Multi-object Generation with Amortized Structural Regularization

Kun Xu, Chongxuan Li, Jun Zhu*, Bo Zhang
Dept. of Comp. Sci. & Tech., Institute for AI, THBI Lab, BNRist Center,
State Key Lab for Intell. Tech. & Sys., Tsinghua University, Beijing, China
{kunxu.thu, chongxuanli1991}@gmail.com, {dcszj, dcszb}@tsinghua.edu.cn

Abstract

Deep generative models (DGMs) have shown promise in image generation. However, most of the existing methods learn a model by simply optimizing a divergence between the marginal distributions of the model and the data, and often fail to capture rich structures, such as attributes of objects and their relationships, in an image. Human knowledge is a crucial element to the success of DGMs to infer these structures, especially in unsupervised learning. In this paper, we propose amortized structural regularization (ASR), which adopts posterior regularization (PR) to embed human knowledge into DGMs via a set of structural constraints. We derive a lower bound of the regularized log-likelihood in PR and adopt the amortized inference technique to jointly optimize the generative model and an auxiliary recognition model for inference efficiently. Empirical results show that ASR outperforms the DGM baselines in terms of inference performance and sample quality.

1 Introduction

Deep generative models (DGMs) [19, 26, 10] have made significant progress in image generation, which largely promotes the downstream applications, especially in unsupervised learning [5, 7] and semi-supervised learning [20, 6]. In most of the real-world settings, visual data is often presented as a scene of multiple objects with complicated relationships among them. However, most of the existing methods [19, 10] lack of a mechanism to capture the underlying structures in images, including regularities (e.g., size, shape) of an object and the relationships among objects. This is because they adopt a single feature vector to represent the whole image and consequently focus on generating images with a single main object [17]. It largely impedes DGMs generalizing to complex scene images. How to solve the problem in an unsupervised manner is still largely open.

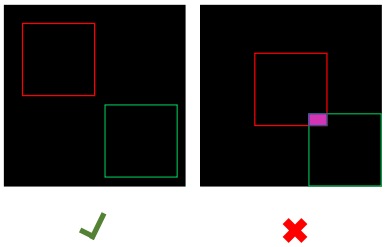


Figure 1: An illustration of the overlapping problem. The first bounding box is in red, and the second one is in green. The overlapping area is in purple. Defining the prior distribution in the auto-regressive manner is still challenging since some locations are not valid even for the first bounding box as shown in the right panel.

The key to address the problem is to model the structures explicitly. Existing work attempts to solve the problem via structured DGMs [8, 24], where a structured prior distribution over latent variables is used to encode the structural information of images and regularize the model behavior under the framework of maximum likelihood estimation (MLE). However, there are two potential limitations of such methods. First, merely maximizing data's log-likelihood of such models often fails to capture the

*Corresponding author.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

structures in an unsupervised manner [21]. Maximizing the marginal likelihood does not necessarily encourage the model to capture the reasonable structures because the latent structures are integrated out. Besides, the optimizing process often gets stuck in local optima because of the highly non-linear functions defined by neural networks, which may also result in undesirable behavior. Second, it is generally challenging to design a proper prior distribution which is both flexible and computationally tractable. Consider the case where we want to uniformly sample several 20×20 bounding boxes in a 50×50 image *without overlap*. It is difficult to define a tractable prior distribution, as shown in Fig.1. Though it is feasible to set the probability to zero when the prior knowledge is violated using indicator functions, other challenges like non-convexity and non-differentiability will be introduced to the optimization problem.

In contrast, the posterior regularization (PR) [9] and its generalized version in Bayesian inference, i.e., regularized Bayesian inference (RegBayes) [30], provide a general framework to embed human knowledge in generative models, which directly regularizes the posterior distribution instead of designing proper prior distributions. In PR and RegBayes, a valid posterior set is defined according to the human knowledge, and the KL-divergence between the true posterior and the valid set (see the formal definition in Sec. 2.2) is minimized to regularize the behavior of structured DGMs. However, the valid set consists of sample-specific variational distributions. Therefore, the number of parameters in the variational distribution grows linearly with the number of training samples, and it requires an inner loop for accurately approximating the regularized posterior [28]. The above computational issue makes it non-trivial to apply PR to large-scale datasets and DGMs directly.

In this paper, we propose a flexible amortized structural regularization (ASR) framework to improve the performance of structured generative models based on PR. ASR is a general framework to properly incorporate structural knowledge into DGMs by extending PR to the amortized setting, and its objective function is denoted as the log-likelihood of the training data along with a regularization term over the posterior distribution. The regularization term can help the model to capture reasonable structures of an image, and to avoid unsatisfactory behavior that violates the constraints. We derive a lower bound of the regularized log-likelihood and use an amortized recognition model to approximate the constrained posterior distribution. By slacking the constraints as a penalty term, ASR can be optimized efficiently using gradient-based methods. We apply ASR to the state-of-the-art structured generative models [8] for the multi-object image generation tasks. Empirical results demonstrate the effectiveness of our proposed method, and both the inference and generative performance are improved under the help of human knowledge.

2 Preliminary

2.1 Iterative generative models for multiple objects

Attend-Infer-Repeat (AIR) [8] is a structured latent variable model, which decomposes an image as several objects. The attributes of objects (i.e., appearance, location, and scale) are represented by a set of random variables $z = \{z_{app}, z_{loc}, z_{scale}\}$. The generative process starts from sampling the number of objects $n \sim p(n)$, and then n sets of latent variables are sampled independently as $z^i \sim p(z)$. The final image is composed by adding these objects into an empty canvas. Specifically, the joint distribution and its marginal over the observed data can be formulated as follows:

$$p(x, z, n) = p(n) \prod_{i=1:n} p(z^i) p(x|z, n), \quad p(x) = \sum_n \int_z p(x, z, n) dz.$$

The conditional distribution $p(x|z, n)$ is usually formulated as a multi-variant Gaussian distribution with mean $\mu = \sum_{i=1:n} f_{dec}(z^i)$, or a Bernoulli distribution with probability $p = \sum_{i=1:n} f_{dec}(z^i)$ for pixels in images. f_{dec} is a decoder network that transfers the latent variables to the image space.

In an unsupervised manner, AIR can infer the number of objects, as well as the latent variables for each object efficiently using amortized variational inference. The latent variables are inferred iteratively, and the number of objects n is represented by z_{pres}: a $n + 1$ binary dimensional vector with n ones followed by a zero. The i-th element of z_{pres} denotes whether the inference process is terminated or not. Then the inference model can be formulated as follows:

$$q(z, n|x) = q(z^{n+1}_{pres} = 0|x, z^{<n}) \prod_{i=1:n} q(z^i|x, z^{<i}) q(z^i_{pres} = 1|x, z^{<i}). \tag{1}$$

The inference model iteratively infers the latent variable z^i of the i-th object conditioning on the previous inferred latent variables $z^{<i}$ and the input image x until $z^{n+1}_{pres} = 0$.

By explicitly modeling the location and appearance of each object, AIR is capable of modeling an image with structural information, rather than a simple feature vector. It is worth noting that the number of steps n and latent variable z^i are pre-defined and cannot be learned from data. In the following, we modify the original AIR by introducing a parametric prior to capture the dependency among objects. Details are illustrated in Sec. 3.1.

2.2 Posterior regularization for structured generative model

Posterior regularization (PR) [9, 30] provides a principled approach to regularize latent variable models with a set of structural constraints. There are some cases where designing a prior distribution for the prior knowledge is intractable, whereas they can be easily presented as a set of constraints [30]. In these cases, PR is more flexible than designing proper prior distributions.

Specifically, a latent variable model is denoted as $p(X, Z; \theta) = p(Z; \theta)p(X|Z; \theta)$ where X is the training data and Z is the corresponding latent variable. θ denotes the parameters of p, and takes value from Θ, which is generally $\mathbb{R}^{|\Theta|}$ with $|\Theta|$ denotes the dimension of the parameter space. PR proposes to regularize the posterior distribution to certain constraints under the framework of maximization likelihood estimation (MLE). Generally, the constraints are defined as the expectation of certain statistics $\psi(X, Z) \in \mathbb{R}^d$, and they form a set of valid posterior distribution Q as follows:

$$Q = \{q(Z)|\mathbb{E}_{q(Z)}[\psi(X, Z)] \leq \mathbf{0}\}, \tag{2}$$

where d is the number of constraints, and $\mathbf{0}$ is a d-dimension zero vector. To regularize the posterior distribution $P(Z|X; \theta)$ to be close to Q, PR proposes to add a regularization term $\Omega(p(Z|X; \theta))$ to the MLE objective. The optimization problem and regularization are given by:

$$\max_{\theta} J(\theta) = \log \int_Z p(X, Z; \theta)dZ - \Omega(p(Z|X; \theta)). \tag{3}$$

$$\Omega(p(Z|X; \theta)) = KL(Q||p(Z|X; \theta)) = \min_{q \in Q} KL(q(Z)||p(Z|X; \theta)). \tag{4}$$

The regularization term is the KL divergence between Q and $p(Z|X; \theta)$ as defined in Eqn. (4). When the regularization term is convex, a close-form solution can be found based on convex analysis [3]. Therefore, the EM algorithm [28] can be applied to optimize the regularized likelihood $J(\theta)$ [9]. However, EM is largely limited when we extend the PR to DGMs because of the highly non-linearity introduced by neural networks. We therefore propose our method by introducing amortized variational inference to efficiently solve the problem.

3 Method

In this section, we first define a variant of AIR which uses a parametric prior distribution to capture the dependency among objects. Then we give a formal definition of the amortized structural regularization (ASR) framework. We mainly follow the notation in Sec. 2, and we abuse the notation when they share the same role in PR and ASR. We illustrate our proposed framework in Fig. 2.

3.1 Generative & inference model

The prior distribution in the vanilla AIR is fixed, and the latent variables of objects are sampled independently. Therefore, the structures, i.e., the attributes and their dependency, cannot be captured by the generative model. We propose to modify the generative model by using a learnable prior. Specifically, an auxiliary variable z_{pres} is used to model the number of objects by denoting whether the generation process is terminated at step t (i.e., $z^t_{pres} = 0$) or not (i.e., $z^t_{pres} = 1$). Besides, the attributes (i.e., latent variables of each object) are sampled conditioned on previously sampled latent variables. Formally, the joint distribution is defined as follows:

$$p(x, z, n; \theta) = p(z^{n+1}_{pres} = 0|z^{\leq n}; \theta) \left(\prod_{t=1}^{n} p(z^t_{pres} = 1|z^{<t}; \theta)p(z^t|z^{<t}; \theta) \right) p(x|z, z_{pres}; \theta), \tag{5}$$

where the θ denotes the parameters for both the prior distribution and conditional distribution and we set $z^0_{pres} = 1$ and $z^0 = \mathbf{0}$. In the following, we omit the θ for simplicity. Following AIR, the

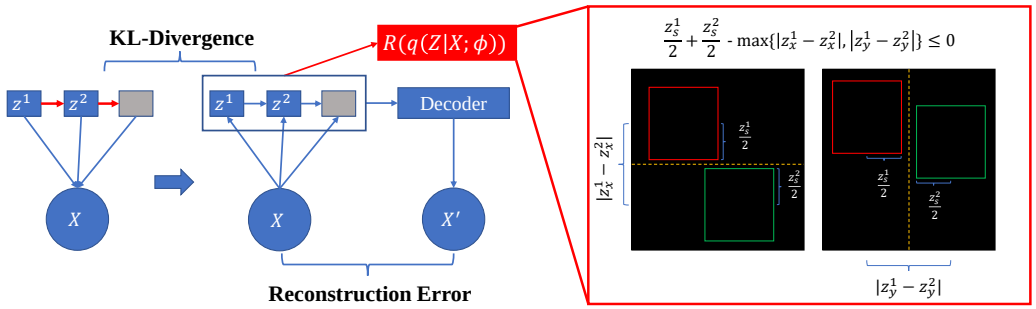


Figure 2: The proposed framework. The blue arrows denote the generative and inference network in AIR. The red arrows highlight the difference between ASR and AIR. The red arrows in the generative model represent the dependency among the latent variables in the generative model. A regularization term is introduced to regularize the generative model, and we use the overlapping term as an example.

conditional distribution $p(x|z, z_{pres})$ is defined as $p(x|z, z_{pres}) = p(x|\sum_{i=1:n} f_{dec}(z^i))$. We use a recurrent neural network (RNN) [11] to model the dependency among the latent variables z, z_{pres}, and use a feed-forward neural network as the decoder to map the latent variables to the image space.

The latent variable z consists of three parts: $z = \{z_{app}, z_{loc}, z_{scale}\}$, which represent the appearances, locations, and scales respectively. The distribution of z^t conditioned on previous $z^{<t}$ is given by:

$$p(z^t|z^{<t};\theta) = p(z^t_{loc}|z^{<t})p(z^t_{scale}|z^{<t}, z^t_{loc})p(z_{app}),$$

where the current scale and location are sampled conditionally on previous sampled results. Since we only consider the spatial relation among the objects, the dependency among the appearances of them is ignored and the appearance variables are independently sampled from a simple prior distribution.

The inference model is defined mainly following the AIR, which is given by:

$$q(z, n|x;\phi) = q(z^{n+1}_{pres} = 0|z^{\leq n}, x;\phi)\prod_{t=1}^{n} q(z^i_{pres} = 1|z^{<t}, x;\phi)q(z^t|z^{<t}, x;\phi), \tag{6}$$

where the $\phi \in \Phi$ denotes the parameters and Φ denotes the parameter space of ϕ. Similar to the generative process, the variational posterior distribution $q(z^t|z^{<t}, x;\phi)$ is given by:

$$q(z^t|z^{<t}, x) = q(z^t_{loc}|z^{<t})q(z^t_{scale}|z^{<t}, z^t_{loc})q(z_{app}|z^t_{loc}, z^t_{scale}).$$

The generative model defined in Eqn. (5) is powerful enough to capture complex structures. However, directly optimizing the marginal log-likelihood (or its lower bound) of training data often stacks at certain local optima, where the model fails to capture the structures. This phenomenon emerges in the baselines as reported in both previous work [21] and our experiments. See details in Sec. 6.1.

3.2 Amortized structural regularization

In original PR, a set of statistics ψ is used to define the valid set Q in Eqn. (2). In ASR, we generalize the constraints as a functional F that maps a distribution defined over the latent space to $\mathbb{R}^d$, with d denoting the number of constraints. The resulted valid set Q is given by:

$$Q = \{q(Z)|F(q(Z)) \leq \mathbf{0}\}, \tag{7}$$

where $\mathbf{0}$ is a d-dimension zero-vector. To train the DGMs using gradient-based methods efficiently, we require that the functional F is differentiable w.r.t. q.

Motivated by PR, ASR regularizes the posterior distribution $P(Z|X;\theta)$ to be close to the valid set Q, by minimizing a regularization term $\Omega(p(Z|X;\theta))$ along with maximizing the likelihood of training data. The objective function is given by:

$$\max_{\theta} J(\theta) = \log \int_Z p(X, Z;\theta)dZ - \Omega(p(Z|X;\theta)). \tag{8}$$

The definition of the regularization term Ω follows original PR as in Eqn. (4). Note that $KL(q(Z)||p(Z|X;\theta)) \geq \Omega(p(Z|X;\theta))$ for all $q(Z) \in Q$. It enables us to obtain a lower bound of $J(\theta)$ by substituting $KL(q(Z)||p(Z|X;\theta))$ for $\Omega(p(Z|X;\theta))$, which is given by:

$$J(\theta) \geq \log \int_Z p(X, Z; \theta) dZ - KL(q(Z)||p(Z|X;\theta)) = J'(\theta, q). \tag{9}$$

Following the variational inference, the lower bound J' can be formulated as the evidence lower bound (ELBO), and Problem (8) is converted as a constrained optimization problem as follows:

$$\max_{\theta, q(Z)\in Q} J'(\theta, q) = \log p(X) - \mathbb{E}_{q(Z)} \log \frac{q(Z)}{p(Z|X;\theta)} = \mathbb{E}_{q(Z)} \log \frac{p(X, Z; \theta)}{q(Z)}.$$

Motivated by amortized variational inference [19], we introduce a recognition model $q(Z|X;\phi)$ to approximate the variational distribution q where ϕ denotes the parameters of the recognition model. Therefore, the lower bound can be optimized w.r.t. θ and ϕ jointly, which is given by:

$$\max_{\theta\in\Theta, \phi\in\Phi, q(Z|X;\phi)\in Q} \mathbb{E}_{q(Z|X;\phi)} \log \frac{p(X, Z; \theta)}{q(Z|X;\phi)}. \tag{10}$$

We abuse the notation $J'(\theta, \phi)$ to denote the amortized version of the lower bound.

Problem (10) is a constrained optimization problem. In order to efficiently solve Problem (10), we propose to slack the constraints as a penalty, and add it to the objective function $J'(\theta, \phi)$ as:

$$\max_{\theta\in\Theta, \phi\in\Phi} J'(\theta, \phi) - R(q(Z|X;\phi)), \tag{11}$$

where $R(q) = \sum_{i=1:d} \lambda_i \max\{F_i(q), 0\}$, and λ_i is the coefficient for the i-th constraint of $F(q)$. For sufficient large λ, Problem (11) is equivalent to Problem (10) and we treat it as a hyperparameter. The training procedure is described in Appendix A.

It is worth noting that we implicitly add another regularization to the generative model when defining q using a parametric model: the posterior distribution $p(Z|X;\theta)$ can be represented by $q(Z|X;\phi)$. This regularization term has the same effect as in VAE [19, 27], which is introduced to make the optimization process more efficient. In contrast, it is the penalty term $R(q(Z|X;\phi))$ that embeds human knowledge into DGMs and regularizes DGMs for desirable behavior.

4 Application on multi-object generation

In the following, we give two examples of applying ASR to image generation with multiple objects. In this section, we mainly focus on regularizing on the number of objects, and the spatial relationships among them. Therefore, the functional F in Eqn. (7) are defined over $q(z_{pres}, z_{loc}, z_{scale})$.

4.1 ASR regularization on the number of objects

In this setting, we consider the case where each image contains a certain number of objects. For example, each image has either 2 or 4 objects, and images of each number of objects appear of the same frequency. We define the possible numbers of objects as $L \subsetneq [K]$, where $[K] = \{0, 1, \cdots, K-1\}$ is the set of all non-negative integer less than K, and K is the largest number of objects we consider. Since we use z_{pres} to denote the number of objects, an image x with n objects is equivalent to the corresponding latent variable $z_{pres}|x = u_n$ with probability one, where u_n is a $n+1$ dimension binary vector with n ones followed by a zero. We further denote q_i as $q_i(z_{pres} = u_j) = \mathbb{1}(i = j)$, where $\mathbb{1}$ is the indicator function. The valid posterior is given by $V_{z_{pres}} = \{q_i\}_{i\in L}$. According to ASR, we regularize our variational posterior $q(Z|X;\phi)$ in the valid posterior set $V_{z_{pres}}$. Besides, we also regularize the marginal distribution to $q_{uni}(z) = \frac{1}{|L|}\sum_{i\in L} q_i$, which is a uniform distribution over $V_{z_{pres}}$. The valid posterior set is given by:

$$Q^{num} = \{q(Z|X)|q(Z|X = x) \in V_{z_{pres}} \ \forall\, x \in \mathcal{D}, \mathbb{E}_{p(X)} q(Z|X) = q_{uni}(Z)\},$$

where $\mathcal{D}$ denotes the set of all training samples. As the constraints are defined in the equality form, and we reformulate it in the inequality form, and the regularization term R_{num} are given by:

$$Q^{num} = \{q(Z|X)|\min_{q_i\in V_{z_{pres}}} KL(q_i||q(Z|X)) \leq 0, KL(q_{uni}(Z)||\mathbb{E}_{p(X)} q(Z|X)) \leq 0\},$$

$$R^{num}(q(Z|X)) = \lambda_1^{num} \min_{q_i\in Q_{num}} KL(q_i||q(Z|X)) + \lambda_2^{num} KL(q_u(Z)||\mathbb{E}_{p(X)} q(Z|X)).$$

The λ_1^{num} and λ_2^{num} are the hyper-parameters to balance the penalty term and the log-likelihood.

4.2 ASR regularization on overlap

In this setting, we focus on the overlap problem, and we introduce several regularization terms to reduce the overlap among objects, which is defined over the location of bounding boxes. The location of a bounding box is determined by its center $z_{loc} = (z_x, z_y)$, and scale z_{scale}, and the functional F^o is defined over these latent variables.

The first set of regularization terms directly penalize the overlap. Given the centers and scales of the i-th and j-th bounding box, they are not overlapped if and only if both of the following constraints are satisfied: $\frac{z^i_{scale}+z^j_{scale}}{2} - |z^i_x - z^j_x| \leq 0$, $\frac{z^i_{scale}+z^j_{scale}}{2} - |z^i_y - z^j_y| \leq 0$. These constraints have a straightforward explanation and are illustrated in Fig. 2.

In the following, we denote $\ell(x) = \max\{x, 0\}$ for simplicity, and we define the functional F^o as:

$$F^o_1(q) = \mathbb{E}_{q(z)} \sum_{i,j<n, i\neq j} \ell(\frac{z^i_{scale} + z^j_{scale}}{2} - \max\{|z^i_x - z^j_x|, |z^i_y - z^j_y|\}) \leq 0,$$

which regularizes each pair of the bounding boxes to reduce overlap.

Simply regularizing the overlap by minimizing F_1 usually results in the fact that the inferred bounding boxes are of different size: a big bounding box that covers the whole image, and several bounding boxes of extremely small size that lie beside the boundary of the image, or out of the image. To overcome this issue, we add another two regularization terms, where the first one regularize the bounding boxes stay within the image, and the second regularize the bounding boxes are of the same size. The first set of regularization terms are formulated as the following four constraints:

$$F^o_2(q) = \mathbb{E}_{q(z)} \sum_{i=1:n} \ell(\frac{z^i_{scale}}{2} - z^i_x) \leq 0, \;\; F^o_3(q) = \mathbb{E}_{q(z)} \sum_{i=1:n} \ell(z^i_x + \frac{z^i_{scale}}{2} - S) \leq 0,$$

$$F^o_4(q) = \mathbb{E}_{q(z)} \sum_{i=1:n} \ell(\frac{z^i_{scale}}{2} - z^i_y) \leq 0, \;\; F^o_5(q) = \mathbb{E}_{q(z)} \sum_{i=1:n} \ell(z^i_y + \frac{z^i_{scale}}{2} - S) \leq 0,$$

and the second set of regularization terms are given by:

$$F^o_6(q) = \mathbb{E}_{q(z)} \sum_{i=1:n} \ell(c_{min} - z^i_{scale}) + \ell(z^i_{scale} - c_{max}) \leq 0,$$

$$F^o_7(q) = \mathbb{E}_{q(z)} \sum_{i,j<n} \ell(|z^i_{scale} - z^j_{scale}| - \epsilon) \leq 0,$$

where S denotes the size of the final image, c_{min}/c_{max} denotes the possible minimum/maximum size of an object, and ϵ denotes the perturbation of the size for objects. Therefore, the regularization for reducing overlap is given by:

$$R^o(q) = \sum_{i=1:7} \lambda^o_i F^o_i(q). \tag{12}$$

5 Related work

Recently, several methods [8, 12, 16, 29, 24] introduce structural information to deep generative models. Eslami et al. [8] propose the Attend-Infer-Repeat (AIR), which defines an iterative generative process to compose an image with multiple objects. Greff et al. [12] further generalize this method to more complicated images, by jointly modeling the background and objects using masks. Li et al. [24] use graphical networks to model the latent structures of an image, and generalize probabilistic graphical models to the context of implicit generative models. Johnson et al. [16] introduce the scene graph as conditional information to generate scene images. Xu et al. [29] use the and-or graph to model the latent structures and use a refinement network to map the structures to the image space.

To embed prior knowledge into structured generative models, posterior regularization (PR) [9] provides a flexible framework to regularize model w.r.t. a set of structural constraints. Zhu et al. [30] generalize this framework to the Bayesian inference and apply it in the non-parametric setting. Shu et al. [27] introduce to regularize the smoothness of the inference model to improve the generalization on both inference and generation and refer it as amortized inference regularization. Li et al. [23] propose to regularize the latent space of a latent variable model with large-margin in the context of amortized variational inference, which can also be considered as a special case of PR. Bilen et al. [2] apply PR to the object detection in a discriminative manner and improve the detection accuracy.

(a) The reconstruction of AIR-13. (b) The reconstruction of AIR-pPrior-13. (c) The reconstruction of AIR-ASR-13.

Figure 3: The reconstruction results of Multi-MNIST on 1 or 3 objects.

6 Experiments

In this section, we present the empirical results of ASR on two dataset: Multi-MNIST [8] and Multi-Sprites [12], which are the multi-object version of MNIST [22] and dSprites [13]. We use AIR-pPrior to denote the variants of AIR proposed in this paper, and AIR-ASR to denote the regularized AIR-pPrior using ASR.

We implement our model using TenworFlow [1] library. In our experiments, the RNNs in both the generative model and recognition model are LSTM [14] with 256 hidden units. A variational auto-encoder [19] is used to encode and decode the appearance latent variables, and both the encoder and decoder are implemented as a two-layer MLP with 512 and 256 units. We use the Adam optimizer [18] with learning rate as 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We train models with 300 epochs with batch size as 64. Our code is attached in the supplementary materials for reproducing.

In this paper, we use four metrics for quantitative evaluation: negative ELBO (nELBO), squared error (SE), inference accuracy (ACC) and mean intersection over union (mIoU). The nELBO is an upper bound of negative log-likelihood, where a lower value indicates a better approximation of data distribution. The SE is the squared error between the original image and its reconstruction, and it is summed over pixels. The ACC is defined as $\mathbb{1}(num_{inf} == num_{gt})$, num_{inf} and num_{gt} are the number of objects inferred by the recognition model and ground truth respectively. This evaluation metric demonstrates whether the inference model can correctly infer the exact number of objects in an image. Besides, we also use another evaluation metric mIoU to evaluate the accuracy of inferred location for each objects. The mIoU of a single image is defined as $\max_\pi \sum_{i=1:min\{num_{inf},num_{gt}\}} IoU(z^{\pi_i}, gt^i)/\max\{num_{inf}, num_{gt}\}$, where π is a permutation of $\{1, 2, \cdots, num_{inf}\}$ and gt^i is the ground truth location for the i-th object.

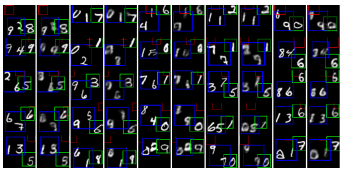
Table 1: Results on regularization on the number of objects. The numbers followed the model name denotes the possible number of objects for a certain image. Results are averaged over 3 runs.

Methods	nELBO	ACC	SE	mIoU
AIR-13	404.41 ± 4.58	0.81 ± 0.23	31.94 ± 4.68	$\mathbf{0.61} \pm 0.13$
AIR-pPrior-13	405.21 ± 1.17	0.48 ± 0.00	49.42 ± 0.24	0.43 ± 0.01
AIR-ASR-13	$\mathbf{360.20} \pm 19.67$	$\mathbf{0.96} \pm 0.00$	$\mathbf{28.84} \pm 1.11$	$\mathbf{0.61} \pm 0.00$
AIR-14	543.44 ± 54.71	0.48 ± 0.03	52.77 ± 4.92	0.43 ± 0.07
AIR-pPrior-14	519.06 ± 5.47	0.50 ± 0.00	68.72 ± 0.55	0.43 ± 0.00
AIR-ASR-14	$\mathbf{441.54} \pm 30.97$	$\mathbf{0.96} \pm 0.01$	$\mathbf{41.05} \pm 7.11$	$\mathbf{0.55} \pm 0.08$
AIR-24	639.49 ± 23.13	0.55 ± 0.09	57.69 ± 4.88	0.46 ± 0.06
AIR-pPrior-24	643.28 ± 8.67	0.00 ± 0.00	83.35 ± 0.44	0.10 ± 0.00
AIR-ASR-24	$\mathbf{495.73} \pm 35.80$	$\mathbf{0.98} \pm 0.01$	$\mathbf{48.54} \pm 5.60$	$\mathbf{0.54} \pm 0.08$

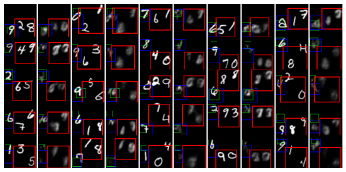
6.1 ASR regularization on the number of objects

When regularizing on the number of objects, we consider three settings on Multi-MNIST: 1 or 3 objects, 1 or 4 objects, and 2 or 4 objects. 40000 training samples are synthesized where 20000 images for each number of objects. 2000 images are used as the test data to evaluate the performance for inference. In this setting, we evaluate our methods with $\lambda_1^{num}, \lambda_2^{num} \in \{1, 10, 100\}$, and we finally set $\lambda_1^{num} = 10$ and $\lambda_2^{num} = 100$.

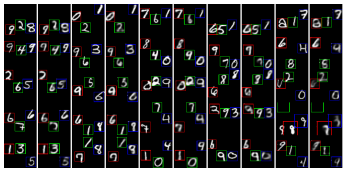
As illustrated in Fig. 3, AIR-pPrior simply treats the whole image as a single object, and fails to identify the objects in an image. With a powerful decoder network, the generative model tends to ignore the latent structures. The ASR can successfully regularize the model towards proper behavior.



(a) The reconstruction of AIR-3.

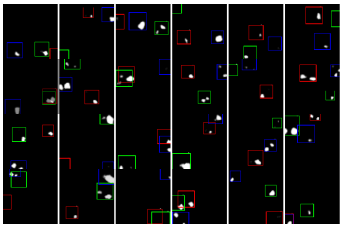


(b) The reconstruction of AIR-pPrior-3.

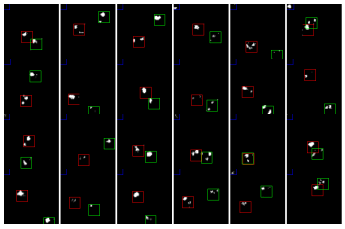


(c) The reconstruction of AIR-ASR-3.

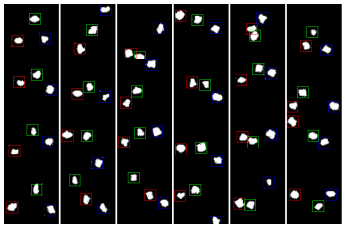
Figure 4: The reconstruction results of Multi-MNIST on 3 objects. There is no overlap among objects in the training data. ASR can successfully infer the underlying structures, and improve the reconstruction results.



(a) The generative results of AIR-3.



(b) The generative results of AIR-pPrior-3.



(c) The generative results of AIR-ASR-3.

Figure 5: The generative results of Multi-dSprites on 3 objects without overlap.

In AIR-ASR, the inference model can successfully identify each object, and the generative model learns the underlying structures. The original AIR has a better performance compared to AIR-pPrior, as the prior distribution can partly regularize the generative model. However, the original AIR still treats two objects close to each other as one object. The performance of these three models on the other two settings shares the same property, i.e., original AIR tends to merge objects and AIR-pPrior stacks at a local optimum. The other reconstruct results are illustrated in the Appendix.

Table 1 presents the quantitative results. AIR-ASR outperforms its baseline and the original AIR on all the evaluation metrics, which demonstrates the effectiveness of our proposed method. Specifically, ASR can significantly regularize the model in terms of the inference steps and achieves the accuracy up to 96% for all the three settings. It is worth noting that introducing a proper regularization will not affect the ELBO which is the objective function of AIR and AIR-pPrior. The main reason is that ASR can encourage the model to avoid the unsatisfactory behavior which violate the structural constraints.

During the training process, all of the three models suffer from sever instability. It results the fact that the nELBO is of large variance. The results largely depend on the initialization and the randomness in the training process. We try to reduce the effect of randomness by fixing the initialization and averaging our results over multiple runs.

Table 2: Experimental Results on regularization over overlap. Results are averaged over 3 runs.

	multi-MNIST			**multi-dSprites**		
Methods	nELBO	SE	mIoU	nELBO	SE	mIoU
AIR	328.5 ± 17.1	37.5 ± 3.8	0.25 ± 0.03	341.5 ± 76.5	34.8 ± 8.9	0.13 ± 0.05
AIR-pPrior	$\mathbf{306.6} \pm 58.8$	41.5 ± 15.4	0.35 ± 0.10	274.3 ± 64.4	29.3 ± 12.1	0.21 ± 0.13
AIR-ASR	337.3 ± 55.1	$\mathbf{36.5} \pm 3.9$	$\mathbf{0.67} \pm 0.05$	$\mathbf{271.8} \pm 18.8$	$\mathbf{20.9} \pm 2.1$	$\mathbf{0.61} \pm 0.03$

6.2 ASR regularization on the overlap

When regularizing the overlap, we evaluate models on both Multi-MNIST and Multi-dSprites data. We use 20000 images with three non-overlapping objects as training data and use 1000 images to evaluate performance. Since the number of objects is fixed, we simply set both the generative and inference steps to 3 for fair comparison. We search the hyper-parameters $\lambda^o_{i=1:7}$ in $\{1, 10, 20, 100\}$, and we set $\lambda^o_1 \sim \lambda^o_5$ to 1, λ^o_6 to 20, and λ^o_7 to 10.

The reconstruction of Multi-MNIST and generative results of Multi-Sprites are demonstrated in Fig. 4 and Fig. 5 correspondingly. In Fig. 4, the original AIR still merges two objects as one, and it cannot

capture the non-overlapping structures. AIR-pPrior has a similar performance. In contrast, AIR-ASR significantly outperforms its baselines, and infers the location of bounding boxes without overlap. In terms of generative results, the sample quality of AIR-ASR surpasses AIR's and AIR-pPrior's, where the AIR-ASR can generate multiple objects without overlap whereas its baseline cannot. It demonstrates that the ASR can embed human knowledge into DGMs.

Table 2 presents the quantitative results. The AIR-ASR surpasses its baselines significantly in terms of mIoU, which indicates that DGMs successfully captures the non-overlapping structures with ASR. It is worth noting that for the Multi-MNIST setting, the nELBO of AIR-pPrior is better than AIR-ASR's. However, AIR-ASR still surpasses AIR-pPrior in terms of the SE and the mIoU, which indicates that AIR-ASR gives better reconstruction results and identifies the location of objects more accurately. This results also verify the claim that simply optimizing the marginal log-likelihood cannot guarantee the generative model to capture the underlying distribution.

7 Conclusion

We present a framework ASR to embed human knowledge to improve the inference and generative performance in structured DGMs for multi-object generation. ASR encodes human knowledge as a set of structural constraints, and the framework can be optimized efficiently. We use the number of objects and the spatial relationships among them as two examples to demonstrate the effectiveness of our proposed method. In Multi-MNIST and Multi-dSprites datasets, ASR significantly improves its baselines and successfully captures the underlying structures of the training data.

ASR is a general framework to properly incorporate structural knowledge into DGMs as long as the knowledge can be quantitatively represented and can be applied to a wide range of structured DGMs. In this paper, we only consider the cases with hard constraints on synthetic datasets. For one thing, it is shown that PR can be extended to "selectively" incorporate uncertain knowledge (e.g., with noise) represented by the general language of first-order logic [25], where highly uncertain knowledge will be dropped according to the faithfulness of fitting the given data. Further, Hu et al. [15] extend PR to the learnable constraints using differentiable neural networks. ASR extends PR to an amortized version for structured generation, thereby inheriting the generality in a principled manner. For another, recently significant progress has been made in structured generative models [12, 4] for more realistic multi-object images. Together with the theoretical generality and the practical progress, ASR can be applied to more complicated applications and we leave it as future work.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2017YFA0700904), NSFC Projects (Nos. 61620106010, 61621136008), Beijing NSF Project (No. L172037), Beijing Academy of Artificial Intelligence (BAAI), Tiangong Institute for Intelligent Computing, the JP Morgan Faculty Research Program and the NVIDIA NVAIL Program with GPU/DGX Acceleration. C. Li was supported by the Chinese postdoctoral innovative talent support program and Shuimu Tsinghua Scholar.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference*, volume 3, 2014.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[6] LI Chongxuan, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *Advances in neural information processing systems*, pages 4088–4098, 2017.

[7] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[8] SM Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems*, pages 3225–3233, 2016.

[9] Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(Jul):2001–2049, 2010.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[12] Klaus Greff, Raphaël Lopez Kaufmann, Rishab Kabra, Nick Watters, Chris Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. *arXiv preprint arXiv:1903.00450*, 2019.

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[15] Zhiting Hu, Zichao Yang, Ruslan R Salakhutdinov, LIANHUI Qin, Xiaodan Liang, Haoye Dong, and Eric P Xing. Deep generative models with learnable knowledge constraints. In *Advances in Neural Information Processing Systems*, pages 10501–10512, 2018.

[16] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[21] Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems*, pages 8606–8616, 2018.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Chongxuan Li, Jun Zhu, Tianlin Shi, and Bo Zhang. Max-margin deep generative models. In *Advances in neural information processing systems*, pages 1837–1845, 2015.

[24] Chongxuan Li, Max Welling, Jun Zhu, and Bo Zhang. Graphical generative adversarial networks. *arXiv preprint arXiv:1804.03429*, 2018.

[25] Shike Mei, Jun Zhu, and Jerry Zhu. Robust regbayes: Selectively incorporating first-order logic domain knowledge into bayesian models. In *International Conference on Machine Learning*, pages 253–261, 2014.

[26] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[27] Rui Shu, Hung H Bui, Shengjia Zhao, Mykel J Kochenderfer, and Stefano Ermon. Amortized inference regularization. In *Advances in Neural Information Processing Systems*, pages 4393–4402, 2018.

[28] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[29] Kun Xu, Haoyu Liang, Jun Zhu, Hang Su, and Bo Zhang. Deep structured generative models. *arXiv preprint arXiv:1807.03877*, 2018.

[30] Jun Zhu, Ning Chen, and Eric P Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1):1799–1847, 2014.