NON-MAXIMUM SUPPRESSION ALSO CLOSES THE VARIATIONAL APPROXIMATION GAP OF MULTI-OBJECT VARIATIONAL AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning object-centric scene representations is crucial for scene structural understanding. However, current unsupervised scene factorization and representation learning models do not reason about scene objects' relations while making an inference. In this paper, we address the issue by introducing a differentiable correlation prior that forces the inference models to suppress duplicate object representations. The extension is evaluated by adding it to three different scene understanding 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 close the approximation gap between the data evidence and the evidence lower bound.

1 INTRODUCTION

Variational autoencoders (VAEs) (Kingma & Welling, 2013) 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 $p(\boldsymbol{x};\boldsymbol{\theta})$ over the observation data $\boldsymbol{x} \in \mathbb{R}^M$ but also it allows to describe and represent the observed scenes in a more compact latent space $\boldsymbol{z} \in \mathbb{R}^D$ ($D \ll M$) for simplicity and efficiency. Recent advances in this area (Burgess et al., 2019; Greff et al., 2019; Anon, 2020) treat a multi-object scene as a composition of scene objects (aka scene components) and show success in scene factorization and object-based representation learning. I.e. a scene representation $\boldsymbol{z}$ is a set of K scene object representations $\boldsymbol{z} = \{\boldsymbol{z}_k\}$ where each object representation $\boldsymbol{z}_k$ explains one and only one object in the observation $\boldsymbol{x}$. These object-based VAE models are often referred to as the *multi-object VAEs*, they are called *component VAEs* (abbr. CompVAEs) in this paper for simplicity.

Inference of the latent representations $\{\boldsymbol{z}_k\}$ in CompVAEs uses variational Bayesian methods that approximate an intractable posterior $p(\{\boldsymbol{z}_k\}|\boldsymbol{x})$ with a variational distribution $q(\{\boldsymbol{z}_k\}|\boldsymbol{x})$. A necessary assumption that the inferred latent object representations are independent given an observation needs to be made in CompVAEs to attain object-wise posteriors: $q(\{\boldsymbol{z}_k\}|\boldsymbol{x}) = q(\boldsymbol{z}_1|\boldsymbol{x})q(\boldsymbol{z}_2|\boldsymbol{x})\dots q(\boldsymbol{z}_k|\boldsymbol{x})$. Because the original posterior is decomposed into independent pieces, one can easily sample each object's approximate posterior and manipulate a single scene object without interfering with the other objects. This is crucial when it comes to model evaluation, statistical criticism and interpretation.

Existing CompVAEs (e.g. MONet (Burgess et al., 2019), IODINE (Greff et al., 2019), MulMON (Anon, 2020)) show impressive results in factoring scenes and learning scene objects, however, we argue that the independence assumption is wrong as it ignores the fact that scene objects are not independent: for example, two objects cannot appear at the same spatial location. Also the trained inference models cannot perform correlation checks and thus allow inferring duplicate object representations. This harms directly the CompVAEs' scene factorization performance—two or more duplicate component representations need to compete with each other to explain the same segment of the observation (see Figure 1). Also, because the independence assumption increases the variational approximation gap (Cremer et al., 2018) between $q(\{\boldsymbol{z}_k\}|\boldsymbol{x})$ and $p(\{\boldsymbol{z}_k\}|\boldsymbol{x})$, the optimization process can get stuck at local minimas and thus produce wrong scene decomposition (see the local-minima example in Figure 1).

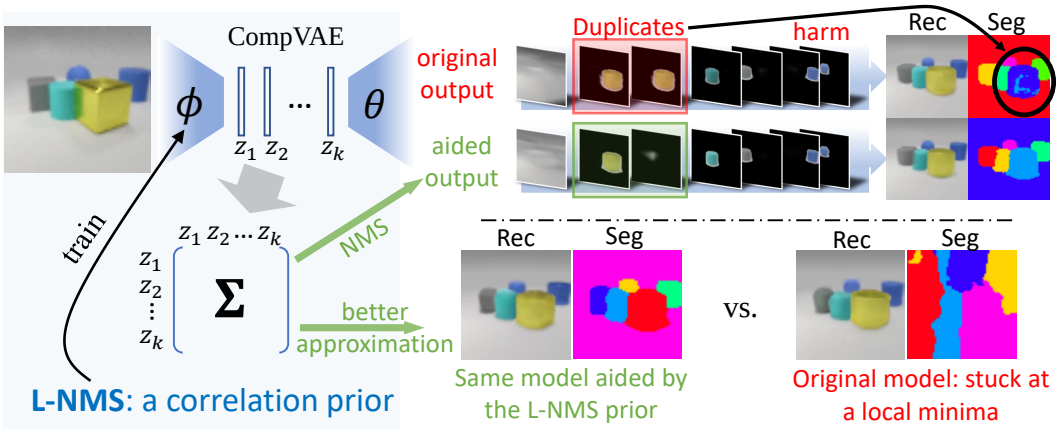

Figure 1: Overview: we propose a correlation prior, namely L-NMS, as an additional training constraint to train a CompVAE's (aka. multi-object VAE) inference model (left side of the Figure). The proposed L-NMS prior is able to not only suppress duplicates (top right of the Figure), but also tackle problems that are related to the CompVAEs' suboptimalities such as the background splitting (bottom right example) problem which is a known issue of IODINE.

The independence assumption, as discussed, it is a key assumption that simplifies an intractable scene factorization. In this paper, to address the aforementioned issues, we weaken the independence assumption during training the inference models by introducing a differentiable correlation prior. This implements the key insight that two identical object representations cannot be inferred for the same object so the inferred duplicates will be penalized by the correlation prior during training. This shares the same spirit with the *non-maximum suppression* (abbr. NMS) technique that is widely used in computer vision. We call our correlation prior the *latent non-maximum suppression* (abbr. L-NMS). We clarify that our goal is not to infer a set of mutually-correlated object representations $\{\boldsymbol{z}_k\}$ but to enable the inference models to reason and resolve correlation while inferring $\{\boldsymbol{z}_k\}$. I.e. with duplicates removed, we expect to train inference models that can infer de-correlated $\{\boldsymbol{z}_k\}$.

In our experiments, we train three representative CompVAE models, i.e. MONet, IODINE and MulMON, with L-NMS as the experimental group and train the same models without L-NMS as the control group. We illustrate the effectiveness of training CompVAEs' with L-NMS in suppressing scene factorization duplicates and closing the approximation gap by comparing the performance of the two groups of models. We claim and demonstrate:
1) Training a CompVAE with the proposed L-NMS prior enables the CompVAE to make an inference taking account of scene objects' correlations and produce better scene factorizations with fewer duplicate objects (see Section 4.1).
2) Training a CompVAE with the proposed L-NMS prior closes the *approximation gap* and thus increases the original evidence lower bound (see Section 4.2).
3) With the *approximation gap* closed, CompVAEs' trained with the proposed L-NMS better overcome local minimas and thus learn better scene representations that supports better scene observation reconstructions (see Section 4.1).

2 METHOD

Our goal is to weaken the independence assumption made in the existing CompVAEs during training so that the trained inference models can handle scene object correlations and therefore infer de-correlated object scene representations. Our approach is to introduce a differentiable correlation prior, i.e. the L-NMS prior, as an additional constraint to train the CompVAEs' inference models. In Section 2.1, we briefly review the general construction of CompVAEs. In Section 2.2, we present the L-NMS prior and how to train a CompVAE model with it. In Section 2.3, we discuss CompVAEs' suboptimality and define a measure for the comparison of two posterior approximations.

2.1 GENERAL CONSTRUCTION OF COMPVAES

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 (Williams & Titsias, 2004; Greff et al., 2017) parametrized by θ (where the variables θ parameterize the generative model):

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

where i indexes a pixel location (M in total) and x_{ik} is the RGB value of the k-th object at the location. RGB values are samples of $\mathcal{N}(x_{ik};\, g_\theta(\boldsymbol{z}_k), \sigma^2)$ 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 $p_\theta(C_i = k|\boldsymbol{z}_k)$, where $C_i = k$ denotes the event of object k's winning. This formulation is similar to that seen in MulMON (Anon, 2020), 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 $\{\boldsymbol{z}_k\}$). 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. Based on the independence assumption about the scene objects, the inference problem is solved by computing a tractable variational approximation:

$$q_\Phi(\{\boldsymbol{z}_k\}|\boldsymbol{x}) = q_\Phi(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k|\boldsymbol{x}) = \prod_{k=1}^{K} q_\Phi(\boldsymbol{z}_k|\boldsymbol{x}), \tag{2}$$

where Φ denotes the trainable amortized parameters (Kingma & Welling, 2013) that parameterize a family of distributions. Note that equation 2 is a general form of a CompVAE inference model, however, the amortization and factorization hold for all existing CompVAE variants.

2.2 LATENT NON-MAXIMUM SUPPRESSION

As discussed in Section 1, the main goal of correlation modeling is to weaken the independence assumption in CompVAEs' training processes so as to produce fewer duplicate object representations during inference. In other words, we want the trained Φ to resolve scene object correlations. Because CompVAEs use fixed numbers (K) of object slots for the inferred latent representations, we can easily construct a fixed-size correlation matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{K\times K}$ using a kernel function. In this paper, we use a simple cosine-similarity function to compute the correlation between any two objects' latent representations in the set $\{\boldsymbol{z}_k\}$. This is computationally equivalent to concatenating the inferred K D-dim object latent representations $\{\boldsymbol{z}_k\}$ to make a matrix $\boldsymbol{Z} \in \mathbb{R}^{K\times D}$ and computing the correlation matrix : $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{Z}^T/(||\boldsymbol{Z}_r|| \cdot ||\boldsymbol{Z}_c^T||)$, where $||\boldsymbol{Z}_r||$ and $||\boldsymbol{Z}_c^T||$ compute the Euclidean norms for matrix $\boldsymbol{Z}$ and $\boldsymbol{Z}^T$'s row and column vectors respectively.

The self-correlations of the inferred objects are captured by the constructed $\boldsymbol{\Sigma}$'s diagonal elements and the mutual correlations are captured by $\boldsymbol{\Sigma}$'s off-diagonal elements. The goal is to reformulate the inference model resolve correlations so as to produce less-correlated $\{\boldsymbol{z}_k\}$. We penalize high off-diagonal values, i.e. by maximizing the L-NMS prior:

$$\mathcal{L}_{L-NMS}(\{\boldsymbol{z}_k\}; \Phi) = \sum_{h=1}^{K} \sum_{,j=1, h\neq j}^{K} \log \mathcal{N}(\boldsymbol{\Sigma}_{h,j};\, 0, \sigma^2). \tag{3}$$

The log normal density regulates its measure to certain range and σ (which models small variation in the correlation values) is 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 $\mathcal{L}_{ELBO}(\boldsymbol{x}; \Phi, \theta)$) w.r.t. the two trainable parameters Φ and θ. Taking a CompVAE

model, we thus train it by maximizing:

$$\mathcal{L}(\boldsymbol{x};\Phi,\theta) = \mathcal{L}_{ELBO}(\boldsymbol{x};\Phi,\theta) + \mathcal{L}_{L-NMS}(\{\boldsymbol{z}_k\};\Phi). \tag{4}$$

A general CompVAE ELBO can be defined using equation 1 and 2 as: $\mathcal{L}_{ELBO}(\boldsymbol{x};\Phi,\theta) = \mathbb{E}_{q_\Phi(\{\boldsymbol{z}_k\}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\{z_k\})] - D_{\mathrm{KL}}(q_\Phi(\{\boldsymbol{z}_k\}|\boldsymbol{x})|p_\theta(\{\boldsymbol{z}_k\}))$ but the exact formulations for a specific CompVAE is model-dependent. Note that although ELBOs are computed by the iterative inference processes of IODINE and MulMON during testing, we use the L-NMS priors only in training.

2.3 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 L-NMS prior) or the control group (original CompVAEs). To validate that after weakening the independence assumption, the obtained variational posterior $q_{\Phi^+}(\{\boldsymbol{z}_k\}|\boldsymbol{x})$ becomes a better approximation than $q_{\Phi^0}(\{\boldsymbol{z}_k\}|\boldsymbol{x})$ with respect to $p(\{\boldsymbol{z}_k\}|\boldsymbol{x})$, we need a measure to quantify approximation qualities and thus support model comparisons.

Through the derivation of VAEs' ELBO (Kingma & Welling, 2013), a gap between the observed evidence $\log p_\theta(x)$ and the ELBO $\mathcal{L}_{ELBO}(\boldsymbol{x};\Phi,\theta)$ is illustrated:

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

The gap shown in equation 5 is further decomposed by Cremer et al. (2018) into two items: the *variational approximation gap* and the *amortization gap*. In this paper, we are interested in only the former and the latter will not be discussed.

The approximation gap for a VAE is defined as: $\mathcal{G} = D_{\mathrm{KL}}(q^\star_\Phi(\boldsymbol{z}|\boldsymbol{x})\|p_\theta(\boldsymbol{z}|\boldsymbol{x}))$, where a superscript $\star$ indicates the optimum. This provides a quantitative measure of how good is an approximation when the optimal is reached: smaller denotes better and 0 is the smallest value. Similarly, we formulate $\mathcal{G} = D_{\mathrm{KL}}(q^\star_\Phi(\{\boldsymbol{z}_k\}|\boldsymbol{x})\|p_\theta(\{\boldsymbol{z}_k\}|\boldsymbol{x}))$ as the approximation gap 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.

Because $\mathcal{G}$ is not computable due to the inaccessibility of $\log p_\theta(\boldsymbol{x})$, to simplify the discussion hereafter, we define a measure *ELBO increment* (denoted as $\Delta\mathcal{L}^+$) using $\mathcal{G}^+$ and $\mathcal{G}^0$:

$$\Delta\mathcal{L}^+ = \mathcal{G}^0 - \mathcal{G}^+ = \mathcal{L}^\star_{ELBO}(\boldsymbol{x};\Phi^+,\theta) - \mathcal{L}^\star_{ELBO}(\boldsymbol{x};\Phi^0,\theta), \tag{6}$$

to tell directly how much the VAE approximation is improved w.r.t. to a model change, e.g. adding the L-NMS prior as in our case. In general, a positive $\Delta\mathcal{L}^+$ suggest a smaller gap is achieved and thus provides better approximation, a negative $\Delta\mathcal{L}^+$ suggests the opposite. In our experiments, we use $\Delta\mathcal{L}^+$ as an important metric for our model suboptimality analysis (see Section 4.2).

3 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 (Eslami et al., 2016) and its variants (Hsieh et al., 2018; Kosiorek et al., 2018) 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 (Burgess et al., 2019; Engelcke et al., 2020; Greff et al., 2019; Anon, 2020) 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 Lin et al. (2020b) as they all use the spatial mixture models (Williams & Titsias, 2004; Greff et al., 2017) to explain the image observations of scenes (see Eq.1 for an example). This allows the models to reason about depth and occlusions which are essential for 3D understanding.

Different from all the aforementioned works, our work targets an unsolved problem that is commonly seen in recent *scene-mixture* models, i.e. the inference models cannot resolve scene correlations and thus produce duplicate object representations. Although there are some unsupervised scene factorization models that handle the relations among the inferred objects, e.g. R-NEM (van

Steenkiste et al., 2018), STOVE (Kossen et al., 2020) and G-SWM (Lin et al., 2020a), they define "relations" as the interactions and scene dynamics of the scene objects and thus differs from what we are trying to solve in this paper.

The proposed work is related to the non-maximum-suppression (or duplicate-removal) idea that is widely used across many computer-vision tasks such as edge detection (Rosenfeld & Thurston, 1971) and feature extraction (Lowe, 2004). Among all the tasks it is applied, NMS's usage in object detection is the closest to ours, where duplicate detection candidates will be removed or suppressed (Rothe et al., 2014; Bodla et al., 2017) 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. For example, these techniques cannot handle the local minima case shown in Figure 1 (bottom right). In fact, such cases are no more a non-maximum suppression problem, it is related to the inference suboptimality of VAEs. Hence, other than a NMS problem, we deal with also the inference suboptimality (Cremer et al., 2018) caused by miss reasoning of the underlying scene correlations. There are other related works (Salimans et al., 2015; Mattei & Frellsen, 2018) that discuss the suboptimality in variational inference in a general manner. In the specific CompVAE cases, we only take the *approximation gap* of Cremer et al. (2018) to define a measure, i.e. the *ELBO increment* $\Delta\mathcal{L}^+$, and use it for model evaluation in our experiments.

4 Experiments

Our experiments are based on two datasets: CLE-MV (Anon, 2020) 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.). We show several data samples from the two data sets in Figure 2. 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 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. For the experiments, we use as baseline three CompVAE models, i.e. MONet, IODINE, and MulMON, and create our experimental group with the three CompVAEs trained with the proposed L-NMS prior. We train all models using the same training specifications as that of the experimental group except for removing the L-NMS 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 model and training specifications.

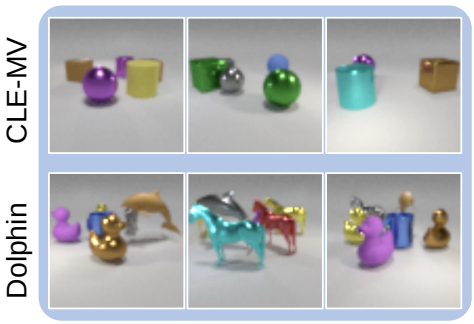

Figure 2: Data samples from the two datasets we use for our experiments.

		CLE-MV		Dolphin	
Models	L-NMS	MSE↓	mIoU↑	MSE↓	mIoU↑
MONet	0	0.0037 ± 0.0000	0.6806 ± 0.0039	0.0060 ± 0.0001	$\star\mathbf{0.6584 \pm 0.0044}$
	+	$\mathbf{0.0024 \pm 0.0000}$	$\mathbf{0.7899 \pm 0.0032}$	$\mathbf{0.0060 \pm 0.0001}$	0.6546 ± 0.0042
IODINE	0	$\mathbf{0.0016 \pm 0.0000}$	0.1907 ± 0.0007	0.0053 ± 0.0001	0.3475 ± 0.0030
	+	0.0020 ± 0.0001	$\mathbf{0.7256 \pm 0.0009}$	$\mathbf{0.0050 \pm 0.0001}$	$\mathbf{0.6257 \pm 0.0024}$
MulMON	0	0.0019 ± 0.0000	0.7823 ± 0.0010	0.0057 ± 0.0001	0.6266 ± 0.0024
	+	$\mathbf{0.0019 \pm 0.0000}$	$\mathbf{0.7903 \pm 0.0009}$	$\mathbf{0.0051 \pm 0.0001}$	$\mathbf{0.6565 \pm 0.0010}$

Table 1: Quantitative comparisons between the experimental group (tagged with "+") and the control group (tagged with "0"). All experiments are run across five different random seeds. $\star$ denotes the most significant case where L-NMS does not generate obvious improvements which we will discuss in the text.

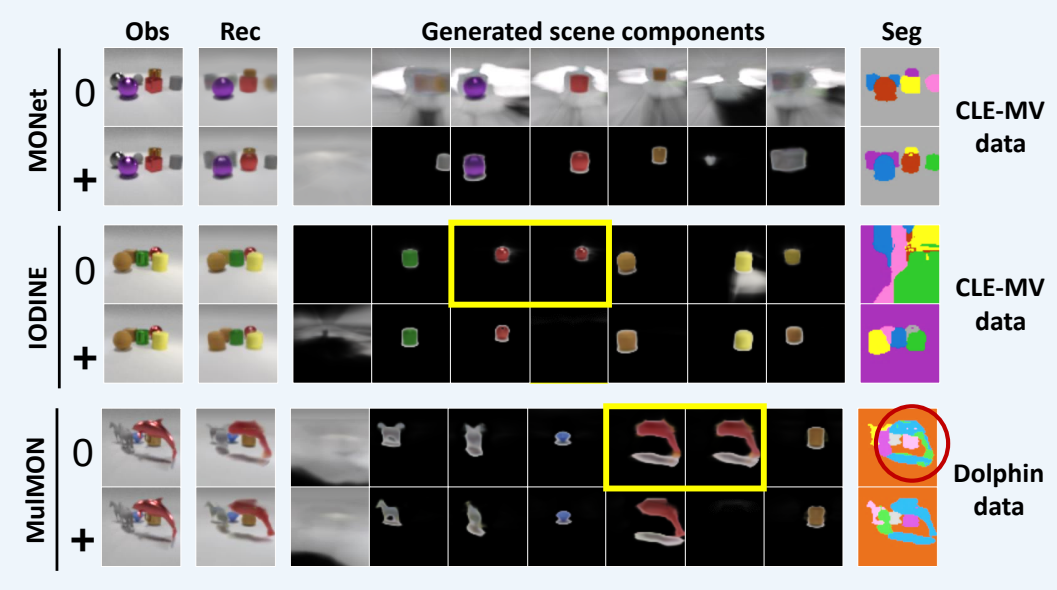

Figure 3: Qualitative comparisons between the experimental group (tagged with "+") and the control group (tagged with "0"). The Obs column is a source image, Rec is the corresponding reconstructed image based on the inferred representation. The next 7 columns show the independent generation of the inferred scene components. The Seg column shows the pixel label for the component with highest probability (the specific color of the pixel is not important). **Top** Training with the proposed L-NMS 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 the proposed L-NMS 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 Greff et al. (2019). **Bottom** Training with the proposed L-NMS allows MulMON to suppress duplicates and thus produce a better segmentation map. (Colored boxes and circles highlight the duplicates and failures caused by them.)

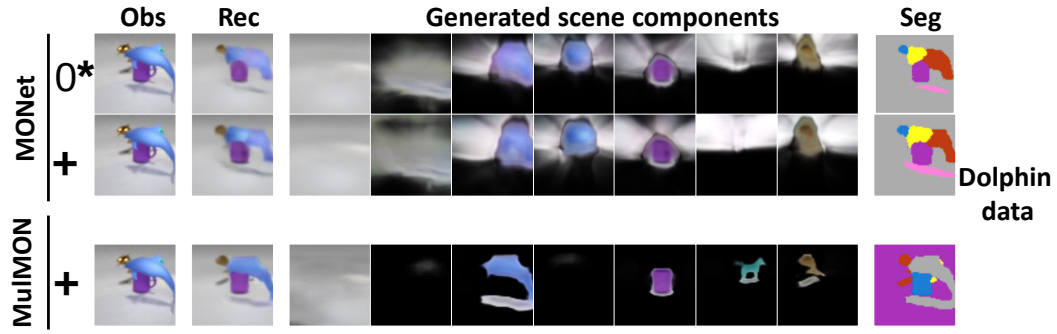

Figure 4: A partial-failure example from the "outlier" model (MONet^{0}) on Dolphin (tagged with "$\star$" in Table 1). **Top** The model produces good factorization but fails badly to learn good-quality object representations and thus show noisy generations. The proposed L-NMS fails to fix it. **Bottom**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

4.1 Task Performance

Scene Factorization The biggest advantage of CompVAEs over traditional VAEs in visual scene understanding is that they can handle unsupervised scene factorization. Therefore, we compare the scene object decomposition performance between the experimental group (CompVAEs trained with L-NMS) 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

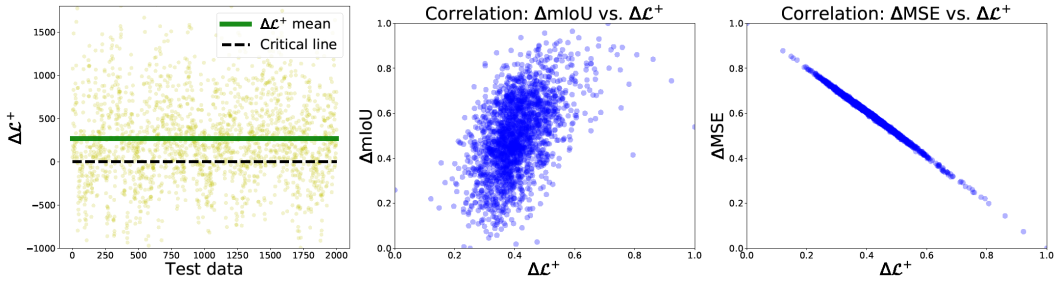

Figure 5: Results of the suboptimality analysis. **Left** Yellow dots represents the $\Delta\mathcal{L}^+$ for each test data sample (2000 test images), and the green line is the mean $\Delta\mathcal{L}^+$, which is the change in the ELBO (evidence lower bound) value from Eqn 6. Positive values are improvements. Observe that most dots lie above the "no improvement" line at 0, demonstrating that L-NMS generally produces improvements. **Middle** The correlation between scene factorization performance difference and the $\Delta\mathcal{L}^+$, which shows a close-to-linear positive correlation, i.e. bigger improvements in the ELBO measure correlate with better object overlap. **Right** The correlation between scene reconstruction performance difference and the $\Delta\mathcal{L}^+$, which shows a perfect linear negative correlation, i.e. improvements in ELBO mean better scene reconstruction.

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 use 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 L-NMS prior, results in similar or improved performance compared to the control group over all models and datasets. Figure 3 demonstrates the effectiveness of the proposed L-NMS prior in reducing duplicates and aiding CompVAEs' local minimas. We also examine the "outlier model", i.e. MONet$^+$ trained on Dolphin, and show some output samples in Figure 4. For the outlier model, even though the quantitative measures are achieved, the model still suffer from the local minima. We also consider this a failure instance of the proposed L-NMS as it does not aid the model like it does to MONet trained on the Dolphin dataset (see Figure 3).

Scene Reconstruction Reconstruction quality reflects the representation-learning quality of a VAE model. Hence, we compare 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 is computed from the RGB vector distances, where color values are on a $[0, 1]$ scale. Table 1 shows that the proposed L-NMS improves not only the scene factorization but also the scene reconstruction. This suggests the proposed L-NMS helps CompVAEs to learn better scene representations.

4.2 Suboptimality Analysis

The suboptimality analysis presented in this section gives a better understanding of how the proposed L-NMS helps to improve the task performance. The experiments illustrate the relationships between: **1)** the variational approximation gap and the proposed L-NMS, **2)** the task performance and the variational approximation gap. We first verify that the proposed L-NMS closes the variational approximation gap. We use the two MONet models trained on CLE-MV for the analysis. As discussed in Section 2.3, closing the variational gap is equivalent to obtaining a positive $\Delta\mathcal{L}^+$ (i.e. the *ELBO increment*). We use the 2000 test images from the CLE-MV dataset and compute the $\Delta\mathcal{L}^+$ for each of them and then average over 2000 samples to obtain the mean $\Delta\mathcal{L}^+$. Figure 5 (left) shows the $\Delta\mathcal{L}^+$ of these 2000 test samples and their mean. As illustrated by Figure 5 (left), MONet trained with the proposed L-NMS produces a positive mean $\Delta\mathcal{L}^+$, which reduces the variational approximation gap and is thus a better approximation than the original model.

To demonstrate the correlations between the task performance and the computed $\Delta\mathcal{L}^+$, we compute also the task performance differences between MONet$^+$ and MONet^0 (also denoted by Δ) for every sample of the CLE-MV test set. Note that we standardize both the task performance and the $\Delta\mathcal{L}^+$ to range $[0, 1]$ for visual clarity. Figure 5 (middle and right) show strong correlations (close-to-linear)

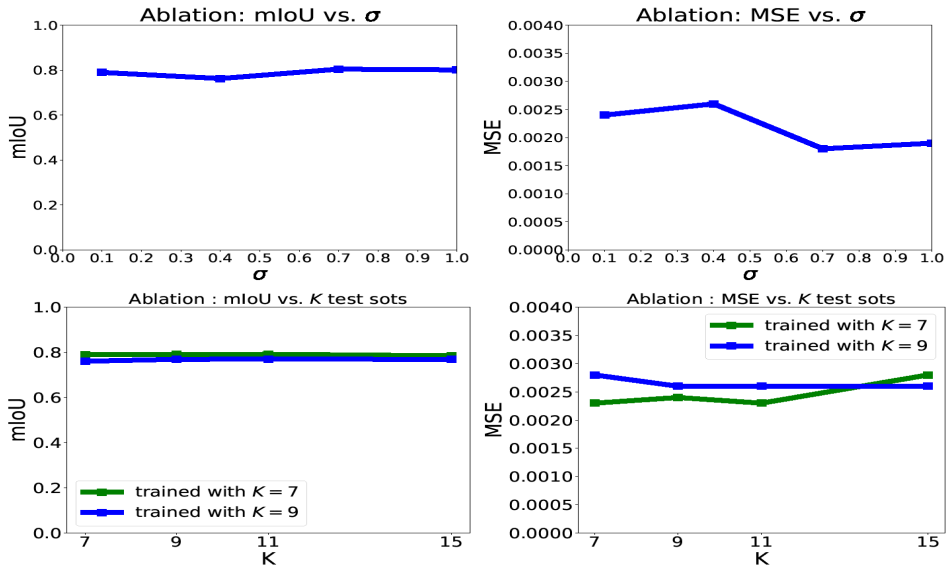

Figure 6: Ablation study results. **Top left** Scene observation reconstruction performance vs. L-NMS prior precision (σ). **Top right** Scene decomposition performance vs. L-NMS prior precision (σ). **Bottom left** Scene observation reconstruction performance vs. the number of object slots used in training and testing (K). **Bottom right** Scene decomposition performance vs. the number of object slots used in training and testing (K).

between the task performance difference and the $\Delta\mathcal{L}^+$, which indicates that the improvements are essentially driven by $\Delta\mathcal{L}^+$, i.e. reaching a better approximation suboptimum. Hence, based on our analysis, we conclude that the proposed L-NMS improves component inference performance by training a CompVAE to reach a better suboptimum, i.e. reduces the approximation gap.

4.3 Ablation Study

The ablation study focuses on two hyperparameters: 1) the standard deviation σ used in the L-NMS prior (see Section 2.2) and 2) the number of object slots K. The former relates to the precision of the correlation modeling and the latter determines the size of the correlation matrix constructed in L-NMS' computation, i.e. it relates to the scalability of L-NMS. 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 6 shows no significant performance loss in tasks by changing σ from the default value, 0.1, to other values. 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 6 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 clear performance drop.

5 Conclusion

In this work, we present a correlation prior 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 of the multi-object VAE models such as inferring duplicates, splitting background, etc. These problems are often related to the independence assumption made in CompVAEs which, as we consider, increases the approximation gap of VAEs or CompVAEs. We thus demonstrate through experiments that the proposed L-NMS solves most of the aforementioned problems by closing the approximation gap of CompVAEs and illustrate the correlations between the approximation gap and the task performance. Regarding the future work, we are particularly interested in basing correlation modeling on causal understanding, i.e. identifying explicitly the inter-object correlations' effect on each dimension of an object's latent representation.

REFERENCES

Anon. Anon. In *Advances in Neural Information Processing Systems*, 2020.

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pp. 5561–5569, 2017.

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.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1078–1086, 2018.

Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: Generative Scene Inference and Sampling of Object-Centric Latent Representations. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

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*, pp. 3225–3233, 2016.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *Advances in Neural Information Processing Systems*, pp. 6691–6701, 2017.

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 *Proceedings of the 36th International Conference on Machine Learning*, pp. 2424–2433, 2019.

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 *Advances in Neural Information Processing Systems*, pp. 517–526, 2018.

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

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*, pp. 8606–8616, 2018.

Jannik Kossen, Karl Stelzner, Marcel Hussing, Claas Voelcker, and Kristian Kersting. Structured object-aware physics prediction for video modeling and planning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1e-kxSKDH.

Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Multi-object representation learning with iterative variational inference. In *Proceedings of the 37th International Conference on Machine Learning*, 2020a.

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 *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=rkl03ySYDH.

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

Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the exact likelihood of deep latent variable models. In *Advances in Neural Information Processing Systems*, pp. 3855–3866, 2018.

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

Rasmus Rothe, Matthieu Guillaumin, and Luc Van Gool. Non-maximum suppression for object detection by passing messages between windows. In *Asian conference on computer vision*, pp. 290–306. Springer, 2014.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pp. 1218–1226, 2015.

Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryH20GbRW.

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.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

Training specifications We refer to Table 2, 3 & 4 to the training configurations of MONet, IODINE and MulMON respectively.

Table 2: Training Configurations For MONet

TYPE	THE TRAININGS OF MONET^0 AND MONET$^+$
OPTIMIZER	RMSPROP
INITIAL LEARNING RATE η_0	$3e^{-4}$
BATCH SIZE	8
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
L-NMS (MONET$^+$ ONLY) STRENGTH	0.5

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

TYPE	THE TRAININGS OF IODINE^0 AND IODINE$^+$
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 (GREFF ET AL., 2019)	5
LOG-NORMAL LIKELIHOOD STRENGTH	1.0
KL (GAUSSIAN PRIOR) STRENGTH β	1.0
L-NMS (IODINE$^+$ ONLY) STRENGTH	1.0
$\star$: SAME SCHEDULER AS GQNS'.	

Table 4: 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 (GREFF ET AL., 2019)	5
LOG-NORMAL LIKELIHOOD STRENGTH	1.0
KL (GAUSSIAN PRIOR) STRENGTH β	1.0
L-NMS (IODINE$^+$ ONLY) STRENGTH	1.0
$\star$: SAME SCHEDULER AS GQNS'.	

Model Architecture Specifications As discussed in the main paper, we use three existing Comp-VAE 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 that of (Greff et al., 2019) for the CLEVR-based datasets as it trains too slow, over one week for one run on two RTX2080TI, we thus reduce the dimension of IODINE to 16 for our IODINE. This is also the only difference of implementation to the original models. As constructing the proposed L-NMS prior requires no model architecture design and architecture parameter tweaking, we refer to the original papers of MONet (Burgess et al., 2019), IODINE (Greff et al., 2019), and MulMON (Anon, 2020) for the architecture details.