Generative Modeling of Infinite Occluded Objects for Compositional Scene Representation (Supplementary Material)

Jinyang Yuan [1] **Bin Li** [1] **Xiangyang Xue** [1]

1. Details of the Lower Bound of ELBO in Section 3.4

We apply the multinomial approximation to $\mathbb{E}_{\boldsymbol{\nu}}[\log(1-\prod_{k'=1}^{k}\nu_{k'})]$ as mentioned in (Doshi et al., 2009) and optimize parameters of neural networks by maximizing a lower bound $\hat{\mathcal{L}}$ of the evidence lower bound (ELBO) $\mathcal{L}$. Detailed expression of $\hat{\mathcal{L}}$ is given by

$$
\begin{aligned}
\hat{\mathcal{L}} = &-\frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{s}}\left[\log 2\pi+\log\hat{\sigma}^2+\frac{1}{\hat{\sigma}^2}\sum_{k=0}^{K}\gamma_{n,k}(\boldsymbol{x}_n-\boldsymbol{a}_{n,k})^2\right]\\
&-\frac{1}{2}\sum_{k=0}^{K}\sum_{m}\mathbb{E}_{\boldsymbol{s}^{\text{stn}}_{\cdot k}}\left[\log\frac{\tilde{\sigma}_m^2}{\sigma_{k,m}^2}+\frac{(\mu_{m,k}-\tilde{\mu}_m)^2+\sigma_{m,k}^2}{\tilde{\sigma}_m^2}-1\right]\\
&-\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{s}^{\text{stn}}_{\cdot k}}\left[\log\frac{\Gamma(\tau_{1,k}+\tau_{2,k})}{\Gamma(\tau_{1,k})\Gamma(\tau_{2,k})}-\log\alpha+(\tau_{1,k}-\alpha)\psi(\tau_{1,k})+(\tau_{2,k}-1)\psi(\tau_{2,k})-(\tau_{1,k}+\tau_{2,k}-\alpha-1)\psi(\tau_{1,k}+\tau_{2,k})\right]\\
&-\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{s}^{\text{stn}}_{\cdot k}}\left[\zeta_k(\log\zeta_k-\kappa_{1,k})+(1-\zeta_k)(\log(1-\zeta_k)-\kappa_{2,k})\right]\\
&-\sum_{k=1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{s}^{\text{shp}}_{\cdot k},\boldsymbol{s}^{\text{stn}}_{\cdot k}}\left[\xi_{k,n}(\log\xi_{n,k}-\log\tilde{\xi}_{n,k})+(1-\xi_{n,k})(\log(1-\xi_{n,k})-\log(1-\tilde{\xi}_{n,k}))\right]
\end{aligned}
$$

where

$$
\begin{aligned}
\gamma_{n,k} &= \begin{cases}\zeta_k\xi_{n,k}\prod_{k'=1}^{k-1}(1-\zeta_{k'}\xi_{n,k'}), & 1\le k\le K\\ \prod_{k'=1}^{K}(1-\zeta_{k'}\xi_{n,k'}), & k=0\end{cases}\\
\kappa_{1,k} &= \sum_{k'=1}^{k}\big(\psi(\tau_{1,k'})-\psi(\tau_{1,k'}+\tau_{2,k'})\big)\\
\kappa_{2,k} &= \sum_{k'=1}^{k}c_{k,k'}\psi(\tau_{2,k'})+\sum_{k'=1}^{k-1}\psi(\tau_{1,k'})\sum_{k''=k'+1}^{k}c_{k,k''}-\sum_{k'=1}^{k}\psi(\tau_{1,k'}+\tau_{2,k'})\sum_{k''=k'}^{k}c_{k,k''}-\sum_{k'=1}^{k}c_{k,k'}\log c_{k,k'}\\
c_{k,k'} &= \frac{\exp\left(\psi(\tau_{2,k'})+\sum_{k''=1}^{k'-1}\psi(\tau_{1,k''})-\sum_{k''=1}^{k'}\psi(\tau_{1,k''}+\tau_{2,k''})\right)}{\sum_{k^*=1}^{k}\exp\left(\psi(\tau_{2,k^*})+\sum_{k''=1}^{k^*-1}\psi(\tau_{1,k''})-\sum_{k''=1}^{k^*}\psi(\tau_{1,k''}+\tau_{2,k''})\right)}\\
\tilde{\xi}_{n,k} &= f_{\text{stn}}\big(f_{\text{shp}}(\boldsymbol{s}^{\text{shp}}_{\cdot k}),\boldsymbol{s}^{\text{stn}}_{\cdot k}\big)_n
\end{aligned}
$$

The Γ and ψ in the above expressions are gamma and digamma functions, respectively.

[1]Shanghai Key Laboratory of Intelligent Information Processing; Fudan-Qiniu Joint Laboratory for Deep Learning; Shanghai Institute of Intelligent Electronics & Systems; School of Computer Science, Fudan University, China. Correspondence to: Bin Li <libin@fudan.edu.cn>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2. Details of Evaluation Metrics in Section 4

- *Adjusted Mutual Information* (AMI) is a normalized version of the mutual information that takes the number of clusters into account. For N-EM and the proposed method, AMI is computed based on the cluster assignment inferred at each pixel. For AIR, similarities between reconstructions of individual objects and the original image are used instead. In order to be consistent with previous work (Greff et al., 2016b;a; 2017), pixels in the background and overlapping regions do not participate in computations of AMI.
- *Mean Squared Error* (MSE) measures the similarities between reconstructions of individual objects and the ground truth. Because MSE is evaluated at all pixels, they provide information about how different methods perform when occlusions of objects exist, which is not reflected by AMI.
- *Object Counting Accuracy* (OCA) measures the quality of estimated number of objects in each image. It is the ratio between the number of images in which objects are correctly counted and the total number of images. In AIR and the proposed method, numbers of objects are modeled by latent variables and can be inferred directly. For N-EM, they are computed heuristically based on posterior probabilities.
- *Object Ordering Accuracy* (OOA) computes the weighted sum of pairwise order estimation accuracies. The weight $w^l_{i,j}$ for each pair of objects i and j in the lth image is proportional to the sum of squared appearance differences in the overlapping region $R^l_{i,j}$, i.e. $w^l_{i,j} \propto \sum_{n \in R_{i,j}} \|\boldsymbol{a}_{n,i} - \boldsymbol{a}_{n,j}\|^2$ and $\sum_{l,i,j} w^l_{i,j} = 1$. Let $r^l_{i,j}$ be 1 if the relative order of objects i and j in the lth image is correctly estimated and 0 otherwise. OOA is computed by $\sum_{l,i,j} w^l_{i,j} r^l_{i,j}$. If two objects do not overlap or share identical appearance in overlapping regions, the estimation of their relative order does not contribute to OOA because the ground truth is not well-defined.

3. Additional Experimental Results

3.1. Performance Comparisons on Datasets Containing No Occlusions

The segregation and counting performance of N-EM, AIR, and the proposed method is presented in Table 1. Comparisons of generalizability are shown in Tables 2, 3 and 4.

3.2. Influence of Figure-Ground Organization

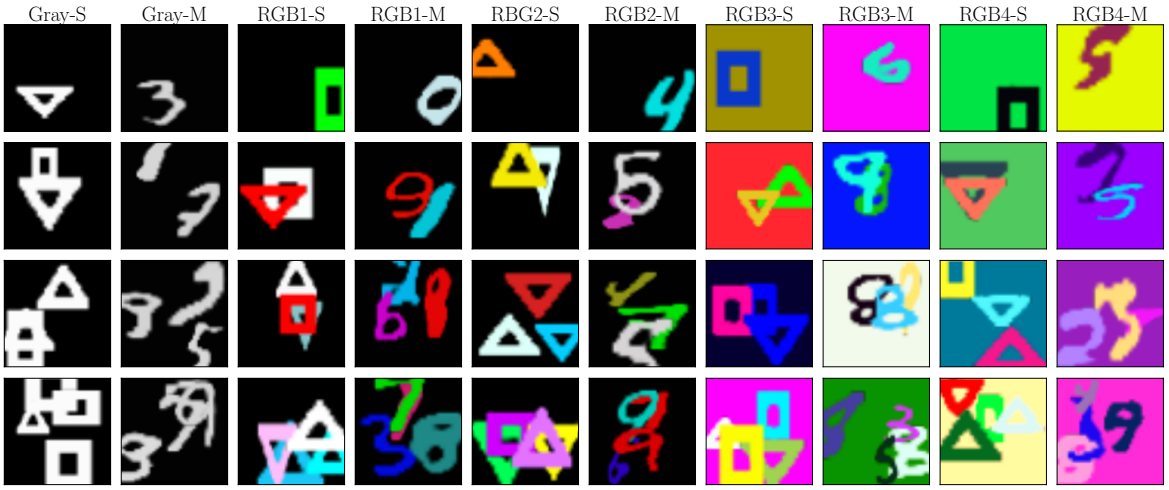
The figure-ground organization is a step in perceptual grouping that distinguishes foreground objects from the background. It is partially solved if the background appearances are known. We evaluate performance of the proposed method in the absence of this information and assess the influence of figure-ground organization. The results is shown in Table 5. Compared with the results obtained when the background appearances are given, whether knowing the background appearances does not influence the performance much.

3.3. Comparison of Log-Likelihood

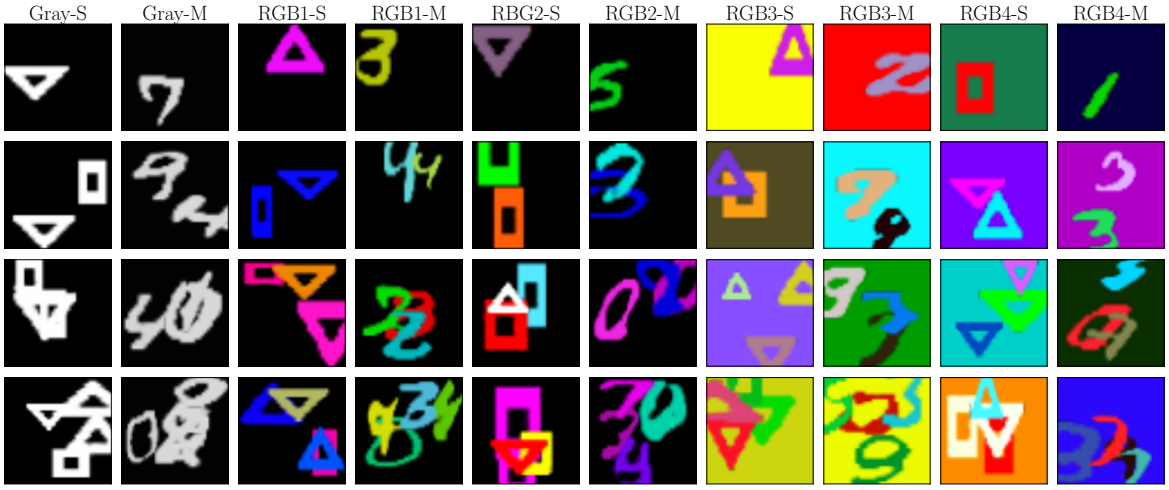
The log-likelihoods of N-EM, AIR, and the proposed method computed with 0.3 as the standard deviation of the normal distribution are presented in Table 6. When no occlusions exist, the proposed method outperforms N-EM and AIR on all but the Gray-S dataset. When objects may be occluded, N-EM achieves highest log-likelihoods on the Gray-M, RGB1-M, and RGB2-M datasets (objects are handwritten digits and backgrounds are black), while the proposed method achieves the best results on the other 7 datasets. Although N-EM outperforms AIR and the proposed method in terms of reconstruction error of the visual scene (higher log-likelihoods) on the Gray-M, RGB1-M, and RGB2-M datasets, individual objects are reconstructed less accurately (higher MSE scores) by N-EM because objects are not well separated (low AMI scores).

3.4. Samples from the Learned Models

Samples from the models learned using images *without* and *with* object occlusions are presented in Figure 1. Background appearances are unknown and latent representations of background $\boldsymbol{s}^{\text{apc}}_{\cdot 0}$ are inferred when training the models. As shown in Figure 1(a), objects in the generated samples may be occluded even though occlusions do not exist in the training images. According to samples from models trained on RGB2 and RGB4 datasets, the colors of generated objects in the same image are not identical although objects in each training image are indistinguishable in appearance. The reason for these two phenomena is that relations between objects except occlusion are not modeled in the proposed method.



(a) Learned using images *without* existence of occlusion



(b) Learned using images *with* existence of occlusion

Figure 1. Samples from the models learned without knowing background appearances.

4. Ablation Study

4.1. Effectiveness of Iteratively Updating Latent Variables

The effectiveness of iteratively updating latent variables is evaluated using models trained without knowing background appearances. If latent variables are not iteratively updated by LSTMs which imitate the behaviors of coordinate ascent, the average AMI/MSE/OCA scores are 96.7%/$1.5e$-3/95.6% on datasets *without* existence of occlusion, and 78.6%/$3.5e$-3/76.9% on datasets *with* existence of occlusion. If latent variables are updated for 2 iterations, the average AMI/MSE/OCA scores are improved to 98.2%/$1.2e$-3/97.1% and 84.2%/$3.3e$-3/81.2% on datasets *without* and *with* existence of occlusion, respectively. Iteratively updating latent variables is beneficial to the performance.

4.2. Effectiveness of Initializing with Neural Networks

The effectiveness of initializing intermediate variables $\boldsymbol{w}_{\cdot k}$ and $\boldsymbol{u}_{\cdot k}$ with neural networks is assessed under the circumstances that background appearances are unknown. If $\boldsymbol{w}_{\cdot k}$ and $\boldsymbol{u}_{\cdot k}$ are initialized with random values and latent variables are updated for 2 iterations, the average AMI/MSE/OCA scores are 98.0%/$1.4e$-3/96.8% on datasets *without* existence of occlusion, and 82.5%/$3.8e$-3/75.8% on datasets *with* existence of occlusion. If the number of iterations for each latent variable is increased

Table 1. Comparison of segregation and counting performance *without* existence of occlusion.

DATA SET	N-EM AMI (%)	N-EM MSE (E-3)	N-EM OCA (%)	AIR AMI (%)	AIR MSE (E-3)	AIR OCA (%)	PROPOSED AMI (%)	PROPOSED MSE (E-3)	PROPOSED OCA (%)
GRAY-S	87.7± 4E-1	6.5± 1E-1	78.7± 3E-1	**100**± 1E-2	**0.5**± 5E-3	99.5± 3E-2	**100**± 8E-3	1.0± 3E-3	**100**± 1E-2
GRAY-M	37.3± 4E-1	22± 7E-2	17.0± 2E-1	97.5± 5E-2	2.4± 8E-3	97.3± 1E-1	**98.1**± 7E-2	**2.3**± 8E-3	**97.4**± 1E-1
RGB1-S	98.5± 1E-1	2.0± 4E-2	94.4± 6E-1	99.7± 2E-2	**0.5**± 1E-2	96.8± 6E-2	**99.8**± 2E-2	0.6± 9E-3	**98.6**± 7E-2
RGB1-M	63.8± 1E-1	8.7± 2E-2	21.2± 8E-1	97.9± 5E-2	**1.2**± 1E-3	94.8± 1E-1	**98.0**± 4E-2	**1.2**± 7E-3	**96.0**± 2E-1
RGB2-S	96.4± 2E-1	1.8± 6E-2	92.6± 3E-1	**100**± 1E-2	**0.3**± 2E-3	99.7± 2E-2	**100**± 6E-3	0.4± 3E-3	**99.9**± 3E-2
RGB2-M	39.3± 1E-1	13± 4E-2	16.3± 1.4	97.1± 3E-2	**1.2**± 3E-3	93.7± 1E-1	**97.7**± 1E-1	**1.2**± 1E-2	**95.3**± 3E-1
RGB3-S	25.8± 3E-1	21± 5E-3	5.63± 5E-1	**99.7**± 2E-2	1.3± 3E-3	95.8± 7E-2	**99.7**± 9E-3	**0.7**± 1E-2	**97.9**± 6E-2
RGB3-M	14.5± 4E-1	24± 1E-1	2.35± 7E-2	97.7± 3E-2	1.5± 3E-3	**97.7**± 9E-2	**97.8**± 7E-2	**1.3**± 9E-3	96.3± 3E-1
RGB4-S	62.9± 8E-1	14± 1E-1	50.2± 9E-1	**99.7**± 1E-2	1.2± 9E-3	94.9± 1E-1	98.7± 4E-2	**1.0**± 1E-2	**97.8**± 2E-1
RGB4-M	6.42± 1E-1	25± 2E-1	3.30± 2E-1	94.1± 4E-2	2.2± 4E-3	93.4± 2E-1	**96.6**± 4E-2	**1.4**± 7E-3	**95.6**± 2E-1

Table 2. Comparison of generalizability *with* existence of occlusion (tested on images containing 4 objects with $K=10$).

DATA SET	N-EM AMI (%)	N-EM MSE (E-3)	N-EM OCA (%)	AIR AMI (%)	AIR MSE (E-3)	AIR OCA (%)	PROPOSED AMI (%)	PROPOSED MSE (E-3)	PROPOSED OCA (%)
GRAY-S	6.03± 4E-2	50± 3E-4	0.01± 0.0	81.8± 3E-2	14± 2E-2	48.9± 3E-1	**94.1**± 1E-1	**5.8**± 7E-2	**73.6**± 4E-1
GRAY-M	32.5± 1E-1	23± 2E-2	15.9± 1.2	56.5± 4E-2	17± 2E-2	17.6± 3E-1	**66.2**± 4E-2	**14**± 1E-2	**49.2**± 6E-1
RGB1-S	39.1± 2E-1	20± 7E-2	15.6± 2.3	92.8± 4E-2	5.0± 7E-3	69.9± 4E-1	**97.4**± 5E-2	**1.4**± 8E-3	**84.2**± 2E-1
RGB1-M	54.2± 1E-1	11± 2E-2	13.7± 1.5	73.7± 5E-2	6.2± 1E-2	49.3± 5E-1	**78.8**± 8E-2	**4.9**± 2E-2	**56.9**± 3E-1
RGB2-S	4.32± 8E-2	29± 4E-3	3.66± 5E-1	81.3± 3E-2	8.3± 6E-3	53.9± 3E-1	**90.4**± 5E-2	**4.5**± 3E-2	**60.5**± 3E-1
RGB2-M	37.2± 1E-1	14± 2E-2	11.2± 8E-1	58.1± 6E-2	9.0± 6E-3	28.1± 3E-1	**63.6**± 1E-1	**8.1**± 2E-2	**42.9**± 4E-1
RGB3-S	41.2± 2E-1	90± 4E-1	3.74± 5E-1	77.3± 1E-1	9.9± 3E-2	42.2± 6E-1	**95.0**± 4E-2	**2.4**± 9E-3	**75.9**± 6E-1
RGB3-M	18.9± 3E-1	20± 2E-2	0.02± 2E-2	61.5± 7E-2	9.0± 8E-3	41.5± 4E-1	**74.0**± 1E-1	**5.8**± 2E-2	**45.2**± 7E-1
RGB4-S	13.6± 1E-1	190± 1.0	14.1± 6E-1	71.1± 5E-2	12± 1E-2	34.0± 2E-1	**88.2**± 8E-2	**5.4**± 4E-2	**53.0**± 5E-1
RGB4-M	4.37± 1E-1	20± 2E-2	0.03± 3E-2	45.1± 3E-2	12± 8E-3	2.5± 1E-1	**62.3**± 6E-2	**8.3**± 1E-2	**34.2**± 4E-1

to 5, the average AMI/MSE/OCA scores are improved to 98.4%/$1.2e$-3/96.9% and 83.6%/$3.6e$-3/78.2% on datasets *without* and *with* existence of occlusion, respectively. By initializing $\boldsymbol{w}_{\cdot k}$ and $\boldsymbol{u}_{\cdot k}$ with neural networks and updating latent variables for 2 iterations, the average AMI/MSE/OCA scores (98.2%/$1.2e$-3/97.1% and 84.2%/$3.3e$-3/81.2% on datasets *without* and *with* existence of occlusion, respectively) are comparable to the results obtained when $\boldsymbol{w}_{\cdot k}$ and $\boldsymbol{u}_{\cdot k}$ are initialized randomly and latent variables are updated for 5 iterations. The effectiveness of initializing with neural networks is more noticeable when objects in images may be occluded.

5. Choices of Hyperparameters

5.1. N-EM

For all datasets, mixture components are chosen to be normal distributions with standard deviation 0.25. When computing inputs of the encoder network, masked uniform noises are added to images. The mean of the Bernoulli distribution that generates the mask is 0.2. Each latent representation is updated for 10 iterations. The batch size is chosen as 32, and models are trained with Adam algorithm (Kingma & Ba, 2015) for 50 epochs.

For datasets containing shapes, the learning rate is chosen to be 1×10^{-3}, and the coefficient of the regularization term in the loss function is set to 1. Inputs of the encoder network are not normalized.

For datasets containing handwritten digits, the learning rate is chosen to be 5×10^{-4}, and the coefficient of the regularization term in the loss function is set to 0.2. Whitening transformation is applied to inputs of the encoder network.

Parameters of the encoder-decoder architecture of N-EM are:

Table 3. Comparison of generalizability *without* existence of occlusion (tested on images containing 4 objects with $K = 4$).

DATA SET	N-EM			AIR			PROPOSED		
	AMI (%)	MSE (E-3)	OCA (%)	AMI (%)	MSE (E-3)	OCA (%)	AMI (%)	MSE (E-3)	OCA (%)
GRAY-S	83.9± 1E-1	9.7± 1E-1	16.7± 41	99.6± 2E-2	**1.0**± 1E-2	**100**± 1E-2	**100**± 7E-3	1.6± 7E-3	**100**± 8E-3
GRAY-M	51.7± 2E-1	29± 4E-2	2.20± 3E-1	94.7± 5E-2	4.3± 2E-2	93.9± 3E-1	**97.7**± 4E-2	**3.7**± 4E-3	**96.2**± 2E-1
RGB1-S	97.7± 1E-1	3.2± 4E-2	20.0± 73	97.7± 3E-2	1.1± 2E-2	**100**± 1E-2	**99.8**± 2E-2	**0.9**± 5E-3	99.9± 4E-2
RGB1-M	75.2± 4E-1	11± 9E-2	1.35± 3E-1	96.6± 6E-2	2.1± 7E-3	**99.0**± 7E-2	**98.3**± 3E-2	**1.7**± 8E-3	98.6± 9E-2
RGB2-S	88.7± 2E-1	4.3± 5E-2	23.1± 57	99.4± 3E-2	0.8± 6E-3	**100**± 0.0	**99.9**± 2E-2	**0.6**± 5E-3	**100**± 3E-2
RGB2-M	53.2± 2E-1	18± 6E-2	2.66± 2E-1	89.6± 6E-2	3.1± 7E-3	93.2± 2E-1	**97.5**± 4E-2	**1.8**± 7E-3	**95.4**± 3E-1
RGB3-S	23.5± 1E-1	23± 1E-2	33.1± 7E-1	97.3± 8E-2	2.6± 2E-2	**100**± 0.0	**99.5**± 5E-2	**1.0**± 2E-2	99.6± 5E-2
RGB3-M	14.8± 1E-1	31± 2E-1	45.2± 4E-1	92.3± 5E-2	2.8± 1E-2	83.6± 2E-1	**98.0**± 4E-2	**2.0**± 7E-3	**97.8**± 9E-2
RGB4-S	59.3± 3E-1	19± 1E-1	4.51± 18	97.1± 1E-1	2.4± 2E-2	**100**± 2E-2	**97.4**± 7E-2	**1.9**± 2E-2	98.3± 1E-1
RGB4-M	5.98± 8E-2	32± 4E-1	37.5± 4E-1	81.3± 7E-2	4.4± 5E-3	54.8± 3E-1	**96.0**± 5E-2	**2.2**± 1E-2	**94.5**± 1E-1

Table 4. Comparison of generalizability *without* existence of occlusion (tested on images containing 4 objects with $K = 10$).

DATA SET	N-EM			AIR			PROPOSED		
	AMI (%)	MSE (E-3)	OCA (%)	AMI (%)	MSE (E-3)	OCA (%)	AMI (%)	MSE (E-3)	OCA (%)
GRAY-S	3.57± 1E-2	50± 2E-3	1.20± 8E-1	99.8± 2E-2	0.8± 1E-2	98.5± 9E-2	**100**± 1E-2	**1.3**± 6E-3	**99.9**± 6E-2
GRAY-M	42.6± 1E-1	20± 7E-2	14.0± 4E-1	95.1± 4E-2	3.7± 9E-3	88.6± 1E-1	**97.1**± 5E-2	**3.2**± 1E-2	**90.0**± 4E-1
RGB1-S	10.7± 5E-2	29± 2E-4	14.4± 14	99.3± 2E-2	0.8± 5E-3	89.2± 9E-2	**99.5**± 2E-2	**0.7**± 8E-3	**95.5**± 1E-1
RGB1-M	59.4± 9E-2	9.1± 3E-2	11.0± 1.4	96.9± 4E-2	1.7± 3E-3	86.2± 4E-1	**97.8**± 3E-2	**1.4**± 5E-3	**91.4**± 7E-2
RGB2-S	10.7± 9E-2	26± 8E-2	1.10± 2E-1	99.7± 3E-2	0.7± 5E-3	98.0± 1E-1	**99.9**± 7E-3	**0.5**± 4E-3	**99.6**± 4E-2
RGB2-M	46.0± 1E-1	12± 4E-2	11.8± 1.2	90.2± 6E-2	2.7± 8E-3	66.5± 5E-1	**96.6**± 4E-2	**1.6**± 6E-3	**86.6**± 4E-1
RGB3-S	40.2± 1E-1	65± 4E-1	5.80± 4E-1	99.1± 5E-2	2.0± 1E-2	84.8± 1E-1	**99.2**± 5E-2	**0.8**± 1E-2	**91.4**± 1E-1
RGB3-M	20.8± 2E-1	18± 8E-2	0.03± 5E-2	92.5± 1E-1	2.5± 9E-3	79.6± 2E-1	**97.6**± 4E-2	**1.6**± 6E-3	**90.9**± 1E-1
RGB4-S	22.6± 2E-1	161± 7E-1	15.1± 2E-1	**98.6**± 6E-2	2.0± 2E-2	85.7± 2E-1	96.9± 4E-2	**1.5**± 1E-2	**87.5**± 2E-1
RGB4-M	6.32± 2E-1	20± 4E-2	0.05± 3E-2	81.3± 5E-2	4.8± 4E-3	54.8± 2E-1	**95.3**± 4E-2	**1.9**± 9E-3	**85.5**± 3E-1

1. 4×4 conv, 32 ELU, stride 2; layer norm
2. 4×4 conv, 64 ELU, stride 2; layer norm
3. 4×4 conv, 128 ELU, stride 2; layer norm
4. fully connected, 32 ELU; layer norm
5. RNN, 16 Sigmoid; layer norm on the output
6. fully connected, 32 ReLU; layer norm
7. fully connected, $6 \times 6 \times 128$ ReLU; layer norm
8. 2x nearest-neighbor upsample; 4×4 conv, 64 ReLU; layer norm
9. 2x nearest-neighbor upsample; 4×4 conv, 32 ReLU; layer norm
10. 2x nearest-neighbor upsample; 4×4 conv, 1 or 3 linear

5.2. AIR

The likelihood function is chosen to be a normal distribution with standard deviation 0.3. The priors for latent variables of relative scales of objects are normal distributions with mean 0 and standard deviation 0.5. These variables are transformed by the *sigmoid* function to relative scales. The priors for latent variables of relative translations of objects are normal

Table 5. Perceptual grouping performance of the proposed method without knowing background appearances.

DATA SET	WITHOUT OCCLUSION				WITH OCCLUSION			
	AMI (%)	MSE (E-3)	OCA (%)	OOA (%)	AMI (%)	MSE (E-3)	OCA (%)	OOA (%)
GRAY-S	100± 1E-2	0.8± 8E-3	100± 2E-2	N/A	96.1± 1E-1	2.2± 4E-2	93.1± 2E-1	N/A
GRAY-M	97.8± 5E-2	2.5± 9E-3	96.5± 2E-1	N/A	70.2± 6E-2	7.5± 3E-2	76.2± 3E-1	N/A
RGB1-S	99.9± 1E-2	0.4± 8E-3	99.4± 7E-2	N/A	98.7± 1E-1	1.4± 3E-2	90.2± 4E-1	97.0± 0.2
RGB1-M	97.8± 9E-2	1.3± 1E-2	93.3± 4E-1	N/A	81.8± 3E-2	3.1± 9E-3	73.2± 3E-1	59.3± 0.9
RGB2-S	99.3± 2E-2	0.8± 6E-3	99.0± 7E-2	N/A	95.0± 4E-2	1.5± 6E-3	91.0± 9E-2	N/A
RGB2-M	98.0± 4E-2	1.2± 9E-3	96.5± 1E-1	N/A	66.0± 1E-1	5.0± 2E-2	68.2± 7E-1	N/A
RGB3-S	99.6± 3E-2	0.7± 1E-2	97.9± 6E-2	N/A	96.5± 6E-2	1.6± 3E-2	91.2± 4E-1	88.5± 0.5
RGB3-M	96.9± 6E-2	1.7± 1E-2	96.6± 1E-1	N/A	75.9± 1E-1	4.1± 1E-2	67.2± 2E-1	53.2± 0.8
RGB4-S	99.1± 4E-2	0.9± 9E-3	98.9± 1E-1	N/A	93.8± 2E-2	1.8± 2E-2	88.3± 2E-1	N/A
RGB4-M	93.3± 7E-2	2.1± 9E-3	92.7± 3E-1	N/A	67.5± 4E-2	4.7± 2E-2	73.6± 3E-1	N/A

Table 6. Comparison of log-likelihood (standard deviation of the normal distribution is 0.3).

DATA SET	WITHOUT OCCLUSION			WITH OCCLUSION		
	N-EM	AIR	PROPOSED	N-EM	AIR	PROPOSED
GRAY-S	**627.73**± 0.7	626.60± 0.2	605.70± 0.1	575.06± 0.9	585.97± 0.3	**615.80**± 0.2
GRAY-M	545.81± 0.2	547.22± 0.3	**547.89**± 0.4	**560.86**± 0.2	512.68± 0.4	536.63± 0.2
RGB1-S	1784.1± 1.2	1884.5± 1.8	**1893.6**± 0.5	1775.8± 0.7	1757.8± 1.3	**1876.7**± 0.5
RGB1-M	1799.4± 1.0	1788.7± 0.6	**1809.3**± 0.7	**1786.6**± 1.0	1716.5± 0.5	1769.9± 0.4
RGB2-S	1882.2± 0.6	1903.2± 0.6	**1910.1**± 0.7	1771.6± 1.1	1841.3± 0.4	**1891.6**± 0.2
RGB2-M	1783.6± 0.9	1798.9± 0.8	**1812.7**± 0.2	**1826.4**± 0.3	1736.8± 0.5	1797.3± 0.2
RGB3-S	-1132± 0.9	1729.9± 1.9	**1891.2**± 0.6	-963.2± 0.1	1545.1± 0.3	**1833.4**± 0.7
RGB3-M	116.73± 2.6	1743.4± 0.3	**1786.4**± 0.2	61.69± 2.0	1484.6± 0.6	**1716.8**± 1.0
RGB4-S	882.22± 2.7	1761.6± 1.0	**1836.3**± 1.1	170.10± 1.3	1674.5± 0.5	**1887.0**± 1.0
RGB4-M	21.29± 2.5	1650.6± 0.3	**1775.2**± 0.1	249.47± 3.4	1608.6± 0.4	**1732.0**± 0.3

distributions with mean 0 and standard deviation 0.5. These variables are transformed by the *tanh* function to relative translations. The priors for latent variables of cropped images of objects are standard normal distributions. The prior for the number of objects (characterized by presences of objects) is a geometric distribution with success probability 1×10^{-6}. Concrete relations (Maddison et al., 2017; Jang et al., 2017) are applied to discrete latent variables because it has been observed (Kosiorek et al., 2018) that using NVIL (Mnih & Gregor, 2014) in AIR does not results in stable performance. The batch size is 128, and the learning rate is 1×10^{-4}. Models are trained with the Adam algorithm for 100 epochs.

Parameters of the neural network that transforms images to hidden states of the LSTM are:

1. 3×3 conv, 8 ReLU; layer norm
2. 3×3 conv, 8 ReLU, stride 2; layer norm
3. 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 16 ReLU, stride 2; layer norm
5. fully connected, 256 ReLU; layer norm
6. LSTM, 256 Tanh

Parameters of the neural network that transforms hidden states of the LSTM to latent variables of presences of objects are:

1. fully connected, 128 ReLU; layer norm

2. fully connected, 64 ReLU; layer norm
3. fully connected, 1 linear for mean of the Bernoulli distribution; Concrete relaxation

Parameters of the neural network that transforms hidden states of the LSTM to latent variables of relative scales and translations of objects are:

1. fully connected, 256 ReLU; layer norm
2. fully connected, 256 ReLU; layer norm
3. fully connected, $4 + 4$ linear for mean+variance of the normal distribution; reparameterization

Parameters of the neural network that transforms crops of images to the corresponding latent variables are:

1. 3×3 conv, 8 ReLU; layer norm
2. 3×3 conv, 8 ReLU, stride 2; layer norm
3. 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 16 ReLU, stride 2; layer norm
5. fully connected, 256 ReLU; layer norm
6. fully connected, $32 + 32$ linear for mean+variance of the normal distribution; reparameterization

Parameters of the neural network that reconstructs crops of images from the corresponding latent variables are:

1. fully connected, 256 ReLU; layer norm
2. fully connected, $5 \times 5 \times 16$ ReLU; layer norm
3. 2x nearest-neighbor upsample; 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 8 ReLU; layer norm
5. 2x nearest-neighbor upsample; 3×3 conv, 8 ReLU; layer norm
6. 3×3 conv, 1 or 3 linear

5.3. The proposed method

Hyperparameters of the likelihood function, priors for latent variables of relative scales and translations of objects are identical to those chosen for AIR. The priors for latent variables of shapes and appearances of objects are standard normal distributions. The α parameter of the beta distribution is 1×10^{-3}. Each latent representation is updated for 2 iterations. The batch size is set to 128, and the learning rate is 1×10^{-4}. Models are trained with the Adam algorithm for 100 epochs.

Parameters of the neural network that initializes the intermediate variables $\boldsymbol{w}_{\cdot k}$ are:

1. 3×3 conv, 8 ReLU; layer norm
2. 3×3 conv, 8 ReLU, stride 2; layer norm
3. 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 16 ReLU, stride 2; layer norm
5. fully connected, 256 ReLU; layer norm

6. LSTM, 256 Tanh
7. fully connected, 256 ReLU; layer norm
8. fully connected, 256 ReLU; layer norm
9. LSTM, 256 Tanh

Parameters of the neural network g_w that updates the intermediate variables $\boldsymbol{w}_{\cdot k}$ are:

1. 3×3 conv, 8 ReLU; layer norm
2. 3×3 conv, 8 ReLU, stride 2; layer norm
3. 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 16 ReLU, stride 2; layer norm
5. fully connected, 256 ReLU; layer norm
6. LSTM, 256 Tanh

Parameters of the neural network that initializes the intermediate variables $\boldsymbol{u}_{\cdot k}$ with $1 \leq k \leq K$ are:

1. 3×3 conv, 8 ReLU; layer norm
2. 3×3 conv, 8 ReLU, stride 2; layer norm
3. 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 16 ReLU, stride 2; layer norm
5. fully connected, 256 ReLU; layer norm
6. LSTM, 256 Tanh

Parameters of the neural network g_u^{obj} that updates the intermediate variables $\boldsymbol{u}_{\cdot k}$ with $1 \leq k \leq K$ are:

1. 3×3 conv, 8 ReLU; layer norm
2. 3×3 conv, 8 ReLU, stride 2; layer norm
3. 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 16 ReLU, stride 2; layer norm
5. fully connected, 256 ReLU; layer norm
6. LSTM, 256 Tanh

Parameters of the neural network that initializes the intermediate variables $\boldsymbol{u}_{\cdot 0}$ are:

1. 3×3 conv, 8 ReLU; layer norm
2. 3×3 conv, 8 ReLU, stride 2; layer norm
3. 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 16 ReLU, stride 2; layer norm
5. fully connected, 256 ReLU; layer norm

6. LSTM, 256 Tanh

Parameters of the neural network g_u^{back} that updates the intermediate variables $\boldsymbol{u}_{\cdot 0}$ are:

1. 3×3 conv, 8 ReLU; layer norm
2. 3×3 conv, 8 ReLU, stride 2; layer norm
3. 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 16 ReLU, stride 2; layer norm
5. fully connected, 256 ReLU; layer norm
6. LSTM, 256 Tanh

Parameters of the neural networks g_τ and g_ζ that transforms $\boldsymbol{w}_{\cdot k}$ and $\boldsymbol{u}_{\cdot k}$ to $\tau_{1,k}$, $\tau_{2,k}$ and ζ_k are:

1. fully connected, 128 ReLU; layer norm
2. fully connected, 64 ReLU; layer norm
3. fully connected, 1 Softplus/1 Softplus/1 Sigmoid for $\tau_{1,k}$/$\tau_{2,k}$/ζ_k

$\xi_{n,k}$ is computed by $\xi_{n,k} = g_\xi(\boldsymbol{s}_{\cdot k}^{\text{shp}}, \boldsymbol{s}_{\cdot k}^{\text{stn}})_n = f_{\text{stn}}\big(f_{\text{shp}}(\boldsymbol{s}_{\cdot k}^{\text{shp}}), \boldsymbol{s}_{\cdot k}^{\text{stn}}\big)_n$ for simplicity.
Parameters of the neural network g_{stn} that transforms $\boldsymbol{w}_{\cdot k}$ to $\boldsymbol{s}_{\cdot k}^{\text{stn}}$ are:

1. fully connected, 256 ReLU; layer norm
2. fully connected, 256 ReLU; layer norm
3. fully connected, $4 + 4$ linear for mean+variance of the normal distribution; reparameterization

Parameters of the neural networks g_{shp} and $g_{\text{apc}}^{\text{obj}}$ that transforms $\boldsymbol{u}_{\cdot k}$ with $1 \leq k \leq K$ to $\boldsymbol{s}_{\cdot k}^{\text{shp}}$ and $\boldsymbol{s}_{\cdot k}^{\text{apc}}$ are:

1. fully connected, 256 ReLU; layer norm
2. fully connected, $32 + 32/32 + 32$ linear for mean+variance of $\boldsymbol{s}_{\cdot k}^{\text{shp}}$/mean+variance of $\boldsymbol{s}_{\cdot k}^{\text{apc}}$; reparameterization

Parameters of the neural network $g_{\text{apc}}^{\text{back}}$ that transforms $\boldsymbol{u}_{\cdot 0}$ to $\boldsymbol{s}_{\cdot 0}^{\text{apc}}$ are:

1. fully connected, 256 ReLU; layer norm
2. fully connected, $32 + 32$ linear for mean+variance of $\boldsymbol{s}_{\cdot 0}^{\text{apc}}$; reparameterization

Parameters of the neural network f_{shp} that transforms $\boldsymbol{s}_{\cdot k}^{\text{shp}}$ to reconstructed shapes of objects are:

1. fully connected, 256 ReLU; layer norm
2. fully connected, $5 \times 5 \times 16$ ReLU; layer norm
3. 2x nearest-neighbor upsample; 3×3 conv, 16 ReLU; layer norm
4. 3×3 conv, 8 ReLU; layer norm
5. 2x nearest-neighbor upsample; 3×3 conv, 8 ReLU; layer norm
6. 3×3 conv, 1 linear

Parameters of the neural network $f_{\text{apc}}^{\text{obj}}$ that transforms $\boldsymbol{s}_{\cdot k}^{\text{apc}}$ with $1 \leq k \leq K$ to reconstructed appearances of objects are:

1. fully connected, 16 ReLU; layer norm
2. fully connected, 16 ReLU; layer norm
3. fully connected, 1 or 3 linear

Parameters of the neural network $f_{\text{apc}}^{\text{back}}$ that transforms $\boldsymbol{s}_{\cdot 0}^{\text{apc}}$ to the reconstructed appearance of background are:

1. fully connected, 16 ReLU; layer norm
2. fully connected, 16 ReLU; layer norm
3. fully connected, 1 or 3 linear

References

Doshi, F., Miller, K., Van Gael, J., and Teh, Y. W. Variational inference for the Indian buffet process. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 137–144, 2009.

Greff, K., Rasmus, A., Berglund, M., Hao, T., Valpola, H., and Schmidhuber, J. Tagger: Deep unsupervised perceptual grouping. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 4484–4492, 2016a.

Greff, K., Srivastava, R. K., and Schmidhuber, J. Binding via reconstruction clustering. In *International Conference on Learning Representations (ICLR) Workshop*, 2016b.

Greff, K., van Steenkiste, S., and Schmidhuber, J. Neural expectation maximization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 6691–6701, 2017.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Kosiorek, A. R., Kim, H., Posner, I., and Teh, Y. W. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 8615–8625, 2018.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*, 2017.

Mnih, A. and Gregor, K. Neural variational inference and learning in belief networks. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 1791–1799, 2014.