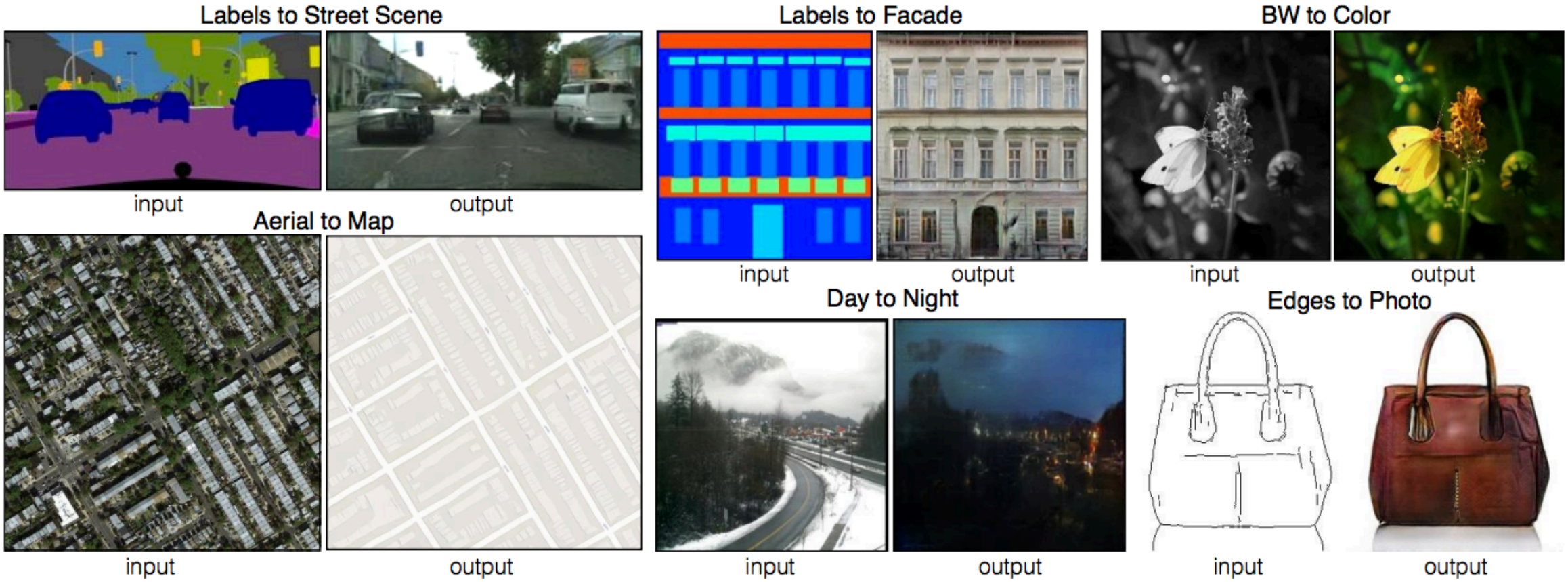
A Variational U-Net for Conditional Apperance and Shape Generation

Patrick Esser, Ekaterina Sutter, Bjorn Ommer
IWR, Heidelberg University, Germany
In CVPR, 2018.

Sharer: Du Ang

2018.05.30

Motivation



pix2pix results – no spatial deformation

Motivation

(a) DeepFashion

(b) Market-1501

PG^2 results – two-stage model, fully-supervised

Motivation

The results of image synthesis deteriorate in case of spatial deformations.

Because they generate images of objects directly, rather than modeling the intricate interplay of their inherent shape and apperance.

Approach

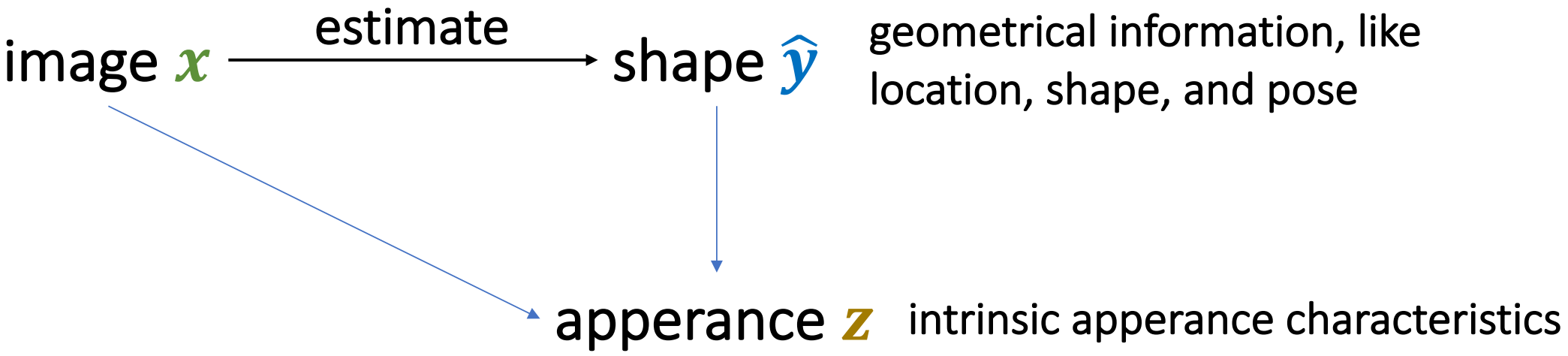
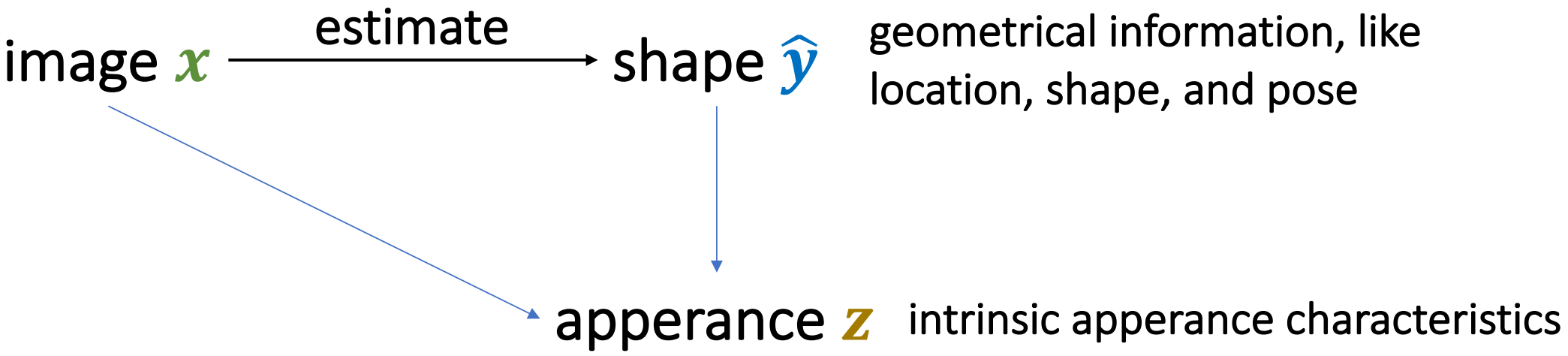


image x
estimate
shape $\hat{y}$
geometrical information, like location, shape, and pose
apperance z
intrinsic apperance characteristics

Approach



Why $\hat{y} \rightarrow z$?

To capture potential interrelations between shape and apperance.

Approach

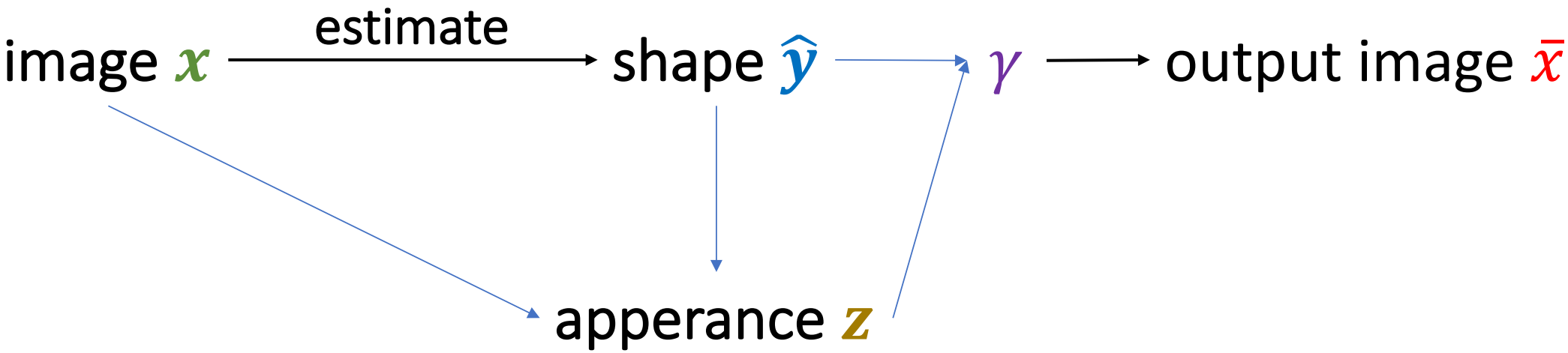
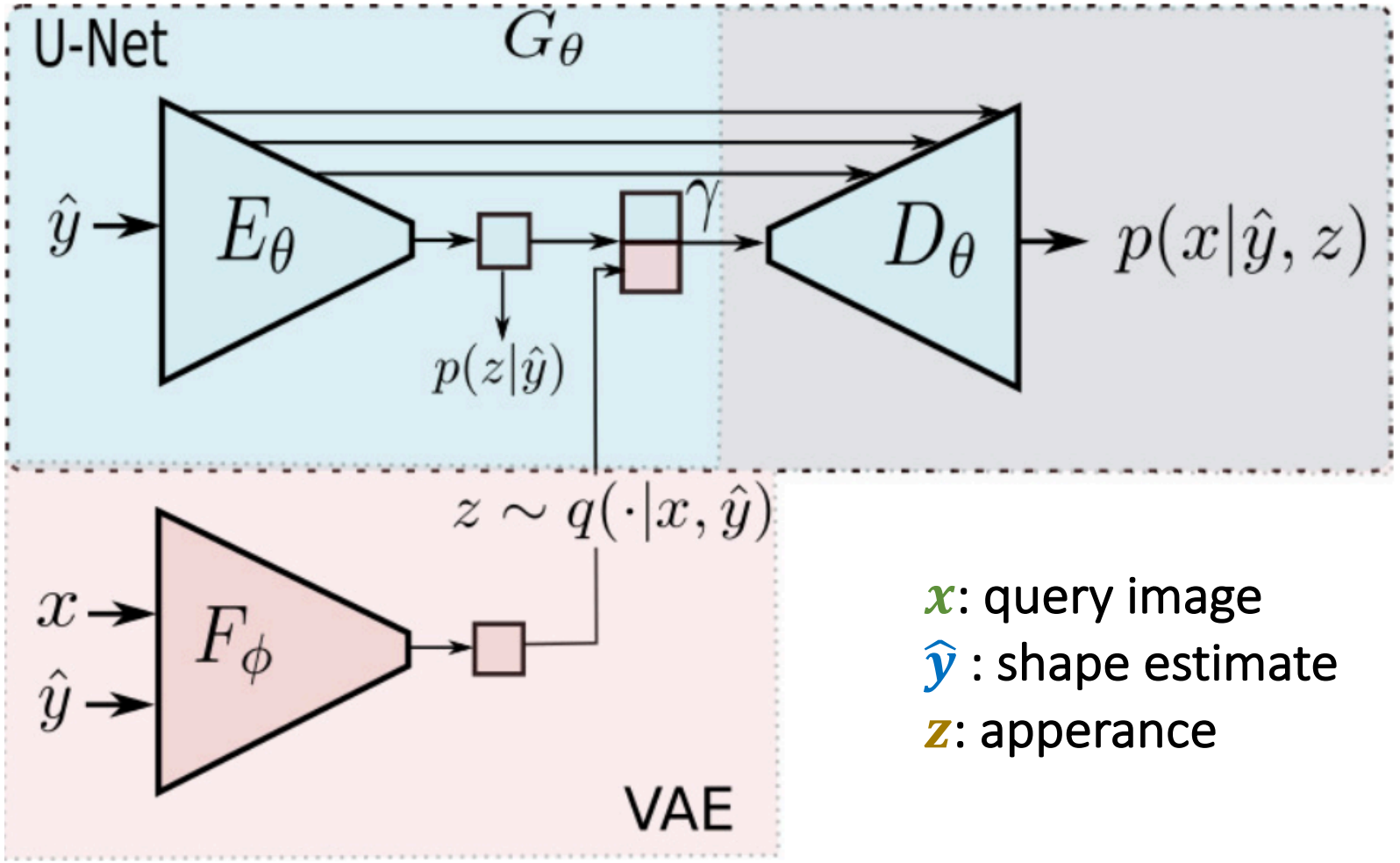


image x
estimate
shape $\hat{y}$
γ
output image $\bar{x}$
apperance z

Approach



$\boldsymbol{x}$: query image

$\hat{\boldsymbol{y}}$: shape estimate

$\boldsymbol{z}$: apperance

Approach

$$\mathcal{L}(x, \theta, \phi) = -KL(q_\phi(z|x, \hat{y})||p_\theta(z|\hat{y})) + \sum_k \lambda_k \|\Phi_k(x) - \Phi_k(G_\theta(\hat{y}, z))\|_1$$

where Φ is a network for measuring perceptual similarity (in our case VGG19 [37]) and λ_k, k are hyper-parameters that control the contribution of the different layers of Φ to the total loss.

Experiments

Datasets

- shoes
- handbags
- Market-1501
- DeepFashion
- COCO

Experiments

Image reconstruction and apperance sampling

Figure 3: Generating images with only the edge image as input (GT image (left) is held back). We compare our approach to pix2pix on the datasets of shoes [43] and handbags [49]. On the right: sampling from our latent appearance distribution.

Experiments

Image reconstruction and apperance sampling

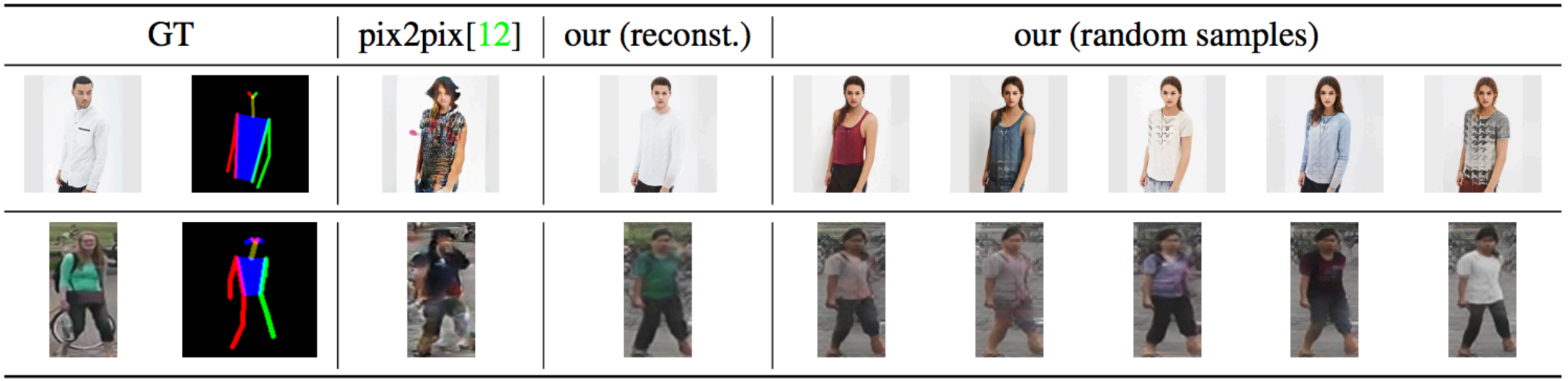


Figure 4: Generating images based only the stickman as input (GT image is held back). We compare our approach with pix2pix [12] on Deepfashion and Market-1501 datasets. On the right: sampling from our latent appearance distribution.

Experiments

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

Table 1: Inception scores (IS) and structured similarities (SSIM) of reconstructed test images on DeepFashion and Market1501 datasets. Our method outperforms both pix2pix [12] and PG^2 [24] in terms of SSIM. As to IS the proposed method performs better than pix2pix and obtains comparable results to PG^2.

Experiments

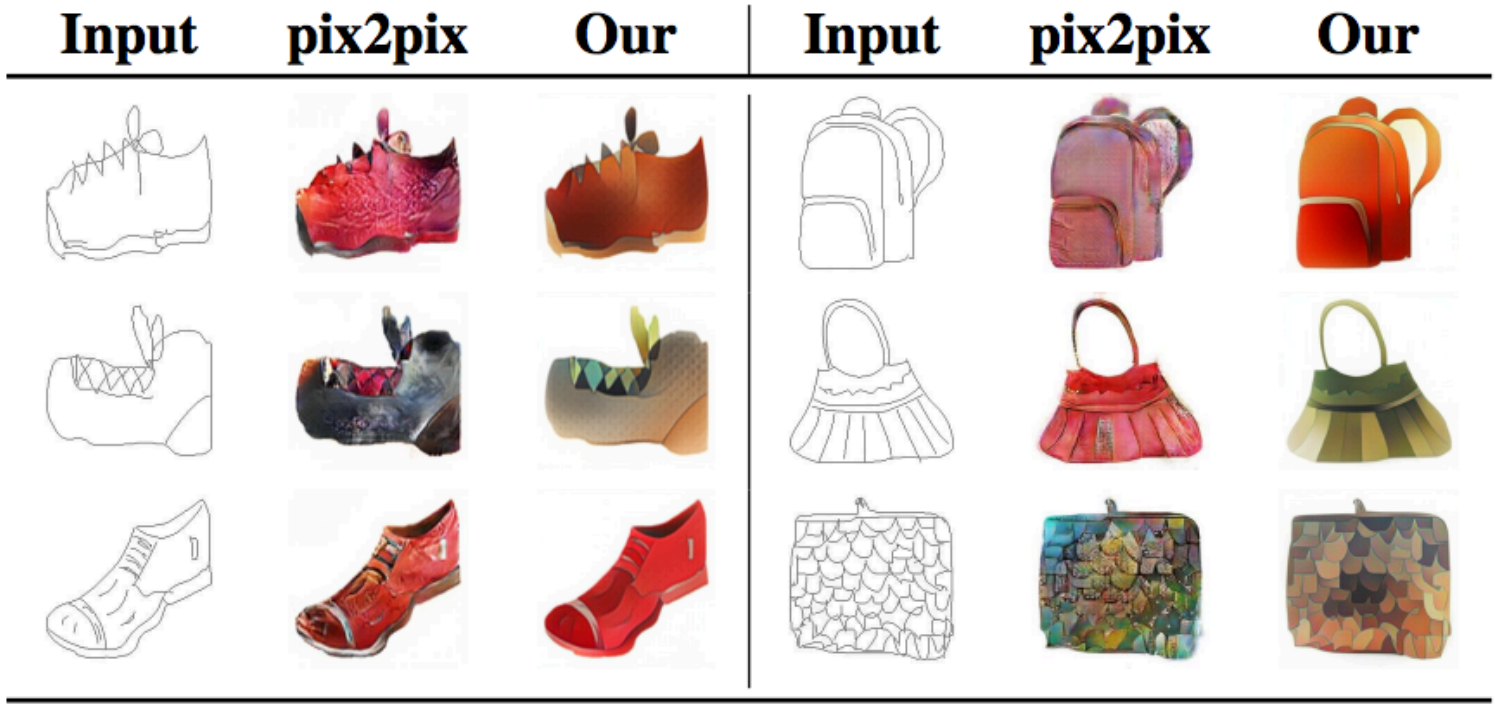


Figure 5: Colorization of sketches: we compare generalization ability of pix2pix [12] and our model trained on real images. The task is to generate plausible appearances for human-drawn sketches of shoes and handbags [9].

Variational U-Net generalizes better to sketchy drawings made by humans.

Experiments

Independent transfer of shape and apperance

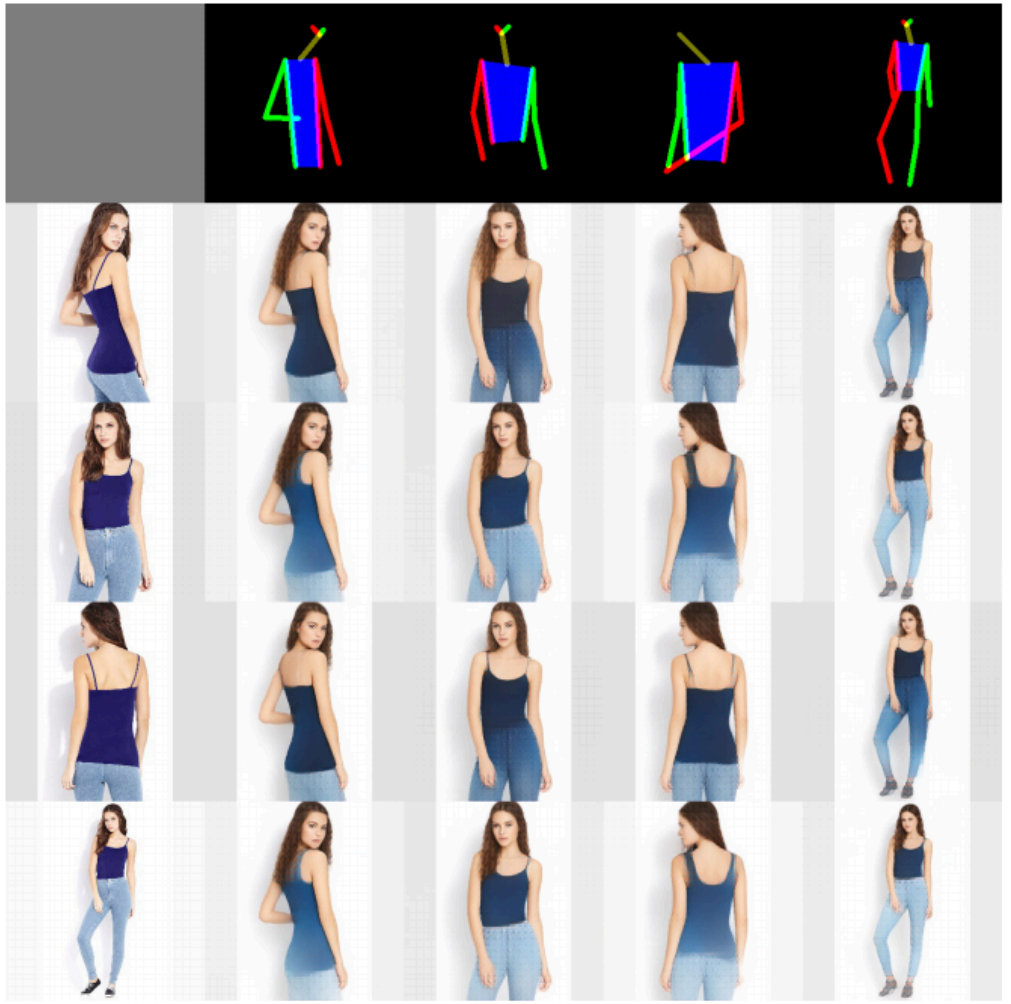


Figure 7: Stability of appearance transfer on DeepFashion. Each row is synthesized using appearance information from the leftmost image and each column is synthesized from the pose in the first row. Notice that inferred appearance remains constant across a wide variety of viewpoints.

Experiments

Independent transfer of shape and apperance

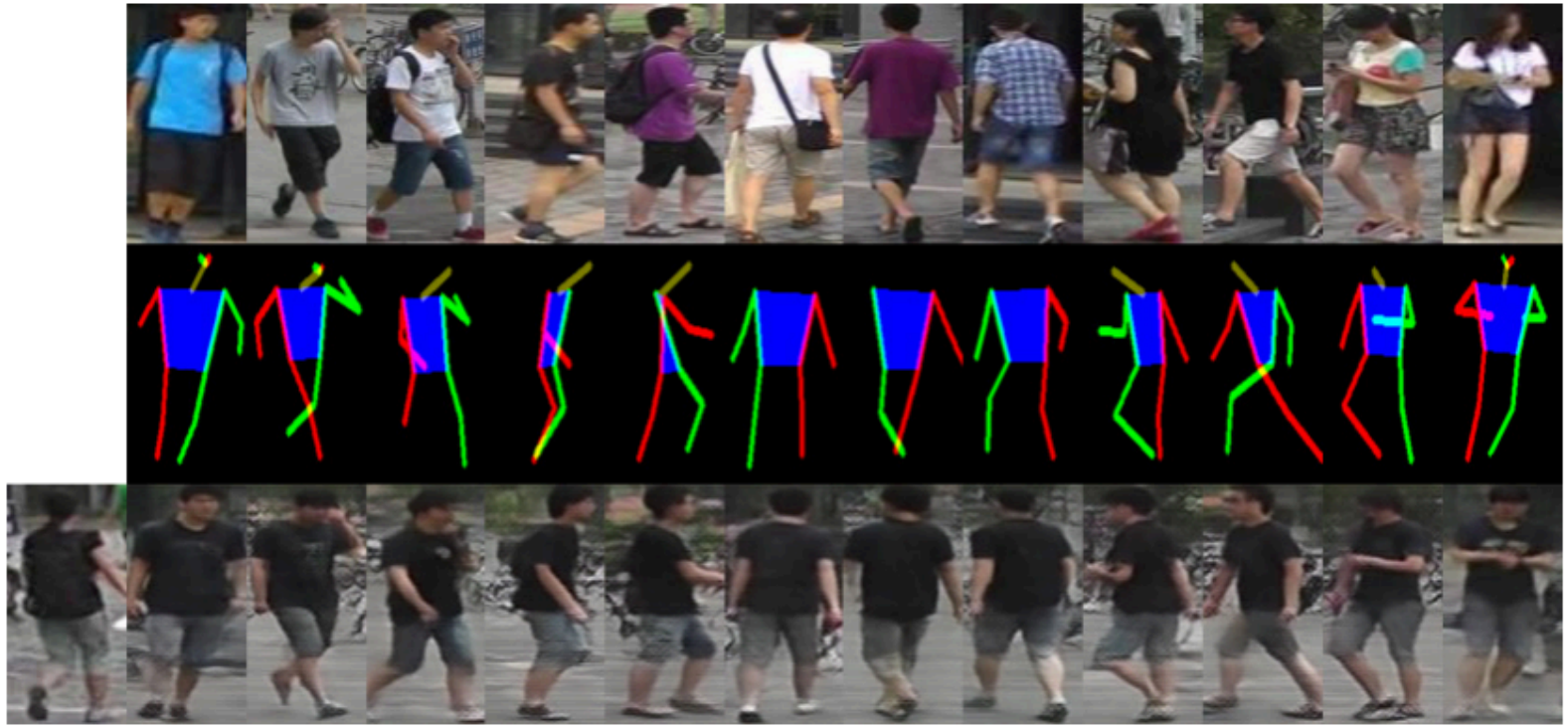


Figure 6: Appearance transfer on Market-1501. Appearance is provided by image on bottom left. $\hat{y}$ (middle) is automatically extracted from image at the top and transferred to bottom.

Experiments

Re-estimate joint points to indicate how good shape is preserved.

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

Table 2: Automatic body joint detection is applied to images of humans synthesized by our method, pix2pix, and PG^2. The L2 error of joint location is presented, indicating how good shape is preserved. The error is measured in pixels based on a resolution of 256×256.

Experiments

Comparing image transfer against PG^2

Figure 8: Comparing image transfer against PG^2. Left: Results on Market. Right: Results on DeepFashion. Appearance is inferred from the conditional image, the pose is inferred from the target image. Note that our method does not require labels about person identity.

Experiments

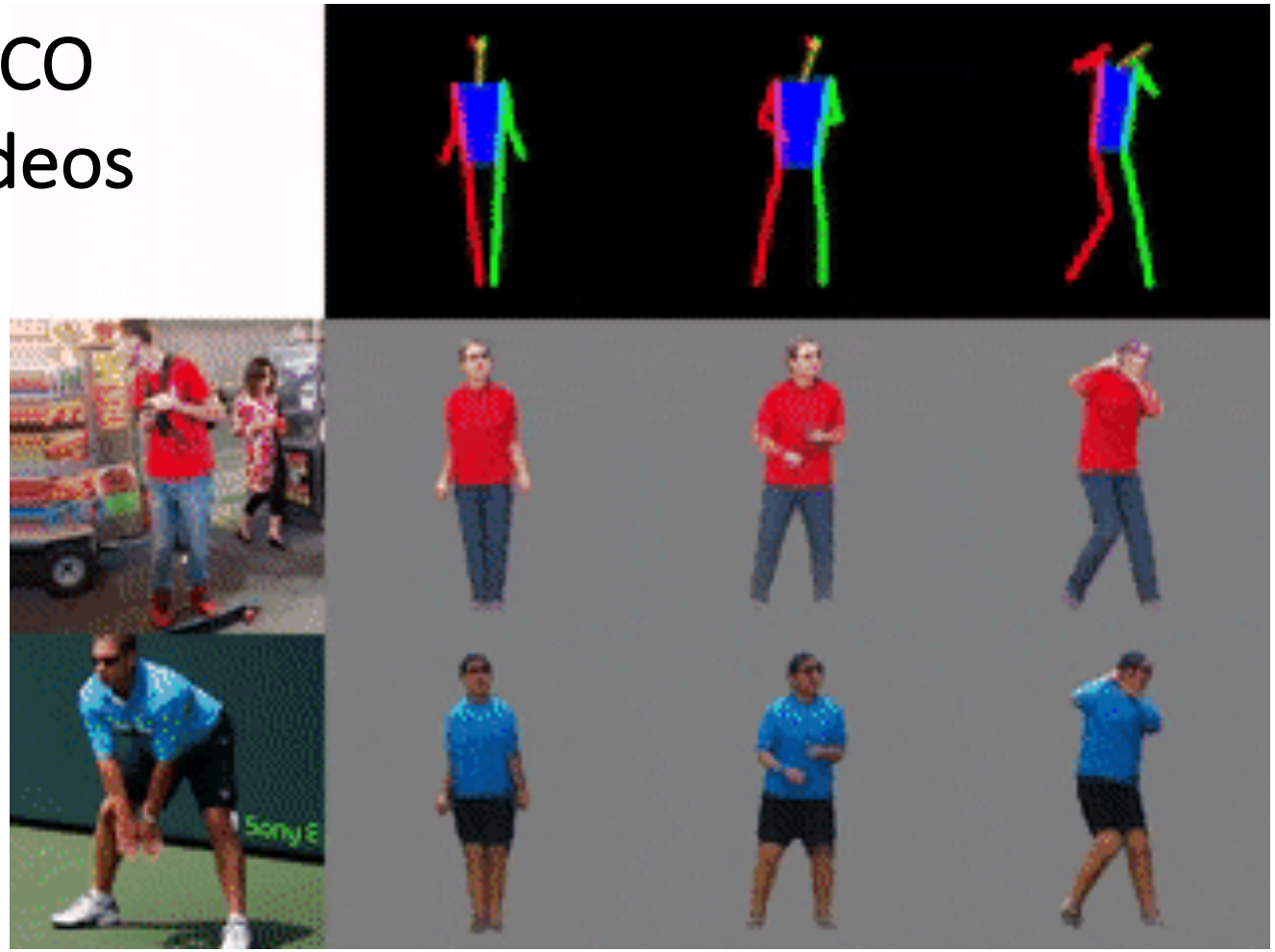
Comparing image transfer against PG^2

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

Table 3: Given an image its appearance is transferred from an image to different target poses. For these synthesized images, the unwanted deviation in appearance is measured using a pairwise perceptual VGG16 loss.

Experiments

Results on COCO synthesizing videos



Experiments

Ablation study

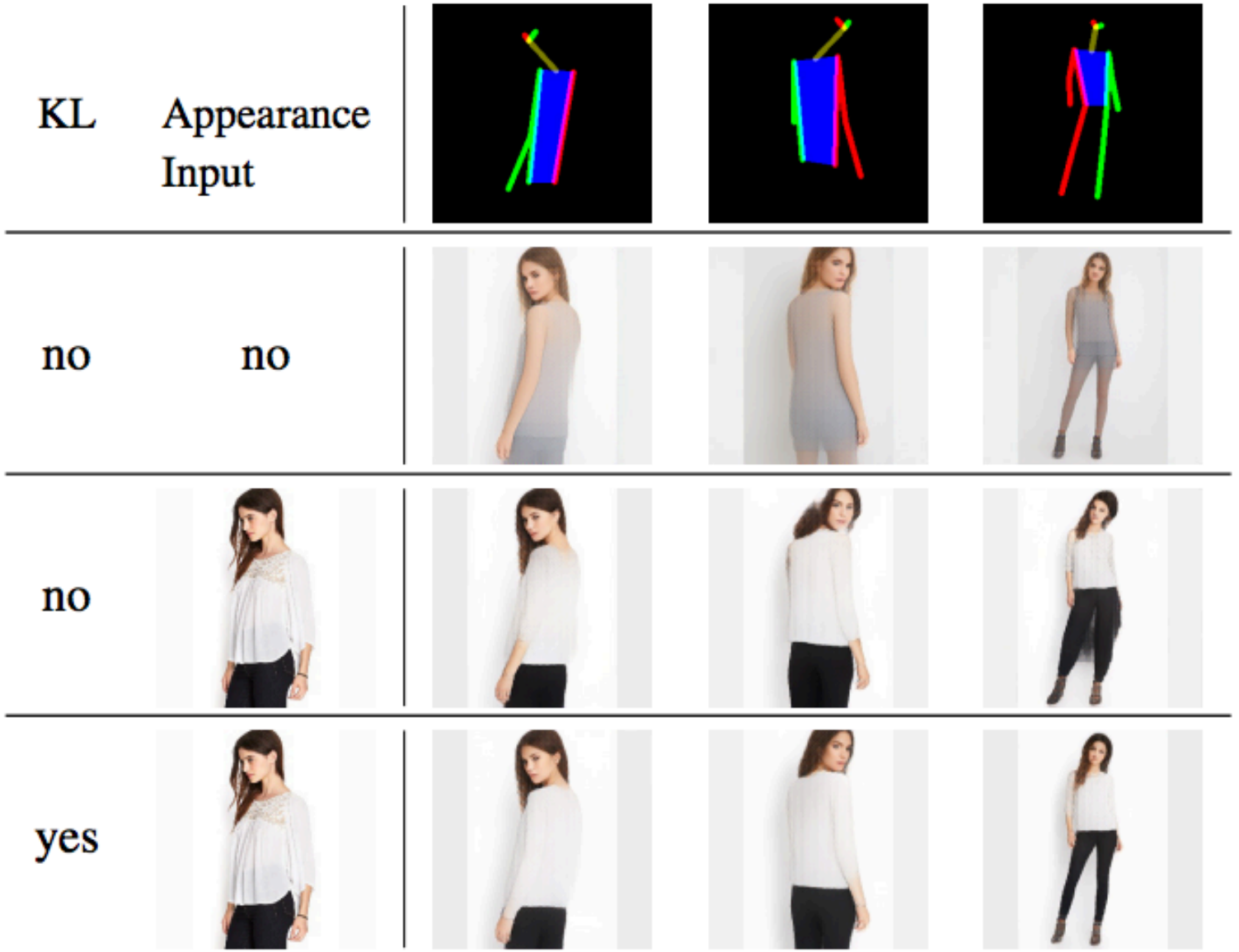


Figure 9: Ablation study on the task of appearance transfer. See Sec. 4.4.

Further reading

- Image-to-Image Translation with Conditional Adversarial Nets. In CVPR, 2017.
- Pose Guided Person Image Generation. In NIPS, 2017.
- Disentangled Person Image Generation. In CVPR, 2018.
- Exemplar Guided Unsupervised Image-to-Image Translation.
  In NIPS (under review), 2018.