation)

(d) AbsScene (generation)

Figure 6: The decomposition and generation results of the models trained on single-object images with supervision. Reconstructed objects are superimposed on black (MNIST) or gray (AbsScene) images.

Dataset	MSE	OCA
MNIST	5.41e-2± 4e-4	0.993± 1e-3
AbsScene	4.76e-3± 1e-4	0.992± 8e-4

Table 1: MSE and OCA scores of the models trained in the first learning stage.

Evaluation Metrics: Four metrics are used to evaluate the performance of the trained models: 1) *Object Reconstruction Error* (MSE) measures the similarity between the reconstructions of discovered objects and the ground truth by mean squared error (MSE). It is evaluated on pixels belonging to the union of estimated and ground truth shapes of objects, and provides information about how accurately the occluded regions of objects are estimated. 2) *Adjusted Mutual Information* (AMI) assesses qualities of segmentations. It is not evaluated in the background regions to better illustrate how accurately different objects are separated. For the MNIST dataset, overlapped regions are also excluded because all objects share similar appearances. 3) *Object Counting Accuracy* (OCA) is the ratio of images in which the numbers of objects are correctly estimated. 4) *Object Ordering Accuracy* (OOA) is the weighted average of accuracies of the estimated pairwise object orderings. OOA scores are not reported for the MNIST datasets because the orderings of digits are indistinguishable. Formal descriptions of evaluation metrics are provided in the supplementary material. All the models are trained once and evaluated for 5 runs.

Dataset	Avg Occl	Config	MSE	AMI	OCA	OOA
MNIST	$0\% \sim 50\%$	Direct	2.19e-1± 6e-4	0.770± 4e-4	**0.738**± 1e-3	N/A
		ADI	**1.02e-1**± 3e-4	**0.833**± 3e-4	0.704± 2e-3	N/A
	$50\% \sim 100\%$	Direct	3.15e-1± 4e-4	0.406± 1e-4	0.399± 2e-3	N/A
		ADI	**1.94e-1**± 5e-4	**0.514**± 5e-4	**0.579**± 4e-3	N/A
AbsScene	$0\% \sim 50\%$	Direct	2.74e-2± 2e-4	0.828± 4e-4	**0.873**± 1e-3	0.876± 1e-3
		ADI	**2.04e-2**± 1e-4	**0.874**± 4e-4	0.845± 1e-3	**0.888**± 2e-3
	$50\% \sim 100\%$	Direct	5.42e-2± 7e-5	0.487± 6e-4	0.213± 2e-3	0.748± 1e-3
		ADI	**1.94e-2**± 1e-4	**0.752**± 5e-4	**0.709**± 2e-3	**0.911**± 9e-4

Table 2: Performance evaluated on images containing $2 \sim 4$ objects. The models are trained on images containing $2 \sim 4$ objects.

Dataset	Avg Occl	Config	MSE	AMI	OCA	OOA
MNIST	$0\% \sim 50\%$	Direct	2.37e-1± 2e-4	0.719± 4e-4	0.303± 3e-3	N/A
		ADI	**1.29e-1**± 1e-4	**0.792**± 2e-4	**0.562**± 4e-3	N/A
	$50\% \sim 100\%$	Direct	3.29e-1± 3e-4	0.541± 3e-4	0.080± 2e-3	N/A
		ADI	**2.13e-1**± 2e-4	**0.630**± 3e-4	**0.423**± 4e-3	N/A
AbsScene	$0\% \sim 50\%$	Direct	2.29e-2± 2e-4	0.832± 4e-4	0.610± 4e-3	0.841± 9e-4
		ADI	**1.60e-2**± 7e-5	**0.867**± 3e-4	**0.661**± 3e-3	**0.864**± 2e-3
	$50\% \sim 100\%$	Direct	5.90e-2± 2e-4	0.562± 3e-4	0.030± 2e-3	0.705± 1e-3
		ADI	**2.29e-2**± 7e-5	**0.772**± 3e-4	**0.450**± 2e-3	**0.868**± 6e-4

Table 3: Performance evaluated on images containing $5 \sim 6$ objects. The models are trained on images containing $2 \sim 4$ objects.

Performance of Supervised Learning

In the first learning stage, the models are trained on images $\mathcal{D}_{\text{single}}$ containing a single object under the supervision of perceived shapes of objects and background (mixture weights of the spatial mixture models). The performance of the trained models is shown in Table **??**. Only MSE and OCA scores are reported, because only one object exists in each image and there is no need to distinguish between or determine the orderings of different objects. Both models trained on the MNIST and AbsScene datasets are able to reconstruct the objects well (low MSE scores) and determine the number of objects in the images accurately (high OCA scores). Qualitative results of decomposed scenes in $\mathcal{D}_{\text{single}}$ and generated scenes used in the second learning stage are illustrated in Figure 6. The models are able to decompose scenes into objects and background accurately, and generate images containing multiple objects with high quality.

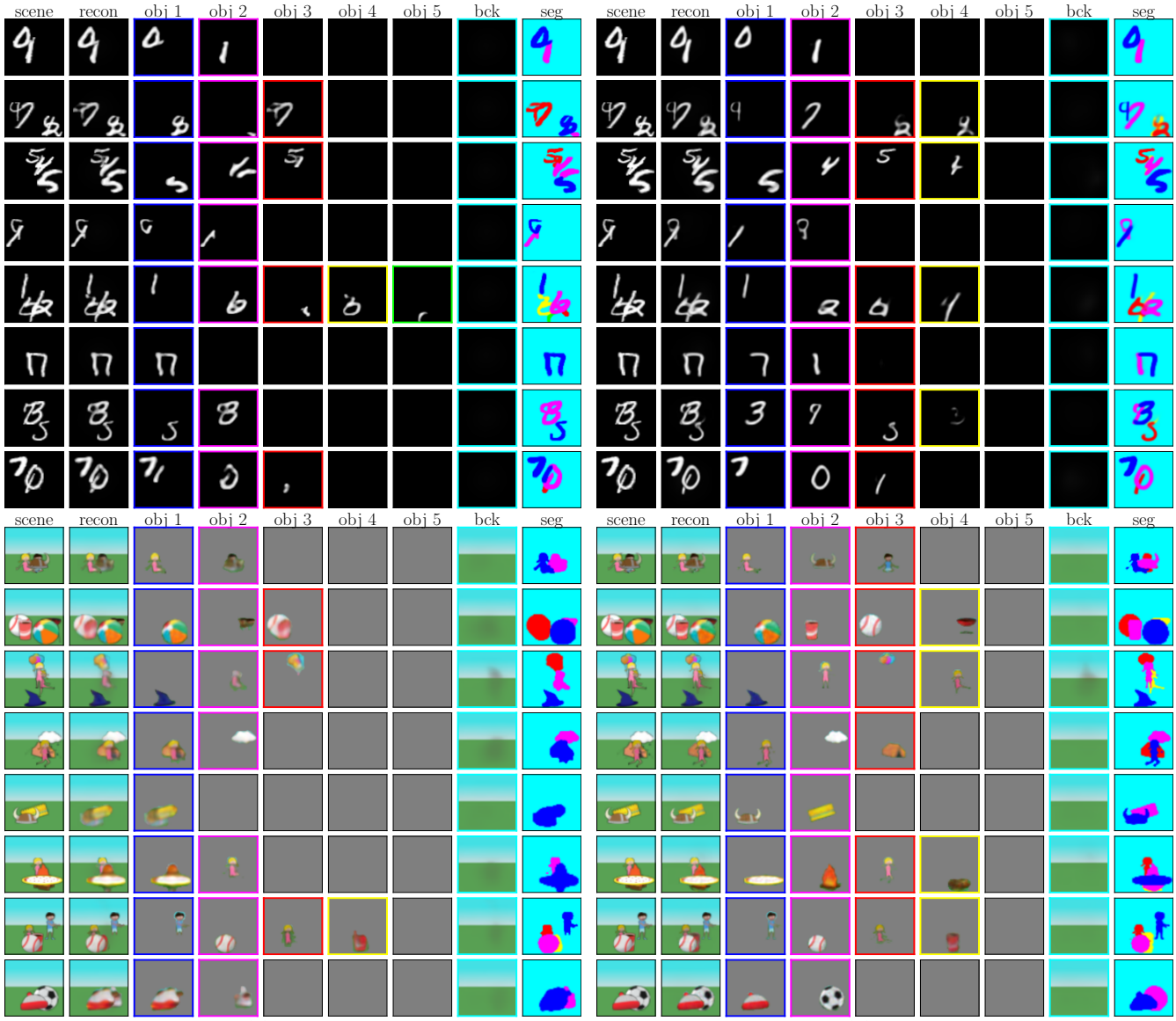
Effectiveness of Acquired Impressions

In the second learning stage of the ADI framework, the models continue to learn from multi-object images $\mathcal{D}_{\text{multi}}$ with the help of the impressions acquired in the first learning stage. To verify that the acquired impressions can assist the learning of visual scenes, we also directly train models using multi-object images and make comparisons between the scene decomposition performance of the models trained with and without the acquired impressions. Quantitative results are shown in Tables **??** and **??**, and some of the qualitative results are demonstrated in Figure 7. Experimental results of influences of hyperparameters and additional qualitative results are included in the supplementary material.

The performance presented in Tables **??** and **??** are of the same models trained on images containing $2 \sim 4$ objects. In Table **??**, the distribution of the test images is identical to the distribution of the training images. In Table **??**, the distribution of the test images is different in that the number of objects per image is within $5 \sim 6$ instead of $2 \sim 4$. On both MNIST and AbsScene datasets, the models trained using acquired impressions (ADI) are able to reconstruct individual objects better (lower MSE scores), and distinguish between and determine the orderings of different objects more accurately (higher AMI and OOA scores) than those directly trained on multi-object images (Direct).

With the help of the previously acquired impressions, the object counting accuracies (OCA scores) are also higher when objects are heavily occluded ($50\% \sim 100\%$ average degrees of occlusion). When the average degrees of occlusion are relatively small ($0\% \sim 50\%$), utilizing acquired impressions increases or decreases the OCA scores on test images contain $2 \sim 4$ or $5 \sim 6$ objects. The possible reason is that the distribution of test images containing $2 \sim 4$ objects is identical to the training images $\mathcal{D}_{\text{multi}}$, which makes determining the numbers of objects in these test images relatively easy. The extra training images provided by the generative replay strategy are drawn from a distribution different from these test images, which improves the generalizabilities of the models on test images contain $5 \sim 6$ but decreases the OCA scores on these test images containing $2 \sim 4$ objects.

Figure 7 presents samples of decomposition results of the models trained with or without the acquired impressions. Compared with the models directly trained on multi-objects images (subfigure (a)), the models trained using pre-



(a) Direct (b) ADI

Figure 7: Decomposition results of the models trained with (ADI) or without (Direct) the acquired impressions. The average degrees of occlusion are $50\% \sim 100\%$. Reconstructed objects are superimposed on black (MNIST) or gray (AbsScene) images.

viously acquired impressions (subfigure (b)) output significantly better decomposition results. Furthermore, novel objects (digits $5 \sim 9$ and abstract objects other than boys and girls) are better discovered and separated from other objects by utilizing the previously acquired impressions, even though they are not observed in the first learning stage and supervisions of them are not available.

Conclusions

In this paper, we have proposed a human-like learning framework called Acquired Deep Impressions (ADI) to facilitate the understanding of novel visual scenes by building impressions of objects with compositional latent representations and the learned decoding and inference neural networks. The proposed ADI framework is complementary to existing methods proposed for compositional scene representation in that it provides a mechanism to effectively acquire and utilize knowledge. An existing compositional deep generative model is adapted to fall within the ADI framework, and extensive experiments are conducted using this model. We have demonstrated that the model achieves significantly better decomposition performance under the guidance of previously acquired impressions (prior knowledge) in most experimental configurations, which has validated our motivation. Incorporating more expressive impressions in the ADI framework by using structured prior distributions of latent representations could be investigated in the future.

Acknowledgments

This research was supported in part by STCSM Projects (20511100400, 18511103104), Shanghai Municipal Science and Technology Major Projects (2017SHZDZX01, 2018SHZDZX01), Shanghai Research and Innovation Functional Program (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 preprint arXiv:1901.11390* .

Crawford, E.; and Pineau, J. 2019. Spatially invariant unsupervised object detection with convolutional neural networks. In *AAAI*, volume 33, 3412–3420.

Dinh, L.; Krueger, D.; and Bengio, Y. 2015. Nice: Nonlinear independent components estimation. In *ICLR Workshop*.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2017. Density estimation using real nvp. In *ICLR*.

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.

Goldstein, E. B.; and Brockmole, J. 2016. *Sensation and perception*. Cengage Learning.

Greff, K.; Kaufman, R. L.; Kabra, R.; Watters, N.; Burgess, C.; Zoran, D.; Matthey, L.; Botvinick, M.; and Lerchner, A. 2019. Multi-Object Representation Learning with Iterative Variational Inference. In *ICML*, 2424–2433.

Greff, K.; Rasmus, A.; Berglund, M.; Hao, T.; Valpola, H.; and Schmidhuber, J. 2016. Tagger: Deep unsupervised perceptual grouping. In *NeurIPS*, 4484–4492.

Greff, K.; Srivastava, R. K.; and Schmidhuber, J. 2016. Binding via reconstruction clustering. In *ICLR Workshop*.

Greff, K.; van Steenkiste, S.; and Schmidhuber, J. 2017. Neural expectation maximization. In *NeurIPS*, 6691–6701.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*.

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

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 10215–10224.

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

Koffka, K. 2013. *Principles of Gestalt psychology*, volume 44. Routledge.

Kosiorek, A. R.; Kim, H.; Posner, I.; and Teh, Y. W. 2018. Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects. In *NeurIPS*, 8615–8625.

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

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.

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

Prémont-Schwarz, I.; Ilin, A.; Hao, T.; Rasmus, A.; Boney, R.; and Valpola, H. 2017. Recurrent Ladder Networks. In *NeurIPS*, 6009–6019.

Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; and Raiko, T. 2015. Semi-supervised learning with ladder networks. In *NeurIPS*, 3546–3554.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *ICML*, 1278–1286.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual learning with deep generative replay. In *NeurIPS*, 2990–2999.

Singh, K. K.; Ojha, U.; and Lee, Y. J. 2019. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 6490–6499.

Stelzner, K.; Peharz, R.; and Kersting, K. 2019. Faster attend-infer-repeat with tractable probabilistic models. In *ICML*, 5966–5975.

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

Ulutan, O.; Iftekhar, A.; and Manjunath, B. S. 2020. VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. In *CVPR*, 13617–13626.

van Steenkiste, S.; Chang, M.; Greff, K.; and Schmidhuber, J. 2018. Relational Neural Expectation Maximization: Unsupervised Discovery of Objects and their Interactions. In *ICLR*.

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

Yang, Y.; Chen, Y.; and Soatto, S. 2020. Learning to Manipulate Individual Objects in an Image. In *CVPR*, 6558–6567.

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*, volume 33, 9135–9142.

Zhao, S.; Song, J.; and Ermon, S. 2017. Learning hierarchical features from deep generative models. In *ICML*, 4091–4099.

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

Zitnick, C. L.; Parikh, D.; and Vanderwende, L. 2013. Learning the visual interpretation of sentences. In *ICCV*, 1681–1688.