Generative Modeling of Infinite Occluded Objects for Compositional Scene Representation

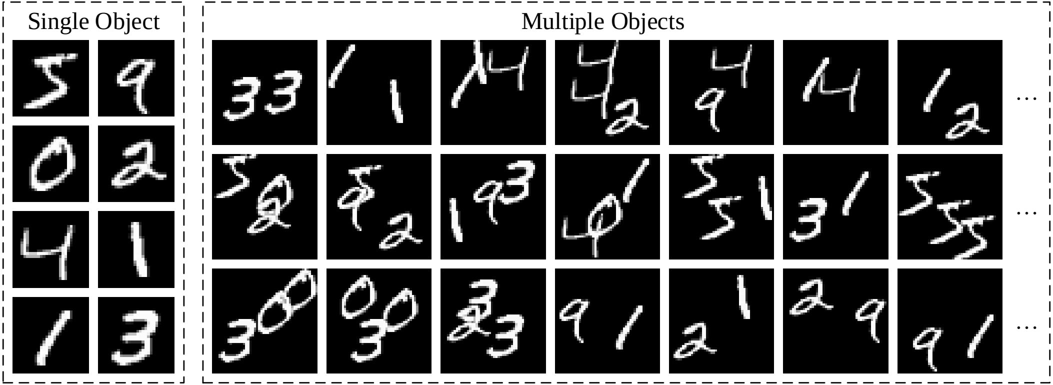
Jinyang Yuan, Bin Li, Xiangyang Xue

Fudan University

{*yuanjinyang, libin, xyxue*}*@fudan.edu.cn*

Jun 12, 2019

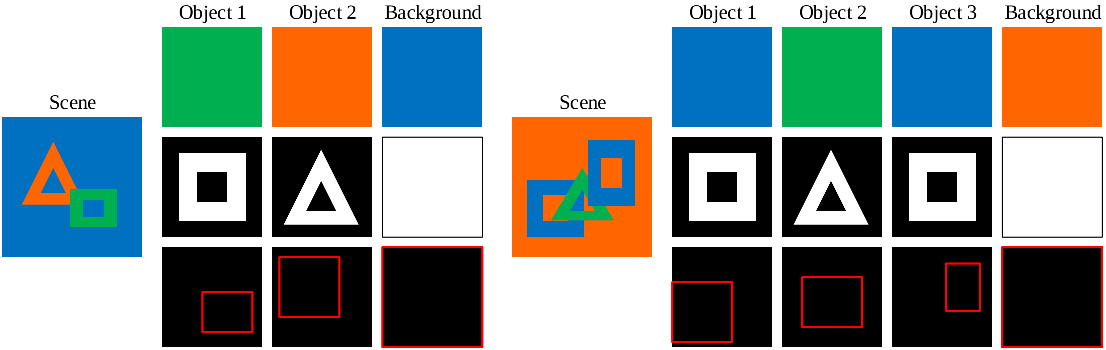
- **Scenes** are composed of **objects** and **background**
- The **combinations** of objects and background are **diverse**
- A **single representation** for the entire scene is **relatively complex**



Compositional Scene Representation

Compositional scene representation is desirable

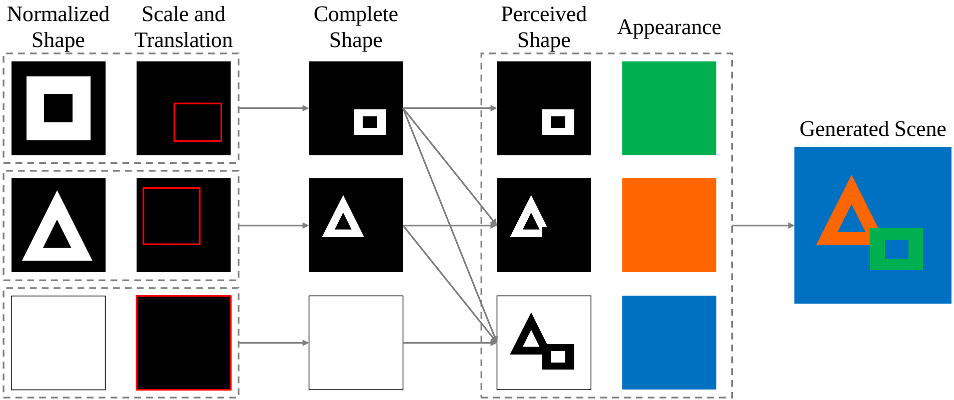
- Lower representation **complexity**
- Higher **generalizability** to novel scenes



Generative Modeling of Infinite Occluded Objects

Two major difficulties

- The **number of objects** is unknown
- The perceived objects may be incomplete due to **occlusions**



Generative Modeling of Infinite Occluded Objects

Background: $k = 0$, Objects: $k \geq 1$

Latent Representation	$\boldsymbol{s}_{\cdot k} \sim \mathcal{N}\left(\tilde{\boldsymbol{\mu}}, \operatorname{diag}(\tilde{\boldsymbol{\sigma}}^2)\right),$	$k \geq 0$
Presence (number of objects)	$\nu_k \sim \operatorname{Beta}(\alpha, 1), \quad z_k^{\text{ind}} \sim \operatorname{Ber}\left(\prod_{k'=1}^{k} \nu_{k'}\right),$	$k \geq 1$
Complete Shape	$z_{n,k}^{\text{dep}} \sim \operatorname{Ber}\left(f_{\text{stn}}(\underbrace{f_{\text{shp}}(\boldsymbol{s}_{\cdot k}^{\text{shp}})}_{\text{normalized shape}}, \underbrace{\boldsymbol{s}_{\cdot k}^{\text{stn}}}_{\text{scale and translation}})_n\right),$	$k \geq 1$

Perceived Shape (occlusions)

$$\rho_{n,k} = \begin{cases} z_k^{\text{ind}} z_{n,k}^{\text{dep}} \prod_{k'=1}^{k-1} \left(1 - z_{k'}^{\text{ind}} z_{n,k'}^{\text{dep}}\right), & k \geq 1 \\ 1 - \sum_{k'=1}^{\infty} \rho_{n,k'}, & k = 0 \end{cases}$$

Appearance

$$\boldsymbol{a}_{n,k} = \begin{cases} f_{\text{apc}}^{\text{obj}}(\boldsymbol{s}_{\cdot k}^{\text{apc}}), & k \geq 1 \\ f_{\text{apc}}^{\text{back}}(\boldsymbol{s}_{\cdot k}^{\text{apc}}), & k = 0 \end{cases}$$

Generated Scene

$$\boldsymbol{x}_n \sim \sum_{k=0}^{\infty} \rho_{n,k} \mathcal{N}(\boldsymbol{a}_{n,k}, \hat{\sigma}^2 \boldsymbol{I})$$

- Parameters are inferred by long short-term memories (**LSTMs**)
- Each object and background are updated **sequentially** and **iteratively**
- The LSTMs imitate the procedure of **coordinate ascent**

$$q(\boldsymbol{h}|\boldsymbol{x}) = q(\boldsymbol{s}_{\cdot 0}^{\mathsf{apc}}) \prod_{k=1}^{K} \left(q(\boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) q(\boldsymbol{s}_{\cdot k}^{\mathsf{shp}}|\boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) q(\boldsymbol{s}_{\cdot k}^{\mathsf{apc}}|\boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) q(\nu_k|\boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) q(z_k^{\mathsf{ind}}|\boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) \prod_{n=1}^{N} q(z_{n,k}^{\mathsf{dep}}|\boldsymbol{s}_{\cdot k}^{\mathsf{shp}}, \boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) \right)$$

$$\begin{aligned}
q(\boldsymbol{s}_{\cdot k}^{*}|\boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) &= \mathcal{N}(\boldsymbol{s}_{\cdot k}^{*}; \boldsymbol{\mu}_{\cdot k}^{*}, \mathrm{diag}({\boldsymbol{\sigma}_{\cdot k}^{*}}^{2})) \\
q(\nu_k|\boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) &= \mathrm{Beta}(\nu_k; \tau_{1,k}, \tau_{2,k}) \\
q(z_k^{\mathsf{ind}}|\boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) &= \mathrm{Ber}(z_k^{\mathsf{ind}}; \zeta_k) \\
q(z_{n,k}^{\mathsf{dep}}|\boldsymbol{s}_{\cdot k}^{\mathsf{shp}}, \boldsymbol{s}_{\cdot k}^{\mathsf{stn}}) &= \mathrm{Ber}(z_{n,k}^{\mathsf{dep}}; \xi_{n,k})
\end{aligned}$$

Experimental Results

Table: Comparison of segregation and counting performance *with* existence of occlusion.

Data set	N-EM [Greff et al., 2017]			AIR [Eslami et al., 2016]			Proposed		
	AMI	MSE	OCA	AMI	MSE	OCA	AMI	MSE	OCA
Gray-S	77.3%	10e-3	56.2%	85.4%	6.5e-3	80.9%	**94.6%**	**2.9e-3**	**90.5%**
Gray-M	30.5%	22e-3	13.5%	62.8%	9.0e-3	66.0%	**71.1%**	**7.5e-3**	**77.6%**
RGB1-S	81.8%	5.6e-3	74.2%	95.3%	2.4e-3	88.8%	**98.3%**	**1.1e-3**	**95.1%**
RGB1-M	57.0%	9.4e-3	16.3%	78.2%	3.5e-3	67.9%	**82.0%**	**3.1e-3**	**74.8%**
RGB2-S	66.2%	9.0e-3	60.8%	85.7%	3.7e-3	84.4%	**92.3%**	**2.2e-3**	**86.3%**
RGB2-M	34.9%	13e-3	12.5%	64.1%	4.8e-3	69.8%	**67.9%**	**4.7e-3**	**71.0%**
RGB3-S	29.6%	21e-3	7.44%	91.3%	3.9e-3	90.3%	**97.4%**	**1.4e-3**	**92.5%**
RGB3-M	15.4%	22e-3	2.30%	67.5%	5.4e-3	60.5%	**77.9%**	**3.8e-3**	**68.6%**
RGB4-S	24.7%	20e-3	10.3%	86.7%	4.0e-3	78.3%	**90.7%**	**2.5e-3**	**83.3%**
RGB4-M	3.82%	32e-3	2.35%	56.9%	6.3e-3	58.2%	**67.9%**	**4.6e-3**	**77.3%**

References

Eslami, S., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Kavukcuoglu, K., and Hinton, G. E. (2016).
Attend, infer, repeat: Fast scene understanding with generative models.
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3225–3233.

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

Thank You!