UNSUPERVISED LEARNING OF SENSORIMOTOR AFFORDANCES BY STOCHASTIC FUTURE PREDICTION

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Rybkin O.*, Pertsch K.*, Jaegle A., Derpanis K. and Daniilidis K.

{oleh, pertsch}@cis.upenn.edu, ajaegle@upenn.edu, kosta@ryerson.ca, kostas@cis.upenn.edu

Ryerson University

Abstract

Intelligent perception must capture not only a scene's static content, but also its **affordances**: how an agent's actions can affect the scene. We propose an unsupervised method to learn an environment's sensorimotor affordances. We use a recurrent latent variable that is

(i) *minimal* in sensitivity to static content and
(ii) *compositional* in nature.

We show these two properties are sufficient to induce representations that are reusable across different scenes with shared degrees of freedom.

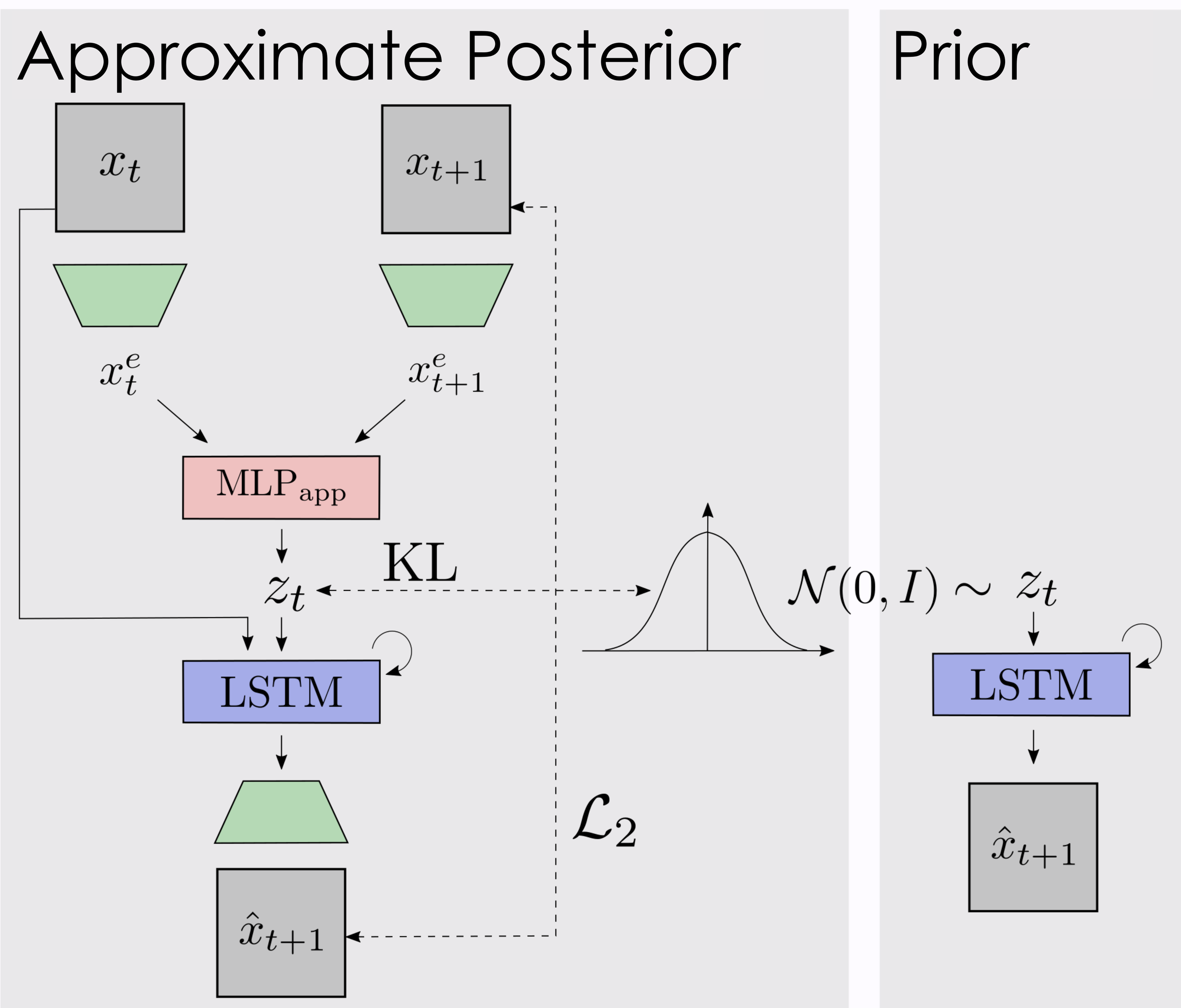
Background

Stochastic video prediction

We use a recurrent latent variable z to capture the distribution of possible future frames [1, 2].

In deterministic environments, z represents the agent's actions.

Balancing the two parts of the objective allows to recover a minimal representation [3].



Variational Information Bottleneck

The Information Bottleneck [4] objective for a representation $\boldsymbol{Z}$, input $\boldsymbol{X}$, output $\boldsymbol{Y}$:

$$\max I(Z, Y) \text{ s.t. } I(X, Z) \leq I_c.$$

VIB [5] optimizes the above using the Lagrangian:

$$\sum_i \left[\mathbb{E}_{p(z|x)} \log q(x_i|Z) - \beta \mathrm{KL}[p(Z|x_i), p(Z)]\right]$$

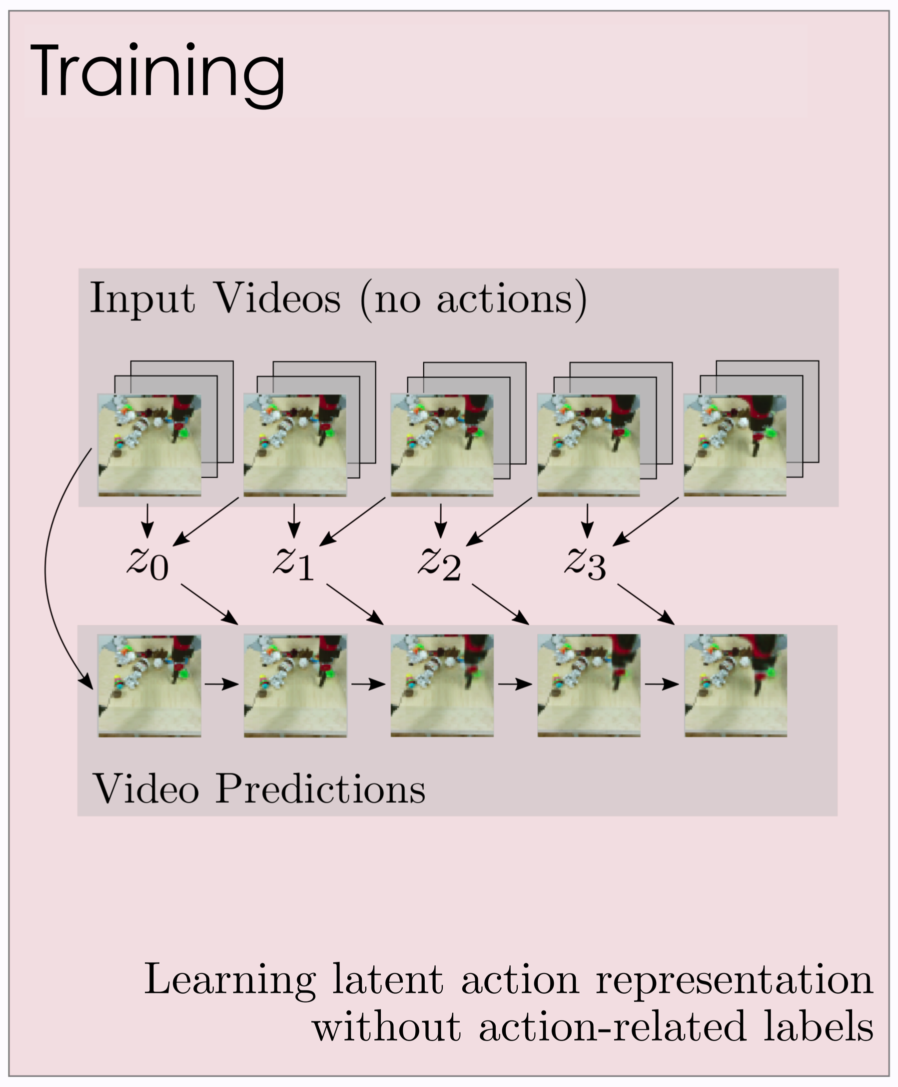
Takeaways

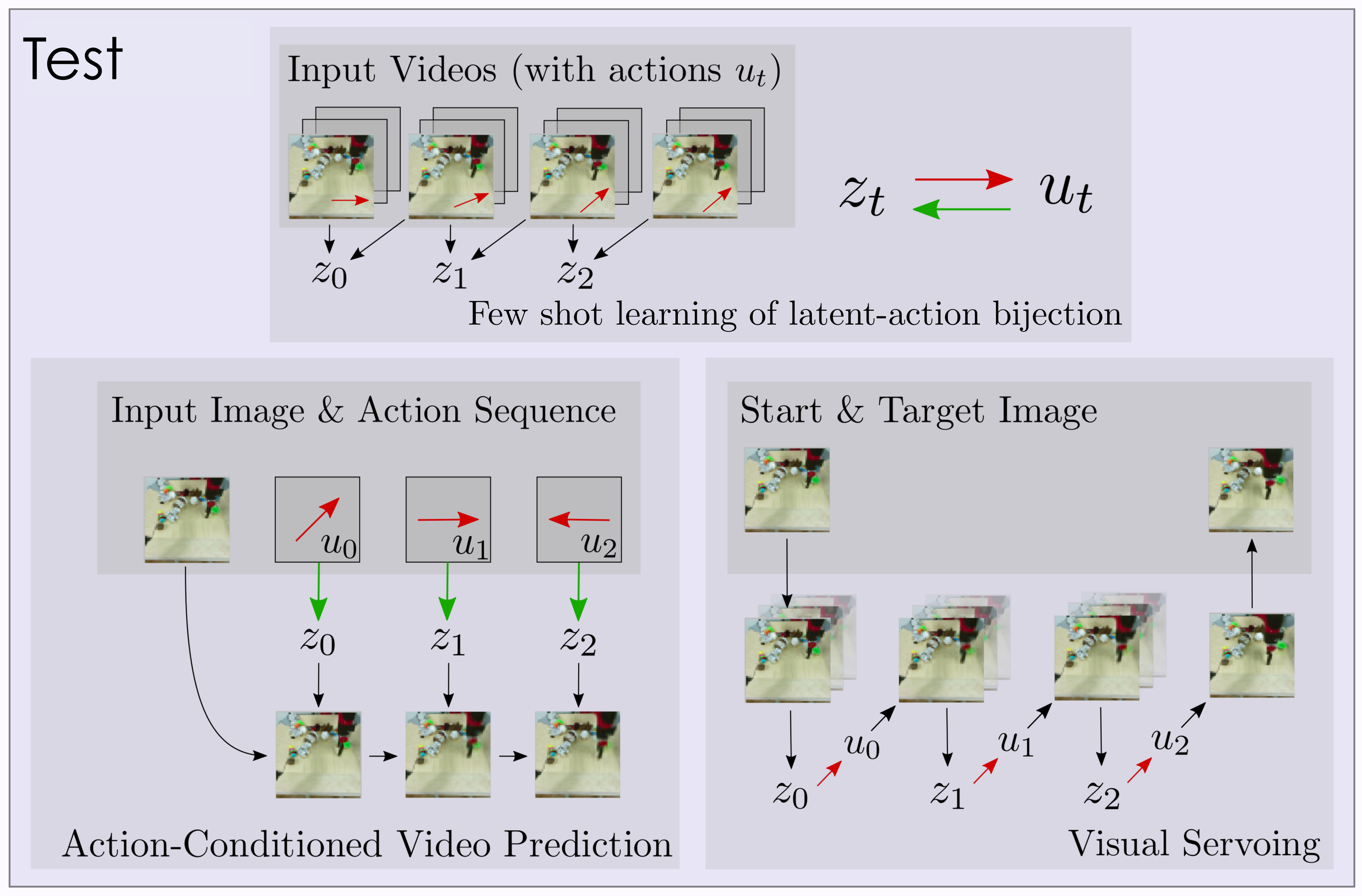
The *inductive biases* of minimality and composability provide sufficient constraints for learning affordances in an unsupervised way.

The learned representation is *disentangled* from the static scene content.

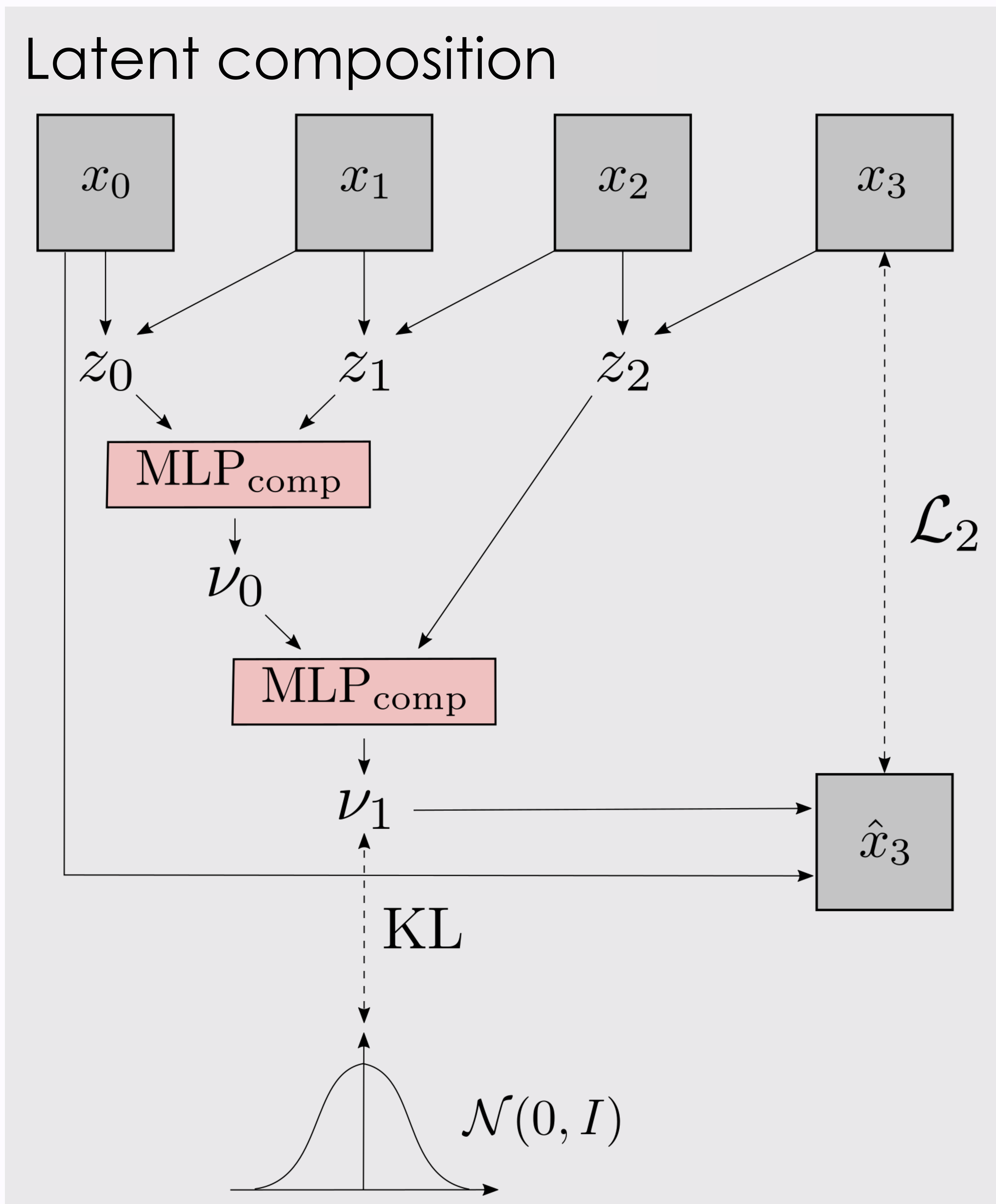
The disentanglement allows the representation to be used for *visual servoing* and *action-conditioned prediction.*

Approach



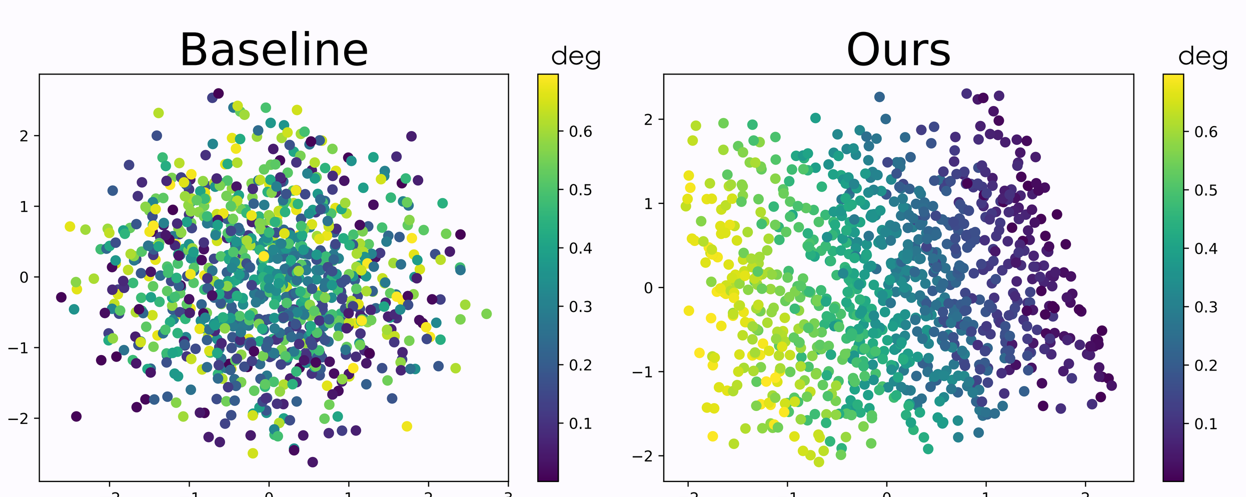


Composability training



The two components of the VIB loss encourage ν to represent the trajectory, while being minimal in the sense of *Information Bottleneck* [4]. In turn, this forces z to be suitable for composition.

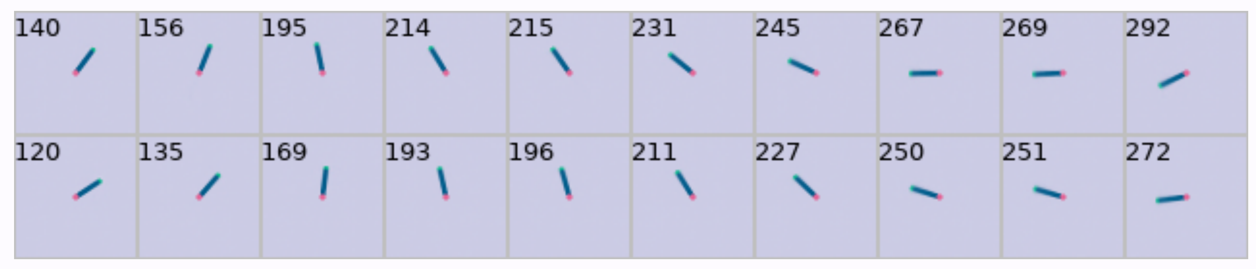
Learned disentanglement



PCA of the latent samples z colored by the value of true action u.

Experiments

Trajectory transplantation



Action-conditioned prediction

Method	Reacher Error [deg]	BAIR Error [px]
Random	26.6 ± 21.5	-
Baseline	22.6 ± 17.7	3.6 ± 4.0
Ours	$\mathbf{2.9 \pm 2.1}$	$\mathbf{3.0 \pm 2.1}$
Supervised	2.6 ± 1.8	2.0 ± 1.3

Visual servoing

References

[1] Denton, E. and Fergus, R., Stochastic Video Generation with a Learned Prior, in *ICML*, 2018.
[2] Lee, A., Zhang, R., Ebert, F., Abbeel, P., Finn, C. and Levine, S., Stochastic Adversarial Video Prediction, *arXiv*:1804.01523, 2018.
[3] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A., β-VAE: Learning basic visual concepts with a constrained variational framework, in *ICLR*, 2017.
[4] Shwartz-Ziv, R. and Tishby, N., Opening the black box of deep neural networks via information, *arXiv*:1703.00810, 2017.
[5] Alemi, A., Fischer, I., Dillon, J. and Murphy, K. Deep variational information bottleneck, in *ICLR*, 2018.