Disentangling Latent Hands for Image Synthesis and Pose Estimation

Linlin Yang (University of Bonn) and Angela Yao (National University of Singapore)

UNIVERSITÄT BONN
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

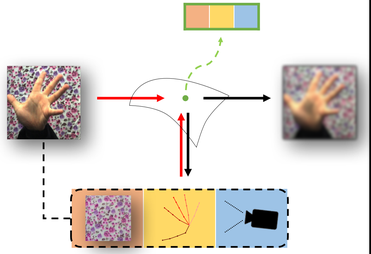
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Motivation

Hand image synthesis and pose estimation from RGB images are both highly challenging tasks due to the large discrepancy between factors of variation ranging from image background content to camera viewpoint. To better analyze these ***factors of variation***, we propose the use of disentangled representations and a disentangled variational autoencoder (dVAE) that allows for ***specific sampling*** and ***inference*** of these factors.

Illustration

Illustration of dVAE. The red lines denote variational approximations while the black lines denote the generative model. With the help of ***labelled factors of variations*** (eg pose, viewpoint and image content), we here learn a disentangled and specifiable representation for RGB hand images in ***a VAE framework***.



Assumption & Steps

Assumption:

(1) The latent variable z is distributed as a multivariate Gaussian (See VAE);

(2) The latent variable z can be deterministically decomposed (See *Graphical models*).

Steps:

(1) Disentangling step: Using the labelled factors of variations to learn a disentangled latent space;

(2) Embedding step: Embedding the image latent space into this disentangled latent space.

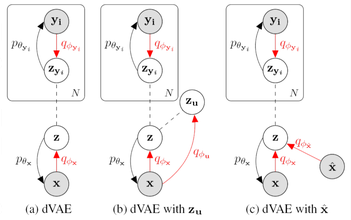
Graphical models

The shaded nodes represent observed variables while un-shaded nodes are latent.

The red solid lines denote variational approximations q_ϕ and the black solid lines denote the generative models p_θ.

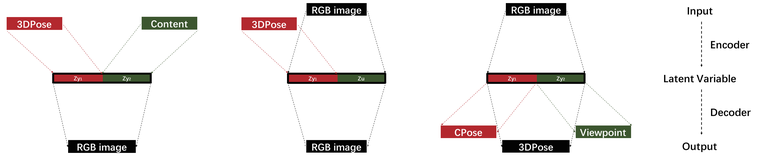
The dashed lines here denote the deterministically constructed variables.

Here (a)(b) are for image synthesis and (c) is for pose estimation.

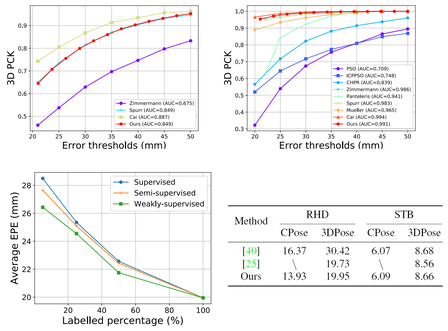


z_{y_i} is directly associated with observed variables y_i. z_u is an extra latent factor which is not independently associated with any observed variables.

Inference models



Pose Estimation

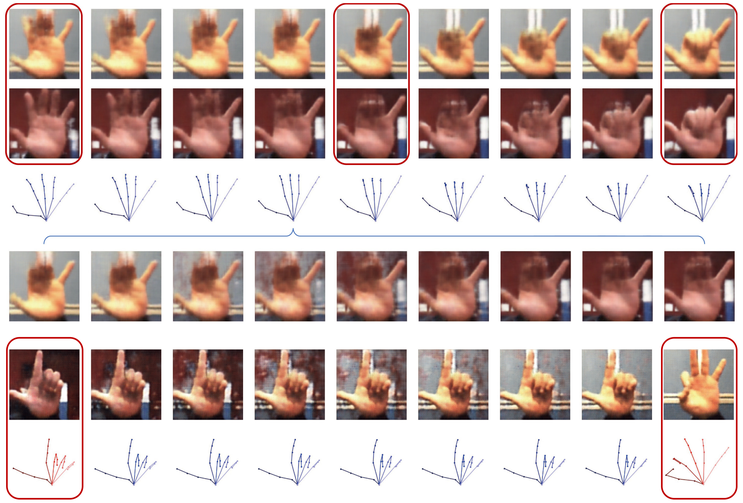


Method	RHD		STB	
	CPose	3DPose	CPose	3DPose
[40]	16.37	30.42	6.07	8.68
[25]	\	19.73	\	8.56
Ours	13.93	19.95	6.09	8.66

Experiments show that our dVAE estimates 3D hand poses from RGB images with accuracy competitive with state-of-the-art on two public benchmarks, RHD and STB.

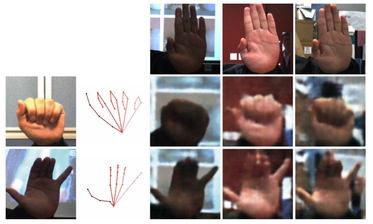
Furthermore, dVAE can be trained respectively for semi- / weakly-supervised setting and makes full use of additional information.

Image Synthesis



We show the synthesized images of latent space walk and pose transfer.

Experiments show that our dVAE can synthesize realistic images of hand specifiable by pose and image background content.



References

[1] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, 2017.

[2] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018.

[3] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. In *ICLR*, 2018.