Duplicate Latent Representation Suppression for Multi-object Variational Autoencoders

Li Nanbo
nanbo.li@ed.ac.uk
Robert B. Fisher
rbf@inf.ed.ac.uk

School of Informatics
University of Edinburgh
Edinburgh, UK

Abstract

Generative object-centric scene representation learning is crucial for structural visual scene understanding. Built upon variational autoencoders (VAEs) [11], current approaches infer a set of latent object representations to interpret a scene observation (e.g. an image) under the assumption that each part (e.g. a pixel) of a scene observation must be explained by one and only one object of the underlying scene. Despite the impressive performance these models achieved in unsupervised scene factorization and representation learning, we show empirically that they often produce duplicate scene object representations which directly harms the scene factorization performance. In this paper, we address the issue by introducing a differentiable prior that explicitly forces the inference to suppress duplicate latent object representations. The extension is evaluated by adding it to three different unsupervised scene factorization approaches. The results show that the models trained with the proposed method not only outperform the original models in scene factorization and have fewer duplicate representations, but also achieve better variational posterior approximations than the original models.

1 Introduction

Variational autoencoders (VAEs) [11] have become a powerful tool for unsupervised visual scene understanding and representation learning. As a particular type of generative model, a VAE model not only inherits the ability to explain scene observations (e.g. images) by learning a distribution $p(x;\theta)$ over the observation data $x \in \mathbb{R}^M$ but also it allows to describe and represent the observed scenes in a more compact latent space $\mathbf{z} \in \mathbb{R}^D$ $(D \ll M)$ for simplicity and efficiency. A rising trend in VAE research is to treat a multi-object scene as a composition of scene objects (aka scene components), i.e. a scene representation $\mathbf{z}$ is a set of K scene object representations $\mathbf{z} = \{z_k\}$, where each z_k corresponds to one and only one object in the scene. These object-based scene representation learning models are often referred to as the *multi-object VAEs*, they are called *component VAEs* (abbr. CompVAEs) in this paper for simplicity.

By making an assumption that each pixel of a scene image observation x must be explained by one and only one object in the scene, recent CompVAE advances [3, 7, 17] show great success in unsupervised image segmentation and object-based representation learning. In these models, this assumption acts as a constraint to force different objects $z_k \in \{z_k\}$ to explain different pixels of an image observation x, which implicitly assumes that the inferred $z \in \{z_k\}$ must be different from each other. However, all three CompVAEs investigated here (i.e. MONET [3], IODINE [7], MulMON [17]) can infer duplicate latent object representations (see Figure 1 for an example), which violates the implicit assumption and thus harms their performance in scene factorization (or image segmentation).

In this paper, we refer to the issues raised by inferring duplicate latent representations as the *uniqueness issues* and the implicit assumption of $\forall z \in \{z_k\}$ being unique as the *uniqueness assumption*. To address the *uniqueness issues*, we propose a differentiable prior, namely for *latent duplicate suppression* (abbr. *LDS*), to train CompVAEs' inference network to suppress duplicates while making inference at test time. The *LDS* prior essentially implements the *uniqueness assumption* — two identical z_k cannot appear in the same scene representation set $\mathbf{z} = \{z_k\}$, i.e. latent object representation pairs that are highly similar will be penalized during training.

In our experiments, we train two representative single-view CompVAEs, i.e. MONET [3] and IODINE [7], and one multi-view CompVAE, i.e. MulMON [17], with *LDS* as the experimental group and train the same models without *LDS* as the control group. We show the effectiveness of training CompVAEs' with *LDS* in suppressing scene factorization duplicates and achieving better variational approximation by comparing the performance of the two groups of models. We claim and demonstrate: **1)** Training a CompVAE with the proposed *LDS* prior enables the CompVAE to produce better scene factorizations with fewer duplicate objects (see Section 4.2). **2)** Training a CompVAE with the proposed *LDS* prior enables the CompVAE to achieve better variational posterior approximation, i.e. decrease the *inference gap* [4] (see Section 4.3). **3)** With better variational approximation achieved, CompVAEs trained with the proposed *LDS* overcome local minima better and thus learn better scene representations that supports better scene observation reconstructions (see Section 4.2).

2 Related Work

Our work lies in the research area of unsupervised scene factorization and representation learning. Earlier works in this area like the Attend-Infer-Repeat (AIR) model [6] and its variants [9, 12] perform object-centric scene factorization by sequentially searching for one object at a time in the image plane until all objects in the image are captured. As these models do not target a 3D understanding of a scene, they cannot resolve occlusions and handle images with complex backgrounds. The problem is overcome by recent advances [3, 5, 7, 17] that the pixel-level compositions of scene objects, i.e. each pixel needs to be explained by one and only one scene component. This line of work is referred to as the *scene-mixture* models by [15] as they all use the spatial mixture models [8, 21] to explain the image observations of scenes. This allows the models to reason about depth and occlusions which are essential for 3D understanding.

Our work is also related to relational reasoning works that are built upon CompVAEs. We discuss them in two categories: implicit and explicit relational reasoning. Although aforementioned works such as [5, 7, 17] do not explicitly reason about relationships, the discovery of scene objects suggests mutual dependence of each other. **These models violate the implicitly-introduced *uniqueness assumption* and thus cannot suppress duplicate object**

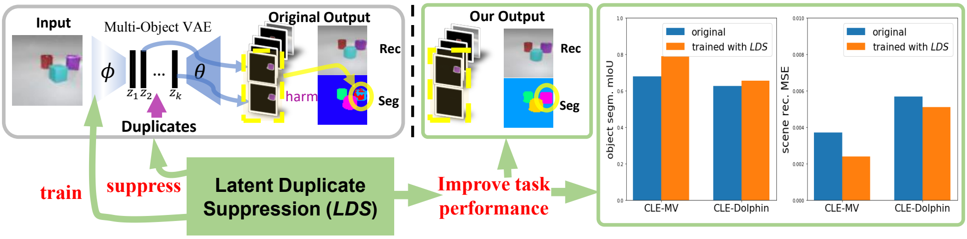


Figure 1: The state-of-the-art unsupervised scene factorization and image segmentation approaches, ie. multi-object VAE models, often infer duplicate latent object representations that harm the scene object segmentation performance (see **top left**). We propose a differentiable latent-duplicate-suppression prior (abbr. *LDS*) to train better multi-object-VAE inference networks that suppress the duplicates (see **bottom left**). **Middle & right** Multi-object VAEs that trained with the proposed *LDS* achieves better scene object segmentation (e.g. higher mIoU on 2 datasets) and observation reconstruction performance (lower MSE).

representations, while we aim at fixing these issues in this work. There are unsupervised scene factorization models that handle explicit relations among the inferred objects, e.g. R-NEM [20], STOVE [13] and G-SWM [14]. They define “relations” as the interactions and scene dynamics of the scene objects and thus differ from the problem solved in this paper, which concerns relations between the inferred representations rather than the scene objects.

The proposed work is related to the duplicate-removal or non-maximum-suppression (abbr. NMS) idea that is widely used across many computer-vision tasks such as edge detection [18] and feature extraction [16]. Among all the applications, NMS's usage in object detection is the closest to ours, where duplicate detection candidates will be removed or suppressed [1, 19] based on a quantifiable criterion, e.g. detection confidence. However, as NMS in these models works as a post-processing technique so it cannot handle the mistakes a model made in the inference stage. Also, the violation of the *uniqueness assumption* by the aforementioned CompVAEs can lead to a worse variational approximation of the VAE posterior [4], which is worse than what the traditional duplicate-removal techniques achieve.

3 Method

Our goal is to train CompVAEs' inference networks that can suppress duplicates when making inferences at test time. Our approach is to introduce a differentiable prior, i.e. the *LDS* prior, as an additional constraint to train the CompVAEs' inference models. In Section 3.1, we briefly review the general construction of CompVAEs. In Section 3.2, we present the *LDS* prior and how to train a CompVAE model with it. In Section 3.3, we discuss CompVAEs' suboptimality and define a measure for the comparison of two posterior approximations.

3.1 Background

Similar to VAEs, a CompVAE model often consists of a generative model and an inference model. The generative likelihood of a scene image observation in a CompVAE is often modeled as a spatial Gaussian mixture [8, 21] parametrized by θ (where the variables θ

parameterize the generative model):

$$p_\theta(x|\{z_k\}) = \prod_{i=1}^{M} \sum_{k=1}^{K} p_\theta(C_i = k|z_k) \cdot \mathcal{N}(x_{ik}; g_\theta(z_k), \sigma^2), \tag{1}$$

where i indexes a pixel location (M in total) and x_{ik} is the RGB value of the k-th object at the location i. RGB values are samples of $\mathcal{N}(x_{ik}; g_\theta(z_k), \sigma^2)$ where $g_\theta(\cdot)$ is a decoder network and the standard deviation σ is set to a fixed value, e.g. $\sigma = 0.1$, for all pixels. The generated K RGB values x_{ik} compete to explain a location i as an instance of object k. The objects and their likelihoods, i.e. the mixing coefficients, are captured by a categorical distribution $p_\theta(C_i = k|z_k)$, where $C_i = k$ denotes the event of object k's winning. This formulation is similar to that used in MulMON [17], but that approach investigated multi-view problems, where viewpoints were taken as conditions.

To tackle the problems of scene factorization and object-centric learning, CompVAEs' inference models infer a joint posterior of all interested factors (i.e. scene objects $\{z_k\}$). Although CompVAEs encode a fixed number (K) of object slots for the inferred object representations, they do not make any assumption about the number of objects in a scene. Ideally, one can use as many object slots as possible. However, in practice, a K that is slightly larger than the number of scene objects is often chosen for efficient computation. Unfortunately, this leads the CompVAEs to use the slots, and thus they may create duplicate representations. Based on the independence assumption about the scene objects, the inference problem is solved by computing a tractable variational approximation:

$$q_\Phi(\{z_k\}|x) = q_\Phi(z_1, z_2, \ldots, z_k|x) = \prod_{k=1}^{K} q_\Phi(z_k|x, *), \tag{2}$$

where Φ denotes the trainable amortized parameters [11] that parameterize a family of distributions and $*$ denotes other conditions (e.g. $z_{1:k-1}$ as in [5]). Note that equation 2 is a general form of a CompVAE inference model, however, the amortization and factorization hold for all existing CompVAE variants.

3.2 Latent Duplicate Suppression

The goal of the proposed *LDS* prior is to penalize duplicates during the training process so the trained model produces fewer duplicate object representations during inference. In other words, we want to train a Φ that better suppresses duplicates. Because CompVAEs use fixed numbers (K) of object slots for the inferred latent representations, we can easily construct a fixed-size pair-wise similarity matrix, $\Sigma \in \mathbb{R}^{K \times K}$ using a kernel function.

In this paper, we use the cosine kernel function to compute the similarities between any two objects' latent representations in the set $\{z_k\}$. This is computationally equivalent to concatenating the inferred K D-dim object latent representations $\{z_k\}$ to make a matrix $\mathbf{Z} \in \mathbb{R}^{K \times D}$ and computing the similarity matrix : $\Sigma = \mathbf{Z}\mathbf{Z}^T/(||\mathbf{Z}_r|| \cdot ||\mathbf{Z}_c^T||)$, where $||\mathbf{Z}_r||$ and $||\mathbf{Z}_c^T||$ compute the Euclidean norms for matrix $\mathbf{Z}$ and $\mathbf{Z}^T$'s row and column vectors respectively.

The self-similarities of the inferred objects are captured by the constructed Σ's diagonal elements and the mutual similarities are captured by Σ's off-diagonal elements. To suppress duplicates, we need to penalize high off-diagonal similarities, i.e. by maximizing the *LDS* prior:

$$\mathcal{L}_{LDS}(\{z_k\};\Phi) = \sum_{h=1}^{K} \sum_{,j=1,h\neq j}^{K} \log \mathcal{N}(\Sigma_{h,j}; 0, \sigma^2). \tag{3}$$

The log normal density regulates its measure to a smaller range and σ (which models small variation in the similarity values) is fixed globally at 0.1. As both VAEs and CompVAEs are variational Bayesian models, their training relies on maximizing their evidence lower bounds (abbr. ELBO, denoted as $\mathcal{L}_{ELBO}(x;\Phi,\theta)$) w.r.t. the two trainable parameters Φ and θ. Taking a CompVAE model, we thus train it by maximizing:

$$\mathcal{L}(x;\Phi,\theta) = \mathcal{L}_{ELBO}(x;\Phi,\theta) + \lambda \cdot \mathcal{L}_{LDS}(\{z_k\};\Phi), \tag{4}$$

where λ is a Lagrange multiplier (set to default: 1). In general, combining Eqn. 1 and 2 leads to a general formulation of CompVAE ELBO: $\mathcal{L}_{ELBO}(x;\Phi,\theta) = \mathbb{E}_{q_\Phi(\{z_k\}|x)}[\log p_\theta(x|\{z_k\})] - D_{\mathrm{KL}}(q_\Phi(\{z_k\}|x)|p_\theta(\{z_k\}))$. However, the exact formulations for a specific CompVAE is model-dependent. It is worth noting that although ELBOs are computed by the iterative inference processes of IODINE and MulMON during testing, we use the *LDS* priors only in training.

3.3 CompVAE Suboptimality Measure

In this paper, we use superscripts $+$ and 0 on a variable to indicate if it is related to the experimental group (CompVAEs trained with *LDS* prior) or the control group (original CompVAEs). To validate that after suppressing duplicate object representations, the CompVAE models less often violate the *uniqueness assumption* and approximates better the variational posterior $p(\{z_k\}|x)$, i.e. $q_{\Phi^+}(\{z_k\}|x)$ becomes a better approximation than $q_{\Phi^0}(\{z_k\}|x)$ with respect to $p(\{z_k\}|x)$, we need a measure to quantify approximation qualities and thus support model comparisons.

Through the derivation of VAEs' ELBO [11], a gap between the observed evidence $\log p_\theta(x)$ and the ELBO $\mathcal{L}_{ELBO}(x;\Phi,\theta)$ is illustrated:

$$D_{\mathrm{KL}}(q_\Phi(z|x)\|p_\theta(z|x)) = \log p_\theta(x) - \mathcal{L}_{ELBO}(x;\Phi,\theta) \geq 0. \tag{5}$$

This is referred to as the *inference gap* of VAEs [4], which provides a quantitative measure of how good is an approximation. Similarly, we formulate $\mathcal{G} = D_{\mathrm{KL}}(q_\Phi(\{z_k\}|x)\|p_\theta(\{z_k\}|x))$ as the approximation quality measure for a CompVAE. Therefore, by comparing $\mathcal{G}^+$ and $\mathcal{G}^0$ we can determine if the experimental group reaches better suboptimality than the control group.

In practice, because $\log p_\theta(x)$ is not inaccessible, $\mathcal{G}$ is not computable (see equation 5). We thus approximate $\log p_\theta(x)$ with a Monte Carlo estimate — the importance weighting estimate [2], where the sample size (denoted as B) is set to 500. Therefore, we can compute the inference gap $\mathcal{G}$ as:

$$\begin{aligned}\mathcal{G} &= D_{\mathrm{KL}}(q_\Phi(\{z_k\}|x)\|p_\theta(\{z_k\}|x)) \\ &= \mathbb{E}_{\mathbf{z}^1,\ldots,\mathbf{z}^b \sim q_\Phi(\mathbf{z}|x)}[\log \frac{1}{B}\sum_{b=1}^{B} \frac{p_\theta(x,\mathbf{z}^b)}{q_\Phi(\mathbf{z}|x)}] - \mathcal{L}_{ELBO}(x;\Phi,\theta) \geq 0.\end{aligned} \tag{6}$$

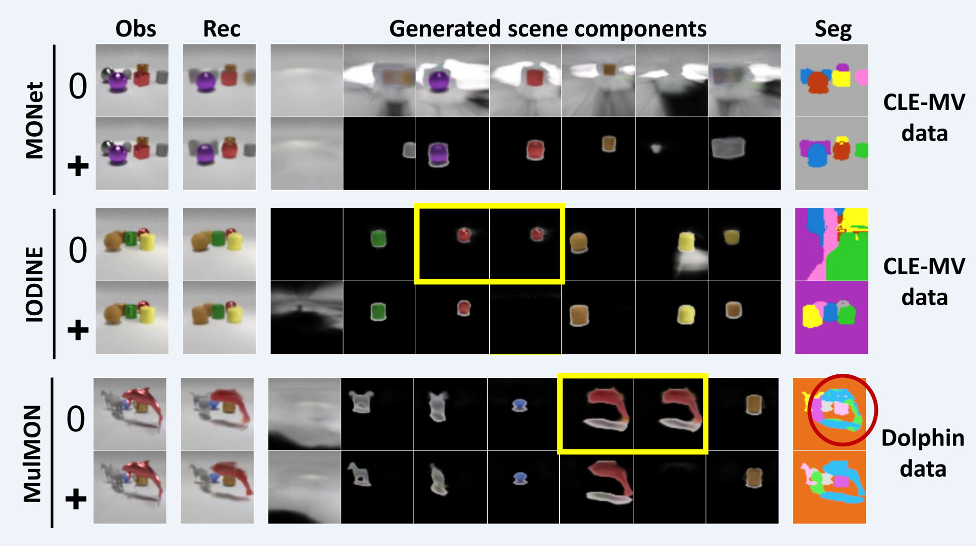


Figure 2: Qualitative comparisons between the experimental group (tagged with "+") and the control group (tagged with "0"). The Obs column is a source image, Rec is the corresponding reconstructed image based on the inferred representation. The next 7 columns show the independent generation of the inferred scene components (order not important). The Seg column shows the pixel label for the component with highest probability (pixel color is not important). **Top** Training with *LDS* aids the original MONet model which suffers from local minima: obtains fair factorization and reconstruction while fails to learn clean object geometries and thus generates noisy scene components whereas MONet$^+$ produces cleaner inferred components. **Middle** Training with *LDS* aids IODINE: resolves duplicates (circled in yellow) and fixes the weak background segmentation, as shown by the large colored regions in the Seg column, which is a known issue of IODINE [7]. **Bottom** Training with *LDS* allows MulMON to suppress duplicates and thus produce a better segmentation map. (Colored boxes and circles highlight the duplicates and failures caused by them.)

to simplify the discussion hereafter, we define a measure *inference gap drop* (denoted as $\Delta\mathcal{G}^+$) using $\mathcal{G}^0$ and $\mathcal{G}^+$: $\Delta\mathcal{G}^+ = \mathcal{G}^0 - \mathcal{G}^+$. In general, a positive $\Delta\mathcal{G}^+$ suggest a smaller gap is achieved and thus provides better approximation, a negative $\Delta\mathcal{G}^+$ suggests the opposite. In our experiments, we use $\Delta\mathcal{G}^+$ as an important metric for our model suboptimality analysis (see Section 4.3).

4 Experiments

Our experiments are based on two datasets: CLE-MV [17] and Dolphin. The Dolphin dataset is synthesized using CLE-MV's graphics engine by adding more complex and general shapes (e.g. dolphins, horses, ducks, etc.). There are in total 1700 and 3631 different scenes in the CLE-MV and the Dolphin datasets respectively and each scene consists of 3-6 objects including the background (a trivial object). As there are 10 image observations (with size 64 × 64) taken from 10 different viewpoints, both the two datasets support multi-view tasks. We thus randomly select 1500 scenes (15000 images) from CLE-MV and 3000 scenes (30000 images) from Dolphin to make the training sets. At test time, we sample 160 unseen scenes (i.e. 1600 images) from CLE-MV and 200 unseen scenes (2000 images) from Dolphin, where “unseen scenes” denote scenes that are not in the training sets. Note that we use multi-view CLEVR datasets instead of the original CLEVR [10] because we want to show that the proposed method works for both single-view and multi-view scenarios. For the experiments, we use three baseline CompVAE models including two single-view models, i.e. MONet and IODINE, and a multi-view model MulMON, and create our experimental group with the three CompVAEs trained with the proposed *LDS* prior. We train all models using the same training specifications as that of the experimental group except for removing the *LDS* prior. We thus study and demonstrate the effectiveness by comparing the two groups in various aspects. We refer the reader to the Appendix for the ablation study and the model specifications.

4.1 Duplicate Suppression

The first set of experiments justify the proposed *LDS* methods by demonstrating its effects on suppressing duplicates. We ran both the control-group and experimental group models on the 200 CLE-MV test scenes (2000 images) to get two quantitative measures: 1) the average pair-wise similarities (see eqn (3), denoted as $\overline{\mathbf{Sim}}$) among all the inferred latent object representations, 2) the percentage of images for which object duplicates were inferred. To better visualize the effect of the proposed *LDS* on reducing latent-object-representation similarities, we used the difference between average pair-wise similarities of the control- and experimental-group models, i.e. $\Delta\overline{\mathbf{Sim}}^{+} = \overline{\mathbf{Sim}}^{0} - \overline{\mathbf{Sim}}^{+}$, where a positive $\Delta\overline{\mathbf{Sim}}^{+}$ suggests positive effect of *LDS* in suppressing latent object replicates. For the second measure, we randomly picked 100 images and counted the total number of image cases where duplicates were produced. The results in Figure 3 suggest that the proposed *LDS* prior works effectively reduces latent-object-representation similarities and suppresses duplicate representations.

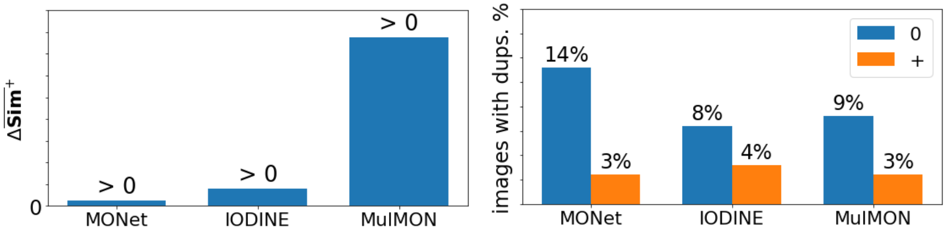


Figure 3: Effectiveness of the proposed *LDS* in duplicate suppression. **Left** All of the three tested CompVAEs give positive $\Delta\overline{\mathbf{Sim}}^{+}$ values, where positive $\Delta\overline{\mathbf{Sim}}^{+}$ suggests smaller similarities (i.e. improvements) of the experimental-group (trained with *LDS*) latent object representations than that of the control group. **Right** Direct comparison between the experimental- (tagged with $^{+}$) and control-groups (tagged with ^{0}) in duplicate suppression. The lower percentages when using LDS mean fewer duplicates and thus effective duplicate suppression.

4.2 Task Performance

Scene Factorization The biggest advantage of CompVAEs over traditional VAEs in visual scene understanding is that they can perform unsupervised scene factorization, which di-

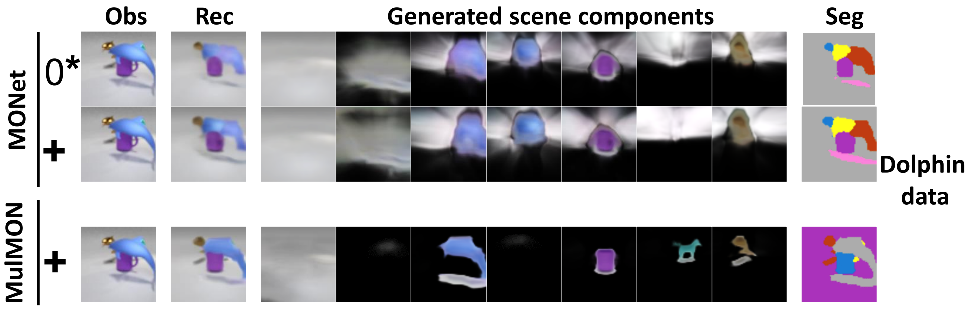


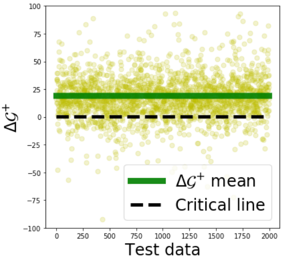
Figure 4: A partial-failure example from the "outlier" model (MONet^0) on Dolphin (tagged with "⋆" in Table 1). **Top** The model produces good factorization but fails badly to learn good-quality object representations and thus shows noisy generations. The proposed *LDS* fails to fix it. **Bottom** A good example shown by a model that achieves similar quantitative performance (MulMON$^+$).

		CLE-MV		Dolphin	
Models	*LDS*	MSE↓	mIoU↑	MSE↓	mIoU↑
MONet	0	0.0037 ± 0.0002	0.6806 ± 0.0072	**⋆0.0059 ± 0.0002**	**⋆0.6620 ± 0.0070**
	+	**0.0024 ± 0.0002**	**0.7899 ± 0.0092**	0.0063 ± 0.0005	0.6567 ± 0.0077
IODINE	0	**0.0016 ± 0.0002**	0.1911 ± 0.0042	0.0054 ± 0.0001	0.3501 ± 0.0043
	+	0.0020 ± 0.0001	**0.7252 ± 0.0054**	**0.0050 ± 0.0002**	**0.6224 ± 0.0052**
MulMON	0	0.0019 ± 0.0001	0.7834 ± 0.0046	0.0055 ± 0.003	0.6246 ± 0.0056
	+	**0.0019 ± 0.0001**	**0.7911 ± 0.0043**	**0.0051 ± 0.0002**	**0.6556 + ±0.0027**

Table 1: Quantitative comparisons between the experimental group (tagged with "+") and the control group (tagged with "0"). All results are averaged over five different random seeds. ⋆ denotes the most significant case where *LDS* does not generate obvious improvements which we will discuss in the text.

rectly links to observation segmentation. Therefore, we compared the scene object decomposition performance between the experimental group (CompVAEs trained with *LDS*) and control group (original CompVAEs) on scene object decomposition task. Because both the CLE-MV and Dolphin datasets are synthesized with the ground-truth segmentation maps, we can thus compute the *mean intersection over union* (mIoU) score as the performance measure. To solve the bipartite matching problem as the output object masks (in a list) are not in the same order as the GT masks, we used the Hungarian matching algorithm to find the best match that maximizes the mIoU score for a scene. Table 1 shows that the experimental group, i.e. CompVAEs trained with the proposed *LDS* prior, results in similar or improved performance compared to the control group over most models and datasets. Figure 2 demonstrates the effectiveness of the proposed *LDS* prior in reducing duplicates and aiding CompVAEs' local minimas. We also examined the "outlier model", i.e. MONet$^+$ trained on Dolphin, and some output samples are shown in Figure 4. For the outlier model, even though the quantitative measures are improved, the model still suffers from the local minima. We also consider this a failure instance of the proposed *LDS* as it does not aid the model like it does to MONet trained on the Dolphin dataset (see Figure 2).

Scene Reconstruction Reconstruction quality reflects the representation-learning qual-



Pearson Correlations Coefficients	
Reconstruction MSE vs. $\Delta\mathcal{G}^+$	-0.2874
Segmentation mIoU vs. $\Delta\mathcal{G}^+$	0.2019

Table 2: **Left Figure**: The effect of *LDS* on the variational approximation quality: yellow dots represents the $\Delta\mathcal{G}^+$ for each test data sample (2000 test images), and the green line is the mean $\Delta\mathcal{G}^+$, which is the change in the ELBO (evidence lower bound) value from Eqn 5. Positive values are improvements. Observe that most dots lie above the "no improvement" line at 0, demonstrating that *LDS* generally produces improvements. **Right Table**: The correlation between the task performance and the *inference gap*. 0 suggests no correlation and +1/-1 denotes the strongest positive/negative correlation. The right table exhibits a positive correlation between the *inference gap drop* $\Delta\mathcal{G}^+$ and the reconstruction errors and a positive correlation between the segmentation accuracy (mIoU) and $\Delta\mathcal{G}^+$.

ity of a VAE model. Hence, we compared the experimental group and the control group also on reconstruction quality using the *mean squared error* (MSE) between the observation image and the reconstruction image as our quantitative measure. The MSE was computed from the RGB vector distances, where color values are on a $[0,1]$ scale. Table 1 shows that the proposed *LDS* improves not only the scene factorization but also the scene reconstruction. This suggests the proposed *LDS* helps CompVAEs to learn better scene representations.

4.3 Suboptimality Analysis

As shown in Fig 2, the proposed *LDS* prior not only suppressed the object replicates, but it also fixed several issues (uniqueness and degenerated inference) that exist in the original CompVAEs, improving scene reconstruction quality. The suboptimality analysis presented in this section gives a better understanding of how the proposed *LDS* helps to improve the task performance. To verify our hypothesis that the proposed *LDS* reduced the violation of the *uniqueness assumption* and thus achieved a better variational approximation of the target posterior $p_{\theta}(\mathbf{z} = \{z\}|x)$ and improved the task performance, we studied: **1)** the effect of the proposed *LDS* on the variational approximation quality, and **2)** the correlation between the task performance (mIoU) and the variational approximation quality. We evaluated the variational approximation quality by computing the *inference gap drop* $\Delta\mathcal{G}^+$ (see Section 3.3) for the 2000 test images from the CLE-MV dataset and averaged the $\Delta\mathcal{G}^+$ over 2000 samples to obtain the mean $\Delta\mathcal{G}^+$. Table 2 (left figure) shows the drop $\Delta\mathcal{G}^+$ of these 2000 test samples and their mean. As illustrated, MONet trained with the proposed *LDS* produces a positive drop $\Delta\mathcal{G}^+$ — the proposed *LDS* reduces the *inference gap* and is thus a better approximation than the original model. We computed the Pearson correlation coefficients between the task performance measures, i.e. MSE (for reconstruction) and mIoU (for segmentation), and $\Delta\mathcal{G}^+$ on the 2000 test samples. As shown in Table 2 (right table) an increased *inference gap drop* $\Delta\mathcal{G}^+$ does indeed decrease the reconstruction error (negative correlation) and increase the segmentation accuracy (positive correlation).

5 Conclusion

In this work, we present a differentiable prior that leverages similarity measures to regulate the object-centric latent representations inferred by multi-object VAEs, i.e. CompVAEs. Despite its simplicity, we demonstrate its effectiveness in fixing known issues, namely the *uniqueness issues*, of the multi-object VAE models — inferring duplicate object representations. We ascribe the *uniqueness issues* to the violation of the *uniqueness assumption* that is implicitly introduced by the scene-mixture-model assumption, i.e. each part of an scene observation (e.g. a pixel) must be explained by one and only scene object. Therefore, we demonstrate through experiments that, by suppressing duplicates, better variational approximation and task performance can be achieved. Regarding the future work, we are particularly interested in modelling more flexible and possibly learnable similarity functions, e.g. a similarity measure that can distinguish explicitly the inter-object correlations' effect on each dimension of an object's latent representation and thus weight them accordingly.

6 Acknowledgements

The authors would like to thank the reviewers for the insightful suggestions.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV, 5561–5569, 2017.*

[2] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *ICLR, 2016.*

[3] 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.

[4] Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *ICML, 1078–1086, 2018.*

[5] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: Generative Scene Inference and Sampling of Object-Centric Latent Representations. In *ICLR, 2020.*

[6] 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 *NeurIPS, 3225–3233, 2016.*

[7] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *ICML, 2424–2433, 2019*, .

[8] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *NeurIPS, 6691–6701,2017*, .

[9] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *NeurIPS, 517–526, 2018.*

[10] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR, 2901–2910, 2017.*

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

[12] Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. In *NeurIPS, 8606–8616,2018.*

[13] Jannik Kossen, Karl Stelzner, Marcel Hussing, Claas Voelcker, and Kristian Kersting. Structured object-aware physics prediction for video modeling and planning. In *ICLR, 2020.*

[14] Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Multi-object representation learning with iterative variational inference. In *ICML, 2020*, .

[15] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. In *ICLR, 2020*, .

[16] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[17] Li Nanbo, Cian Eastwood, and Robert B Fisher. Learning object-centric representations of multi-object scenes from multiple views. In *NeurIPS, 2020*.

[18] Azriel Rosenfeld and Mark Thurston. Edge and curve detection for visual scene analysis. *IEEE Transactions on computers*, 100(5):562–569, 1971.

[19] Rasmus Rothe, Matthieu Guillaumin, and Luc Van Gool. Non-maximum suppression for object detection by passing messages between windows. In *ACCV, 290–306, 2014*.

[20] Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *ICLR, 2018*.

[21] Christopher K I Williams and Michalis K Titsias. Greedy learning of multiple objects in images using robust statistics and factorial learning. *Neural Computation*, 16(5): 1039–1062, 2004.

Supplementary Material for: Duplicate Latent Representation Suppression for Multi-object Variational Autoencoders

Li Nanbo
nanbo.li@ed.ac.uk
Robert B. Fisher
rbf@inf.ed.ac.uk

School of Informatics
University of Edinburgh
Edinburgh, UK

A. Implementation Details

Training specifications We refer to Table 1, 2 & 3 to the training configurations of MONet, IODINE and MulMON respectively. Note that 1) for IODINE adn MulMON that use iterative inference modules, we apply LDS per iterative step to compute their ELBOs during training, and 2) for all CompVAEs, we apply LDS only in their training times.

Model Architecture Specifications As discussed in the main paper, we use three existing CompVAE models as our baselines and build our contributions on top of these architectures. It is important to use the same architectures as the that of the original papers. However, we found it difficult to use a latent dimension of 64 as in [5] for the CLEVR-based datasets as it trains too slow, over one week for one run on two RTX2080TI, we thus reduced the dimension of IODINE to 16 for our IODINE. As constructing the proposed *LDS* prior requires no model architecture design and architecture parameter tweaking, we refer to the original papers of MONet [2], IODINE [5], and MulMON [6] for the architecture details.

Table 1: Training Configurations For MONet

Type	the trainings of MONet^0 and MONet$^+$
Optimizer	RMSprop
Initial learning rate η_0	$3e^{-4}$
Batch size	40 (unit: images)
Learning rate at step s	N/A
Total gradient steps	$600k$
Gradient-norm clipping	5.0
log-normal likelihood strength	1.0
KL (Gaussian prior) strength β	0.5
KL (attention prior) strength	0.5
LDS (MONet$^+$ only) strength	0.5

Table 2: Training Configurations of IODINE^0 and IODINE$^+$

Type	the trainings of IODINE^0 and IODINE$^+$
Optimizer	Adam
Initial learning rate η_0	$1e^{-4}$
Batch size	8
Learning rate at step s	$\star\max\{0.1\eta_0 + 0.9\eta_0 \cdot (1.0 - s/1e^6), 0.1\eta_0\}$
Total gradient steps	$600k$
Gradient-norm clipping	5.0
inference iterations [5]	5
log-normal likelihood strength	1.0
KL (Gaussian prior) strength β	1.0
LDS (IODINE$^+$ only) strength	1.0

$\star$: same scheduler as GQNs'.

Table 3: Training Configurations of MulMON^0 and MulMON$^+$

Type	the trainings of MulMON^0 and MulMON$^+$
Optimizer	Adam
Initial learning rate η_0	$2e^{-4}$
Batch size	8
Learning rate at step s	$\star\max\{0.1\eta_0 + 0.9\eta_0 \cdot (1.0 - s/1e^6), 0.1\eta_0\}$
Total gradient steps	$600k$
Gradient-norm clipping	5.0
inference iterations [5]	5
log-normal likelihood strength	1.0
KL (Gaussian prior) strength β	1.0
LDS (IODINE$^+$ only) strength	1.0

$\star$: same scheduler as GQNs'.

B. CompVAE Rendering Process

Figure 1 shows the CompVAE rendering process we used to produce all qualitative results presented in this paper. **Importantly, we used softmax functions to compute the compositional probabilities of each components, i.e. the mixing probabilities in Eqn.(1), to render the whole scene, and sigmoid functions to render independent objects**. However, one might also see independent component rendering with other functions in the related literature, e.g. IODINE [5] uses a linear mapping of x_k to render independent components.

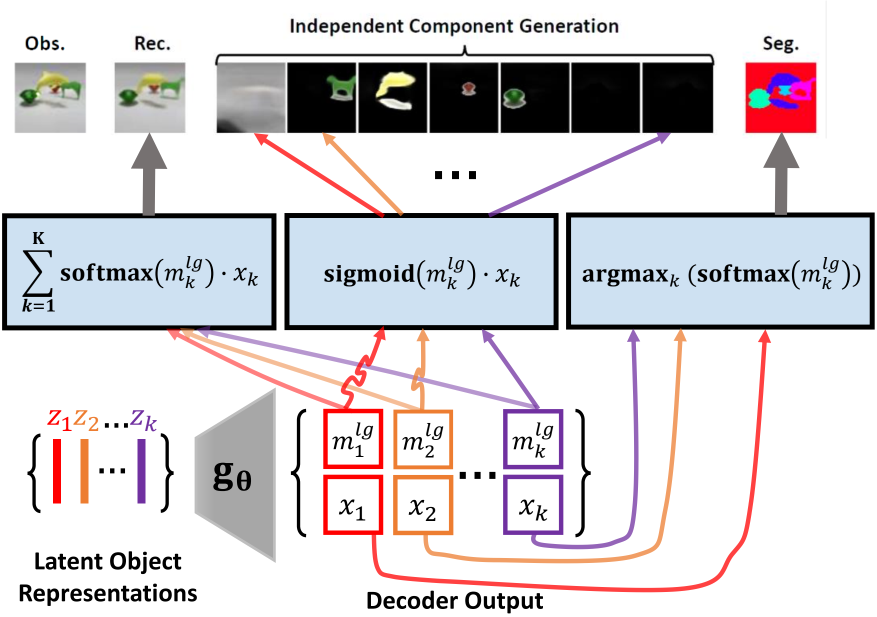

Obs.
Rec.
Independent Component Generation
Seg.
...
$\sum_{k=1}^{K} \mathbf{softmax}(m_k^{lg}) \cdot x_k$
$\mathbf{sigmoid}(m_k^{lg}) \cdot x_k$
$\mathbf{argmax}_k\ (\mathbf{softmax}(m_k^{lg}))$
$z_1 z_2 \ldots z_k$
g_θ
m_1^{lg}
m_2^{lg}
m_k^{lg}
x_1
x_2
x_k
Latent Object Representations
Decoder Output

Figure 1: Overview of a CompVAE rendering process. The rendering process starts by inputting a set of inferred latent object representations (**Bottom left**) into the generator network g_θ. The generator g_θ outputs a raw mask ($m_k^{lg} \in \mathbb{R}^{H \times W \times 1}$) and a color pool ($x_k \in \mathbb{R}^{H \times W \times 3}$) (**Bottom middle**). The decoder output is then passed into three different functions (**Middle row**) to get different render results (**Top row**). All computations are defined pixel-wise but executed in parallel.

C. Additional Results

C.1 Ablation Study

The ablation study focuses on two hyperparameters: 1) the standard deviation σ used in the *LDS* prior (see Section 3.2 of the main paper) and 2) the number of object slots K. The former relates to the precision of the similarity measure and the latter determines the size of the similarity matrix constructed in the *LDS* computation, i.e. it relates to the scalability of *LDS*. We do the ablation study with only MONet and on only the CLE-MV dataset for computation efficiency. We select 4 different σ to train MONet and compare their performance on the scene reconstruction and the scene factorization tasks. Figure 2 shows no significant performance loss in tasks by changing σ from the default value, 0.1, to other values. A future investigation will be further increasing σ until it is sufficiently close to a uniform distribution and thus breaks the *LDS* prior. Moreover, the performance might get boosted in some cases. For the object-slot quantity K, we first train MONet with $K = 7$ and $K = 9$ respectively and test them with 7,9, 11, 15 object slots. Figure 2 shows: 1) the models trained with $K = 7$ and $K = 9$ have very similar performance in both tasks and 2) testing with a different K does not cause a significant performance drop.

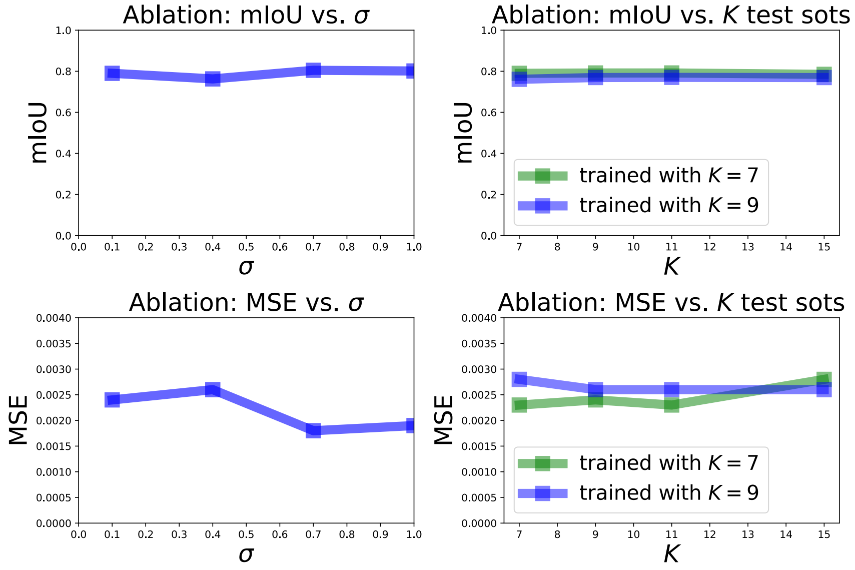


Figure 2: Ablation study results. **Top left** Scene decomposition performance vs. *LDS* prior precision (σ). **Top right** Scene decomposition performance vs. the number of object slots used in training and testing (K). **Bottom left** Scene observation reconstruction performance vs. *LDS* prior precision (σ). **Bottom right** Scene observation reconstruction performance vs. the number of object slots used in training and testing (K).

C.2 GENESIS on the CLE-MV Data

We tested GENESIS [4] on the CLE-MV data to assess how well the inference redundancy problems are handled by the autoregressive model of GENESIS. The experiment was conducted on top of the official implementation of GENESIS [1] with strict abidance of its original hyperparameter configurations. However, as shown in Figure 3, GENESIS failed to factorise CLE-MV scenes correctly—it treats a CLE-MV scene observation (i.e. an image) as a big and flat object that contains all the content. As a result, it produces wrong image segmentation. A possible reason could be that GENESIS represents the autoregressive conditioning of object discovery in the latent space (i.e. $z_k \,|\, z_{1:k-1}$) instead of the image space as that of MONet—a successive object mask conditions directly on all the previous obtained masks (i.e. $m_k \,|\, m_{1:k-1}$). According to [3], this could introduce more severe global information leaking issue. In general, future study is needed to better understand the practical limitations and their causes in GENESIS.

C.3 Real-image Experiments

To demonstrate that the proposed LDS can efficiently perform duplicate suppression on real images, we conducted comparison experiments between CompVAEs that are trained with and without LDS priors on the a collected real-image dataset.

Real-image Dataset We created such dataset by randomly placing 2−4 cubes (of different colours) on white table top and taking photos with a webcam that is mounted on a moving robot arm. We created 109 scenes in total and for each scene we captured 20−30

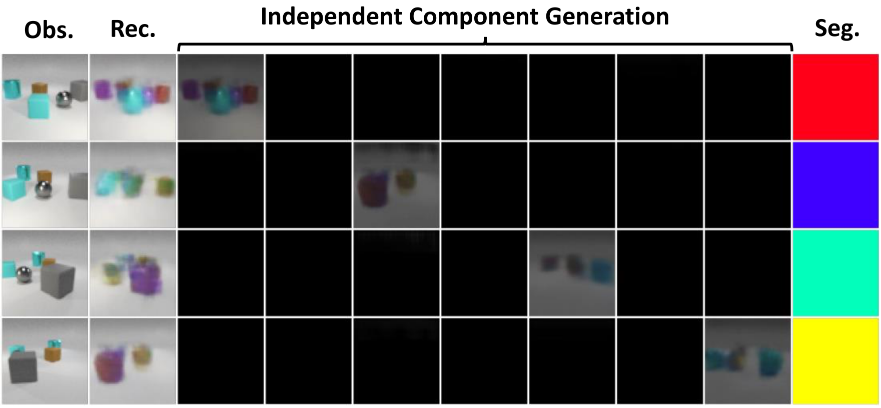


Figure 3: Qualitative results of GENESIS on the CLE-MV dataset.

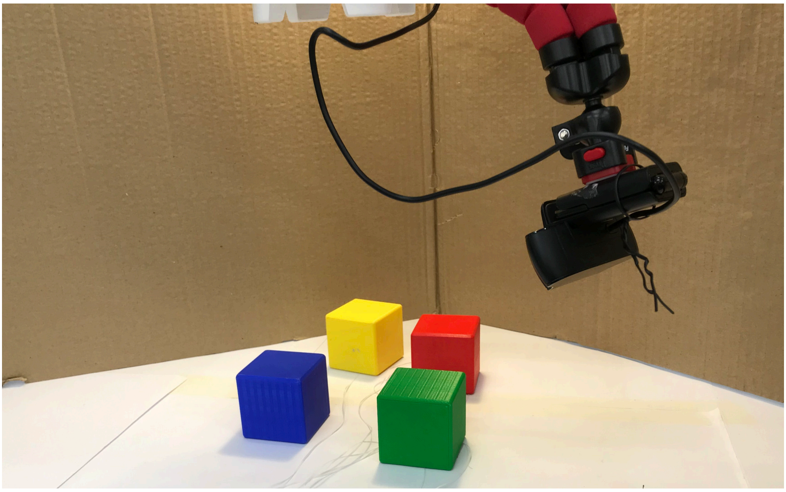


Figure 4: Hardware platform for real-image dataset recording.

images from different viewing angles. We show the hardware platform setup in Figure 4.

Results Figure 5 shows that the original MONet$^{0^\star}$ infers redundant white table components. Although MONet$^+$ demonstrates a slight performance drop in handling occlusions (e.g. renders the independent table component worse than MONet$^{0^\star}$), it does suppress the duplicate table finding issues of MONet$^{0^\star}$. Also, we see that MONet$^+$ produces cleaner segmentation results than MONet$^{0^\star}$. Compared with synthetic data, real images often come from complex distributions and thus exhibit significant larger pixel variances (due to uncontrolled lighting, materials, etc.), complicating the training of a generative model. This also explains why neither MONet$^{0^\star}$ nor MONet$^+$ model the independent table (always partially occluded) distribution properly. In conclusion, LDS is an effective addition to CompVAEs on real data and can potentially serve as a useful tool in some real applications.

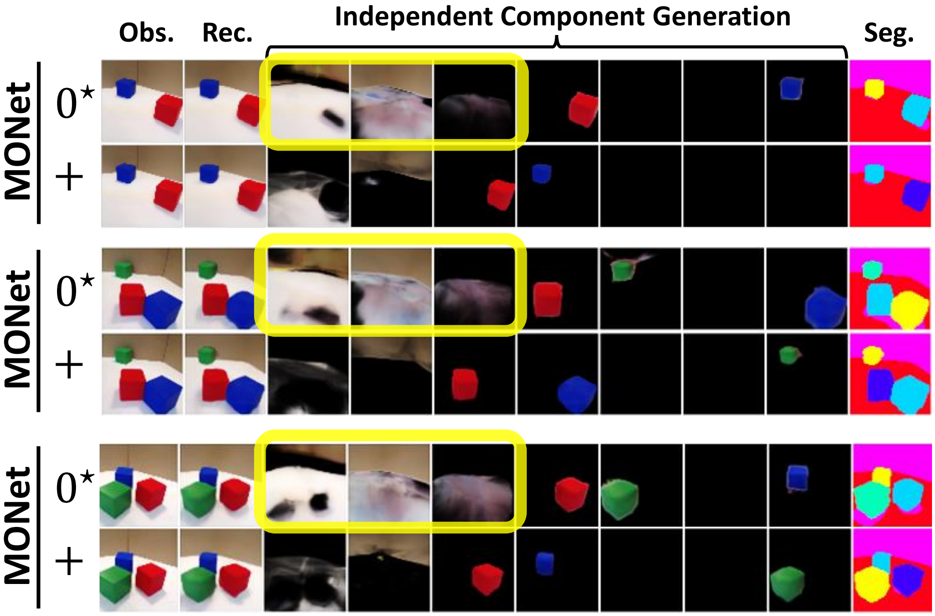


Figure 5: Qualitative results of MONet on real images. Symbols "0*" and "+" tag models that trained with and without LDS respectively. Yellow circles highlight duplicated or partially duplicated components.

References

[1] GENESIS official implementation. https://github.com/applied-ai-lab/genesis.git. Accessed: 2021-07-01.

[2] 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.

[3] Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations. In *ICML, 2021*.

[4] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: Generative Scene Inference and Sampling of Object-Centric Latent Representations. In *ICLR, 2020*.

[5] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *ICML, 2424–2433, 2019*.

[6] Li Nanbo, Cian Eastwood, and Robert B Fisher. Learning object-centric representations of multi-object scenes from multiple views. In *NeurIPS, 2020*.