# A Semi-supervised Deep Generative Model for Human Body Analysis

Rodrigo de Bem[1,2(✉)], Arnab Ghosh[1], Thalaiyasingam Ajanthan[1], Ondrej Miksik[1], N. Siddharth[1], and Philip Torr[1]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
{rodrigo,arnabg,ajanthan,omiksik,nsid,phst}@robots.ox.ac.uk
[2] Center of Computational Sciences, Federal University of Rio Grande, Rio Grande, Brazil

**Abstract.** Deep generative modelling for human body analysis is an emerging problem with many interesting applications. However, the latent space learned by such models is typically not interpretable, resulting in less flexible models. In this work, we adopt a structured semi-supervised approach and present a deep generative model for human body analysis where the body pose and the visual appearance are disentangled in the latent space. Such a disentanglement allows independent manipulation of pose and appearance, and hence enables applications such as pose-transfer without being explicitly trained for such a task. In addition, our setting allows for semi-supervised pose estimation, relaxing the need for labelled data. We demonstrate the capabilities of our generative model on the Human3.6M and on the DeepFashion datasets.

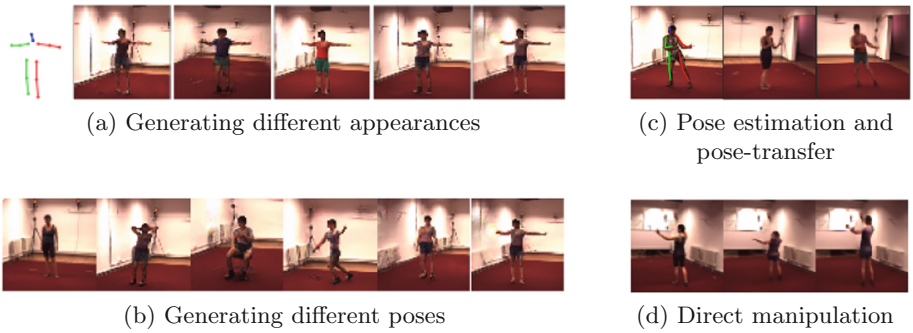
**Keywords:** Deep generative models · Variational autoencoders · Semi-supervised learning · Human pose estimation · Analysis-by-synthesis

## 1 Introduction

Human-body analysis has been a long-standing goal in computer vision, with many applications in gaming, human-computer interaction, shopping and healthcare [1,29,30,37]. Typically, most approaches to this problem have focused on supervised learning of discriminative models [4–6,41], to learn a mapping from given visual input (images or videos) to a suitable abstract form (*e.g.* human pose). While these approaches do exceptionally well on their prescribed task, as evidenced by performance on pose estimation benchmarks, they fall short due to: (a) reliance on fully-labelled data, and (b) the inability to generate novel data from the abstractions.

The former is a fairly onerous requirement, particularly when dealing with real-world visual data, as it requires many hours of human-annotator time and effort to collect. Thus, being able to relax the reliance on labelled data is a highly desirable goal. The latter addresses the ability to manipulate the abstractions


L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11130, pp. 500–517, 2019.
https://doi.org/10.1007/978-3-030-11012-3_38



(a) Generating different appearances

(c) Pose estimation and pose-transfer

(b) Generating different poses

(d) Direct manipulation

**Fig. 1.** Sampled results from our deep generative model for natural images of people. (a) For a given pose (first image), we show some samples of appearance. (b) For a given appearance (first image), samples of different poses. (c) For an estimated pose (first image) and an estimated appearance (second image), we show a generated sample combining the pose of the first image with the appearance of the second. (d) For a given pose and appearance (first image), by the direct manipulation of pose, we can modify the body size, while the appearance is kept the same.

directly, with a view to *generating* novel visual data; *e.g.* moving the pose of an arm results in generation of images or videos where that arm is correspondingly displaced. Such *generative* modelling, in contrast to discriminative modelling, enables an *analysis-by-synthesis* approach to human-body analysis, where one can generate images of humans in combinations of poses and clothing unseen during training. This has many potential applications. For instance, it can be used for performance capture and reenactment of RGB videos, as already possible for faces [34], and still incipient for human bodies. It can also be used to generate images in a user specified pose to enhance datasets with minimal annotation effort. Such an approach is typically tackled using deep generative models (DGMs) [9,18,27] – an extension of standard generative models that incorporate neural networks as flexible function approximators. Such models are particularly effective in complex perceptual domains such as computer vision [19], language [25], and robotics [40], effectively delegating bottom-up feature learning to neural networks, while simultaneously incorporating top-down probabilistic semantics into the model. They solve both the deficiencies of discriminative approach discussed above by (a) employing unsupervised learning, thereby removing the need for labels, and (b) embracing a fully generative approach.

However, DGMs introduce a new problem – the learnt abstractions, or latent variables, are not *human-interpretable*. This lack of interpretability is a by-product of the unsupervised learning of representations from data. The learnt latent variables, typically represented as some smooth high-dimensional manifold, do not have consistent semantic meaning – different sub-spaces in this manifold can encode arbitrary variations in the data. This is particularly unsuitable for our purposes as we would like to view and manipulate the latent variables, *e.g.* the body pose.

In order to ameliorate the aforementioned issue, while still eschewing reliance on fully-labelled data, we rely on the *structured semi-supervised* variational autoencoder (VAE) framework [17,32]. Here, the model structure is assumed to be *partially specified*, with consistent semantics imposed on some interpretable subset of the latent variables (*e.g.* pose), and the rest is left to be non-interpretable, although referred by us here as *appearance*. Weak (semi) supervision acts as a means to constrain the pose latent variables to actually encode the pose. This gives us the full complement of desirable features, allowing (a) semi-supervised learning, relaxing the need for labelled data, (b) generative modelling through stochastic computation graphs [28], and (c) interpretable subset of latent variables defined through model structure.

In this work, we introduce a structured semi-supervised VAEGAN [20] architecture, Semi-DGPose, in which we further extend previous structured semi-supervised models [17,32] with a discriminator-based loss function [9,20]. We show some results on human pose in Fig. 1. It is formulated in a principled, unified probabilistic framework. To our knowledge, it is the first structured semi-supervised deep generative model of people in natural images, directly learned in the image space. In contrast to previous work [21,23,24,31,38], it directly enables: (i) *semi-supervised pose estimation* and (ii) *indirect pose-transfer* across domains without explicit training for such a task, both of which are tested and verified by experimental evidence.

In summary, our main contributions are: (i) a real-world application of structured semi-supervised deep generative model of natural images, separating pose from appearance in the analysis of the human body; (ii) a quantitative and qualitative evaluation of the generative capabilities of such model; and (iii) a demonstration of its utility in performing semi-supervised pose estimation and indirect pose-transfer.

## 2 Preliminaries

Deep generative models (DGMs) come in two broad flavours – Variational Autoencoders (VAEs) [18,27], and Generative Adversarial Networks (GANs) [9]. In both cases, the goal is to learn a generative model $p_\theta(\mathbf{x}, \mathbf{z})$ over data $\mathbf{x}$ and latent variables $\mathbf{z}$, with parameters $\theta$. Typically the model parameters $\theta$ are represented in the form of a neural network.

VAEs learn the parameters $\theta$ that maximise the marginal likelihood (or evidence) of the model denoted as $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})dz$. They introduce a conditional probability density $q_\phi(\mathbf{z}|\mathbf{x})$ as an approximation to the unknown and intractable model posterior $p_\theta(\mathbf{z}|\mathbf{x})$, employing the variational principle in order to optimise a surrogate objective $\mathcal{L}_{\text{VAE}}(\phi, \theta; \mathbf{x})$, called the evidence lower bound (ELBO), as

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}_{\text{VAE}}(\phi, \theta; \mathbf{x}) \ = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]. \tag{1}$$

The conditional density $q_\phi(\mathbf{z}|\mathbf{x})$ is called the recognition or inference distribution, with parameters $\phi$ also represented in the form of a neural network.

To enable structured semi-supervised learning, one can factor the latent variables into unstructured or non-interpretable variables $\mathbf{z}$ and structured or interpretable variables $\mathbf{y}$ without loss of generality [17,32]. For learning in this framework, the objective can be expressed as the combination of supervised and unsupervised objectives. Let $\mathcal{D}_u$ and $\mathcal{D}_s$ denote the unlabelled and labelled subset of the dataset $\mathcal{D}$, and let the joint recognition network factorise as $q_\phi(\mathbf{y}, \mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{y}|\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$. Then, the combined objective summed over the entire dataset corresponds to

$$\mathcal{L}_{\text{SS}}(\theta, \phi; \mathcal{D}) = \sum_{\mathbf{x}_u \in \mathcal{D}_u} \mathcal{L}_u(\theta, \phi; \mathbf{x}_u) + \gamma \sum_{(\mathbf{x}_s, \mathbf{y}_s) \in \mathcal{D}_s} \mathcal{L}_s(\theta, \phi; \mathbf{x}_s, \mathbf{y}_s), \tag{2}$$

where $\mathcal{L}_u$ and $\mathcal{L}_s$ are defined as

$$\mathcal{L}_u(\theta, \phi; \mathbf{x}_u) = \mathcal{L}_{\text{VAE}}(\theta, \phi; \mathbf{x}_u), \tag{3}$$

$$\mathcal{L}_s(\theta, \phi; \mathbf{x}_s, \mathbf{y}_s) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_s, \mathbf{y}_s)} \left[ \log \frac{p_\theta(\mathbf{x}_s, \mathbf{z}|\mathbf{y}_s)}{q_\phi(\mathbf{z}|\mathbf{x}_s, \mathbf{y}_s)} \right] + \alpha \log q_\phi(\mathbf{y}_s|\mathbf{x}_s). \tag{4}$$

Here, the hyper-parameter $\gamma$ (Eq. 2) controls the relative weight between the supervised and unsupervised dataset sizes, and $\alpha$ (Eq. 4) controls the relative weight between generative and discriminative learning.

Note that by the factorisation of the generative model, VAEs require the specification of an explicit likelihood function $p_\theta(\mathbf{x}|\mathbf{z})$, which can often be difficult. GANs [9] on the other hand, attempt to sidestep this requirement by learning a surrogate to the likelihood function, while avoiding the learning of a recognition distribution. Here, the generative model $p_\theta(\mathbf{x}, \mathbf{z})$, viewed as a mapping $G : \mathbf{z} \mapsto \mathbf{x}$, is setup in a two-player minimax game with a "discriminator" $D : \mathbf{x} \mapsto \{0, 1\}$, whose goal is to correctly identify if a data point $\mathbf{x}$ came from the generative model $p_\theta(\mathbf{x}, \mathbf{z})$ or the true data distribution $p(\mathbf{x})$. Such objective is defined as

$$\mathcal{L}_{\text{GAN}}(D, G) = \mathbb{E}_{p(\mathbf{x})} \left[ \log D(\mathbf{x}) \right] + \mathbb{E}_{p_\theta(\mathbf{z})} \left[ 1 - \log D(G(\mathbf{z})) \right]. \tag{5}$$

In fact, in our structured model, generation is defined as a function of pose and appearance as $G(\mathbf{y}, \mathbf{z})$. Crucially, learning a customised approximation to the likelihood can result in a much higher quality of generated data, particularly for the visual domain [15].

A more recent family of DGMs, VAEGANs [20], bring together these two different approaches into a single objective that combines both the VAE and GAN objectives directly as

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{GAN}}. \tag{6}$$

This marries better the likelihood learning with the inference-distribution learning, providing a more flexible family of models.

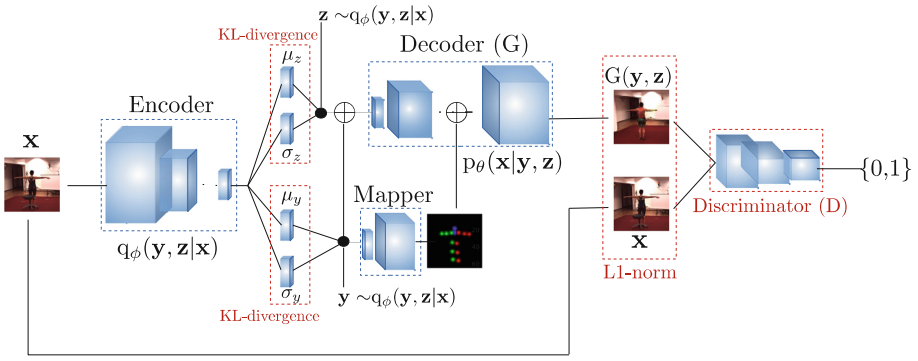



**Fig. 2. Semi-DGPose architecture.** The Encoder receives $\mathbf{x}$ as input. The KL-divergence losses between the Gaussian distribution $q_\phi(\mathbf{y}, \mathbf{z}|\mathbf{x})$ and the weak Gaussian priors $p(\mathbf{y})$ and $p(\mathbf{z})$ works as a regulariser for unsupervised training samples (see Eq. 3). The sampling of appearance and pose is done using the reparametrization trick [18] and propagated to the Decoder. For the supervised training (not shown above for simplicity, see Eq. 4), a regression loss between the estimated pose and the pose ground-truth label substitutes the KL-divergence over the pose distribution. In both, supervised and unsupervised training, the low-dimensional pose vector $\mathbf{y}$ is mapped to a heatmap representation by the Mapper module. The L1-norm and the Discriminator losses are computed over the reconstructed $\mathrm{G}(\mathbf{y}, \mathbf{z})$ and the original $\mathbf{x}$ images. G denotes Generator (see Eq. 5).

## 3 Semi-DGPose Network

Our structured semi-supervised VAEGAN model consists of two tasks: (i) learning of a recognition network (Encoder) estimating pose $\mathbf{y}$ and appearance $\mathbf{z}$ from a given RGB image $\mathbf{x}$ and (ii) learning of a generative network (Decoder) combining pose and appearance to generate corresponding RGB images. Overview of our model is shown in Fig. 2. In our model, Eq. 2 is used the aforementioned tasks, while Eq. 5 learns to discriminate between real and generated images. In contrast to the standard VAEGAN objective (Eq. 6), the structured semi-supervised VAEGAN objective is given by

$$\mathcal{L} = \mathcal{L}_{\mathrm{SS}} + \mathcal{L}_{\mathrm{GAN}}. \tag{7}$$

**Pose Representation and the Mapper Module.** Pose can be represented either using the 2D $(x, y)$ positions of the joints themselves in vector form, or using Gaussian heatmaps of the joints, which is a preferred variant successfully used in many discriminative pose estimation approaches [2,6,26,35,41]. The heatmaps $\mathbf{y} \in \mathcal{R}^{P \times H \times W}$ consists of $P$ channels, each one corresponding to a distinct body part, where $H = 64$ and $W = 64$ are the heatmaps' height and width, respectively. As the set of joints are sparse discrete points in the image, we use heatmaps for $J$ joints, $R$ rigid parts and $B = 1$ whole body, such that $P = J + R + B$ (see Appendix A). It covers the entire area of the body

in the image, as in [2]. Our preliminary experiments showed that heatmaps led to better quality results, in contrast to the vector-based representation. On the other hand, a low-dimensional representation is more suitable and desirable as a latent variable, since human pose lies in a low-dimensional manifold embedded in the high-dimensional image space [7,8].

To cope with this mismatch, we introduce the Mapper module, which maps 2D pose-vectors to heatmaps. Ground-truth heatmaps are constructed from manually annotated ground-truth 2D joints labels, by means of a simple weak annotation strategy described in [2]. The Mapper module is then trained to map 2D joints to heatmaps, minimizing the L2-norm between predicted and ground-truth heatmaps. This module is trained separately with the same training hyper-parameters used for our full architecture, described later in Sect. 4. In the training of the full Semi-DGPose architecture, the Mapper module is integrated to it with its weights fixed, since the mapping function has been learned already. As it is illustrated in Fig. 2, the Mapper allows us to keep a low-dimensional representation in the latent space, at the same time that a dense high-dimensional "spatial" heatmap representation facilitates the generation of accurate images by the Decoder. As it is fully differentiable, the module allows the gradients to be backpropagated normally from the Decoder to the Encoder, when it is required during the training of the full architecture.

We have extensively tested several architectures of our model. All of its modules are deep CNNs and their details are in Tables 2 and 1 (Sect. A, Appendix).

**Training.** The terms of Eq. 2 correspond to two training routines which are alternately employed, according to the presence of ground-truth labels. In the *unsupervised case*, when no label is available, it is similar to the standard VAE (see Eq. 3). Specifically, given the image $\mathbf{x}$, the Encoder estimates the posterior distribution $q_\phi(\mathbf{y}, \mathbf{z}|\mathbf{x})$, where both appearance $\mathbf{z}$ and pose $\mathbf{y}$ are assumed to be independent given the image $\mathbf{x}$. Then, pose and appearance are sampled from the posterior, while the KL-divergences between the posterior and the prior distributions, $\mathrm{KL}[q_\phi(\mathbf{y}|\mathbf{x})|p(\mathbf{y})]$ and $\mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})|p(\mathbf{z})]$, are used as regularisers. The samples $\mathbf{y}$ and $\mathbf{z}$ are passed through the Decoder to generate a reconstructed image. Finally, the unsupervised loss function minimized during training is composed of the L1-norm reconstruction loss, the KL-divergences and the cross-entropy Discriminator loss. In the *supervised case*, when the pose label is available, the KL-divergence between the posterior pose distribution and the pose prior, $\mathrm{KL}[q_\phi(\mathbf{y}|\mathbf{x})|p(\mathbf{y})]$, is replaced with a regression loss between the estimated pose and the given label (see Eq. 5). Now, only the appearance $\mathbf{z}$ is sampled from the posterior distribution and passed to the Decoder, along with the ground-truth pose label. Finally, the supervised loss function minimized during training is composed of the L1-norm reconstruction loss, the KL-divergence over the appearance distribution, the regression loss over the pose vector and the cross-entropy Discriminator loss. In this case, gradients are not backpropagated from the Decoder to the Encoder, through the pose posterior distribution, since pose was not estimated. In both *unsupervised* and *supervised* cases, the Mapper

module, which is trained *offline*, is used to map the 2D pose-vector in the latent space to a dense heatmap representation, as illustrated in Fig. 2.

**Reconstruction.** At test time, only an image $\mathbf{x}$ is given as input, and the reconstructed image $G(\mathbf{y}, \mathbf{z})$ is obtained from the Decoder. In the reconstruction process, *direct manipulation* of the pose representation $\mathbf{y}$ allows image generations with varying body pose and size while the appearance is kept the same (see Fig. 8, Sect. 4.1) (Fig. 3).

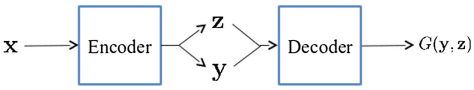



**Fig. 3.** Reconstruction at test time.

**Indirect Pose-Transfer.** Our method allows us to do *indirect* pose-transfer without explicit training for such task. In this case, (i) an image $\mathbf{x}_1$ is first passed through the Encoder network, from which the target pose $\mathbf{y}_1$ is kept. (ii) In the second step, another image $\mathbf{x}_2$ is propagated through the Encoder, from which the appearance encoding $\mathbf{z}_2$ is kept. (iii) Finally, $\mathbf{z}_2$ and $\mathbf{y}_1$ are jointly propagated through the Decoder, and an image $\mathbf{x}_3$ is reconstructed, containing a person in the pose $\mathbf{y}_1$ estimated from the first image, but with the appearance $\mathbf{z}_2$ defined by the second image. This is a novel application that our approach enables; in contrast to prior art, our network neither rely on any external pose estimator nor on conditioning labels to perform pose-transfer (see Fig. 13, Sect. 4.1) (Fig. 4).

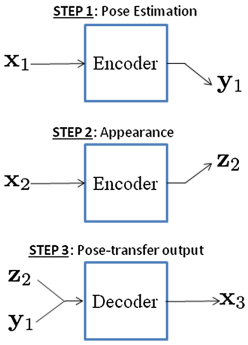



**Fig. 4.** Indirect pose-transfer at test time.

**Sampling.** When no image is given as input, we can jointly or separately sample pose $\mathbf{y}$ and appearance $\mathbf{z}$ from the posterior distribution. They may be sampled at the same time or one may be kept fixed while the other distribution is sampled. In all cases, the encodings are passed through the Decoder network to generate a corresponding RGB image (Fig. 5).

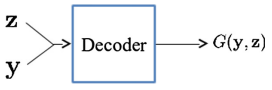



**Fig. 5.** Sampling at test time.

**Pose Estimation.** One of the main differences between our approach and prior art is the ability of our model to estimate human-body pose as well. In our model, given an input image $\mathbf{x}$, it is possible to perform pose estimation by regressing to the pose representation vector $\mathbf{y}$. In this case, the appearance encoding $\mathbf{z}$ is disregarded and the Decoder, Mapper and Discriminator networks are not used (Fig. 6).

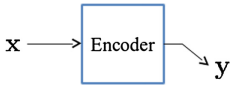



**Fig. 6.** Pose estimation at test time.

## 4 Experiments and Discussion

In this section, we present the datasets, metrics and training hyper-parameters used in our work. Finally, quantitative and qualitative results show the effectiveness and novelty of our Semi-DGPose architecture.

**Human3.6M Dataset.** Human3.6M [11] is a widely used benchmark for human body analysis. It contains 3.6 million images acquired by recording 5 female and 6 male actors performing a diverse set of motions and poses corresponding to 15 activities, under 4 different viewpoints. We followed the standard protocol and used sequences of 2 out of 11 actors as our test set, while the rest of the data was used for training. We use a subset of 14 (out of 32) body joints represented by their $(x, y)$ 2D image coordinates as our ground-truth data, neglecting minor body parts (*e.g.* fingers). Due to the high frequency of the video acquisition (50 Hz), there is a considerable level of practically redundant images. Thus, out of images from all 4 cameras, we subsample frames in time, producing subsets for training and test, with $317,989$ and $1,280$ images, respectively. All the images have resolution of $1000 \times 1000$ pixels.

**DeepFashion Dataset.** The DeepFashion dataset (In-shop Clothes Retrieval Benchmark) [22] consists of 52,712 images of people in a variety of clothing and poses. We follow [23], using their joints' annotations obtained with an off-the-shelf pose estimator [5], and divide the dataset into training (44,950 images) and test (6,560 images) subsets. Images with wrong pose estimations were suppressed, with all original images having $256 \times 256$ pixels. Note, we aim to learn a complete generative model of people in natural images, which is significantly more complex, compared to models focusing on a particular task, such as pose-transfer. For this reason, we do not restrict our training set to pairs of images of the same person and use individual images, in contrast to [23,31].

**Metrics.** Since our model explicitly represents *appearance* and *body pose* as separate variables, we evaluate its performance with respect to two different aspects: (i) *image quality* of reconstructions, evaluated using the standard Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [39] metrics and (ii) *accuracy of pose estimation*, obtained by the Semi-DGPose model, measured using the Percentage of Correct Keypoints (PCK) metric [43], which computes the percentage of 2D joints correctly located by a pose estimator, given the *ground-truth* and a normalized distance threshold corresponding to the size of the person's torso.

**Training Parameters.** All models were trained with mini-batches consisting of 64 images. We used the Adam optimizer [16] with initial learning rate set to $10^{-4}$. The weight decay regulariser was set to $5 \times 10^{-4}$. Network weights were initialized randomly for fully-connected layers and with robust initialization [10]

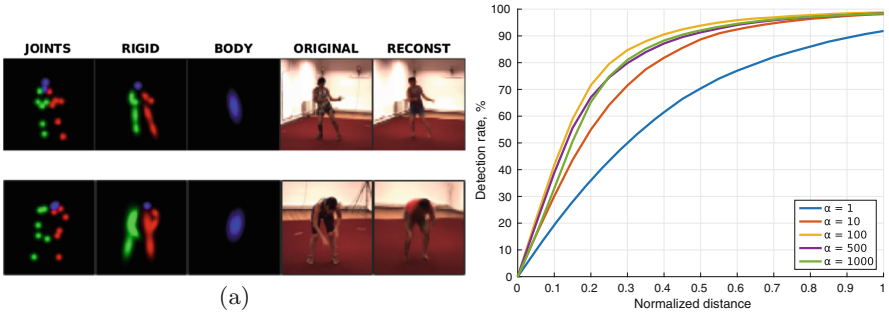
for convolutional and transposed-convolutional layers. Except when stated differently, for all images and all models, we used a $64 \times 64$ pixel crop, centring the person of interest. We did not use any form of data augmentation or preprocessing except for image normalisation to zero mean and unit variance. All models were implemented in Caffe [14] and all experiments ran on an NVIDIA Titan X GPU.

### 4.1 Semi-DGPose Results

Here we evaluate our Semi-DGPose model on the Human3.6M [11] and on the DeepFashion [22] datasets. The Human3.6M is well-suited for pose estimation evaluation, since it has joints' annotations obtained in studio by mean of an accurate motion capture system. We show quantitative and qualitative results, focusing particularly on pose estimation and on *indirect* pose transfer capabilities, described later in this section. We show qualitative experiments on the DeepFashion, comparing reconstructions with original images. Our experiments and results show the effectiveness of the Semi-DGPose method.

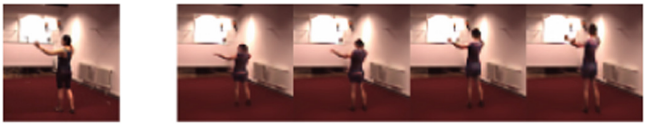
**Results on Human3.6M.** To evaluate the efficacy of our model, we perform a "relative" comparison. In other words, we first train our model with full supervision (*i.e.* all data points are labelled) to evaluate performance in an ideal case and then we train the model with other setups, using labels only for 75%, 50% and 25% data points. Such an evaluation allows us to decouple the efficacy of the model itself and the semi-supervision to see how the gradual decrease in the level of supervision affects the final performance of the method on the same validation set. We first cross-validated the hyper-parameter $\alpha$ which weights the regression loss (see Eq. 4, in Sect. 2) and found that $\alpha = 100$ yields the best results, as shown in Fig. 7b. Following [32], we keep $\gamma = 1$ in all experiments (see Eq. 2, in Sect. 2). In Fig. 7a, we show reconstructed images along with the heatmap pose representation. When pose representation is *directly manipulated* during the reconstruction process, appearance can be kept the same while the body pose can modified, as shown in Fig. 8.

We evaluated it across different levels of supervision, with the PSNR and SSIM metrics and show results in Fig. 9a. We also evaluated the pose estimation accuracy of the Semi-DGPose model. It achieves 93.85% PCK score, normalized at 0.5, in the fully-supervised setup (100% of supervision over the training data). This pose estimation accuracy is on par with the state-of-the-art pose estimators on unconstrained images [42]. However, since the Human3.6M was captured in a controlled environment, a standard (discriminative) pose estimator can be expected to perform better. The overall PCK curves corresponding to each percentage of supervision in the training set is shown in Fig. 9b. Note that, even with 25% supervision, our model still obtains 88.35% PCK score, normalized at 0.5, showing the effectiveness of the semi-supervised approach. Finally, we show the pose estimation accuracy for different samples in Fig. 10. In Fig. 11, we show reconstructed images obtained with different levels of supervision. It




(a) (b)

**Fig. 7.** (a) Qualitative reconstructions with full supervision. (b) PCK scores for the cross-validation adjustment of the regression loss weight $\alpha$.

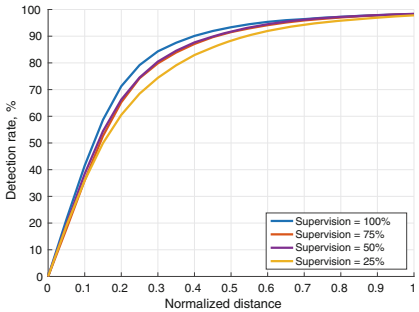


(a) (b) (c) (d) (e)

**Fig. 8. Direct manipulation.** Original image (**a**), followed by reconstructions in which the person's height was changed to a percentage of the original, as: (**b**) 80%, (**c**) 95%, (**d**) 105% and (**e**) 120%. The same procedure may be applied to produce different changes in the body size and aspect ratio.

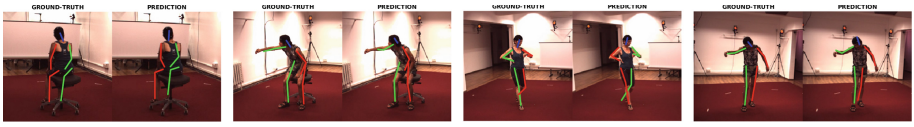
| Level of supervision | PSNR | SSIM |
|---|---|---|
| 100% | 22.27 | 0.89 |
| 75% | 21.36 | 0.86 |
| 50% | 21.49 | 0.87 |
| 25% | 20.06 | 0.83 |

(a)



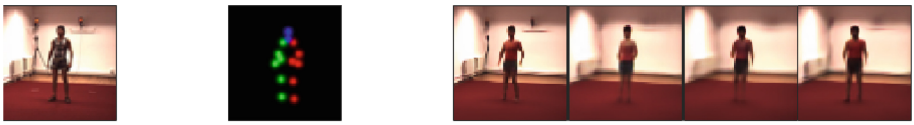

(b)

**Fig. 9. Quantitative evaluations of Semi-DGPose on Human3.6M:** (**a**) PSNR and SSIM measures for different levels of supervision, (**b**) PCK scores for different levels of supervision. Note that, even with 25% supervision, our Semi-DGPose obtains 88.35% PCK score, normalized at 0.5.



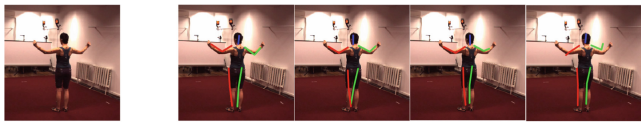

(a) PCK=92.9% (b) PCK=100.0% (c) PCK=96.4% (d) PCK=100.0%

**Fig. 10.** PCK scores for 100% of supervision, normalized at 0.5, for ground-truth (left) and prediction (right) pairs, superimposed on the original images. Each pair correspond to one of the 4 cameras from the Human3.6M dataset.



(a) (b) (c) (d) (e) (f)

**Fig. 11. Semi-DGPose reconstructions:** (**a**) original image, and (**b**) heatmap pose representation, followed by reconstructions with different levels of supervision: (**c**) 100%, (**d**) 75%, (**e**) 50%, (**f**) 25%.



(a) (b) (c) (d) (e)

**Fig. 12. Pose estimation.** Original image (**a**), followed by estimations, over the original image, with: (**b**) 100%, (**c**) 75%, (**d**) 50% and (**e**) 25% of supervision.

allows us to observe how image quality is affected when we gradually reduce the availability of labels. Following that, we evaluate results on pose estimation and on *indirect* pose transfer. Regarding **semi-supervised pose estimation**, we complement the previous quantitative evaluation with the results shown in Fig. 12. We highlight this distinctive capability of our Semi-DGPose generative model. Again, we aimed to analyse how the gradual decrease of supervision in the training set affects the quality of pose estimation on the test images. Concerning ***indirect pose-transfer***, as both latent variables corresponding to pose and appearance can be inferred by the model's Encoder (recognition network) at test time, latent variables extracted from different images can be combined in a subsequent step, and employed together as inputs for the Decoder (generative network). The result of that is a generated image combining appearance and body pose, extracted from two different images. The process is done in three phases, as illustrated in Fig. 13: (i) the latent pose representation $\mathbf{y}_1$ is estimated from the first input image through the Encoder; (ii) the latent *appearance* representation $\mathbf{z}_2$ is estimated from a second image, also through the Encoder, (iii) $\mathbf{y}_1$ and $\mathbf{z}_2$ are propagated through the Decoder, and a new image is generated,

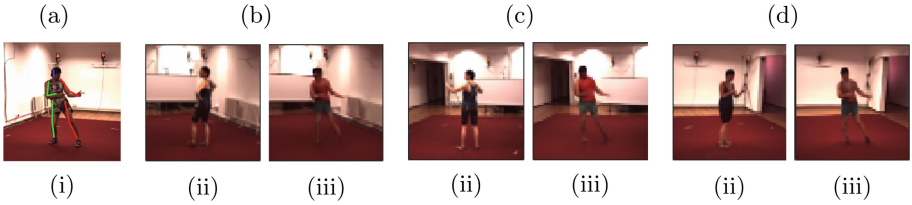



**Fig. 13.** ***Indirect*** **pose transfer:** (**i**) the latent target pose representation $\mathbf{y}_1$ is estimated (Encoder). The pairs (**b**), (**c**) and (**d**), show (**ii**) the image from which the latent *appearance* $\mathbf{z}_2$ is estimated (Encoder); (**iii**) the output image generated as a combination of $\mathbf{y}_1$ and $\mathbf{z}_2$ (Decoder). The person's outfit in the output images (**iii**) is approximated to the ones in images (**ii**), however restricted by the low diversity of outfits observed in Human3.6M training data. Backgrounds of images (**ii**) are reproduced in the output images (**iii**) and all them differ from the one in image (**i**).

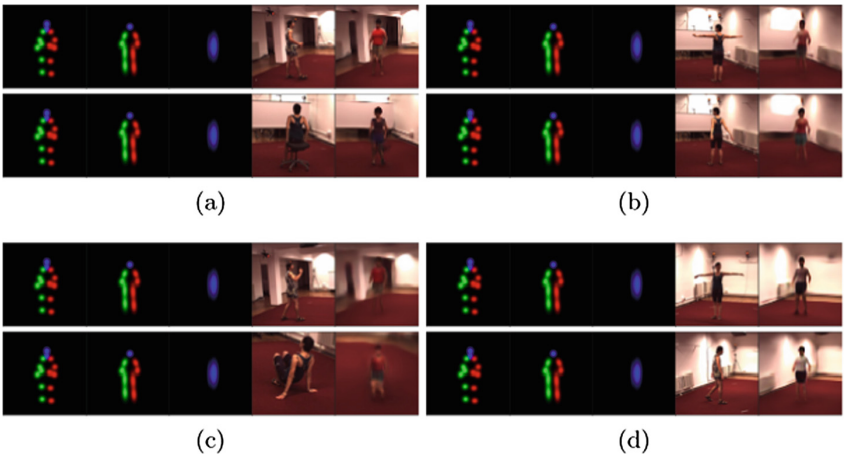



**Fig. 14.** Indirect pose-transfers with different levels of supervision: (a) 100%; (b) 75%; (c) 50%; (d) 25%.

combining body pose and appearance, respectively, from the first and second *encoded* images. We evaluate qualitatively the effects of semi-supervision over the indirect pose-transfer in Fig. 14.

**Results on DeepFashion.** To show the generality of the Semi-DGPose model we show in Fig. 15 reconstructed images on the DeepFashion dataset. The same hyper-parameters described before were used in training. Related methods in the literature [23,31] focus only on pose-transfer, training on pairs of images from the same person, which is a simpler task in comparison to ours. Such difference prevents a direct fair comparison with these methods.

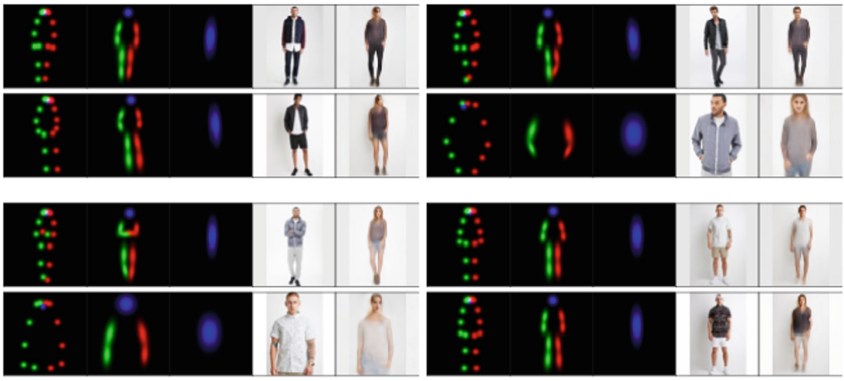


**Fig. 15. Semi-DGPose DeepFashion reconstructions** with 100% of supervision during training. Heatmaps are only shown as references, since the only input of the Semi-DGPose is the original image. At test time, as pose is estimated in the latent space, discrepancies between the original and reconstructed poses may be observed. Reconstructed images have $64 \times 64$ pixels. Best viewed if zoomed in digital version.

## 5 Related Work

Generative modelling for human body analysis has a long history in computer vision [13,33]. However, in the past years, deep generative models have been far less investigated compared to their discriminative counterparts [4–6,41]. Recently, Lassner *et al.* [21] presented a deep generative model based on a CVAE conditioned on human pose which allowed generating images of segmented people and their clothing. However, this model does not encode pose using raw image data but only low dimensional (binary) segmentation masks and an "image-to-image" transfer network [12] is used to generate realistic images. In contrast, we learn the generative model directly on the raw image data without the need of body parts segmentation. A closely related model is introduced in [3], but it is again a conditional model which does not allow for pose estimation neither semi-supervision. Difficulty of generating poses and detailed appearance simultaneously in an end-to-end fashion is admitted by Ma *et al.* [23]. In order to tackle this issue, they proposed a two stage image-to-image translation model. However, their model does not allow sampling, thus in its essence it is not a generative model, which is again in contrast to our approach.

In a concurrent work to ours, Siarohin *et al.* [31] improves approach of [23] by making it single-stage and trainable end-to-end. While this approach is relatively similar to ours, the key difference is that the human body joints (keypoints) are given to the algorithm (detected by another off-the-shelf discriminative method) while our method learns to encode them directly from the raw image data. Hence, our model allows sampling of different poses independent of appearance. Finally, Ma *et al.* [24] proposed a model for learning image embeddings of foreground, background and pose variables encoded as interpretable variables. However, this model has to rely on an off-the-shelf pose estimator to perform pose-transfer

but our model can perform pose estimation even in a semi-supervised setting in addition to image generation. The existing approaches do not have the flexibility to manipulate pose independently of appearance and they have to be explicitly trained with pairs of images to allow pose transfer. This is in sharp contrast to our approach, where we learn pose *estimation* and pose transfer is a by-product.

Apart from this, Walker *et al.* [38] proposed a hybrid VAEGAN architecture for forecasting future poses in a video. Here, a low-dimensional pose representation is learned using a VAE and once the future poses are predicted, they are mapped to images using a GAN generator. Following [20], we use a discriminator in our training to improve the quality of the generated images, however, in contrast to [20], the latent space of our approach is interpretable which enables us to sample different poses and appearance. Considering GAN based generative models, Tulyakov *et al.* [36] presents a GAN network that learns motion and content in two separate latent spaces in an unsupervised manner. However it does not allow an explicit manipulation over the human pose.

## 6 Conclusions

In this paper we have presented a deep generative model for human pose analysis in natural images. To this end, we have proposed a structured semi-supervised VAEGAN approach. Our model allows independent manipulation of pose and appearance and hence enables applications such as pose-transfer without being explicitly trained for such a task. In addition to that, the semi-supervised setting relaxes the need for labelled data. We have systematically evaluated our model on the Human3.6M and DeepFashion datasets, showing applications such as indirect pose-transfer and semi-supervised pose estimation.

**Acknowledgements.** Rodrigo Andrade de Bem is a CAPES Foundation scholarship holder (Process no: 99999.013296/2013-02, Ministry of Education, Brazil). Ondrej Miksik is currently with Emotech Labs. This work was supported by the EPSRC, ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1 and EPSRC/MURI grant EP/N019474/1.

## A Semi-DGPose Architecture

The heatmaps correspond to: **(i)** 14 joints (head top, neck, right{shoulder, elbow, wrist, hip, knee, ankle}, left{shoulder, elbow, wrist, hip, knee, ankle}); **(ii)** 9 rigid parts (head, right{upper arm, lower arm, upper leg, lower leg}, left{upper arm, lower arm, upper leg, lower leg}); **(iii)** 1 whole body position. In the DeepFashion dataset, extra facial keypoints are used [23].

**Table 1.** Semi-DGPose architecture for $64 \times 64$ input images. Abbreviations: N for number of kernels/neurons, K for kernel size, S for stride and P for zero padding. CONCAT means concatenation layer, CONV means convolutional layer, BN means batch normalization layer with running average coefficient $\beta = 0.9$ and learnable affine transformation, DECONV means transpose convolutional layer, FC means fully connected layer, SUM corresponds to element-wise sum layer and RESIDUAL denotes a residual block (Table 2). The additional layers can be clearly understood.

| Encoder | |
|---|---|
| **Input:** *images*(batch_size=64, channels=3, height=64, width=64) | |
| **Layer** | **Definition** |
| 1 | CONV-(N64, K7, S2, P1), LeakyReLU(0.01) |
| 2 | CONV-(N128, K3, S2, P1), BN, ReLU |
| 3 | CONV-(N256, K3, S2, P1), BN, ReLU |
| 4-6 | CONV-(N512, K3, S2, P1), BN, ReLU |
| 7-9 | RESIDUAL-(N512, K3, S1, P1) |
| 10 | RESIDUAL-(N512, K3, S1, P1), SIGMOID |
| $\mu_z$ | FC-(N100) |
| $\sigma_z$ | FC-(N100) |
| $\mu_y$ | FC-(N48) |
| $\sigma_y$ | FC-(N48) |
| **Mapper** | |
| **Input:** *pose_vector*(batch_size=64, channels=48) | |
| **Layer** | **Definition** |
| 1 | RESHAPE(batch_size=64, channels=48, height=1, width=1) |
| 2 | DECONV-(N512, K4, S1, P0), BN, LeakyReLU(0.2) |
| 3 | DECONV-(N256, K4, S2, P1), BN, LeakyReLU(0.2) |
| 4 | DECONV-(N128, K4, S2, P1), BN, LeakyReLU(0.2) |
| 5 | DECONV-(N64, K4, S2, P1), BN, LeakyReLU(0.2) |
| *pose_heatmaps* | DECONV-(N24, K4, S2, P1), SIGMOID |
| **Decoder** | |
| **Input:** *sample*(batch_size=64, channels=100); *pose_heatmaps*(batch_size=64, channels=24, height=64, width=64); | |
| **Layer** | **Definition** |
| 1 | RESHAPE(batch_size=64, channels=100, height=1, width=1) |
| 2 | DECONV-(N512, K4, S1, P0), BN, LeakyReLU(0.2) |
| 3 | DECONV-(N256, K4, S2, P1), BN, LeakyReLU(0.2) |
| 4 | DECONV-(N128, K4, S2, P1), BN, LeakyReLU(0.2) |
| 5 | DECONV-(N64, K4, S2, P1), BN, LeakyReLU(0.2) |
| 6 | DECONV-(N128, K4, S2, P1), BN, LeakyReLU(0.2) |
| 7 | CONCAT(*deconv6_output*, *pose_heatmaps*) |
| 8 | CONV-(N512, K5, S1, P2), BN, LeakyReLU(0.2) |
| 9 | CONV-(N256, K5, S1, P2), BN, LeakyReLU(0.2) |
| 10-11 | CONV-(N128, K5, S1, P2), BN, LeakyReLU(0.2) |
| $G(\mathbf{y}, \mathbf{z})$ | CONV-(N3, K5, S1, P2), TANH |

(*continued*)

**Table 1.** (*continued*)

| **Discriminator** | |
|---|---|
| **Input:** *decoder_output*(batch_size=64, channels=3, height=64, width=64); *images*(batch_size=64, channels=3, height=64, width=64) | |
| **Layer** | **Definition** |
| 1 | CONV-(N64, K4, S2, P1), LeakyReLU(0.2) |
| 2 | CONV-(N128, K4, S2, P1), BN, LeakyReLU(0.2) |
| 3 | CONV-(N256, K4, S2, P1), BN, LeakyReLU(0.2) |
| 4 | CONV-(N512, K4, S2, P1), BN, LeakyReLU(0.2) |
| 5 | CONV-(N1, K4, S1, P0), SIGMOID |

**Table 2.** Architecture of the residual block employed in the Semi-DGPose encoder.

| RESIDUAL Layer | |
|---|---|
| **Input:** *previous_layer_output* | |
| **Layer** | **Definition** |
| 1 | CONV-(N512, K3, S1, P1), BN, ReLU |
| 2 | CONV-(N512, K3, S2, P1), BN |
| 3 | SUM(*conv2_output*, *previous_layer_output*) |

## References


1. Achilles, F., Ichim, A.-E., Coskun, H., Tombari, F., Noachtar, S., Navab, N.: Patient MoCap: human pose estimation under blanket occlusion for hospital monitoring applications. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9900, pp. 491–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46720-7_57
2. de Bem, R., Arnab, A., Sapienza, M., Golodetz, S., Torr, P.: Deep fully-connected part-based models for human pose estimation. In: ACML (2018)
3. de Bem, R., Ghosh, A., Ajanthan, T., Siddharth, N., Torr, P.: A conditional deep generative model of people in natural images. In: WACV (2019)
4. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 717–732. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_44
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
6. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: CVPR (2017)
7. Elgammal, A., Lee, C.S.: Inferring 3D body pose from silhouettes using activity manifold learning. In: CVPR (2004)
8. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning. MIT press, Cambridge (2016)
9. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS (2014)

10. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: ICCV (2015)
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. TPAMI **36**, 1325–1339 (2014)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
13. Jaeggli, T., Koller-Meier, E., Van Gool, L.: Learning generative models for monocular body pose estimation. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007. LNCS, vol. 4843, pp. 608–617. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76386-4_57
14. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: ACMMM (2014)
15. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: ICLR (2018)
16. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
17. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: NIPS (2014)
18. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR (2014)
19. Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.: Deep convolutional inverse graphics network. In: NIPS (2015)
20. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: ICML (2016)
21. Lassner, C., Pons-Moll, G., Gehler, P.V.: A generative model for people in clothing. In: ICCV (2017)
22. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016)
23. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. In: NIPS (2017)
24. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: CVPR (2018)
25. Massiceti, D., Siddharth, N., Dokania, P., Torr, P.H.: FlipDial: a generative model for two-way visual dialogue. In: CVPR (2018)
26. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
27. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: ICML (2014)
28. Schulman, J., Heess, N., Weber, T., Abbeel, P.: Gradient estimation using stochastic computation graphs. In: NIPS (2015)
29. Seemann, E., Nickel, K., Stiefelhagen, R.: Head pose estimation using stereo vision for human-robot interaction. In: FG (2004)
30. Shotton, J., et al.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011)
31. Siarohin, A., Sangineto, E., Lathuiliere, S., Sebe, N.: Deformable GANs for pose-based human image generation. In: CVPR (2018)
32. Siddharth, N., et al.: Learning disentangled representations with semi-supervised deep generative models. In: NIPS (2017)
33. Sigal, L., Balan, A., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NIPS (2008)

34. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: real-time face capture and reenactment of RGB videos. In: CVPR (2016)
35. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In: NIPS (2014)
36. Tulyakov, S., Liu, M., Yang, X., Kautz, J.: MoCoGAN: decomposing motion and content for video generation. In: CVPR (2018)
37. von Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: Sparse inertial poser: automatic 3D human pose estimation from sparse IMUs. Eurographics (2017)
38. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: video forecasting by generating pose futures. In: ICCV (2017)
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**, 600–612 (2004)
40. Wang, Z., Merel, J.S., Reed, S.E., de Freitas, N., Wayne, G., Heess, N.: Robust imitation of diverse behaviors. In: NIPS (2017)
41. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
42. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: ICCV (2017)
43. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)