GENESIS-V2: Inferring Unordered Object Representations without Iterative Refinement

Anonymous Author(s)
Affiliation
Address
email

Abstract

Advances in object-centric generative models (OCGMs) have culminated in the development of a broad range of methods for unsupervised object segmentation and interpretable object-centric scene generation. These methods, however, are limited to simulated and real-world datasets with limited visual complexity. Moreover, object representations are often inferred using RNNs which do not scale well to large images or iterative refinement which avoids imposing an unnatural ordering on objects in an image but requires the *a priori* initialisation of a fixed number of object representations. In contrast to established paradigms, this work proposes an embedding-based approach in which embeddings of pixels are clustered in a differentiable fashion using a stochastic stick-breaking process. Similar to iterative refinement, this clustering procedure also leads to randomly ordered object representations, but without the need of initialising a fixed number of clusters *a priori*. This is used to develop a new model, GENESIS-V2, which can infer a variable number of object representations without using RNNs or iterative refinement. We show that GENESIS-V2 outperforms previous methods for unsupervised image segmentation and object-centric scene generation on established synthetic datasets as well as more complex real-world datasets.

1 Introduction

Reasoning about discrete objects in an environment is foundational to how agents perceive their surroundings and act in it. For example, autonomous vehicles need to identify and respond to other road users (e.g. [1, 2]) and robotic manipulation tasks involve grasping and pushing individual objects (e.g. [3]). While supervised methods can identify selected objects (e.g. [4, 5]), it is intractable to manually collect labels for every possible object category. Furthermore, we often desire the ability to predict, or *imagine*, how a collection of objects might behave (e.g. [6]). A range of works on object-centric generative models (OCGMs) have thus explored unsupervised segmentation and object-centric generation in recent years (e.g. [7–35]). These models are often formulated as variational autoencoders (VAEs) [36, 37] which allow the joint learning of inference and generation networks to identify objects in images and to generate scenes in an object-centric fashion (e.g. [14, 16, 27]).

Such models require a differentiable mechanism for separating objects in an image. While some works use spatial transformer networks (STNs) [38] to crop out windows that contain objects (e.g. [7–14]), others directly predict pixel-wise instance segmentation masks (e.g. [15–26]). The latter avoids the use of fixed-size sampling grids which are ill-suited for objects of varying size. Instead, object representations are inferred either by iteratively refining a set of randomly initialised representations (e.g. [18–23]) or by using a recurrent neural networks (RNN) (e.g. [15–17]). One particularly interesting model of the latter category is GENESIS [16] which performs both scene segmentation and generation by capturing relationships between objects with an autoregressive prior.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

As noted in Novotny et al. [39], however, using RNNs for instance segmentation requires processing high-dimensional inputs in a sequential fashion which is computationally expensive and does not scale well to large images with potentially many objects. We also posit that recurrent inference is not only problematic from a computational point of view, but that it can also inhibit the learning of object representations by imposing an unnatural *ordering* on objects. In particular, we argue that this leads to different *object slots* receiving gradients of varying magnitude which provides a possible explanation for OCGMs collapsing to a single object slot during training (see [17]). While iterative refinement instead infers unordered object representations, it requires the *a priori* initialisation of a fixed number of object slots even though the number of objects in an image is unknown.

In contrast, our work takes inspiration from the literature on supervised instance segmentation and adopts an *instance colouring* approach (e.g. [39–42]) in which pixel-wise embeddings—or colours—are clustered into attention masks. Typically, either a supervised learning signal is used to obtain cluster seeds (e.g. [39, 40]) or clustering is performed as a non-differentiable post-processing operation (e.g. [41, 42]). Neither of these approaches is suitable for unsupervised, end-to-end learning of segmentation masks. We hence develop an *instance colouring stick-breaking process* (IC-SBP) to cluster embeddings in a differentiable fashion. This is achieved by stochastically sampling cluster seeds from the pixel embeddings to perform a soft grouping of the embeddings into a set of randomly ordered attention masks. It is therefore possible to infer object representations both without imposing a fixed ordering or performing iterative refinement.

Inspired by GENESIS [16], we leverage the IC-SBP to develop GENESIS-V2, a novel OCGM that learns to segment objects in images without supervision and that uses an autoregressive prior to generate scenes in an interpretable, object-centric fashion. GENESIS-V2 is comprehensively benchmarked against recent prior art [15, 16, 23] on established synthetic datasets—ObjectsRoom [43] and ShapeStacks [44]—where it outperforms the baseline methods by significant margins. We also evaluate GENESIS-V2 on more challenging real-world images from the Sketchy [45] and the MIT-Princeton Amazon Picking Challenge (APC) 2016 Object Segmentation datasets [46], where it again achieves considerably better performance than previous methods. Code and pre-trained models will be released.

2 Related Work

OCGMs are typically formulated either as autoencoders (e.g. [7–27]) or generative adversarial networks (GANs) (e.g. [28–35]). Typical GANs are able to generate images, but lack an associated inference mechanism and often suffer from training instabilities (see e.g. [47, 48]). A comprehensive review and discussion of the subject is provided in Greff et al. [49]. STNs [38] can explicitly disentangle object location by cropping out a rectangular region from an input, allowing object appearance to be modelled in a canonical pose (e.g. [7–14]). This operation, however, relies on a fixed-size sampling grid which is not well-suited if objects vary broadly in terms of scale. In addition, gradients are usually obtained via bi-linear interpolation and are therefore limited to the extent of the sampling grid which can impede training: for example, if the sampling grid does not overlap with any object, then its location cannot be updated in a meaningful way. In contrast, purely segmentation based approaches [15–26]) often use RNNs (e.g. [15–17]) or iterative refinement (e.g. [18–23]) to infer object representations from an image. Other works either use a fixed number of slots [24, 25] or group pixels in a non-differentiable fashion [26]. RNN based models need to learn a fixed strategy that sequentially attends to different regions in an image, but this imposes an unnatural ordering on objects in an image. Avoiding such a fixed ordering leads to a *routing problem*. One way to address this is by randomly initialising a set of object representations and iteratively refining them. More broadly, this is also related to Deep Set Prediction Networks [50], where a set is iteratively refined in a gradient-based fashion. The main disadvantage of iterative refinement is that it is necessary to initialise a fixed number of clusters *a priori*, even though ideally we would like the number of clusters to be input-dependent. This is directly facilitated by the proposed IC-SBP.

Unlike some other works which learn unsupervised object representations from video sequences (e.g. [12, 13]), our work considers the more difficult task of learning such representations from individual images alone. GENESIS-V2 is most directly related to GENESIS [16] and SLOT-ATTENTION [23]. Like GENESIS, the model is formulated as a VAE to perform both object segmentation and object-centric scene generation, whereby the latter is facilitated by an autoregressive prior. Similar to SLOT-ATTENTION, the model uses a shared convolutional encoder to extract a feature map from which

features are pooled via an attention mechanism to infer object representations with a random ordering. In contrast to SLOT-ATTENTION, however, the attention masks are obtained with a parameter-free clustering algorithm that does not require iterative refinement or a predefined number of clusters. Both GENESIS and SLOT-ATTENTION are only evaluated on synthetic datasets. In this work, we use synthetic datasets for quantitative benchmarking, but we also perform experiments on two more challenging real-world datasets.

3 GENESIS-V2

An image $\mathbf{x}$ of height H, width W, with C channels, and pixel values in the interval $[0, 1]$ is considered to be a three-dimensional tensor $\mathbf{x} \in [0, 1]^{H \times W \times C}$. This work is only concerned with RGB images where $C = 3$, but other input modalities with a different number of channels could also be considered. Assigning individual pixels to object-like scene components can be formulated as obtaining a set of *object masks* $\pi \in [0, 1]^{H \times W \times K}$ with $\sum_k \pi_{i,j,k} = 1$ for all pixel coordinate tuples (i, j) in an image, where K is the number of scene components. This can be achieved by modelling the image likelihood $p_\theta(\mathbf{x}|\mathbf{z}_{1:K})$ as an SGMM (e.g. [15-24]) where θ are learnable parameters. Assuming pixels are independent and a fixed standard deviation σ_x that is shared across object slots, this likelihood can be written as

$$p_\theta(\mathbf{x}_{i,j} \mid \mathbf{z}_{1:K}) = \sum_{k=1}^{K} \pi_{i,j,k}(\mathbf{z}_{1:K})\, \mathcal{N}(\mu_{i,j}(\mathbf{z}_k),\, \sigma_x^2)\,. \tag{1}$$

The summation in Equation (1) implies that $p_\theta(\mathbf{x}_{i,j} \mid \mathbf{z}_{1:K})$ is *permutation-invariant* to the order of the object representations $\mathbf{z}_{1:K}$ (see e.g. [51, 52]) provided that $\pi_{i,j,k}(\mathbf{z}_{1:K})$ is also permutation-invariant, therefore allowing the generative model to accommodate for a possibly variable number of object representations.

To segment objects in images and to generate synthetic images in an object-centric fashion requires the formulation of appropriate inference and generative models, i.e. $q_\phi(\mathbf{z}_{1:K} \mid \mathbf{x})$ and $p_\theta(\mathbf{x} \mid \mathbf{z}_{1:K})\, p_\theta(\mathbf{z}_{1:K})$, respectively, where ϕ are also learnable parameters. In the generative model, it is necessary to model relationships between object representations to facilitate the generation of coherent scenes. Inspired by GENESIS [16], this is facilitated by an autoregressive prior

$$p_\theta(\mathbf{z}_{1:K}) = \prod_{k=1}^{K} p_\theta(\mathbf{z}_k \mid \mathbf{z}_{1:k-1})\,. \tag{2}$$

GENESIS uses two separate sets of latent variables to encode object masks and object appearances. In contrast, GENESIS-V2 uses a single set of latent variables $\mathbf{z}_{1:K}$ to encode both and increase parameter sharing. The graphical model of GENESIS-V2 is shown in comparison to other similar models in Appendix A.

While GENESIS relies on a recurrent mechanism in the inference model to predict segmentation masks, GENESIS-V2 instead infers latent variables without imposing a fixed ordering and assumes object latents $\mathbf{z}_{1:K}$ to be conditionally independent given an input image $\mathbf{x}$, i.e., $q_\phi(\mathbf{z}_{1:K}|\mathbf{x}) = \prod_k q_\phi(\mathbf{z}_k|\mathbf{x})$. Specifically, GENESIS-V2 first extracts an encoding with a deterministic UNet backbone. This encoding is used to predict a map of *semi-convolutional* pixel embeddings $\zeta \in \mathbb{R}^{H \times W \times D_\zeta}$ (see [39]). These are the inputs to the IC-SBP to obtain a set of normalised *attention masks* $\mathbf{m} \in [0, 1]^{H \times W \times K}$ with $\sum_k m_{i,j,k} = 1$ via a distance kernel ψ. Semi-convolutional embeddings are introduced in Novotny et al. [39] to facilitate the prediction of unique embeddings for multiple objects of identical appearance. The embeddings are obtained by the element-wise addition of relative pixel coordinates to two dimensions of the embeddings, which implicitly induces attention masks to be spatially localised. We leverage this structure to derive principled initialisations for the scaling-factor of different IC-SBP distance kernels ψ in Section 3.2. Inspired by Locatello et al. [23], GENESIS-V2 uses the attention masks $\mathbf{m}_{1:K}$ to pool a feature vector for each scene component from the deterministic image encoding. A set of object latents $\mathbf{z}_{1:K}$ is then computed from these feature vectors. This set of latents is decoded in parallel to compute the statistics of the SGMM in Equation (1). The object masks π are normalised with a softmax operator. The inference and generation models can be jointly trained as a VAE as illustrated in Figure 1. Further architecture details are described in Appendix B.

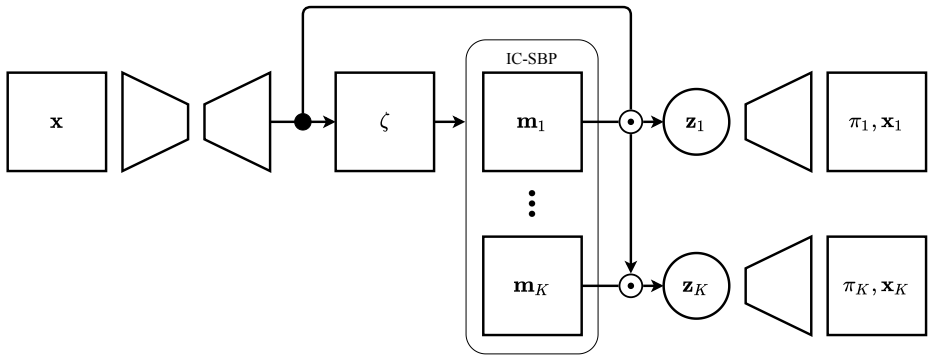


Figure 1: GENESIS-V2 overview. The image $\mathbf{x}$ is passed into a deterministic backbone. The resulting encoding is used to compute the pixel embeddings ζ which are clustered into attention masks $\mathbf{m}_{1:K}$ by the IC-SBP. Features are pooled according to these attention masks to infer the object latents $\mathbf{z}_{1:K}$. These are decoded into the object masks $\pi_{1:K}$ and reconstructed components $\mathbf{x}_{1:K}$.

3.1 Instance Colouring Stick-Breaking Process

The IC-SBP is a stochastic, differentiable algorithm that clusters pixel embeddings $\zeta \in \mathbb{R}^{H \times W \times D_\zeta}$ into a variable number of soft attention masks $\mathbf{m} \in [0,1]^{H \times W \times K}$. This is achieved by stochastically selecting pixel embeddings as cluster seeds, which are used to group pixels into a set of *randomly ordered* soft clusters. Inspired by Burgess et al. [15], a *scope* $\mathbf{s} \in [0,1]^{H \times W}$ is initialised to a matrix of ones $\mathbb{1}^{H \times W}$ to track the degree to which pixels have been assigned to clusters and a matrix of *seed scores* is initialised by sampling from a uniform distribution $\mathbf{c} \sim U(0,1) \in \mathbb{R}^{H \times W}$. These scores determine which embeddings are selected as cluster seeds. A single embedding vector $\zeta_{i,j}$ is selected at the spatial location (i,j) which corresponds to the argmax of the element-wise multiplication of the seed scores and the current scope. This ensures that cluster seeds are sampled from pixel embeddings that have not yet been assigned to clusters. An alpha mask $\alpha_k \in [0,1]^{H \times W}$ is computed as the distance between the cluster seed embedding $\zeta_{i,j}$ and all individual pixel embeddings according to a distance kernel ψ. The output of the kernel ψ is one if two embeddings are identical and decreases to zero as the distance between a pair of embeddings increases. The associated attention mask $\mathbf{m}_k$ is obtained by the element-wise multiplication of the alpha masks by the current scope to ensure that the final set of attention masks is normalised. The scope is then updated by an element-wise multiplication with the complement of the alpha masks and the process is repeated until a stopping condition is satisfied. At this point, the final scope is added as an additional mask to explain any remaining pixels. This is described formally in Algorithm 1.

Algorithm 1: Instance Colouring Stick-Breaking Process

Input: embeddings $\zeta \in \mathbb{R}^{H \times W \times D_\zeta}$
Output: masks $\mathbf{m}_{1:K}$ with $\mathbf{m}_k \in [0,1]^{H \times W}$
Initialise: masks $\mathbf{m} = \varnothing$, scope $\mathbf{s} = \mathbb{1}^{H \times W}$, seed scores $\mathbf{c} \sim U(0,1) \in \mathbb{R}^{H \times W}$

while *not StopCondition*($\mathbf{m}$) **do**
  i, j = argmax($\mathbf{s} \odot \mathbf{c}$);
  α = DistanceKernel(ζ, $\zeta_{i,j}$);
  $\mathbf{m}$.append($\mathbf{s} \odot \alpha$);
  $\mathbf{s} = \mathbf{s} \odot (1 - \alpha)$;
end
$\mathbf{m}$.append($\mathbf{s}$)

When the masks are non-binary, several executions of the algorithm lead to different sets of attention masks, whereby the attention mask for a certain object will also be slightly different between executions. If the mask values are discrete and exactly equal to either zero or one, however, the set of cluster seeds and the set of attention masks are uniquely defined apart from their ordering. This can be seen from the fact that if at each step of the IC-SBP produces a binary mask, then embeddings associated with this mask can never be sampled again as cluster seeds due to the masking of the seed

scores by the scope. A different cluster with an associated binary mask is therefore created at every step until all embeddings are uniquely clustered. This could possibly be facilitated, for example, by applying recurrent mean-shift clustering [53] to the embeddings before running the IC-SBP. Another interesting modification would be to make the output of the IC-SBP *permutation-equivariant* with respect to the ordering of the cluster seeds. This could be achieved either by directly using the cluster seeds for downstream computations or by separating the mask normalisation from the stochastic seed selecting; while the SBP formulation facilitates the selection of a diverse set of cluster seeds, the masks could be normalised separately after the cluster seeds are selected by using a softmax operation or similar. The exploration of such modifications, however, is left for future work.

In contrast to GENESIS, the stochastic ordering of the masks implies that it is not possible for GENESIS-V2 to learn a fixed sequential decomposition strategy. While this does not strictly apply to the last mask which is set equal to the remaining scope, we find empirically that models learn a strategy where the final scope is either largely unused or where it corresponds to a generic background cluster with foreground objects remaining unordered as desired. Unlike as in iterative refinement where a fixed number of clusters needs to be initialised *a priori*, the IC-SBP can infer a variable number of object representations by using a heuristic that considers the current set of attention masks at every iteration in Algorithm 1. While we use a fixed number of K masks during training for efficient parallelism on GPU accelerators, we demonstrate that a flexible number of masks can be extracted at test time with minimal impact on segmentation performance.

3.2 Kernel Initialisation with Semi-Convolutional Embeddings

For semi-convolutional embeddings to be similar according to a kernel ψ, the model needs to predict a delta vector for each embedding to a specific pixel location to negate the additive relative pixel coordinates. The model could for example learn to predict a delta vector for each pixel that describes the location of the pixel relative to the centre of the object that it belongs to. Conversely, if the predicted delta vectors are small, then clustering the embeddings based on their distance to each other only depends on the relative pixel coordinates which results in spatially localised masks. This structure implies that it is possible to derive a meaningful initialisation for free parameters in the distance kernel ψ of the IC-SBP. Established candidates for ψ from the literature are the Gaussian ψ_G [39], Laplacian ψ_L [39], and Epanechnikov ψ_E kernels [53] with

$$\psi_G = \exp\left(-\frac{||\mathbf{u}-\mathbf{v}||^2}{\sigma_G}\right), \quad \psi_L = \exp\left(-\frac{||\mathbf{u}-\mathbf{v}||}{\sigma_L}\right), \quad \psi_E = \max\left(1 - \frac{||\mathbf{u}-\mathbf{v}||^2}{\sigma_E}, 0\right), \tag{3}$$

whereby $\mathbf{u}$ and $\mathbf{v}$ are two embeddings of equal dimension. Each kernel contains a scaling factor $\sigma \in \mathbb{R}^+$ which can be initialised so that attention masks at the beginning of training correspond to approximately equally-sized circular patches. In particular, we initialise these scaling factors as

$$\sigma_G^{-1} = K \ln 2, \quad \sigma_L^{-1} = \sqrt{K} \ln 2, \quad \sigma_E^{-1} = K/2\,, \tag{4}$$

which is derived and illustrated in Appendix C. After initialisation, σ is jointly optimised along with the other learnable parameters of the model.

3.3 Training

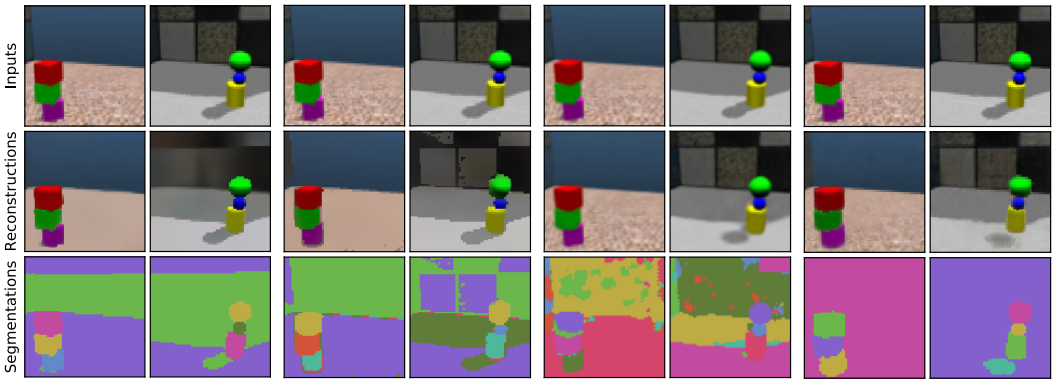
Following Engelcke et al. [16], GENESIS-V2 is trained by minimising the GECO objective [54], which can be written as a loss function of the form

$$\mathcal{L}_g = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\ln p_\theta(\mathbf{x} \mid \mathbf{z})] + \beta_g \cdot \mathrm{KL}[q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p_\theta(\mathbf{z})]\,. \tag{5}$$

The relative weighting factor $\beta_g \in \mathbb{R}^+$ is updated at every training iteration separately from the model parameters according to

$$\beta_g = \beta_g \cdot \mathrm{e}^{\eta(C-E)} \qquad \text{with} \qquad E = \alpha_g \cdot E + (1-\alpha_g) \cdot \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\ln p_\theta(\mathbf{x} \mid \mathbf{z})]\,. \tag{6}$$

$E \in \mathbb{R}$ is an exponential moving average of the negative image log-likelihood, $\alpha_g \in [0, 1]$ is a momentum factor, $\eta \in \mathbb{R}^+$ is a step size hyperparameter, and $C \in \mathbb{R}$ is a target reconstruction error. Intuitively, the optimisation decreases the weighting of the KL regularisation term as long as the reconstruction error is larger than the target C. The weighting of the KL term is increased again once the target is satisfied.



(a) MONET-G (b) GENESIS (c) SLOT-ATTENTION (d) GENESIS-V2

Figure 2: GENESIS-V2 learns better reconstructions and segmentations on ShapeStacks.

We also conduct experiments with an additional auxiliary mask consistency loss that encourages attention masks $\mathbf{m}$ and object segmentation masks π to be similar which might be desirable in some applications. This leads to a modified loss function of the form

$$\mathcal{L}'_g = E + \beta_g \cdot \left(\mathrm{KL}[q_\phi(\mathbf{z} \mid \mathbf{x}) \mid\mid p_\theta(\mathbf{z})] + \mathrm{KL}[\mathbf{m} \mid\mid \mathrm{nograd}(\pi)] \right), \tag{7}$$

in which $\mathbf{m}$ and π are interpreted as pixel-wise categorical distributions. Preliminary experiments indicated that stopping the gradient propagation through the object masks π helps to achieve segmentation quality comparable to using the original loss function in Equation (5).

4 Experiments

This section presents results on two simulated datasets—ObjectsRoom [43] and ShapeStacks [44]—as well as two real-world datasets—Sketchy [45] and APC [46]—which are described in Appendix D. GENESIS-V2 is compared against three recent baselines: GENESIS [16], MONET [15], and SLOT-ATTENTION [23]. Even though SLOT-ATTENTION is trained with a pure reconstruction objective and is not a generative model, it is an informative and strong baseline for unsupervised scene segmentation. The other models are trained with the GECO objective [54] following the protocol from Engelcke et al. [16] for comparability. We refer to MONET trained with GECO as MONET-G to avoid conflating the results with the original settings. Further training details are described in Appendix E. Following prior art (e.g. [16, 17, 21, 23]), segmentation quality is quantified using the Adjusted Rand Index (ARI) and the Mean Segmentation Covering (MSC) of pixels belonging to foreground objects. These are averaged across 320 images from the respective test sets. Similarly, generation quality is measured using the Fréchet Inception Distance (FID) [55] which is computed from 10,000 samples and 10,000 test set images using the implementation from Seitzer [56]. When models are trained with multiple random seeds, we always show qualitative results for the seed that achieves the highest ARI.

4.1 Benchmarking in Simulation

Figure 2 shows qualitative results on ShapeStacks and additional qualitative results are included in Appendix F. Unlike the baselines, GENESIS-V2 cleanly separates the foreground objects from the background and the reconstructions. Moreover, Table 1 summarises quantitative segmentation results on ObjectsRoom and ShapeStacks. Here, GENESIS-V2 outperforms GENESIS and MONET-G across datasets on all metrics, showing that the IC-SBP is indeed suitable for learning object-centric representations. GENESIS-V2 outperforms SLOT-ATTENTION in terms of the ARI on both datasets. SLOT-ATTENTION manages to achieve a better mean MSC, but the variation of those metrics is much larger, which indicates training is not as stable. While the ARI indicates the models ability to separate objects, it does not penalise the undersegmentation of objects (see [16]). The MSC, in contrast, is an IOU based metric and sensitive to the exact segmentation masks. We conjecture that SLOT-ATTENTION manages to predict slightly more accurate segmentation masks given that is trained on a pure reconstruction objective and without KL regularisation, thus leading to a slightly better mean MSC. A set of ablations for GENESIS-V2 is also included in Appendix F.

Table 1: Means and standard deviations of the segmentation metrics from three seeds.

Model	Generative	ObjectsRoom		ShapeStacks	
		ARI-FG	MSC-FG	ARI-FG	MSC-FG
MONET-G	Yes	0.54±0.00	0.33±0.01	0.70±0.04	0.57±0.12
GENESIS	Yes	0.63±0.03	0.53±0.07	0.70±0.05	0.67±0.02
GENESIS-V2	Yes	**0.84±0.01**	**0.58±0.03**	**0.81±0.00**	**0.68±0.01**
SLOT-ATTENTION	No	0.79±0.02	**0.64±0.13**	0.76±0.01	**0.70±0.05**

Table 2: Means and standard deviations of the segmentation metrics on ShapeStacks from three seeds for GENESIS-V2 with a fixed or flexible number of object slots.

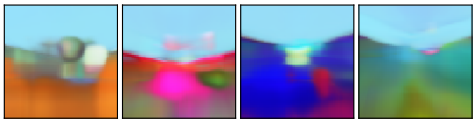
Training	Slots	Avg. K	ARI-FG	MSC-FG
No mask loss	Fixed	9.0±0.0	**0.81±0.00**	**0.68±0.01**
	Flexible	**6.3±0.4**	0.77±0.00	0.63±0.01
With mask loss	Fixed	9.0±0.0	**0.81±0.01**	**0.68±0.01**
	Flexible	**5.6±0.1**	**0.81±0.00**	0.67±0.01

We also examine whether the IC-SBP can indeed be used to extract a variable number of object representations. This is done by terminating the IC-SBP according to a manual heuristic and setting the final mask to the remaining scope. Specifically, we terminate the IC-SBP when the sum of the current attention mask values is smaller than twenty pixels. The results when using a fixed number and a variable number slots after training GENESIS-V2 on ShapeStacks are summarised in Table 2. With the original training objective in Equation (5), this heuristic only requires 6.3 rather than nine object slots on average, but also leads to a decrease in segmentation performance. When using the auxiliary mask loss as in Equation (7), only 5.6 object slots are needed on average at almost the same segmentation accuracy. This heuristic works well on ShapeStacks, but it did not perform well on ObjectsRoom. Here, the models sometimes predict non-empty attention masks which are associated with empty object segmentation masks, thus making it impossible to terminate the IC-SBP appropriately. Even though this should be overcome by the mask consistency loss, we found that this auxiliary loss term often caused models to get stuck in local minima where images are segmented into vertical stripes. These limitations can hopefully be addressed in future work.

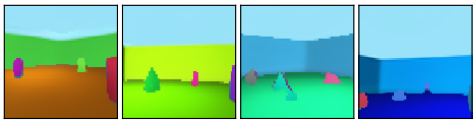
Regarding scene generation, Table 3 summarises the FID scores for GENESIS-V2 and the baselines on ObjectsRoom and ShapeStacks. GENESIS-V2 consistently performs best with a particularly significant improvement on ShapeStacks. Results on ShapeStacks, however, are not as good as for ObjectsRoom, which is likely caused by the increased visual complexity of the images in the ShapeStacks dataset. Qualitative results for scene generation are shown in Figures 3 and 4. Both GENESIS and GENESIS-V2 produce reasonable samples after training on ObjectsRoom. For ShapeStacks, samples from GENESIS contain significant distortions. In comparison, samples from GENESIS-V2 are more realistic, but also still show room for improvement.

Table 3: Means and standard deviations of FID scores from three seeds.

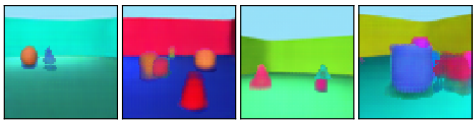
Model	ObjectsRoom	ShapeStacks
MONET-G	205.7±7.6	197.8±5.2
GENESIS	62.8±2.5	186.8±18.0
SLOT-ATT.	—	—
GENESIS-V2	**52.6±2.7**	**112.7±3.2**



(a) MONET-G



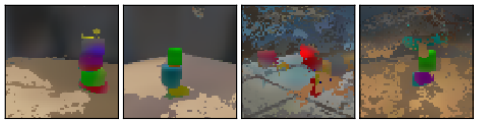
(b) GENESIS



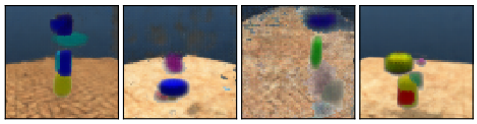
(c) GENESIS-V2

Figure 3: ObjectsRoom samples.

(a) MONET-G

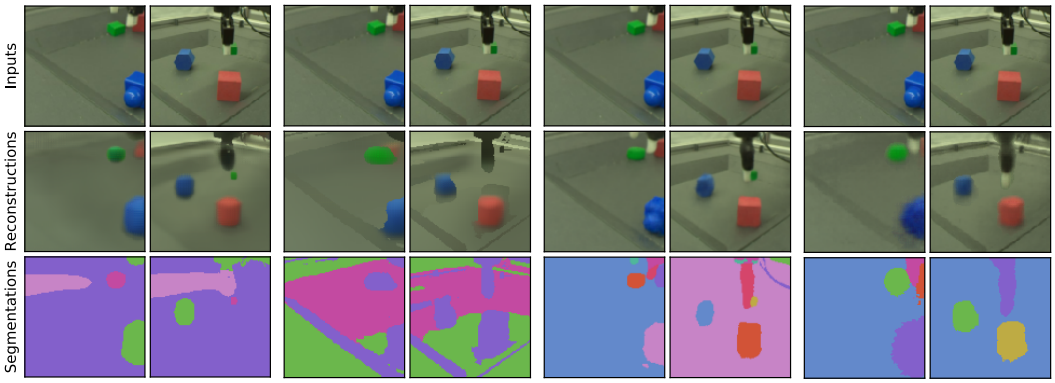


(b) GENESIS



(c) GENESIS-V2

Figure 4: ShapeStacks samples.



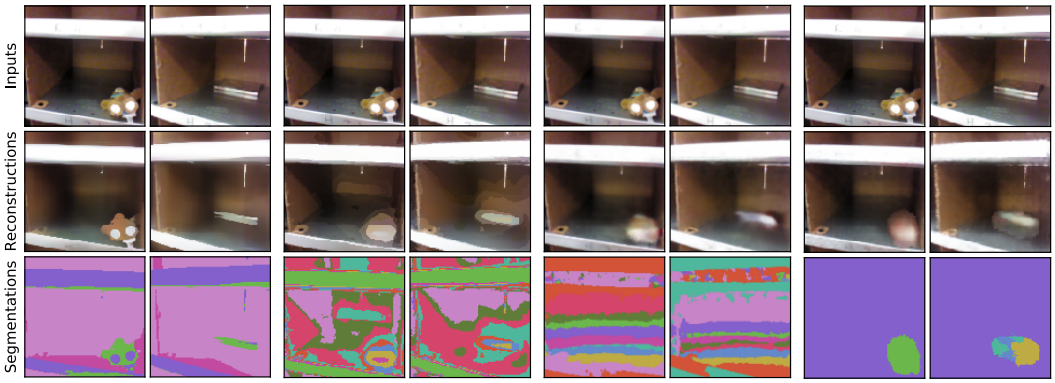
(a) MONET-G (b) GENESIS (c) SLOT-ATTENTION (d) GENESIS-V2

Figure 5: In contrast to MONET-G and GENESIS, GENESIS-V2 as well as SLOT-ATTENTION are able to learn reasonable object segmentations on the more challenging Sketchy dataset.

4.2 Real-World Applications

After validating GENESIS-V2 on two simulated datasets, this section present experiments on the Sketchy [45] and APC datasets [46]; two significantly more challenging real-world datasets collected in the context of robot manipulation. Reconstructions and segmentations after training GENESIS-V2 on Sketchy and APC are shown in Figures 5 and 6. For Sketchy, it can be seen that GENESIS-V2 and SLOT-ATTENTION disambiguate the individual foreground objects and the robot gripper fairly well. SLOT-ATTENTION produces slightly more accurate reconstructions, which is likely facilitated by the pure reconstruction objective that the model is trained with. However, GENESIS-V2 is the only model that separates foreground objects from the background in APC images. Environment conditions in Sketchy are highly controlled and SLOT-ATTENTION appears to be unable to handle the more complex conditions in APC images. Nevertheless, GENESIS-V2 also struggles to capture the fine-grained details and oversegments one of the foreground objects into several parts, leaving room for improvement in future work. Additional qualitative results are included in Appendix G.

There are no ground-truth object masks available for Sketchy, but segmentation performance can be evaluated quantitatively on APC for which annotations are available with the caveat that there is only a single foreground object per image. In this case, when computing the segmentation metrics from foreground pixels alone, the ARI could be trivially maximised by assigning all image pixels



(a) MONET-G (b) GENESIS (c) SLOT-ATTENTION (d) GENESIS-V2

Figure 6: GENESIS-V2 is the only model that separates the foreground objects in images from APC.

Table 4: Segmentation metrics on APC.

	ARI	MSC
MONET-G	0.11	0.48
GENESIS	0.04	0.29
SLOT-ATT.	0.03	0.25
GENESIS-V2	**0.55**	**0.67**

Table 5: FID scores on Sketchy and APC.

	Sketchy	APC
MONET-G	294.3	269.3
GENESIS	241.9	**183.2**
SLOT-ATT.	—	—
GENESIS-V2	**208.1**	245.6

to the same component and the MSC will be equal to the largest IOU of the predicted masks with the foreground objects. When considering all pixels, the optimal solution for both metrics is to have exactly one set of pixels assigned to the foreground object and another set of pixels being assigned to the background. Table 4 therefore reports ARI and MSC scores computed from all pixels GENESIS-V2 and the baselines. GENESIS-V2 performs much better than the baselines on both metrics, corroborating that GENESIS-V2 takes a step in the right direction in terms of learning unsupervised object-representations from *real-world* datasets.

FID scores for generated images are summarised in Table 5 and qualitative results are included in Appendix G. GENESIS-V2 achieves the best FID on Sketchy, but it is outperformed by GENESIS on APC. Both models consistently outperform MONET-G. All of the FID scores, however, are fairly large which is not surprising given the much higher visual complexity of these images. It is therefore difficult to draw strong conclusions from these beyond a rough sense of sample quality. Further work is required to generate high-fidelity images after training on real-world datasets.

5 Conclusions

This work develops GENESIS-V2, a novel object-centric latent variable model of scenes which is able to both decompose visual scenes into semantically meaningful constituent parts while at the same time being able to generate coherent scenes in an object-centric fashion. GENESIS-V2 leverages a differentiable clustering algorithm for grouping pixel embeddings into a variable number of attention masks which are used to infer an unordered set of object representations. This approach is validated empirically on two established simulated datasets as well as two additional real-world datasets. The results show that GENESIS-V2 takes a step towards learning better object-centric representations without labelled supervision from real-world datasets. Yet, there is still room for improvement in terms of reconstruction, segmentation, and sample quality. It would also be interesting to investigate the ability of GENESIS-V2 and the IC-SBP to handle out-of-distribution images that contain more objects than seen during training.

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Coline Devin, Pieter Abbeel, Trevor Darrell, and Sergey Levine. Deep Object-Centric Representations for Generalizable Robot Learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *International Conference on Computer Vision (ICCV)*, 2017.

[6] Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. Greedy Hierarchical Variational Autoencoders for Large-Scale Video Prediction. *arXiv preprint arXiv:2103.04174*, 2021.

[7] Jonathan Huang and Kevin Murphy. Efficient Inference in Occlusion-Aware Generative models of Images. *arXiv preprint arXiv:1511.06362*, 2015.

[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 (NeurIPS)*, 2016.

[9] Kun Xu, Chongxuan Li, Jun Zhu, and Bo Zhang. Multi-Objects Generation with Amortized Structural Regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[10] Eric Crawford and Joelle Pineau. Spatially Invariant Unsupervised Object Detection with Convolutional Neural Networks. In *AAAI Conference on Artificial Intelligence*, 2019.

[11] 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 (ICLR)*, 2020.

[12] 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 (NeurIPS)*, 2018.

[13] Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. SCALOR: Generative World Models with Scalable Object Representations. In *International Conference on Learning Representations (ICLR)*, 2020.

[14] Jindong Jiang and Sungjin Ahn. Generative Neurosymbolic Machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] 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.

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

[17] Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. Reconstruction Bottlenecks in Object-Centric Generative Models. *ICML Workshop on Object-Oriented Learning*, 2020.

[18] Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hao, Harri Valpola, and Jürgen Schmidhuber. Tagger: Deep Unsupervised Perceptual Grouping. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[19] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Neural Expectation Maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[20] Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational Neural Expectation Maximization: Unsupervised Discovery of Objects and their Interactions. *arXiv preprint arXiv:1802.10353*, 2018.

[21] 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. In *International Conference on Machine Learning (ICML)*, 2019.

[22] Rishi Veerapaneni, John D Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua Tenenbaum, and Sergey Levine. Entity Abstraction in Visual Model-Based Reinforcement Learning. In *Conference on Robot Learning (CoRL)*, 2020.

[23] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[24] Adam R Kosiorek, Sara Sabour, Yee Whye Teh, and Geoffrey E Hinton. Stacked Capsule Autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[25] Yanchao Yang, Yutong Chen, and Stefano Soatto. Learning to Manipulate Individual Objects in an Image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[26] Daniel M Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li Fei-Fei, Jiajun Wu, Joshua B Tenenbaum, et al. Learning Physical Graph Representations from Visual Scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[27] Titas Anciukevicius, Christoph H Lampert, and Paul Henderson. Object-Centric Image Generation with Factored Depths, Locations, and Appearances. *arXiv preprint arXiv:2004.00642*, 2020.

[28] Sjoerd van Steenkiste, Karol Kurach, and Sylvain Gelly. A Case for Object Compositionality in Deep Generative Models of Images. *NeurIPS Workshop on Modeling the Physical World: Learning, Perception, and Control*, 2018.

[29] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised Object Segmentation by Redrawing. *arXiv preprint arXiv:1905.13539*, 2019.

[30] Adam Bielski and Paolo Favaro. Emergence of Object Segmentation in Perturbed Generative Models. *arXiv preprint arXiv:1905.12663*, 2019.

[31] Relja Arandjelović and Andrew Zisserman. Object Discovery with a Copy-Pasting GAN. *arXiv preprint arXiv:1905.11369*, 2019.

[32] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional GAN: Learning Image-Conditional Binary Composition. *arXiv preprint arXiv:1807.07560*, 2019.

[33] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. BlockGAN: Learning 3D Object-aware Scene Representations from Unlabelled Images. *arXiv preprint arXiv:2002.08988*, 2020.

[34] Sebastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner, Niloy Mitra, and Andrea Vedaldi. RELATE: Physically Plausible Multi-Object Scene Synthesis Using Structured Latent Spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[35] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. *arXiv preprint arXiv:2011.12100*, 2020.

[36] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[37] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on Machine Learning (ICML)*, 2014.

[38] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[39] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Semi-Convolutional Operators for Instance Segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

[40] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin P Murphy. Semantic Instance Segmentation via Deep Metric Learning. *arXiv preprint arXiv:1703.10277*, 2017.

[41] Min Bai and Raquel Urtasun. Deep Watershed Transform for Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[42] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic Instance Segmentation with a Discriminative Loss Function. *arXiv preprint arXiv:1708.02551*, 2017.

[43] Rishabh Kabra, Chris Burgess, Loic Matthey, Raphael Lopez Kaufman, Klaus Greff, Malcolm Reynolds, and Alexander Lerchner. Multi-Object Datasets, 2019. URL https://github.com/deepmind/multi-object-datasets/.

[44] Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. ShapeStacks: Learning Vision-Based Physical Intuition for Generalised Object Stacking. In *European Conference on Computer Vision (ECCV)*, 2018.

[45] Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, et al. Scaling Data-Driven Robotics with Reward Sketching and Batch Reinforcement Learning. *arXiv preprint arXiv:1909.12200*, 2019.

[46] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker Jr, Alberto Rodriguez, and Jianxiong Xiao. Multi-view Self-supervised Deep Learning for 6D Pose Estimation in the Amazon Picking Challenge. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[47] 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 (NeurIPS)*, 2014.

[48] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[49] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the Binding Problem in Artificial Neural Networks. *arXiv preprint arXiv:2012.05208*, 2020.

[50] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Deep Set Prediction Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[51] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep Sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[52] Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A Osborne. On the Limitations of Representing Functions on Sets. In *International Conference on Machine Learning (ICML)*, 2019.

[53] Shu Kong and Charless C Fowlkes. Recurrent Pixel Embedding for Instance Grouping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[54] Danilo Jimenez Rezende and Fabio Viola. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.

[55] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[56] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.1.1.

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[58] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[59] Yuxin Wu and Kaiming He. Group Normalization. In *European Conference on Computer Vision (ECCV)*, 2018.

[60] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[61] Nicholas Watters, Loic Matthey, Christopher P Burgess, and Alexander Lerchner. Spatial Broadcast Decoder: A Simple Architecture for Learning Disentangled Representations in VAEs. *arXiv preprint arXiv:1901.07017*, 2019.

[62] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 1997.

[63] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[64] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic Differentiation in PyTorch. 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Limitations with respect to reconstruction, segmentation, and generation quality as well as with respect to using the IC-SBP for inferring a variable number of object representations are discussed in the experimental section and the conclusions.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix H.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Training and evaluation code as well as pre-trained model checkpoints will be provided with the camera-ready version.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Experimental settings are described in sufficient detail that should enable the independent reproduction of the results, but we do not always describe in detail how they were chosen—some choices are based on practitioner's experience, some are based on manual tuning, etc.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Means and standard deviations from three random seeds are reported for experiments on the simulated datasets. Experiments on the real-world datasets were only conducted with a single random seed as they are considerably more computation intensive.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] Approximate runtimes on a specified GPU type are described in Appendix E, but we did not track the total amount of compute that was used for this research project.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix I.
   (b) Did you mention the license of the assets? [Yes] See Appendix I.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] Training and evaluation code as well as pre-trained model checkpoints will be provided with the camera-ready version.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix I.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix I.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]