Learning Descriptor Networks for 3D Shape Synthesis and Analysis

Jianwen Xie[1,*], Zilong Zheng[2,*], Ruiqi Gao[2], Wenguan Wang[2,3], Song-Chun Zhu[2], Ying Nian Wu[2] (*equal contribution)

[1] Hikvision Research Institute, [2] University of California, Los Angeles, [3] Beijing Institute of Technology

CVPR 2018 SALT LAKE CITY • JUNE 18-22

Abstract

This paper proposes a 3D shape descriptor network, which is a deep convolutional energy-based model, for modeling volumetric shape patterns. The maximum likelihood training of the model follows an "analysis by synthesis" scheme and can be interpreted as a mode seeking and mode shifting process. The model can synthesize 3D shape patterns by sampling from the probability distribution via MCMC such as Langevin dynamics. The model can be used to train a 3D generator network via MCMC teaching. The conditional version of the 3D shape descriptor net can be used for 3D object recovery and 3D object super-resolution. Experiments demonstrate that the proposed model can generate realistic 3D shape patterns and can be useful for 3D shape analysis.

3D shape descriptor network (3D DescriptorNet)

Probability density

The model is a 3D deep convolutional energy-based model defined on the 3D data Y, which is in the form of exponential tilting of a reference distribution:

$$p(Y;\theta) = \frac{1}{Z(\theta)} \exp\left[f(Y;\theta)\right] p_0(Y), \tag{1}$$

where $p_0(Y)$ is the reference distribution such as Gaussian white noise $p_0(Y) \propto \exp\left(-\|Y\|^2/2s^2\right)$, $f(Y;\theta)$ is a bottom-up 3D volumetric ConvNet whose parameters are denoted by θ. $Z(\theta) = \int \exp\left[f(Y;\theta)\right] p_0(Y) dY$ is the normalizing constant that is analytically intractable.

Analysis by synthesis

Suppose we observe 3D training examples $\{Y_i, i = 1, ..., n\}$ from an unknown distribution. The maximum likelihood learning seeks to maximize the log-likelihood function $L_p(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i;\theta)$. The gradient of $L_p(\theta)$ is

$$L_p'(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} f(Y_i;\theta) - \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta} f(Y;\theta)\right], \tag{2}$$

where the expectation E_θ is analytically intractable and has to be approximated by MCMC, e.g., Langevin dynamics, which iterates the following step:

$$Y_{\tau+\Delta\tau} = Y_\tau - \frac{\Delta\tau}{2}\left[\frac{Y_\tau}{s^2} - \frac{\partial}{\partial Y} f(Y_\tau;\theta)\right] + \sqrt{\Delta\tau}\,\epsilon_\tau, \tag{3}$$

where τ indexes the time steps, $\Delta\tau$ is the discretized step size, and $\epsilon_\tau \sim \mathrm{N}(0, I)$ is the Gaussian white noise. Suppose we draw $\tilde{n}$ samples $\{\tilde{Y}_i, i = 1, ..., \tilde{n}\}$ from $p(Y;\theta)$ according to (3), $L_p'(\theta)$ in (2) can be approximated by

$$L_p'(\theta) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} f(Y_i;\theta) - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \frac{\partial}{\partial\theta} f(\tilde{Y}_i;\theta). \tag{4}$$

Adversarial interpretation / Mode seeking and mode shifting

If we rewrite model (1) in the form of energy-based model $p(Y;\theta) \propto \exp\left(-\mathcal{E}(Y;\theta)\right)$, then the energy function $\mathcal{E}(Y;\theta) = \|Y\|^2/(2s^2) - f(Y;\theta)$. We rewrite equation (4) in the form of

$$L_p'(\theta) \approx \underbrace{\frac{\partial}{\partial\theta}\left[\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \mathcal{E}(\tilde{Y}_i;\theta) - \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}(Y_i;\theta)\right]}_{V(\{\tilde{Y}_i\};\theta)}. \tag{5}$$

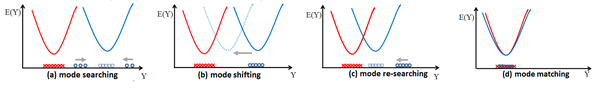
Adversarial interpretation:

The resulting algorithm approximately solves the minimax problem below

$$\max_{\theta} \min_{\{\tilde{Y}_i\}} V(\{\tilde{Y}_i\};\theta) \tag{6}$$

- The sampling step finds $\{\tilde{Y}_i\}$ to decrease V, since it searches for low energy modes in the landscape defined by $\mathcal{E}(Y;\theta)$ via stochastic gradient descent.
- The learning step finds θ to increase V, which can be interpreted as mode shifting by shifting the low energy modes from the synthesized examples $\{\tilde{Y}_i\}$ toward the observed examples $\{Y_i\}$.

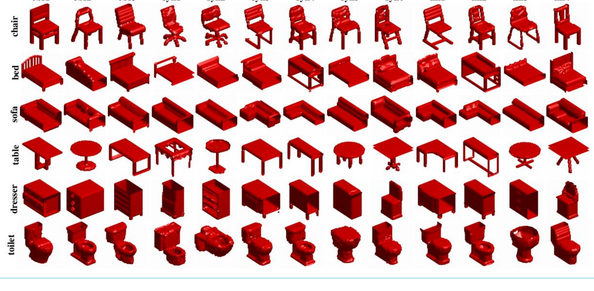
Mode seeking and mode shifting:



Red curve: true distribution; blue curve: learned distribution

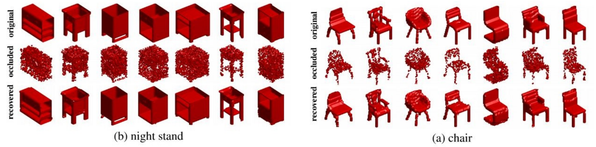
Application 1: 3D object synthesis

Each row displays one experiment, where the first 3 3D objects are some observed examples, columns 4, 5, 6, 7, 8, and 9 are 6 of the synthesized 3D objects. The nearest neighbors retrieved from the training set are shown in columns 10,11, 12, and 13 for the last four synthesized objects.



Application 2: 3D object recovery

We can perform recovery on occluded data by sampling from $p(Y_M|Y_{\tilde{M}},\theta)$, which is learned from fully observed training pairs $\{(Y_M^i, Y_{\tilde{M}}^i), i = 1, ..., n\}$. The sampling is accomplished by Langevin dynamics, which is the same as the one that samples from $p(Y;\theta)$, except that we fix the unmasked part $Y_{\tilde{M}}$ and only update the masked part Y_M through the Langevin dynamics.



(b) night stand (a) chair

Application 3: 3D object super-resolution

We can perform super-resolution (4×) on a low resolution (e.g., $16 \times 16 \times 16$) 3D objects by sampling from $p(Y_{high}|Y_{low},\theta)$, which is learned from fully observed training pairs $\{(Y_{high}^i, Y_{low}^i), i = 1, ..., n\}$.

In each iteration, we first up-scale Y_{low} by expanding each voxel into a $d \times d \times d$ block (where d is the scaling ratio) of constant intensity to obtain an up-scaled version Y'_{high} of Y_{low} and then run Langevin dynamics staring from Y'_{high}.

Application 4: 3D object classification

We first train a single model on all categories of the training set of ModelNet10 dataset in an unsupervised manner. Then we use the model as a feature extractor. We train a multinomial logistic regression classifier from labeled data based on the extracted feature vectors for classification.

Method	Accuracy
Geometry Image	88.4%
PANORAMA-NN	91.1%
ECC	90.0%
3D ShapeNets	83.5%
DeepPano	85.5%
SPH	79.8%
VConv-DAE	80.5%
3D-GAN	91.0%
3D DescriptorNet (ours)	**92.4%**

Conclusion

(1) We propose the 3D DescriptorNet for volumetric objects.
(2) We propose the conditional 3D DescriptorNet for 3D object recovery and 3D object super resolution.
(3) The model can be used to train a 3D generator via cooperative training.
(4) The model is useful for semi-supervised learning in 3D classification.

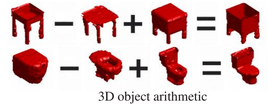
Application 5: Teaching 3D generator net

The 3D generator model $Y = g(Z;\alpha) + \epsilon$, where $Z \sim \mathrm{N}(0, I_d)$ can be trained simultaneously with the 3D DescriptorNet in a cooperative training scheme:

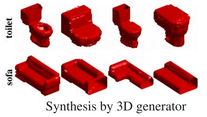
- **Input**: training data $\{Y_i, i = 1, ..., n\}$, and number of learning steps T.
- **Output**: model parameters θ and α, and synthetic data $\{\tilde{Y}_i, \tilde{Y}_i, i = 1, ..., \tilde{n}\}$

1. Let $t \leftarrow 0$, initialize θ and α.
2. Repeat
3. **Initializing mode seeking**: For $i = 1, ..., \tilde{n}$, generate $Z_i \sim \mathrm{N}(0, I_d)$, and generate $\hat{Y}_i = g(Z_i;\alpha^{(t)}) + \epsilon_i$.
4. **Mode seeking**: For $i = 1, ..., \tilde{n}$, starting from $\hat{Y}_i$, run l steps of Langevin dynamics to obtain $\tilde{Y}_i$, each step following equation (3).
5. **Mode shifting**: Update $\theta^{(t+1)} = \theta^{(t)} + \gamma_t L_p'(\theta^{(t)})$, where $L_p'(\theta^{(t)})$ is computed according to (4).
6. **Learning from mode seeking**: Update $\alpha^{(t+1)} = \alpha^{(t)} - \eta_t {L_q}'(\alpha^{(t)})$, where ${L_q}'(\alpha^{(t)})$ is computed by $\frac{\partial}{\partial\alpha}[\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \|\tilde{Y}_i - g(Z_i;\alpha)\|^2]$.
7. Let $t \leftarrow t + 1$
8. Until $t = T$

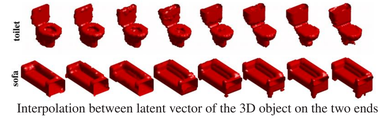
We evaluate a 3D generator trained by a 3D DescriptorNet by experiments on generator synthesis, latent space interpolation and 3D object arithmetic.



3D object arithmetic



Synthesis by 3D generator



Interpolation between latent vector of the 3D object on the two ends

Webpage

http://www.stat.ucla.edu/~jxie/3DDescriptorNet/3DDescriptorNet.html

References

- Jianwen Xie*, Yang Lu*, Song-Chun Zhu, Ying Nian Wu. A theory of generative convnet, *ICML* 2016. (* equal contributions)
- Jianwen Xie, Yang Lu, Ruiqi Gao, Ying Nian Wu. Cooperative Learning of Energy-Based Model and Latent Variable Model via MCMC Teaching, *AAAI* 2018.