A Variational U-Net for Conditional Appearance and Shape Generation

Patrick Esser* Ekaterina Sutter* Björn Ommer

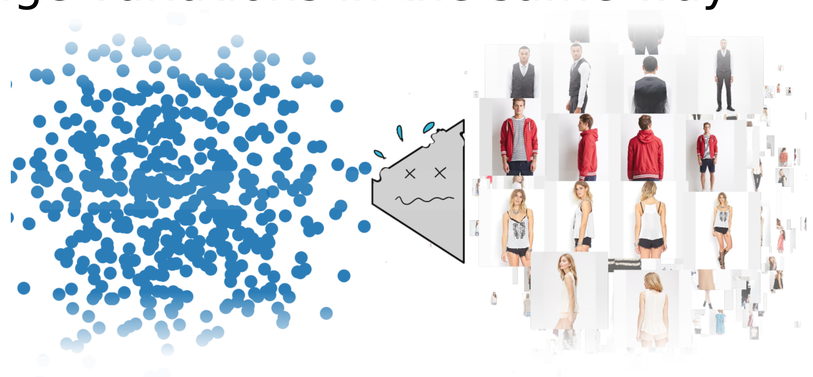
{firstname.lastname}@iwr.uni-heidelberg.de

HCI, IWR, Heidelberg University

*equal contribution

Motivation

Many previous methods try to generate all image variations in the same way.



However, we can identify two groups of variations with different characteristics:

Variations in

Appearance
color, texture, identity, ...

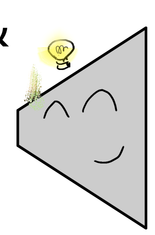
Shape
shape, pose, viewpoint, ...

a linear interpolation between these changes works well

deformations cannot be explained in the same way

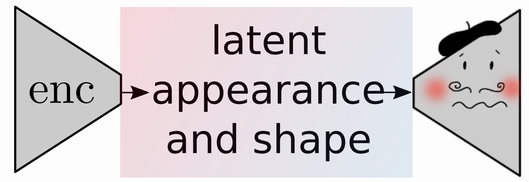
A model which understands this difference must be able to control appearance and shape seperately.

How can we disentangle & generate appearance and shape?



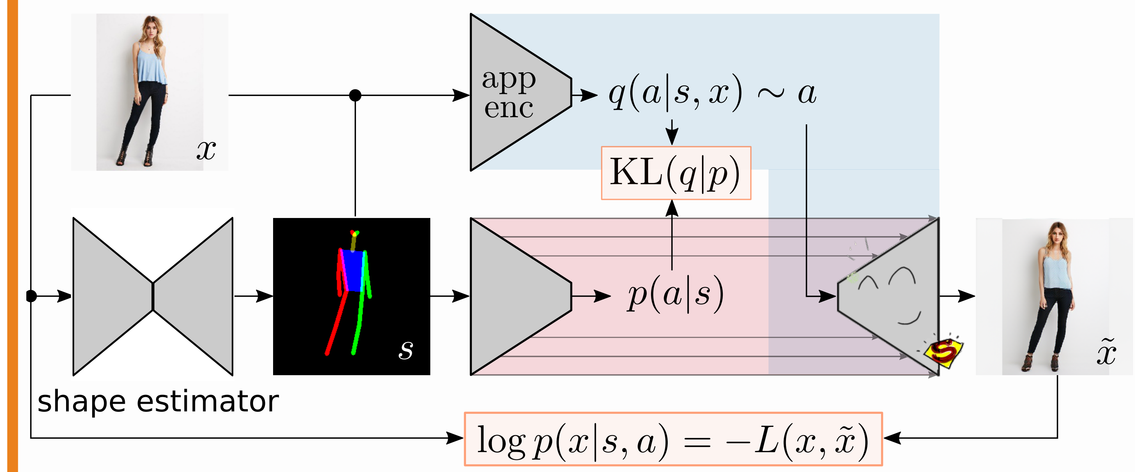
Method

Model images as being generated by two underlying factors: appearance and shape



when modeling both factors as latent variables with a VAE, we cannot disentangle them

$\Longrightarrow$ use an estimate of the shape and then learn to disentangle appearance

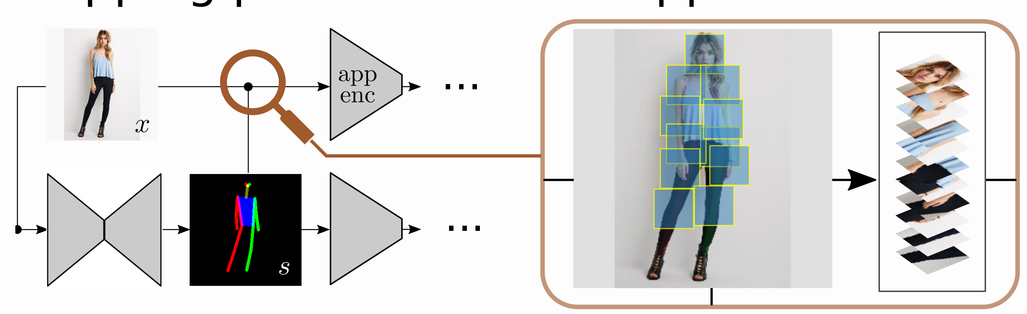


Fit model using shape conditional ELBO:

$$\log p(x|s) \geq \mathbb{E}_{a\sim q(a|x,s)} \log p(x|s,a) \quad \longleftarrow \text{reconstruction}$$
$$- \mathrm{KL}(q(a|s,x)|p(a|s)) \quad \longleftarrow \text{disentanglement}$$

- perceptual loss of VGG19 for reconstruction
- retain spatial information of shape using skip-connections from shape estimate to generator
- KL term penalizes redundancies between appearance and shape $\Longrightarrow$ disentangling of appearance

Cropping parts to localize appearance:



crop part boxes and stack them channelwise as input for appearance encoder

Results

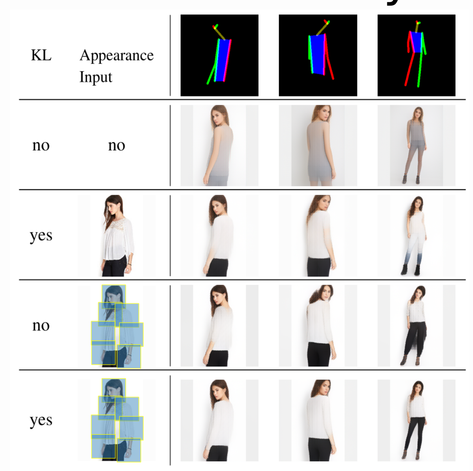
Disentangled synthesis

Variations in appearance and shape can be synthesized independently.

Appearance and shape can be transferred between images.

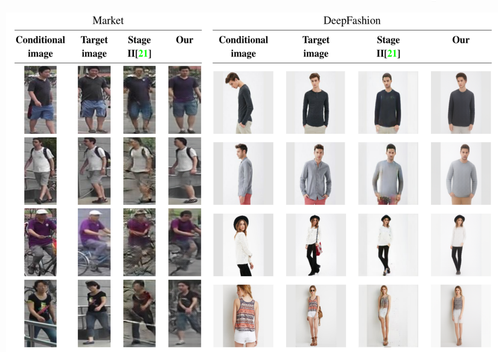
Other sources of shape

Ablation study



Edges can be used as shape estimates, too. Ablation experiments show that disentanglement with both the KL term and object parts is required for highly articulated objects.

Qualitative and quantitative comparisons



method	Market1501				DeepFashion			
	IS		SSIM		IS		SSIM	
	mean	std	mean	std	mean	std	mean	std
real data	3.678	0.274	1.000	0.000	3.415	0.399	1.000	0.000
PG^2 G1-poseMaskedLoss	3.326	–	0.340	–	2.668	–	0.779	–
PG^2 G1+D	**3.490**	–	0.283	–	**3.091**	–	0.761	–
PG^2 G1+G2+D	3.460	–	0.253	–	3.090	–	0.762	–
pix2pix	2.289	0.0489	0.166	0.060	2.640	0.2171	0.646	0.067
our	3.214	0.119	**0.353**	0.097	3.087	0.2394	**0.786**	0.068

(a)

(b)

method	our	pix2pix	PG^2
COCO	**23.23**	59.26	–
DeepFashion	**7.34**	15.53	19.04
Market1501	**54.60**	59.59	59.95

(c)

dataset	Our		PG^2	
	$\|\|std\|\|$	max pairwise dist	$\|\|std\|\|$	max pairwise dist
market1501	**55.95**	**125.99**	67.39	155.16
deepfashion	**59.24**	**135.83**	69.57	149.66
deepfashion	**56.24**	**121.47**	59.73	127.53

Qualitative and quantitative comparisons show improvements over previous methods. (a) SSIM scores are improved, (b) keypoints are more accurately recovered from synthetic images and (c) appearance embeddings are more compact.

Video, more results, code and data

https://compvis.github.io/vunet/