Supplementary materials for Paper Submission 1449:
A Variational U-Net for Conditional Appearance and Shape Generation

A. Network structure

The parameter n of residual blocks in the network (see section 4) may vary for different datasets. For all experiments in the paper the value of n was set to 7. Below we provide a detailed visualization the architecture of the model that generates 128×128 images and has $n = 8$ residual blocks.

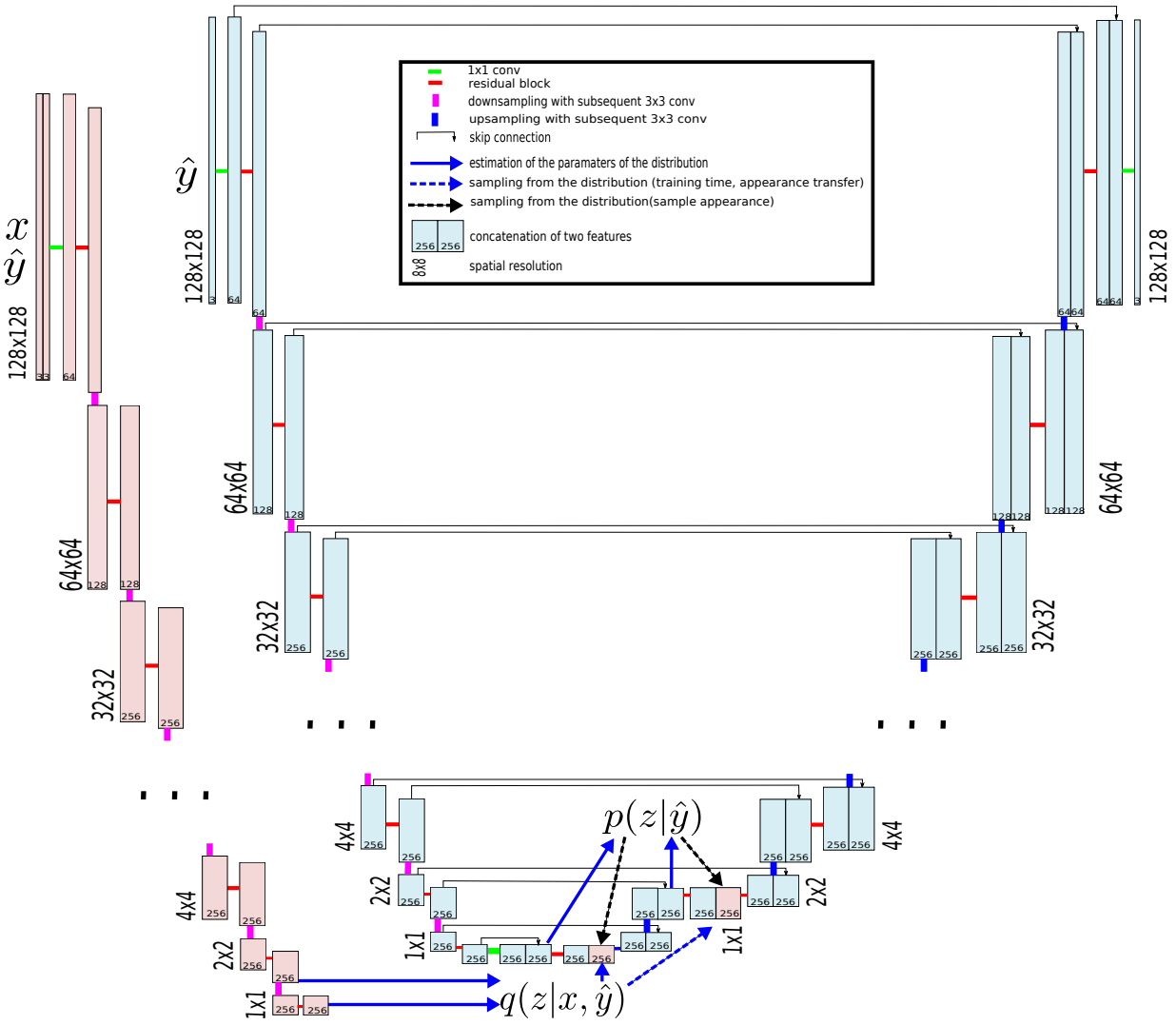

Figure 10: Network architecture with 8 residual blocks for 128×128 images.

B. Examples of appearance sampling in different datasets

We show more examples highlighting the ability of our model to produce diverse samples similar to the results shown in Fig. 3 and 4. In Fig. 11 we condition on edge images of shoes and handbags and sample the appearance from the learned prior. We also run pix2pix multiple times to compare the diversity of the produced samples. A similar experiment is shown in Fig. 12, where we condition on human body joints instead of edge images.

Figure 11: Generating images based only on the edge image as input (GT original image and corresponding edge image are held back). We compare our approach with pix2pix [12]. On the right: each odd row shows images synthesized by pix2pix, each even row presents samples generated by our model. Here again our first image (column 2) is a generation with original appearance, whereby for the 5 following images we sample appearance from the learned prior distribution. The GT images are taken from shoes [43] and handbags [49].

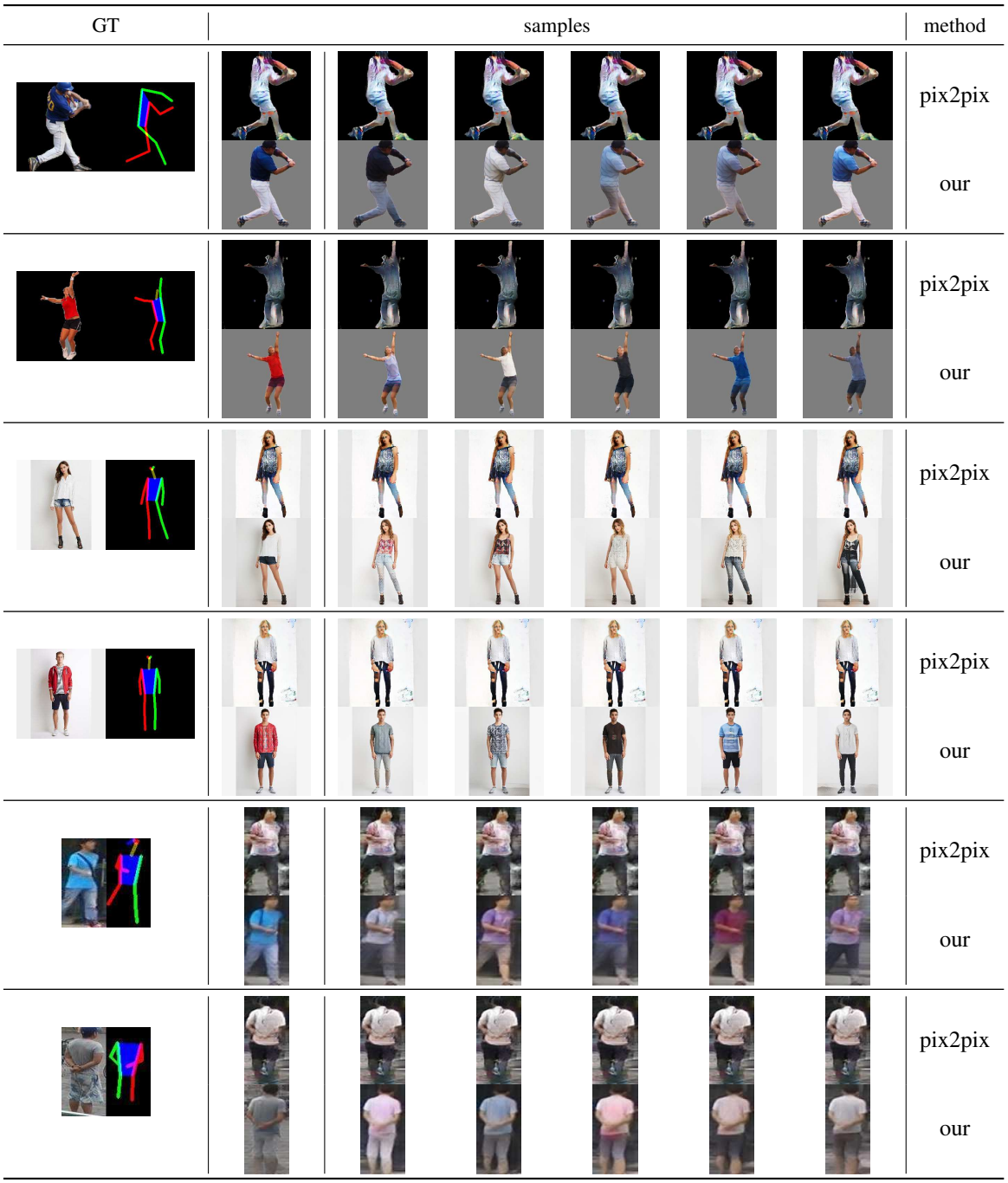

Figure 12: Generating images based only on the stickman as input (GT original image and corresponding stickman are held back). We compare our approach with pix2pix [12]. On the right: each odd row shows images synthesized by pix2pix, each even row presents samples generated by our model. Here again our first image (column 2) is a generation with original appearance, whereby for the 5 following images we sample appearance from the learned prior distribution. The GT images are taken from COCO [20], DeepFashion [21, 23] and Market-1501 [47].

C. Transfer of shape and appearance

We show additional examples of transferring appearances to different shapes and vice versa. We emphasize again that our approach does not require labeled examples of images depicting the same appearance in different shapes. This enables us to

apply it on a broad range of datasets as summarized in Table 4.

Figure	Shape Estimate	Appearance Source	Shape Target
Fig. 13	Edges	Handbags	Shoes
Fig. 14	Edges	Shoes	Handbags
Fig. 15	Body Joints	COCO	COCO
Fig. 16	Body Joints	DeepFashion	DeepFashion
Fig. 17	Body Joints	Market	Market
Fig. 18	Body Joints	COCO	Penn Action

Table 4: Overview of transfer experiments.

Figure 13: Examples of shape and appearance transfer between two datasets: appearance is taken from the shoes and is used to generate matching handbags based on their desired shape. *On the left*: original images from the shoe dataset. *On the top*: edge images of the desired handbags. *Single row*: transfer of fixed appearance to different shapes. *Single column*: transfer of fixed shape to different appearances.

Figure 14: Examples of shape and appearance transfer between two datasets: appearance is taken from the handbags and is used to generate matching shoes based on their desired shape. *On the left*: original images from the handbags dataset. *On the top*: edge images of the desired shoes. *Single row*: transfer of fixed appearance to different shapes. *Single column*: transfer of fixed shape to different appearances.

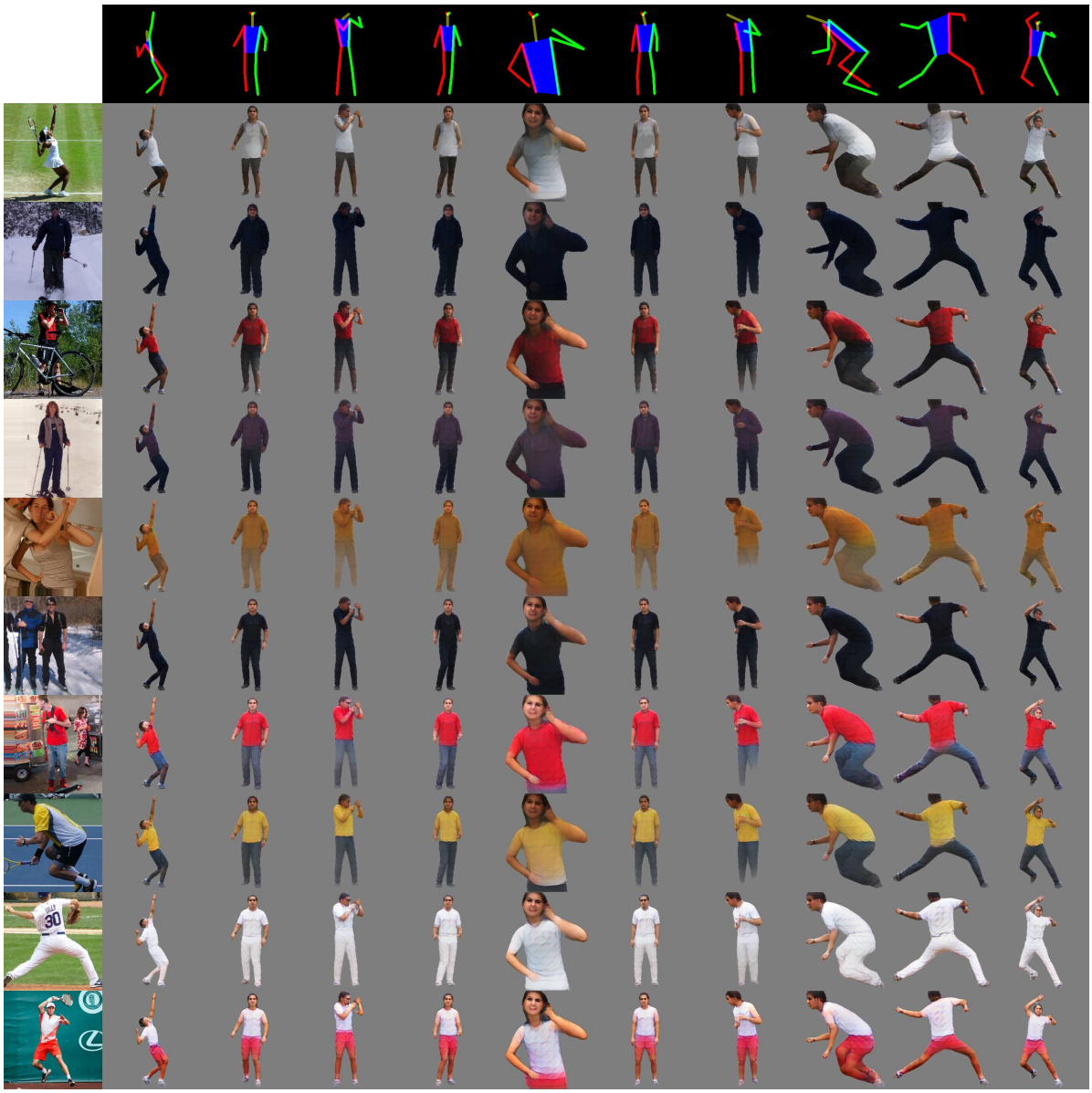

Figure 15: Examples of shape and appearance transfer on COCO dataset. *On the left*: original images from the test split. *On the top*: corresponding stickmen. *Single row*: transfer of fixed appearance to different shapes. *Single column*: transfer of fixed shape to different appearances.

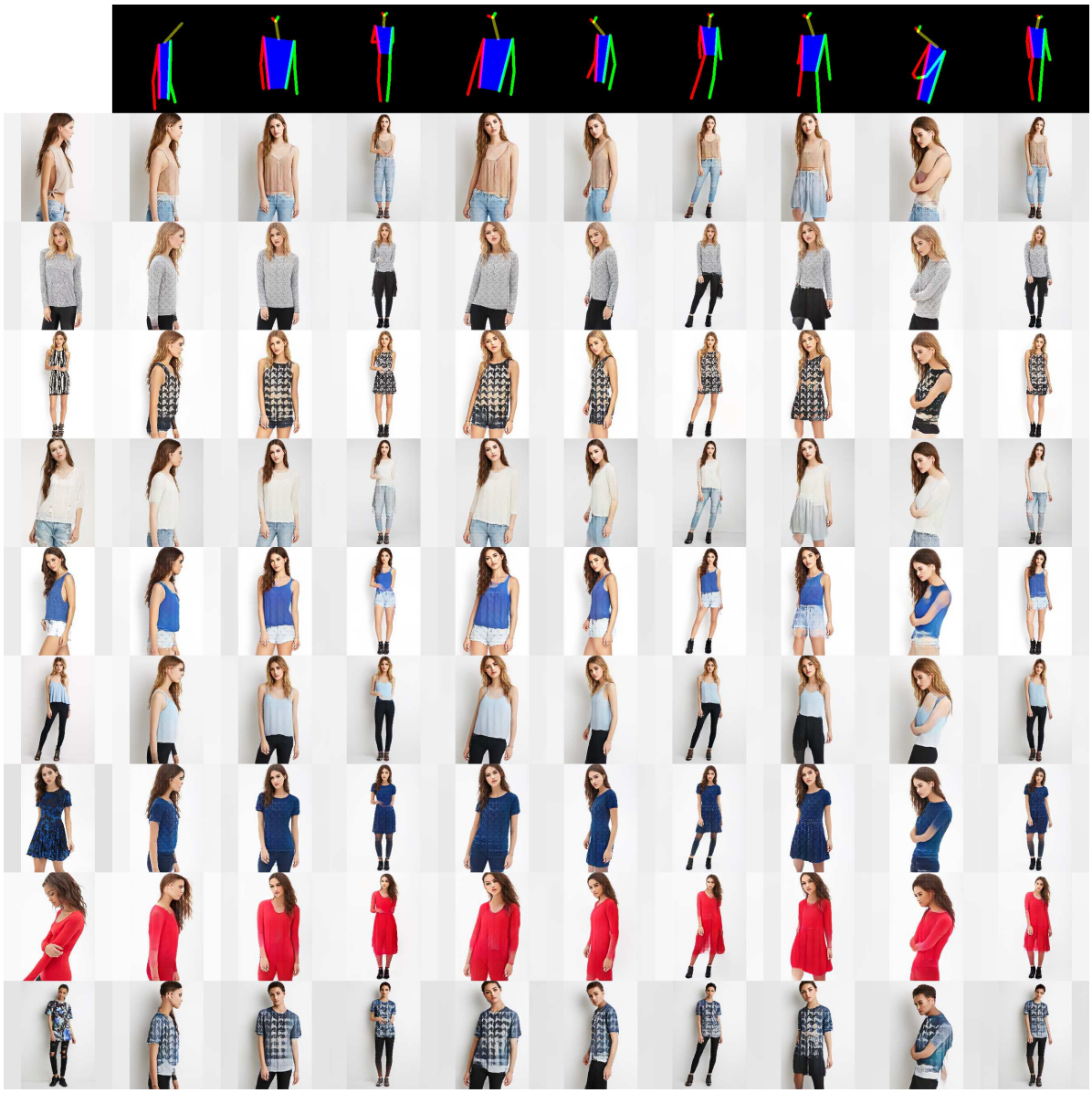

Figure 16: Examples of shape and appearance transfer on DeepFashion dataset. *On the left*: original images from the test split. *On the top*: corresponding stickmen. *Single row*: transfer of fixed appearance to different shapes. *Single column*: transfer of fixed shape to different appearances.

Figure 17: Examples of shape and appearance transfer on Market-1501. *On the left*: original images from the test split. *On the top*: corresponding stickmen. *Single row*: transfer of fixed appearance to different shapes. *Single column*: transfer of fixed shape to different appearances.

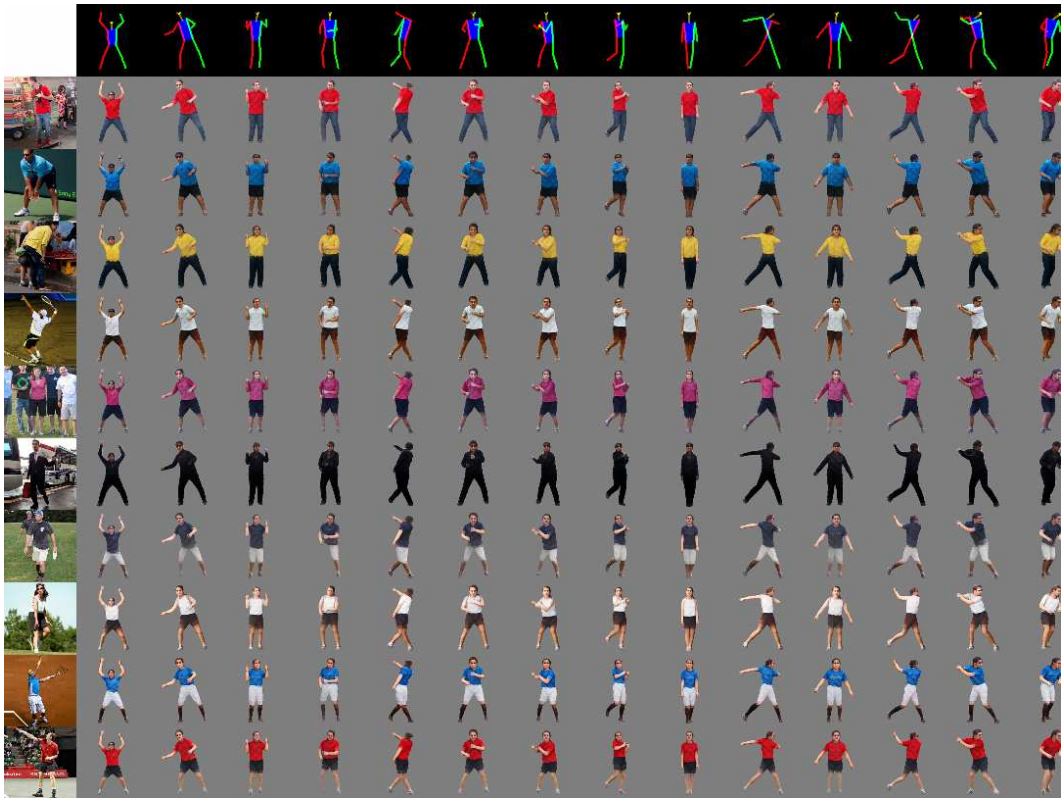

Figure 18: Examples of shape and appearance transfer in video. Appearance is inferred from COCO and target shape is estimated from Penn Action sequences. An animated version can be found at https://compvis.github.io/vunet. Note, that we generate the video independently frame by frame without any temporal smoothing etc.

D. Quantitative results for the ablation study

We have included quantitative results for the ablation study (see section 4.4) in Table 5. The positive effect of the KL-regularization cannot be quantified by the Inception Score and thus we presented the qualitative results in Fig. 9.

method	Reconstruction		Transfer	
	IS		IS	
	mean	std	mean	std
our (no appearance)	2.211	0.080	2.211	0.080
our (no kl)	3.168	0.296	3.594	0.199
our (proposed)	3.087	0.239	3.504	0.192

Table 5: Inception scores (IS) for ablation study. The positive effect of the KL-regularization as seen in Fig. 9 cannot be quantified by the IS.

E. Limitations

The quality of the generated images depends highly on the dataset used for training. Our method relies on appearance commonalities across the dataset that can be used to learn efficient, pose-invariant encodings. If the dataset provides sufficient support for appearance details, they are faithfully preserved by our model (e.g. hats in DeepFashion, see Fig. 8, third row).

The COCO dataset shows large variance in both visual qualities (e.g. lighting conditions, resolutions, clutter and occlusion) as well as in appearance. This leads to little overlap of appearance details in different poses and the model focuses on aspects of appearance that can be reused for a large variety of poses in the dataset.

We show some failure cases of our approach in Fig. 19. The first row of Fig. 19 shows an example of rare data: children are underrepresented in COCO [20]. A similar problem occurs in Market-1501 [47] where most of the images represent a tight crop around a person and only some contain people from afar. This is shown in the second row which also contains an incorrect estimate for the left leg. Sometimes, estimated pose correlates with some other attribute of a dataset (e.g., gender as in DeepFashion [21, 23], where male and female models use very characteristic yet distinct set of poses). In this case our model morphs this attribute with the target appearance, e.g. generates a woman with definitely male body proportions (see row 3 in Fig. 19). Under heavy viewpoint changes, appearance can be entirely unrelated, e.g. front view showing a white t-shirt which is totally covered from the rear view (see fourth row of Fig. 19). The algorithm however assumes that the appearance in both views is related. As the example in the last row of Fig. 19 shows, our model is confused if occluded body parts are annotated since this is not the case for most training samples.

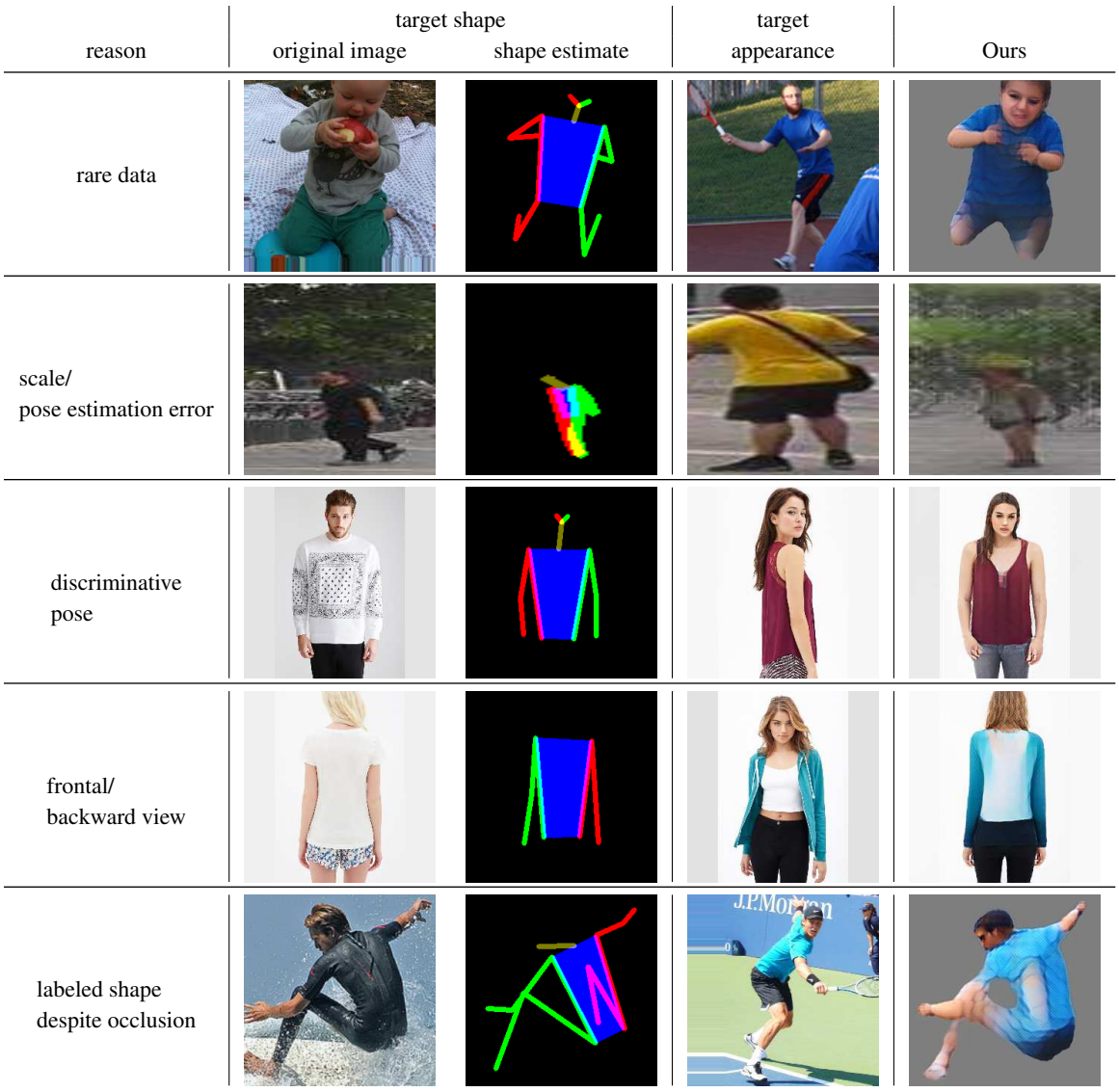

Figure 19: Examples of failure cases. As most of the errors are dataset specific we show a collection of cases over different datasets.