International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

DGPose: Deep Generative Models for Human Body Analysis

Rodrigo de Bem[1,2] · **Arnab Ghosh**[1] · **Thalaiyasingam Ajanthan**[1] · **Ondrej Miksik**[1] · **Adnane Boukhayma**[1] · **N. Siddharth**[1] · **Philip Torr**[1]

Received: date / Accepted: date

Abstract Deep generative modelling for human body analysis is an emerging problem with many interesting applications. However, the latent space learned by such approaches is typically not interpretable, resulting in less flexibility. In this work, we present deep generative models for human body analysis in which the body pose and the visual appearance are disentangled. Such a disentanglement allows independent manipulation of pose and appearance, and hence enables applications such as pose-transfer without specific training for such a task. Our proposed models, the Conditional-DGPose and the Semi-DGPose, have different characteristics. In the first, body pose labels are taken as conditioners, from a fully-supervised training set. In the second, our structured semi-supervised approach allows for pose estimation to be performed by the model itself and relaxes the need for labelled data. Therefore, the Semi-DGPose aims for the joint *understanding* and *generation* of people in images. It is not only capable of mapping images to interpretable latent representations but also able to map these representations back to the image space. We compare our models with relevant baselines, the ClothNet-Body and the Pose Guided Person Generation networks, demonstrating their merits on the Human3.6M, ChictopiaPlus and DeepFashion benchmarks.

[1]Department of Engineering Science
University of Oxford
Oxford, UK
E-mail:
{rodrigo, arnabg, ajanthan,
omiksik, adnane, nsid, phst}@robots.ox.ac.uk

[2]Center of Computational Sciences
Federal University of Rio Grande
Rio Grande, Brazil
E-mail: rodrigobem@furg.br

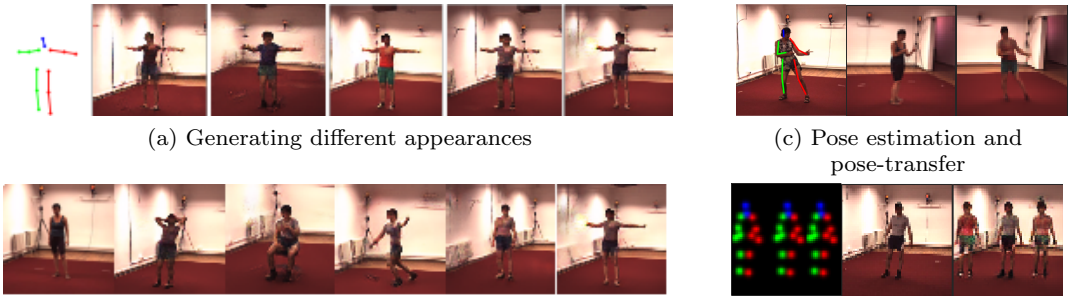
Keywords Deep generative models, Semi-supervised learning, Human pose estimation, Variational autoencoders, Generative adversarial networks

1 Introduction

Human body analysis has been a long-standing goal in computer vision, with many applications in human-machine interaction, health-care, shopping, sports, entertainment and gaming [2,64,80,82,97]. Popular approaches to this problem have focused on supervised learning of discriminative models [12,13,15,103], which map visual inputs (images or videos) to suitable abstract representations (e.g. human body pose). While these approaches do exceptionally well on their prescribed task, as evidenced by their performance on pose estimation benchmarks [3,37,41], 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 shortcoming, particularly when one is dealing with real-world visual data, as it requires a substantial amount of human time and effort to annotate. Thus, being able to relax the reliance on labelled data is a highly desirable goal. The latter, states a rather significant limitation, the incapacity to manipulate abstractions directly with the aim of generating novel visual data. For instance, changes in the pose of an arm cannot be used for the generation of images or videos in which that arm is correspondingly displaced.

Generative models, in contrast to discriminative ones, enable the *analysis-by-synthesis* of the human body. With them, ideally, one could generate images of humans in diverse combinations of body poses and appearances, i.e. clothing, skin colours, hairstyles, and



(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 models for 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 a given appearance (second image), we show a generated sample combining the pose of the first image with the appearance of the second. (d) For manipulated poses (first image) and a given appearance (second image), it can *hallucinate* people in the scene.

scenarios. This has many potential applications. For instance, it can be used for performance capture and reenactment of RGB videos, as already showcased for faces [90], and still incipient for human bodies [4,14]. It can also be used to generate images in user-specified poses, to enhance and augment datasets with minimal annotation effort.

Recently, such approaches have been commonly formulated as deep generative models (DGMs) [29,47,72] – an extension of standard generative models that incorporate neural networks as flexible function approximators. These models are particularly effective in complex perceptual domains such as computer vision [49], language [62], and robotics [102], 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 methods discussed above by a) employing unsupervised learning, thereby removing the need for labels, and b) embracing a fully generative modelling.

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, usually represented as a smooth high-dimensional manifold, do not have the consistent semantic meaning as 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 issue mentioned above, while still eschewing reliance on fully-labelled data, we rely on a structured semi-supervised variational autoencoder (VAE) framework [46,84]. 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 [79], and c) interpretable subset of latent variables defined through the model structure.

In this context, we present a structured semi-supervised VAEGAN architecture, the Semi-DGPose, in which we further extend structured semi-supervised models [46,84] with a discriminator-based loss function from generative adversarial networks (GANs) [29,50], formulating it as a principled and unified probabilistic framework. To our knowledge, it is the first structured semi-supervised deep generative model of people directly learned in the *natural image* (or *natural scene*) space. This allows the method to directly learn the intricacies in the formation of natural (i.e. real) images. However, it is important to mention that natural images, in contrast to artificial visual stimuli (e.g. segmentation masks, binary masks, or pose vectors), have complex statistical structure and are much more challenging to parameterised [27,44,85]. Consequently, methods that work well with the latter may not succeed when tackling the former [22,48]. In contrast to previous work [51,56,57,83,98], our model directly enables: i) *semi-supervised pose estimation*; and ii) *indirect pose-transfer* without specific training for such a task, both of which are tested and verified by experimental evidence.

Additionally, as an intermediate step in the investigation towards our main contribution, we propose a conditional-VAEGAN model, dubbed Conditional-DGPose. It is less distinct from previous art [51, 56], however, still differently from earlier work in the literature, it has: i) allowed pose manipulation on extreme cases, e.g. by performing *cross-domain pose-transfer* and by *hallucinating* multiple people, in a variety of unseen or even unrealistic poses; and ii) achieved state-of-the-art results on image reconstruction conditioned on pose, outperforming the closest related comparable baseline [51]. We illustrate some capabilities of our models in Fig. 1.

The present paper builds upon our previous approaches [6,7] with further theoretical and technical details, evaluation, and discussion. Here, we present in full our comprehensive deep generative model framework for human body analysis in images. Along with an overview of VAEGAN models, this enables us to shed light on differences and similarities between conditional-VAEGANs and structured semi-supervised VAEGANs. More precisely, we provide additional evaluations of our Conditional-DGPose and Semi-DGPose models on the most relevant benchmarks in the literature, the Human3.6M [37], the ChictopiaPlus [51], and the DeepFashion [54] datasets. We also provide new qualitative and quantitative comparisons with the Pose Guided Person Generation (PG^2) baseline [56]. The application of our models to real images and the results obtained are essential to show the relevance of interpretable and structured modelling. This emphasise the effectiveness of the proposals, despite the significant challenge of jointly aim for *understanding* and *generating* people in images. In summary, our main contributions are:

i) a comprehensive framework for the joint *understanding* and *generation* of people in images, not only capable of mapping images to interpretable latent representations but also capable of mapping these representations back to the image space;

ii) a real-world application of structured deep generative models of images, disentangling pose from appearance in the analysis of the human body;

iii) a thorough quantitative and qualitative evaluation of the capabilities of our models; and

iv) a demonstration of its principal utilities by performing semi-supervised pose estimation, pose-transfer and pose manipulation.

2 Related Work

2.1 Analysing Humans in Images: Overview

The analysis of people in visual data has been actively investigated as a computer vision and machine learning topic lately [4,14,30,31,74,75,90,96,99]. Historically, the process of synthesising *virtual humans* [33,59,60] is a computer graphics undertake since its origins in the '60s, with Boeing's "first man" [9,23]. Therefore, the geometric and photometric intricacies in the formation of digital images depicting people are well-known in computer graphics, as demonstrated by the existence of many commercial and academic specialised engines [1, 61,63,65,70,94]. Nonetheless, the unconstrained creation of truly realistic RGB images is still reasonably dependent upon manual intervention [34]. Moreover, to produce accurate images of people is harder since humans seem to be very familiarised to corporal traits (e.g. faces) even since their early ages [58,95].

Over time, the generation of humans in images was also embraced by the computer vision community. Aiming for less manual intervention, image-based techniques were successfully adopted on matters like rendering and modelling [8,11,20,42,87]. For instance, a large body of work has relied on geometric 3D models for generating synthetic images of faces [35, 69], bodies [10,88], and hands [16,76,78]. Despite that, to automatically synthesise artificial images indistinguishable from real ones may be considered as equivalent to succeed in a *visual Turing test* [81]. Hence, a substantially complicated and consequently yet unsolved challenge [21].

Another line of approaches, following the machine learning methodologies closely, had modelled the image formation by designing and learning probabilistic generative models [18,24,25,26,52,100,106]. However, it is highly complex and constrained due to intractable probability distributions and the high variability of latent factors. Often, simplifying assumptions are made in practice, such as independence between different factors of variation, leading to weak generative models that fail to capture statistical subtleties.

Recently, the advent of the deep generative models (DGMs) [29,47,72] somehow gathers the three lines of methods mentioned above. Bringing together characteristics from computer graphics, computer vision, and machine learning makes the DGMs a powerful *analysis-by-synthesis* framework. We discuss the DGM-based approaches related to our work in the following section.

2.2 Analysing Humans in Images with DGMs

Generally, in *classical DGMs*, such as standard VAEs and GANs, pose representation is non-interpretable and unsupervised, entangled with the visual appearance in the latent space. This is similarly employed by some *image-to-image translation networks*, however, in contrast to the relatively low-dimensional manifolds learned by the DGMs, in the latter case high-dimensional abstractions are learned and used strictly for direct mapping from and to the image space. On the other hand, *conditional DGMs* usually define part of the abstract data representation, i.e. body pose, to be an interpretable and observable random variable, while the rest of the representation (visual appearance) is kept non-interpretable and latent, still subjected to unsupervised learning. Finally, in *structured DGMs* approaches, as the Semi-DGPose, the latent space can be simultaneously composed by interpretable and non-interpretable random variables. In the former case, the variables may be fully or semi-supervised, while in the latter group they are still maintained unsupervised. Below, we describe related literature gathering the methods according to their adopted type of approach.

Image-to-image networks. Ma et al. [56] introduce the Pose Guided Person Generation Network (PG^2), a two stage image-to-image translation model which is trained on pairs of images of the same person in different poses, scales and points of view. The authors admit the difficulty of generating poses and detailed appearance simultaneously in an end-to-end fashion. Their model, which is conditioned on images rather than poses, does not allow sampling, thus in its essence, it is not a generative model, which is again in contrast to our single-stage approaches. In a second proposal, Ma et al. [57] present a GAN-based model for learning image embeddings of foreground, background and pose variables encoded as interpretable variables. The method is still limited to training and testing with cross-pose/scale pairs for pose-transfer, however, it allows sampling, differently from the PG^2. In contrast to our Semi-DGPose model, it is not capable of performing either pose estimation or semi-supervised learning, relying on off-the-shelf pose estimators to perform pose-transfer.

Recently, Esser et al. [19] present a conditional image-to-image translation network based on the U-Net [77]. The model is conditioned on an appearance encoding obtained using a VAE architecture. It is more versatile than [56,57], although still not capable of producing either an interpretable encoding of pose (pose estimation) or performing semi-supervised learning. Similarly, Balakrishnan et al. [4] also propose a U-Net-based approach. In this case, the authors make use of three U-Nets which tackle foreground segmentation and synthesis, as well as background synthesis. The model is trained with video sequences of the same person performing a limited set of activities. Therefore, it is limited to translating images of the same person to different poses. Other very recent approaches [14,67] have to be explicitly trained for pose-transfer, i.e. using images pairs, and do not have the capability of predicting pose. This is in sharp contrast to our Semi-DGPose approach, in which we learn pose estimation, while pose-transfer is achieved as a by-product. In the method by Trumble et al. [92], pose is estimated from multiple views, although it does not allow semi-supervised learning.

Rhodin et al. [73] learn 3D pose estimation from multi-view images of the same person acquired from synchronised and calibrated cameras. In contrast to our approach, their method explicitly uses the rotation matrix between cameras during training for the unsupervised learning of a geometry-aware latent representation. From such representation, the 3D pose is estimated posteriorly with a shallow network. The authors do not define their method as a generative model, but as a 3D pose estimator, although it can perform novel viewpoint synthesis. Another work by Zanfir et al. [107] focus uniquely on the specific task of appearance transfer, also based on 3D pose. In contrast, our closely related task of pose-transfer is just one among all the tasks our DGMs can perform (e.g. sampling, pose estimation, direct manipulation) employing only 2D pose representations. Lastly, Zhang et al. [108] focus on a slightly different task. They propose the unsupervised discovery of 2D landmarks using optical flows from Human3.6M videos as a short-term self-supervision. Such landmarks are an intermediate representation of pose since they do not correspond explicitly to specific body parts. In contrast, we employ single still images using directly and explicitly interpretable pose representations.

Finally, it is essential to differentiate such image-to-image translation methods from our DGMs. The former depends upon input images at test time, while the latter effectively allow sampling from the latent structured representations learned during training. This subtle difference means that such structured representations are responsible for learning the underlying factors of variations in image generation, without relying on information from input images for generating outputs at test time.

Classical DGMs. Lassner et al. [51] have proposed the ClothNet-full model, in which a VAE model is used to learn a latent representation of segmentation masks of people in given poses. The reconstructed masks are mapped back to the image space by an image-to-image translation module based on [38]. In contrast, we learn our generative models directly on the raw image data without the need for body parts segmentation. Moreover, pose is interpretable in both of our methods. Siarohin et al. [83] propose a GAN model with skip connection in the generator and a discriminator conditioned on pose. Similarly to [56], the model is restricted to pose-transfer on pairs of images of the same person. The body pose is always given to the model and non-interpretable in the learned latent encoding. Apart from this, Walker et al. [98] proposed a hybrid architecture, associating a VAE and a GAN 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. Considering GAN based generative models, Tulyakov et al. [93] present a GAN network that learns motion and content in two separate latent spaces in an unsupervised manner. However, it does not allow explicit manipulation over the human pose.

Conditional DGMs. Lassner et al. [51] present a second model, the ClothNet-Body, which is a CVAE conditioned on human pose. This model is closely related to our Conditional-DGPose, but it also uses low-dimensional segmentation masks and an auxiliary image-to-image transfer network, based on [38], to generate realistic images. Pumarola et al. [71] propose an unsupervised image synthesis based on a conditional GAN method, yet it is also not capable of performing pose prediction.

In summary, there are methods in the literature closely related to our Conditional-DGPose, mainly due to its conditional nature. Although, to our knowledge, no other method gathers the capabilities of our Semi-DGPose as a *structured DGM*. The novelty in the Semi-DGPose largely relies on how the body pose is handled, differing it from related work. Moreover, the capacity for performing pose estimation, indirect pose-transfer, and semi-supervised learning, while aiming for joint *understanding* and *generation* of people in images is peculiar to our model. Following Larsen et al. [50], we use a discriminator in our training to improve the quality of the generated images. However, in contrast to [50], the latent space of our approach is interpretable, which enables us to sample different poses and appearances.

3 Preliminaries

Deep generative models (DGMs) come in two broad flavours – Variational Autoencoders (VAEs) [47,72], and Generative Adversarial Networks (GANs) [29]. 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 θ. Typically the model parameters θ are represented in the form of a neural network.

VAEs express an objective to learn the parameters θ 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 $\mathcal{L}(\phi, \theta; \mathbf{x})$, called the evidence lower bound (ELBO), as

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

The conditional density $q_\phi(\mathbf{z}|\mathbf{x})$ is called the recognition or inference distribution, with parameters ϕ also represented in the form of a neural network. Lastly, VAEs also admit an extension to *conditional* generative models (CVAEs) [86], simply by incorporating a conditioning variable $\mathbf{y}$, to derive

$$\begin{aligned} \log p_\theta(\mathbf{x}|\mathbf{y}) &\geq \mathcal{L}_{\text{CVAE}}(\phi, \theta; \mathbf{x}|\mathbf{y}) \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z}|\mathbf{y})}{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})}\right]. \end{aligned} \tag{2}$$

On the other hand, in the context of 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 [46,84]. 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

$$\begin{aligned} \mathcal{L}_{\text{SS}}(\theta, \phi; \mathcal{D}) = &\sum_{\mathbf{x}_u \in \mathcal{D}_u} \mathcal{L}_u(\theta, \phi; \mathbf{x}_u) \\ &+ \gamma \sum_{(\mathbf{x}_s, \mathbf{y}_s) \in \mathcal{D}_s} \mathcal{L}_s(\theta, \phi; \mathbf{x}_s, \mathbf{y}_s) \end{aligned} \tag{3}$$

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

$$\mathcal{L}_u(\theta, \phi; \mathbf{x}_u) = \mathcal{L}_{\text{VAE}}(\theta, \phi; \mathbf{x}_u), \text{ and} \tag{4}$$

$$\mathcal{L}_s(\theta, \phi; \mathbf{x}_s, \mathbf{y}_s) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_s, \mathbf{y}_s)} \left[\log \frac{p_\theta(\mathbf{x}_s, \mathbf{z}|\mathbf{y}_s)}{q_\phi(\mathbf{z}|\mathbf{x}_s, \mathbf{y}_s)} \right] + \alpha \log q_\phi(\mathbf{y}_s|\mathbf{x}_s), \tag{5}$$

respectively. Here, the hyper-parameter γ (Eq. 3) controls the relative weight between the supervised and unsupervised dataset sizes, and α (Eq. 5) controls the relative weight between generative and discriminative learning.

Note that by the factorisation of the generative model, VAEs necessitate the specification of an explicit likelihood function $p_\theta(\mathbf{x}|\mathbf{z})$, which can often be difficult. GANs, 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

$$\mathcal{L}_{\text{GAN}}(D, G) = \mathbb{E}_{p(\mathbf{x})} \left[\log D(\mathbf{x}) \right] + \mathbb{E}_{p_\theta(\mathbf{z})} \left[1 - \log D(G(\mathbf{z})) \right]. \tag{6}$$

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 [43].

A more recent family of DGMs, VAEGANs [50], bring together these two different approaches into a single objective that combines both the VAE and GAN objectives directly as

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{GAN}}. \tag{7}$$

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

4 Our Approach

As set out in the preliminaries (Sec. 3), we use the VAEGAN framework as the basis for our generative models [50]. Note that, in incorporating semi-supervised learning, the semi-supervised VAEGAN includes two distinct tasks. First, it involves learning a recognition network that can estimate pose $\mathbf{y}$ and *appearance* $\mathbf{z}$ for any given RGB image $\mathbf{x}$. Second, it involves learning a generative network that combines a given pose with an appearance to generate visual data (RGB image) corresponding to those variables.

From discriminative modelling, we know that the first task, i.e. predicting pose, is eminently plausible up to learning an appearance model. However, learning the full generative model is something that can be fraught with difficulties. For one, pose and appearance can exhibit a large degree of information imbalance – pose can be distilled into a set of (x, y) coordinates, whereas appearance can encode a vast swathe of information (e.g. texture, colour, shapes) about the given input.

Given a generative model that takes both appearance $\mathbf{z}$ and pose $\mathbf{y}$ as inputs to produce an RGB image $\mathbf{x}$, a reasonable first step can be just to evaluate the performance of a conditional generative model, where the conditioning variable is taken to be the interpretable pose $\mathbf{y}$. We refer to this setup as Conditional-DGPose, with reference to the fact that it is a conditional-VAEGAN model. Its lower bound is given by Eq. 2, and its final objective function is defined as

$$\mathcal{L} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{GAN}}, \tag{8}$$

in contrast to the standard VAEGAN objective (Eq. 7). Here, all data is "labelled" with pose, but the goals were: i) primarily, to verify qualitatively if a low-dimensional conditioning variable would affect the conditional generative model; ii) secondly, to evaluate the accuracy of the reconstructed images quantitatively w.r.t. the human body poses and the image quality.

Once verified through experiments that the conditional approach works, we could then proceed towards our structured semi-supervised VAEGAN, referred to as Semi-DGPose, as its main difference from the previous setup is that the encoding distribution is no longer conditioned on the pose, but instead predicts it as per Eq. 3–6. In contrast to the standard VAEGAN objective (Eq. 7), the structured semi-supervised VAEGAN final objective function is given by,

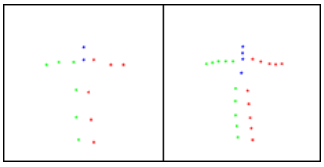
$$\mathcal{L} = \mathcal{L}_{\text{SS}} + \mathcal{L}_{\text{GAN}}. \tag{9}$$

We describe the details and implementations of our models in the rest of this section. Next, we start defining the adopted pose representations, which are common for both, the Conditional-DGPose and the Semi-DGPose architectures.

4.1 Pose Representation

In our DGMs, the random variable $\mathbf{y}$ corresponds to an abstraction of the human body pose. Therefore a suitable concrete representation must be adopted in the

implementation of the models. As mentioned in our literature review, many methods which define a generative model in the *pose space* would simply encode J joints defining the body as a vector $\mathbf{y}_v$, such that $\mathbf{y}_v \in \mathcal{R}^{2J}$. Others employ extended versions of it, in which positions of R rigid parts and B whole body are derived from the annotated joints [105], such that $\mathbf{y}_v \in \mathcal{R}^{2(J+R+B)}$. Both cases are illustrated in Fig. 2.



(a) (b)

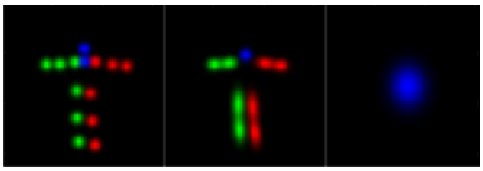
Fig. 2 Vector representation. (a) $J = 14$ joints which compose a 2D pose vector $\mathbf{y}_v \in \mathcal{R}^{2J}$. (b) An extended 2D vector composed by 24 body parts ($J = 14$ annotated joints, $R = 9$ intermediate points between joints and $B = 1$ central point), such that $\mathbf{y}_v \in \mathcal{R}^{2(J+R+B)}$.

On the other hand, the mapping of 2D joints positions to heatmaps has shown to be very effective in several pose estimation approaches [15,68,91,103]. The Gaussian heatmaps represent the underlying probability distribution of body parts' locations. In our method, the heatmap representation $\mathbf{y}_h$ consists of P body elements, in a way that $\mathbf{y}_h \in \mathcal{R}^{P\times H\times W}$, where H and W are the heatmap height and width, respectively. In the simplest case $P = J$, however, as the set of joints is reasonably sparse, to cover the entire area of the bodies, joints, rigid parts and the whole body might be used as an extended case, in which $P = J + R + B$ [5], as illustrated in Fig. 3. In this way, each body element p is represented using a 2D Gaussian around its centre $\boldsymbol{\mu}_p = (i_p, j_p)$, with diagonal covariance matrix $\Sigma_p = R_p \begin{bmatrix} \sigma^2_{p,i} & 0 \\ 0 & \sigma^2_{p,j} \end{bmatrix} R_p^\top$, computed as follows:

Joints. Since joints have a limited spatial extent, we follow previous approaches [15,68,91,103] in modelling them as isotropic Gaussians that are centred at the ground-truth joint location and have a small standard deviation (e.g. $\sigma_{p,i} = \sigma_{p,j} = 1.5$ pixel for a 64×64 heatmap).

Rigid Parts. The centre $\boldsymbol{\mu}_p$ of a rigid part p is defined as the mean point of the centres $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_l$ of the joints it connects. We orient the Gaussian representing the rigid part to align its i axis with the line connecting $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_l$. We define $\sigma_{p,i}$ to be proportional to $|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l|$, and set $\sigma_{p,j} = \kappa_p \sigma_{p,i}$, where κ_p is a part-specific ratio, inspired by anthropometric measurements [66].

Body. The body centre is defined to be the mean of the annotated joint centres. Principal component analysis (PCA) of the joint centres is used to obtain the orientation of the body in the image plane. We define $\sigma_{p,i}$ and $\sigma_{p,j}$ to be proportional to the distance between the extreme projections of the joint centres onto, respectively, the principal and secondary axes of variation.



(a) (b) (c)

Fig. 3 Heatmap representation. Heatmaps superimposed corresponding to (a) $J = 14$ annotated joints, (b) $R = 9$ rigid parts, and (c) $B = 1$ whole body; such that $\mathbf{y}_h \in \mathcal{R}^{P\times H\times W}$. Right, left and central body parts are denoted by the colours green, blue and red, respectively, in the person-centric representation.

In our both models, as detailed in the next sections, we make use of both forms of pose representation, taking advantage of their particular characteristics in each case. In the Conditional-DGPose, only the heatmap representation $\mathbf{y}_h$ is employed, since, as shown later in our experiments, it can be seamlessly concatenated to feature maps, helping on the generation of accurate output images. On the other hand, in the Semi-DGPose model, we additionally employ the vector-based form $\mathbf{y}_v$, as a way of maintaining a low-dimensional latent representation of pose.

4.2 DGPose Architectures

We have tested several variations of deep CNN architectures for implementing our models, culminating in our best performing ones, which are described here. All its modules are deep CNNs, and full implementation definitions are given in the appendix (Sec. A) and referred adequately in the text. Due to the generality of generative models, the architectures may be employed in different ways according to the aimed tasks. Thus, we describe separately training and test phases, dividing the latter into *reconstruction*, *pose-transfer*, *sampling* and *pose-estimation*, for both models. Thus, the Conditional-DGPose and the Semi-DGPose are described following.

4.2.1 Conditional-DGPose

Our conditional-VAEGAN model learns the parameters of four deep CNN networks simultaneously: i) a recog-

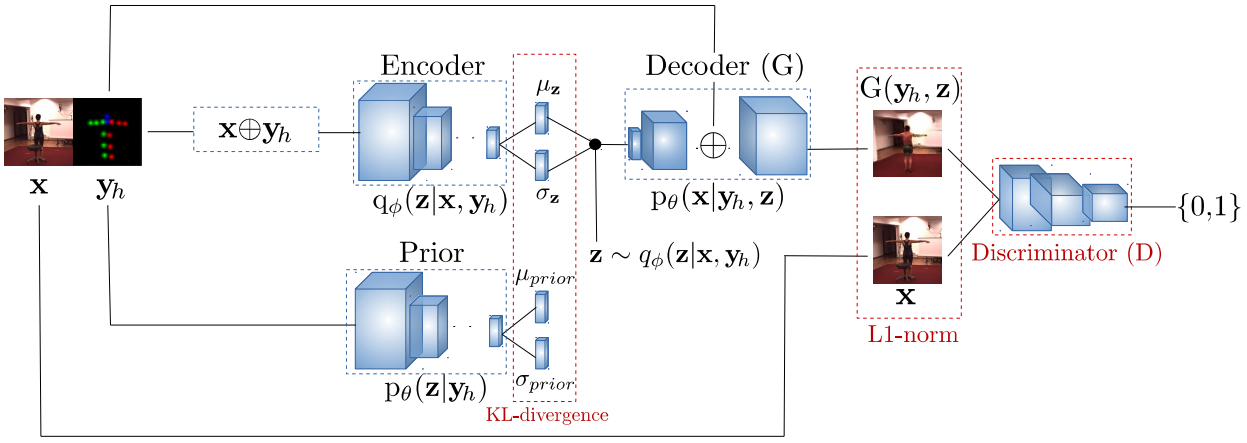


Fig. 4 Conditional-DGPose architecture. At the training, the Encoder receives $\mathbf{x} \oplus \mathbf{y}_h$ as input and learns the posterior $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}_h)$. The Prior module receives $\mathbf{y}_h$ alone and learns the distribution $p_\theta(\mathbf{z}|\mathbf{y}_h)$. Appearance is sampled $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}_h)$, using the reparametrization trick [47], and passed to the Decoder, as well as the conditioning pose $\mathbf{y}_h$, which is concatenated to the Decoder feature maps. The Decoder then generates a reconstructed image $G(\mathbf{y}_h, \mathbf{z})$. The loss function (see Eq. 8, Sec. 4) is composed by the following terms, highlighted in red: the L1-norm $L1(\mathbf{x}, G(\mathbf{y}_h, \mathbf{z}))$ which is computed between the original and the reconstructed image; the KL-divergence $\mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}_h)||p_\theta(\mathbf{z}|\mathbf{y}_h)]$, which is used to regularise the posterior distribution; and the GAN Discriminator cross-entropy loss used to learn how to discern between real and generated images.

nition network (Encoder), which estimates appearance $\mathbf{z}$ conditioned to pose $\mathbf{y}_h$ and to a given RGB image $\mathbf{x}$; ii) a Prior network, which estimates appearance $\mathbf{z}$ conditioned to pose $\mathbf{y}_h$ alone; iii) a generative network (Decoder), which combines appearance $\mathbf{z}$ and the conditioning pose $\mathbf{y}_h$, to generate corresponding RGB images $G(\mathbf{y}_h, \mathbf{z})$; and iv) a Discriminator network, which differentiates between real images $\mathbf{x}$ and generated images $G(\mathbf{y}_h, \mathbf{z})$. Learning is pursued by the minimisation of the loss function $\mathcal{L} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{GAN}}$ (Eq. 8, Sec. 4), composed by the CVAE evidence lower bound (ELBO) $\mathcal{L}_{\text{CVAE}}$ and by the GAN cross-entropy discriminator loss $\mathcal{L}_{\text{GAN}}$. An overview of our model is shown in Fig. 4 and implementation details are provided in Tab. A2 (appendix). Below, we describe further the training and the test phases, dividing the latter into *reconstruction*, *pose-transfer* and *sampling*.

Training. Given an image $\mathbf{x}$, the corresponding heatmap labels (conditioning pose) are concatenated to it as per $\mathbf{x} \oplus \mathbf{y}_h$ (Encoder, Layer 1, Tab. A2). Then, the Encoder estimates the conditional posterior distribution $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}_h)$. The heatmap labels $\mathbf{y}_h$ alone are the input of the Prior module, which estimates the distribution $p_\theta(\mathbf{z}|\mathbf{y}_h)$. Appearance is sampled from the posterior $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}_h)$, using the reparametrisation trick [47]. The sample $\mathbf{z}$, along with the conditioning pose $\mathbf{y}_h$ (Decoder, Layer 7, Tab. A2), are passed through the Decoder which generates a reconstructed image $G(\mathbf{y}_h, \mathbf{z})$. This reconstructed image, along with the real image $\mathbf{x}$, are still used as inputs for the Discriminator module, which learns how to discern between them. Finally, the overall loss function minimised during training is composed of the L1-norm reconstruction loss $L1(\mathbf{x}, G(\mathbf{y}_h, \mathbf{z}))$; the KL-divergence, which acts as a regulariser, between the posterior and the prior distributions, $\mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}_h)|p_\theta(\mathbf{z}|\mathbf{y}_h)]$; and the cross-entropy Discriminator loss (Eq. 6, Sec. 3).

Reconstruction and Direct Pose-transfer. At test time, when an image $\mathbf{x}_1$ and its corresponding pose $\mathbf{y}_{h_1}$ are given as input, the reconstructed image $G(\mathbf{y}_{h_1}, \mathbf{z}_1)$ is obtained as the Decoder output. However, if $\mathbf{x}_1$ is used as input along with a different pose $\mathbf{y}_{h_2}$, the person in the reconstructed image $G(\mathbf{y}_{h_2}, \mathbf{z}_1)$ will keep the appearance of $\mathbf{x}_1$, with the body pose defined by $\mathbf{y}_{h_2}$, as illustrated in Fig. 5. Similarly, as shown later in our experiments, the same procedure may be adopted to *directly manipulate* the reconstructed image, such as changing body size and aspect ratio, moving or suppressing body parts or even hallucinating multiple people.

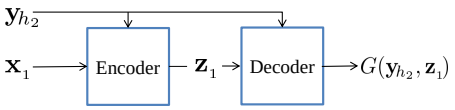


Fig. 5 Conditional-DGPose direct pose-transfer and manipulation at test time.

Sampling. At test time, sampling is obtained when no RGB image is given as input. In this case, as illustrated in Fig. 6, only a conditioning pose $\mathbf{y}_h$ is given as

the input of the Prior module, which defines $p_\theta(\mathbf{z}|\mathbf{y}_h)$. From this Prior distribution, the sampled appearance $\mathbf{z}$ and the conditioning pose $\mathbf{y}_h$ are passed to the Decoder network. In this manner, for a given pose, different appearances can be randomly created from the learned generative model.

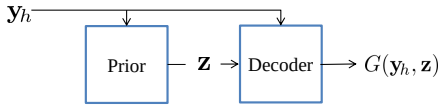


Fig. 6 Conditional-DGPose sampling at test time.

4.2.2 Semi-DGPose

Differently from the Conditional-DGPose, our structured semi-supervised VAEGAN model (Fig. 7) learns the parameters of three deep CNN networks simultaneously: i) a recognition network (Encoder), which estimates appearance $\mathbf{z}$ and pose $\mathbf{y}_v$ from a given RGB image $\mathbf{x}$; ii) a generative network (Decoder), which combines appearance $\mathbf{z}$ and pose $\mathbf{y}_v$, to generate corresponding RGB images $G(\mathbf{y}_v, \mathbf{z})$; and iii) a Discriminator network, which differentiates between real images $\mathbf{x}$ and generated images $G(\mathbf{y}_v, \mathbf{z})$. Learning is pursued by the minimisation of the loss function $\mathcal{L} = \mathcal{L}_{\text{SS}} + \mathcal{L}_{\text{GAN}}$ (Eq. 9, Sec. 4), composed by the structured semi-supervised VAE evidence lower bound (ELBO) $\mathcal{L}_{\text{SS}}$ and by the GAN cross-entropy discriminator loss $\mathcal{L}_{\text{GAN}}$. A fourth module, called Mapper, is introduced by us to overcome a peculiarity caused by the inclusion of pose in the latent space. Such a module, trained separately, is described next.

The Mapper Module. Our preliminary experiments with the Conditional-DGPose showed that heatmaps led to better quality reconstructions, 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 [17,28]. To cope with this mismatch, we introduce the Mapper module, which maps pose-vectors $\mathbf{y}_v$ to heatmaps $\mathbf{y}_h$. Ground-truth heatmaps are constructed from manually annotated 2D joints labels, using a simple weak annotation strategy [5]. The Mapper module is then trained to map 2D joints to heatmaps, minimising 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 Sec. 5.5. In the training of the full Semi-DGPose architecture, the Mapper module is integrated to it with its weights kept fixed, since the mapping function has been learned already. The Mapper allows us to keep a low-dimensional representation $\mathbf{y}_v$ in the latent space, at the same time that a dense high-dimensional "spatial" heatmap representation $\mathbf{y}_h$ 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.

In the rest of this section, we describe further the training and the test phases, dividing the latter into *reconstruction*, *indirect pose-transfer*, *sampling* and *pose estimation*. An overview of our model is shown in Fig. 7 and implementation details are provided in Tab. A3 (appendix).

Training. The terms of Eq. 3 (Sec. 3) correspond to two training routines which are alternately employed, according to the presence or absence of ground-truth labels.

In the *unsupervised case*, when no label is available, it is similar to the standard VAE (see Eq. 4, Sec. 3). Accurately, given the image $\mathbf{x}$, the Encoder estimates the posterior distribution $q_\phi(\mathbf{y}_v, \mathbf{z}|\mathbf{x})$, where both appearance $\mathbf{z}$ and pose $\mathbf{y}_v$ are assumed to be independent given the image $\mathbf{x}$. Then, pose $\mathbf{y}_v$ and appearance $\mathbf{z}$ are sampled from the posterior, using the reparametrization trick [47], and passed to the Decoder to generate a reconstructed image. Finally, the unsupervised loss function minimised during training is composed of the L1-norm reconstruction loss $L1(\mathbf{x}, G(\mathbf{y}_v, \mathbf{z}))$; the KL-divergences, which act as regularisers, between the posterior and the prior distributions, $\text{KL}[q_\phi(\mathbf{y}_v|\mathbf{x})|p(\mathbf{y}_v)]$ and $\text{KL}[q_\phi(\mathbf{z}|\mathbf{x})|p(\mathbf{z})]$; and the cross-entropy Discriminator loss (Eq. 6, Sec. 3).

In the *supervised case*, when the pose label is available, the KL-divergence between the posterior pose distribution and the pose prior, $\text{KL}[q_\phi(\mathbf{y}_v|\mathbf{x})|p(\mathbf{y}_v)]$, is replaced with a regression loss between the estimated pose and the given label (see Eq. 5, Sec. 3). 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 minimised 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.

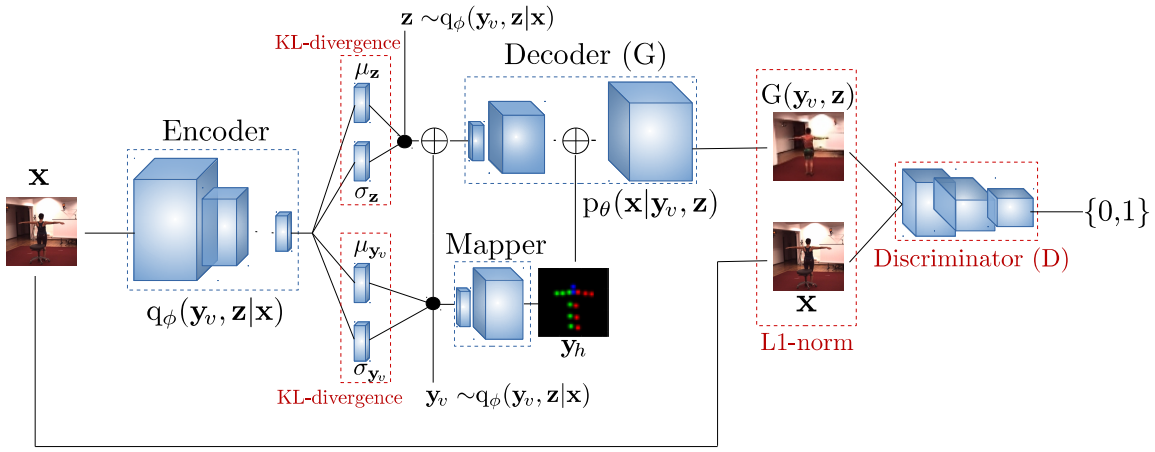


Fig. 7 Semi-DGPose architecture. At the training, the Encoder receives $\mathbf{x}$ as input and learns the posterior distribution $q_\phi(\mathbf{y}_v, \mathbf{z}|\mathbf{x})$. In the *unsupervised* routine, samples of appearance $\mathbf{z}$ and pose $\mathbf{y}_v$ are obtained using the reparametrisation trick [47]. These samples are passed to the Decoder, which generates a reconstructed image $G(\mathbf{y}_v, \mathbf{z})$. The unsupervised loss function is composed by the following terms, highlighted in red: the L1-norm $L1(\mathbf{x}, G(\mathbf{y}_v, \mathbf{z}))$ between the original and the reconstructed images; the KL-divergence losses between the posterior distribution $q_\phi(\mathbf{y}_v, \mathbf{z}|\mathbf{x})$ and the weak priors $p(\mathbf{y}_v)$ and $p(\mathbf{z})$, which work as regularisers (see Eq. 4, Sec. 3); and the cross-entropy Discriminator loss (Eq. 6, Sec. 3). In the *supervised* routine (not shown above for simplicity), the only difference is that a regression loss between the estimated pose and the pose ground-truth label substitutes the KL-divergence over the pose posterior distribution (see Eq. 5, Sec. 3). In both, supervised and unsupervised training routines, the low-dimensional pose vector $\mathbf{y}_v$ is mapped to a heatmap representation $\mathbf{y}_h$ by the Mapper module and concatenated to the Decoder. Eq. 3 (Sec. 3) shows the overall loss function.

In both *unsupervised* and *supervised* cases, the Mapper module, which is trained *offline*, is used to map the pose-vector $\mathbf{y}_v$ in the latent space to a dense heatmap representation $\mathbf{y}_h$, as illustrated in Fig. 7.

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

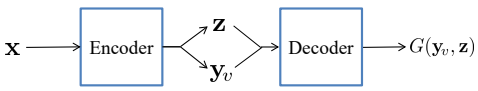


Fig. 8 Semi-DGPose reconstruction at test time.

Indirect Pose-transfer. Our method allows us to do *indirect pose-transfer* without specific training for such a task. As illustrated in Fig. 9, an image $\mathbf{x}_1$ is first passed through the Encoder network, from which the target pose $\mathbf{y}_{v_1}$ is estimated and kept. In the second step, another image $\mathbf{x}_2$ is propagated through the Encoder, from which the appearance encoding $\mathbf{z}_2$ is kept. Finally, $\mathbf{z}_2$ and $\mathbf{y}_{v_1}$ are jointly propagated through the Decoder, and an image $\mathbf{x}_3$ is reconstructed, containing a person in the pose $\mathbf{y}_{v_1}$ 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 the prior art, our network neither relies on any external pose estimator nor on conditioning labels to perform pose-transfer.

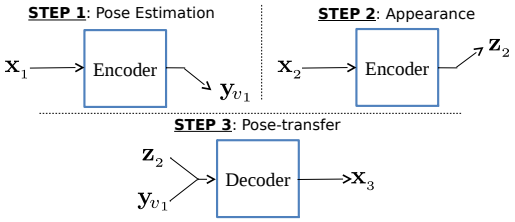


Fig. 9 Semi-DGPose indirect pose-transfer at test time.

Sampling. When no image is given as input, we can jointly or separately sample pose $\mathbf{y}_v$ 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, as illustrated in Fig. 10.

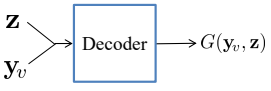


Fig. 10 Semi-DGPose sampling at test time.

Pose Estimation. One of the main differences between our approach and the prior art is the ability of our model to estimate human-body pose as well. In this case, as illustrated in Fig. 11, given an input image $\mathbf{x}$, it is possible to perform pose estimation by regressing to the pose representation vector $\mathbf{y}_v$. Thus, the appearance encoding $\mathbf{z}$ is disregarded, and the Decoder, Mapper, and Discriminator networks are not used.

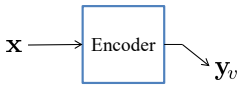


Fig. 11 Semi-DGPose pose estimation at test time.

5 Experiments and Results

We have performed a large number of experiments to evaluate our models. 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 Conditional-DGPose and Semi-DGPose architectures.

5.1 Human3.6M Dataset

Human3.6M [37] 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 video acquisition (50Hz), 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 testing, with $317,989$ and $1,280$ images, respectively. All the original images have a resolution of 1000×1000 pixels.

5.2 ChictopiaPlus Dataset

ChictopiaPlus [51] is an extension of the Chictopia dataset [53]. It augments the original per-pixel annotations for body parts with pose annotation [36], 3D shape [55], and facial segmentation. In contrast to the Human3.6M dataset, in which each actor always wears the same outfit, it contains $23,011$ training, $2,913$ validation, and $2,873$ testing images of segmented people (without background) dressed in a great variety of clothes. All the images have an original resolution of 286×286 pixels.

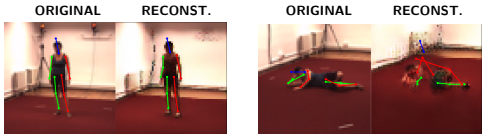
5.3 DeepFashion Dataset

The DeepFashion dataset (In-shop Clothes Retrieval Benchmark) [54] consists of 52,712 images of people in a variety of clothing and poses. We follow Ma et al. [56], using their joints' annotations obtained with an off-the-shelf pose estimator [13], and divide the dataset into training (44,950 images) and testing (6,560 images) subsets. Images with wrong pose estimations were suppressed and all original images have 256×256 pixels. Importantly, we aim to learn a complete generative model of people in images, which is significantly more complex, compared to models focusing on a particular task, such as pose-transfer. For this reason, we use images individually in our training set, instead of employing pairs of images of the same person as in [56, 83].

5.4 Metrics

Quantitative evaluation of generative models is inherently difficult [89]. Since our models explicitly represent *appearance* and *body pose* as separate variables, we evaluate their performance w.r.t. three different aspects. i) **Image quality** of reconstructions is evaluated using the standard Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics [101]. ii) **Accuracy of the reconstructed poses** is evaluated using a protocol introduced by us as follow. To set a common ground for comparing an original test set, with a reconstructed one, we start using a well-established (discriminative) human pose estimator [68], and initially estimating all 2D poses in the original test set. In our protocol, we assume that such estimations are the *ground-truth* poses of the test set. Subsequently, we apply the same discriminative estimator over the reconstructed test images, produced by the trained generative models. Finally, we use of the Percentage of Correct Keypoints (PCK) metric [105], which computes the percentage of 2D joints correctly located by a pose estimator, given the *ground-truth* and a normalised distance threshold corresponding to the size of the person's torso. Thus, we assume that any degradation in

the PCK metric is caused by imperfections on the reconstructed images, since a PCK score of 100% would correspond to having all the estimated joints, in the original and the reconstructed images, at the same locations, up to the distance threshold. We illustrate this metric in Fig. 12. iii) **Accuracy of pose estimation**, obtained by the Semi-DGPose model, is measured using the PCK metric with *real* 2D annotated labels as ground-truths.



(a) (b)

Fig. 12 Accuracy of the reconstructed poses. Samples illustrating best and worst pose reconstructions on the Human3.6M dataset. Each pair of images shows the pose estimation over the original image (left) and the reconstructed image (right). Lines connect the estimated joints for visualisation purposes. Right limbs, left limbs, and head are shown, respectively, by green, red and blue lines. (a) It illustrates the best reconstructed poses, with PCK@0.5 = 1.00. (b) It illustrates the worst reconstructed poses, with PCK@0.5 = 0.00. All images are 64×64 pixels.

5.5 Training

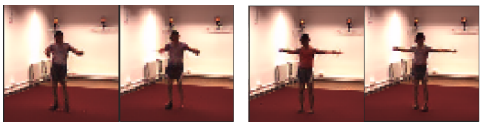
All models were trained with mini-batches consisting of 64 images. We used the Adam optimiser [45] with an initial learning rate set to 10^{-4}. The weight decay regulariser was set to 5×10^{-4}. Network weights were initialised randomly for fully-connected layers and with robust initialisation [32] for convolutional and transposed-convolutional layers. Except when stated differently, for all images and all models, we used a 64×64 pixels 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 [40], and all experiments ran on an NVIDIA Titan X GPU.

5.6 Conditional-DGPose

As mentioned earlier (Sec. 4), the Conditional-DGPose is taken by us as an intermediate step in the investigation towards our Semi-DGPose model. To better evaluate and understand its capabilities, we start our experiments by validating it qualitatively with the Human3.6M benchmark, since this dataset is composed of images in a controlled environment. Initially, in Sec. 5.6.1, we evaluate different pose representations, with the best performance presented by the heatmap representation. In Sec. 5.6.2, we show the effectiveness of the Conditional-DPGose architecture, illustrating *reconstruction* and *sampling* tasks. Besides that, we particularly stress the effects of pose manipulation, by performing *pose-transfer* and *hallucinating* multiple people in a variety of unseen or even unrealistic poses, still on the Human3.6M dataset. After that, we present qualitative and quantitative results on the ChictopiaPlus dataset [51]. The Conditional-DGPose outperforms the closest related comparable baseline, the ClothNetBody [51], achieving state-of-the-art results on the ChictopiaPlus. Finally, qualitative and quantitative experiments on the DeepFashion dataset [54] are shown. On this dataset, our baseline is the image-to-image translation architecture by Ma et al. [56], which is trained on pairs of images showing the same person in different poses. Although our Conditional-DGPose method tackles a significantly more complex problem, i.e. learning a generative model and its latent representation in the high-dimensional image space, instead of mapping one image to another, it presents reasonable results in comparison with the ones from [56].

5.6.1 Pose Representation

We perform experiments with the two pose representations mentioned in Sec. 4.1 and with their respective extensions. We executed end-to-end training with the Conditional-DGPose architecture, which converged in approximately 15 epochs. The qualitative evaluation was performed by the inspection of the reconstructed images, shown in Fig. 13. As can be observed, the vector representations, even the extended one, fail to capture some parts of the body. This problem is particularly evident concerning the extremities of the limbs. On the



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

Fig. 13 Reconstructed images, obtained with each one of the four representations of human pose evaluated: (a) 2D vector, (b) 2D vector extended, (c) heatmaps and (d) heatmaps extended. We highlight the difficult for capturing the spatial extent of some body parts, particularly extremities far from the torso, when the vector representations are adopted. In this example, the use of joints' heatmaps is already sufficient to improve the reconstruction. However, the extended version (with rigid parts and body) turns the model more robust to more complex poses, since the 14 joints are fairly sparse.

other hand, the additional heatmaps for rigid parts and whole body have shown a positive impact in the reconstructions. The quantitative measurements, shown in Tab. 1, support our qualitative evaluation. In all experiments, the heatmaps had the same dimension of the images (64×64).

Pose representation	L1-Norm
2D vector (14 joints)	14.52
2D vector extended (28 joints)	13.91
Heatmaps (14 joints)	13.55
Heatmaps extended	
⌞ (14 joints + 9 rigid parts + 1 whole body)	**13.41**

Table 1 Average reconstruction errors obtained with the Conditional-DGPose architecture using L1-norm for our validation set.

5.6.2 Conditional-DGPose Results on Human3.6M

Initially, in Fig. 14, we show our heatmap pose representation along with reconstructions, to demonstrate that realistic images with accurate poses can be generated. Furthermore, we illustrate *sampling* in Fig. 15, in which the separation between pose and appearance is made evident by the independent change of each variable.

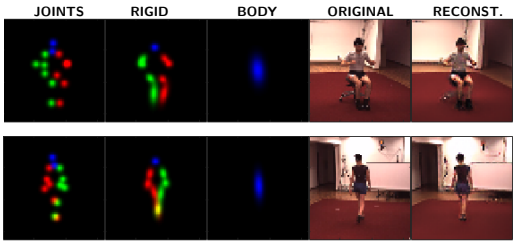


Fig. 14 Reconstructions on Human3.6M. From the left to right columns we have: joints, rigid parts and body heatmaps; original image and finally, the reconstructed image. In the heatmaps, right parts are shown in green, left parts in red and central parts in blue. Human3.6M images are 64×64 pixels.

Next, we stress the pose-transfer and compositionality capabilities of the model, pushing it beyond what is usually done in related methods. Regarding *pose-transfer*, we demonstrate the capability of our model to learn pose and appearance as separate variables which allows direct control over the two at test time. To this end, we generate images in which we maintain the appearance of the input image, yet the generated person is "moved" into the required target pose. The target pose may be composed manually, extracted from another image with an off-the-shelf pose estimator or

Fig. 15 Sampling on Human3.6M. Results obtained by randomly changing pose and appearance independently.

provided interactively by a user. This is illustrated in Fig. 16, in which we employ target poses from the LSP dataset [41], that have completely different poses in a drastically different environment compared to our training set. The quality of the generations shows that our generative model could disentangle pose and appearance and generate images with poses that do not exist in the training data.

Concerning manipulation, we show in Fig. 17 how our model can be used to "compose" images that have never been seen in the training data. For instance, we can generate images with multiple people in the same (replicated) pose simply by conditioning on a respective heatmap. In fact, we can go one step further and generate an image where all people are in the same pose, but one of them is, e.g. *shorter* and another *thinner*, as shown in Fig. 18a. In an extreme case, we can even generate "unreal" images containing only certain body parts (e.g. heads) or disconnecting them from the rest of the body, as in Figs. 18b and 18c, respectively. Note that the training dataset is composed of only single person images. Thus the model has never seen an image with multiple people or only some separate body parts. This demonstrates that the learned latent space of our model is indeed disentangled. To the best of our knowledge, this capability has not been demonstrated by any other work in the literature.

5.6.3 Conditional-DGPose Results on ChictopiaPlus

We compare our method with Lassner et al. [51], the closest related work from the literature. We employ the PSNR and the SSIM metrics to evaluate image quality, and the PCK metric to provide a quantitative evaluation of pose reconstructions, as described previously (see Sec. 5.4). In Tab. 2, we initially show that our method outperforms the ClothNet-body net-

Fig. 16 Cross-domain pose-transfer on Human3.6M. Here we illustrate the *pose-transfer* capability of our Conditional-DGPose. On the leftmost column, we show test images from the LSP dataset [41], along with their corresponding ground-truth 2D pose annotations, composed of 14 joints. These are taken as conditioners (*target-poses*) on our model for the generation of the reconstructions, shown from the third to the rightmost column. As can be observed, the *target-poses* are transferred to the output images, while the latter maintain their original appearances. We highlight the fact that neither the LSP images nor their poses were part of the training set.

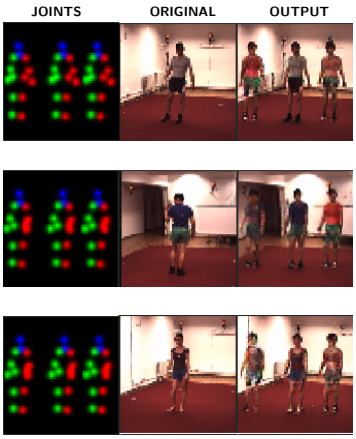


Fig. 17 Hallucinating multiple people on Human3.6M. The Conditional-DGPose model was trained with images containing only one person. The output images are generated keeping the appearance of the original images but conditioned to the manipulated heatmap pose representation (left). Heatmaps of rigid parts and whole body are not shown for simplicity.

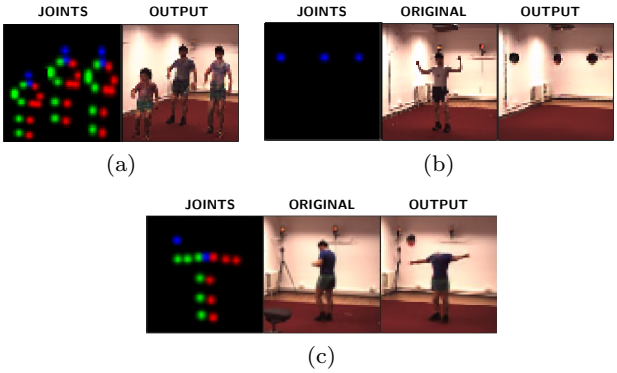


Fig. 18 Generating "unreal" images on Human3.6M. We illustrate the versatility of the model extrapolating the generation of images to unseen scenes. (**a**) Sampled image in which the pose representation in the centre was manually translated and scaled, producing two additional bodies: one shorter and chunkier (*left*) and one taller and thinner (*right*). (**b**) Reconstructed image in which all the body parts were suppressed, except the head. (**c**) Pose-transfer in which the position of the head was manually changed, disconnecting it from the rest of the body. Heatmaps of rigid parts and whole body are not shown for simplicity.

work [51] regarding both, the PSNR and the SSIM metrics. Moreover, our model reports 95.14% of accuracy, with PCK score at 0.5, and again outperforms [51] by a large margin, which reports 70.89%. The overall PCK curve is shown in Fig. 20. Finally, qualitative results are shown in Fig. 19. Our results demonstrate the good quality of our reconstructions w.r.t. image quality and the human pose. The better performance, in comparison with [51], can be particularly noticed in the extremities of body limbs, which we hypothesise as a benefit of the single stage end-to-end Conditional-DGPose model,

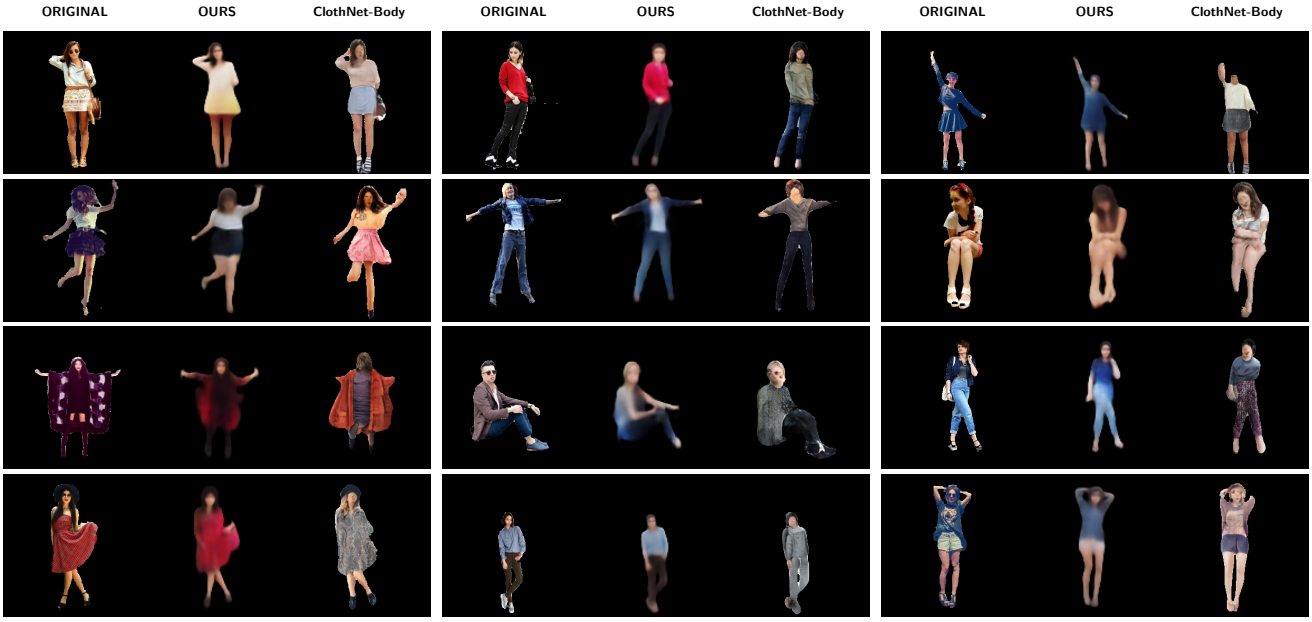


Fig. 19 Reconstructions on ChictopiaPlus. In each trio of images we have, respectively: original image (256 × 256), Conditional-DGPose and ClothNet-body [51] reconstructions. Notice that the images generated by our model are much closer to the originals in terms of appearance (colours). Moreover, in general, the Conditional-DGPose captures the body parts' locations more accurately, resulting in better pose reconstructions (see Fig. 20). Best viewed if zoomed in digital version.

in contrast to the multiple stages of training and testing in [51].

	PSNR	SSIM
Conditional-DGPose	**21.33**	**0.88**
ClothNet-body [51]	16.89	0.82

Table 2 Image Quality on ChictopiaPlus. Quantitative evaluation w.r.t. image quality, showing that our method outperforms [51] considering both metrics, the PSNR and the SSIM.

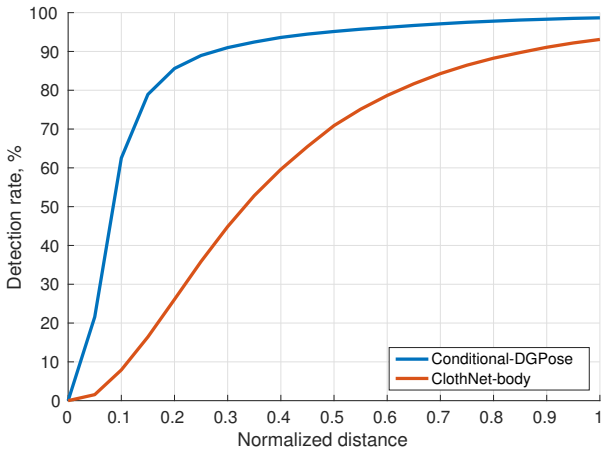


Fig. 20 Accuracy of Poses on ChictopiaPlus. The PCK scores over reconstructed images of our Conditional-DGPose (*blue*) significantly outperforms the ClothNet-body [51] (*red*). Detection rate represents the percentage of joints correctly relocated in the reconstructions.

5.6.4 Conditional-DGPose Results on DeepFashion

Here we show qualitative and quantitative experiments on the DeepFashion dataset [54]. The baseline on this dataset is the image-to-image pose guided generation (PG^2) by Ma et al. [56]. Thus, we use their same training and test sets. However, as our model is not an image-to-image translation architecture, we do not use pairs of images for training. Instead, we use individually 44,950 training images and 6,560 test images.

Again, we employ the PSNR and the SSIM metrics to evaluate image quality, and the PCK metric to provide a quantitative evaluation of pose reconstructions, as described previously (see Sec.5.4). In Table 3, we initially show that even not being trained on images pairs and tackling the significantly more complex task of learning a generative model, instead of executing image-to-image translation, our method achieves scores only slightly below the ones by the PG^2 network on image reconstruction. A similar observation can be done regarding pose reconstruction, since our model reports 74.94% of accuracy, with PCK score at 0.5, against 78.27% from Ma et al. [56]. The overall PCK curve is shown in Fig. 23.

Concretely, the learning of a full generative model, instead of image-to-image translation, allows for the execution of tasks, such as sampling from the learned

Fig. 21 Conditional-DGPose Appearance Manifold. Illustration of the appearance manifold learned on the DeepFashion dataset. We smoothly traverse the manifold for a given pose, causing changes in the visual appearance of the person in the image. No image is used as input, only our heatmap pose representation, evidencing that a truly generative model of images was learned, in which pose and appearance are disentangled. Best viewed if zoomed in digital version.

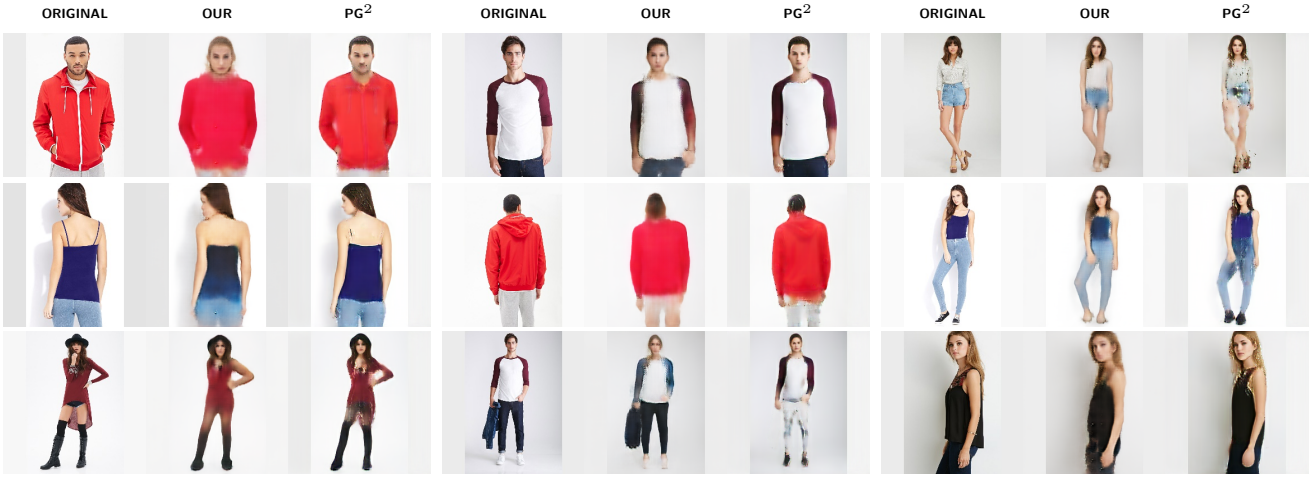


Fig. 22 Reconstructions on DeepFashion . In each trio of images, we have, respectively: original image, Conditional-DGPose and PG^2 [56] reconstructions. All images have 256 × 256 pixels. Although tackling a more complex task than [56], our results are still reasonable. Best viewed if zoomed in digital version.

latent space, which are just not feasible with architectures purely trained on image pairs. To illustrate this, in Fig. 21 we traverse the appearance manifold learned on the DeepFashion dataset. Using only our heatmap pose representation as input, for a given pose, we smoothly vary the values of the latent appearance representation, generating samples with different visual aspect for the same body posture. Such kind of direct sampling is not feasible with the PG^2 [56] architecture.

Finally, the Conditional-DGPose performs 3.06% and 4.82% worse than the PG^2 [56] regarding, respectively, the PSNR and the SSIM metrics (see Table 3). Despite that, it produces reasonable results in comparison with the ones from [56]. A qualitative evaluation is shown in Fig. 22.

	PSNR	SSIM
Conditional-DGPose	18.38	0.79
PG^2 [56]	**18.96**	**0.83**

Table 3 Image Quality on DeepFashion. Quantitative evaluation w.r.t. image quality, showing that our method presents a performance only slightly below the baseline [56], considering both metrics, the PSNR and the SSIM, despite the fact it tackles a significantly more complex task than image-to-image translation.

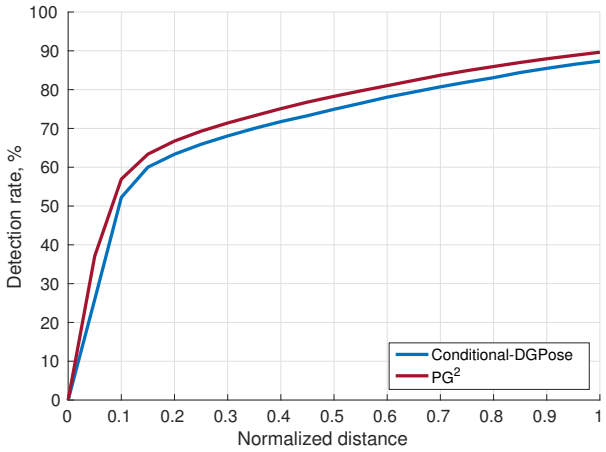


Fig. 23 Accuracy of Poses on DeepFashion. The PCK scores over reconstructed images of our Conditional-DGPose (*blue*) performs only slightly below the PG^2 network [56] (*red*), despite the fact it is tackling a significantly more complex problem than image-to-image translation. Detection rate represents the percentage of joints correctly relocated in the reconstructions.

5.7 Semi-DGPose

Here, we initially evaluate our Semi-DGPose model on the Human3.6M [37] dataset. The Human3.6M is more suitable than both, the ChictopiaPlus and the DeepFashion, for pose estimation evaluations, since the former has joints' annotations obtained by an accurate motion capture system. While the two other datasets are augmented with 2D pose labels obtained using an *off-the-shelf* pose estimator, consequently resulting in more errors in the ground-truth annotations. We show quantitative and qualitative results, focusing particularly on the pose estimation and the *indirect pose-transfer* capabilities, described later in this section. Our experiments and results show the effectiveness of the Semi-DGPose method on the Human3.6M.

To show the generality of the model, we present additional results on the DeepFashion dataset. We now use our Conditional-DGPose architecture and the image-to-image translation network PG^2 [56] as baselines, despite to their relevant differences with the Semi-DGPose. However, to our knowledge, there are no closer related methods in the literature, i.e. that simultaneously pursue the *understanding* and the *generation* of people directly in the image space. Since our Conditional-DGPose method outperforms the ClothNet-body [51] architecture, we do not carry out a direct comparison with the latter.

5.7.1 Semi-DGPose 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.

With full supervision, we first cross-validated the hyper-parameter α which weights the regression loss (see Eq. 5, in Sec. 3) and found that $\alpha = 100$ yields the best results, as shown in Fig. 24a. Following [84], we keep $\gamma = 1$ in all experiments (see Eq. 3, in Sec. 3). In Fig. 24b, we show reconstructed images along with the heatmap pose representation, which are realistic and comparable with the ones obtained with the Conditional-DGPose (see Fig. 14). *Direct manipulation*, when pose representation is changed during the reconstruction process while appearance

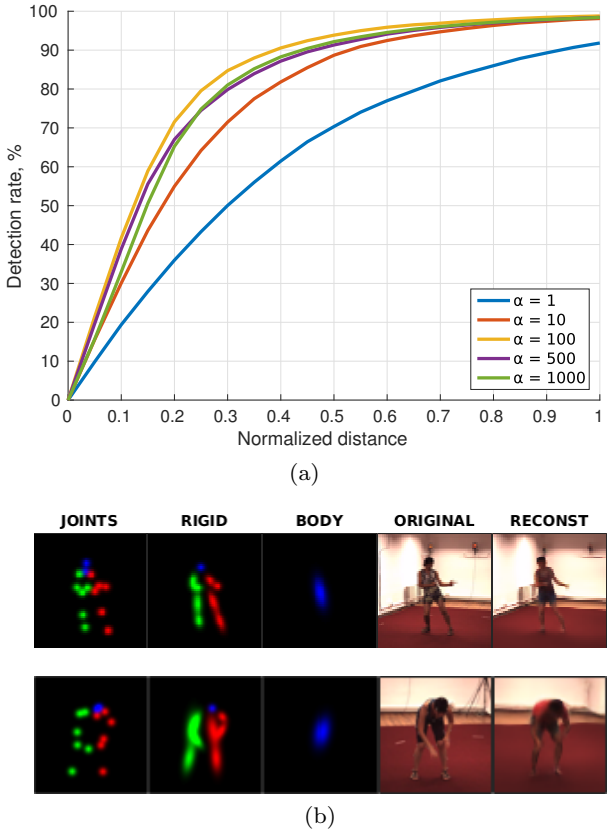


Fig. 24 (a) PCK scores for the cross-validation adjustment of the regression loss weight α. (b) Qualitative reconstructions with full supervision.

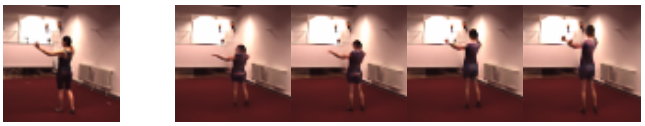
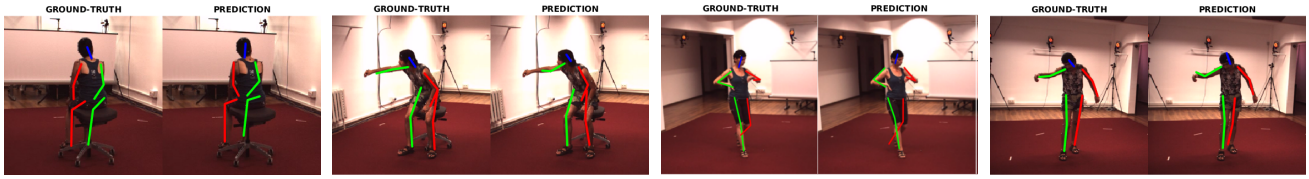


Fig. 25 Direct manipulation. Original image (**a**), followed by reconstructions in which the person's height was changed to a percentage of the original, as: (**b**) 80%, (**c**) 95%, (**d**) 105% and (**e**) 120%. The same procedure may be applied to produce different changes in the body size and aspect ratio.

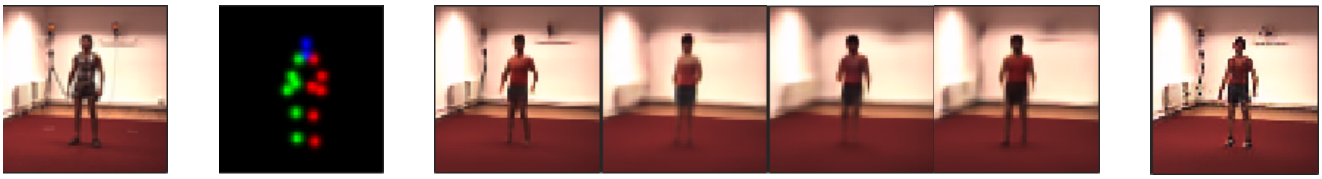
is kept the same, is illustrated in Fig. 25. Still with full supervision, we show the pose estimation accuracy for different samples in Fig. 26. The Semi-DGPose achieves 93.85% PCK score, normalised at 0.5, in the fully-supervised setup (see Fig. 28). This pose estimation accuracy is on par with the state-of-the-art pose estimators on unconstrained images [104]. However, since the Human3.6M was captured in a controlled environment, a standard (discriminative) pose estimator is expected to perform better.

Subsequently, we evaluate it across different levels of supervision, with the PSNR and SSIM metrics and show results in Tab. 4. In Fig. 27, we show reconstructed



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

Fig. 26 Pose Estimation on Human3.6M. Pairs of ground-truth and predicted joints superimposed on the original images. Below each pair, we show the PCK score normalised at 0.5 times the torso size, as usual for the PCK metric. Such normalised distance explains the high scores despite the existence of minor differences between ground-truth and predicted positions. Results were obtained with 100% of supervision during training, and each pair correspond to one of the 4 cameras from the Human3.6M dataset.



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

Fig. 27 Reconstructions on Human3.6M. (**a**) Original image. (**b**) Heatmap pose representation (rigid parts and body suppressed in the illustration for simplicity), followed by reconstructions with different levels of supervision: (**c**) 100%, (**d**) 75%, (**e**) 50%, (**f**) 25%, and (**g**) Conditional-DGPose.

images obtained with such different levels. It allows us to observe how image quality is affected when we gradually reduce the availability of labels. Furthermore, we also evaluate the pose estimation accuracy with semi-supervision. The overall PCK curves corresponding to each percentage of supervision in the training set is shown in Fig. 28. Note that, even with only 25% of labels available, our model still obtains 88.35% PCK score, normalised at 0.5, showing the effectiveness of the semi-supervised approach. Qualitative samples are shown in Figure 29. Again, aiming to illustrate how the gradual decrease of supervision in the training set affects the quality of pose estimation on the test images.

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

Table 4 Image Quality on Human3.6M. Quantitative evaluations of the Semi-DGPose with different levels of supervision using the PSNR and SSIM metrics.

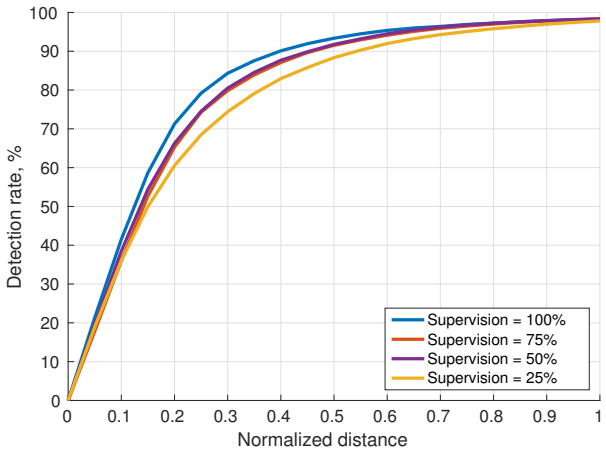
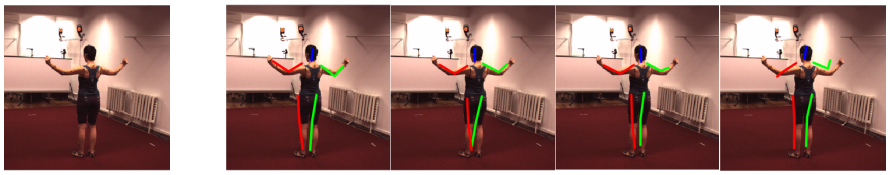


Fig. 28 Accuracy of Poses on Human3.6M. Quantitative evaluations of Semi-DGPose for different levels of supervision using the PCK scores. Note that, even with 25% supervision, our Semi-DGPose obtains 88.35% PCK score, normalised at 0.5.

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. 30. Firstly, the latent pose representation $\mathbf{y}_{v_1}$ is estimated from the first input image through the Encoder. Secondly, the latent *appearance* representation $\mathbf{z}_2$ is estimated from a second image, also through the Encoder. Lastly, $\mathbf{y}_{v_1}$ and $\mathbf{z}_2$ are propagated through the Decoder, and a new image is generated, combining body pose and appearance, respectively, from the first and second



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

Fig. 29 Qualitative results of semi-supervised pose estimation. Original image (**a**), followed by predictions, over the original image, with: (**b**) 100%, (**c**) 75%, (**d**) 50% and (**e**) 25% of supervision. The figure aims to illustrate how the decrease in supervision affects pose estimation. The results are similar, yet it is possible to observe some important discrepancies. For instance, due to the shortage of labelled training data, the pose estimation result in (**e**) is worse than the one shown in (**b**), particularly regarding the location of arms' extremities.

encoded images. We evaluate qualitatively the effects of semi-supervision over the indirect pose-transfer in Fig. 31.

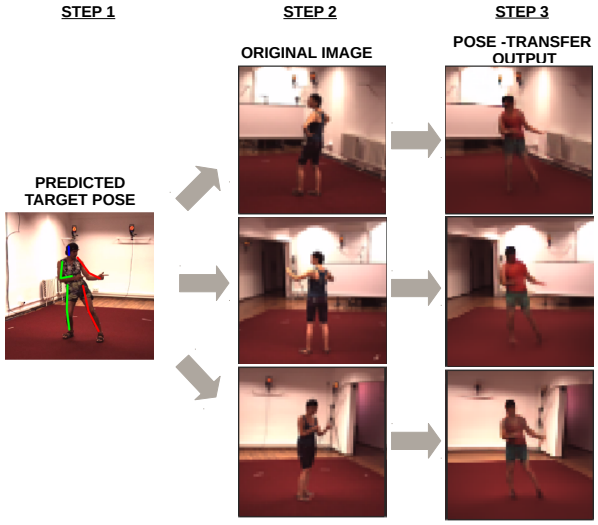


Fig. 30 ***Indirect pose-transfer*** **on Human3.6M. Step 1**: the latent target pose representation $\mathbf{y}_{v_1}$ is estimated (Encoder). **Step 2**: the image from which the latent *appearance* $\mathbf{z}_2$ is estimated (Encoder). **Step 3**: the output image generated as a combination of $\mathbf{y}_{v_1}$ and $\mathbf{z}_2$ (Decoder). The people's outfits in the output images are approximated to the ones in the original images. However, restricted by the low diversity of outfits observed in Human3.6M training data. Note that, to highlight the separation of appearance and pose, we chose the image on Step 1 to be from camera 2, while the original images are from cameras, 1, 3 and 4, respectively. As can be seen, the background scene is totally defined by the original images.

5.7.2 Semi-DGPose Results on DeepFashion

To show the generality of the Semi-DGPose, model we present additional results on the DeepFashion dataset, using our Conditional-DGPose architecture and the image-to-image translation network PG^2 [56] as baselines. The same hyper-parameters reported previously were used in training. In Tab. 5, we compare the image quality of reconstructions, while in Fig. 32 we show the comparison concerning the quality of pose reconstructions. Although the Semi-DGPose presents less accurate results, it is important to highlight that it is also tackling the pose estimation task, which is not performed by either one of the other two methods, i.e. the Conditional-DGPose and the PG^2. To pursue, simultaneously, the *understanding*, i.e. estimation of pose and appearance in the latent space, and the *generation* of people directly in images, shows to be indeed a significantly more complex task. Nevertheless, the justification for seeking such a challenging goal, as mentioned before, mainly lie on its important capability of allowing for semi-supervised learning, that is not present in the comparable methods.

	PSNR	SSIM
Semi-DGPose	16.84	0.76
Conditional-DGPose	18.38	0.79
PG^2 [56]	**18.96**	**0.83**

Table 5 Image Quality on DeepFashion. Quantitative evaluation of Semi-DGPose using PSNR and SSIM measures comparing the image quality of reconstructions. The Semi-DGPose shows less accurate results, yet in contrast to the other methods, it performs a significantly more complex task, simultaneously executing pose estimation, and also allowing for semi-supervised learning.

In Fig. 33, we show comparisons between input and reconstructed images. In some of the samples, we can observe small differences between the original and the reconstructed body postures, mainly regarding the positions of the limbs. This illustrates the higher complexity involved in simultaneously estimating pose and appearance in our latent space. For instance, inaccurate predictions of pose, performed by the Encoder, may have effects into the final reconstructed appearance, and vice-versa, when the latent representations are mapped back to the image space, by the Decoder. Such interdependency does not exist when pose is a

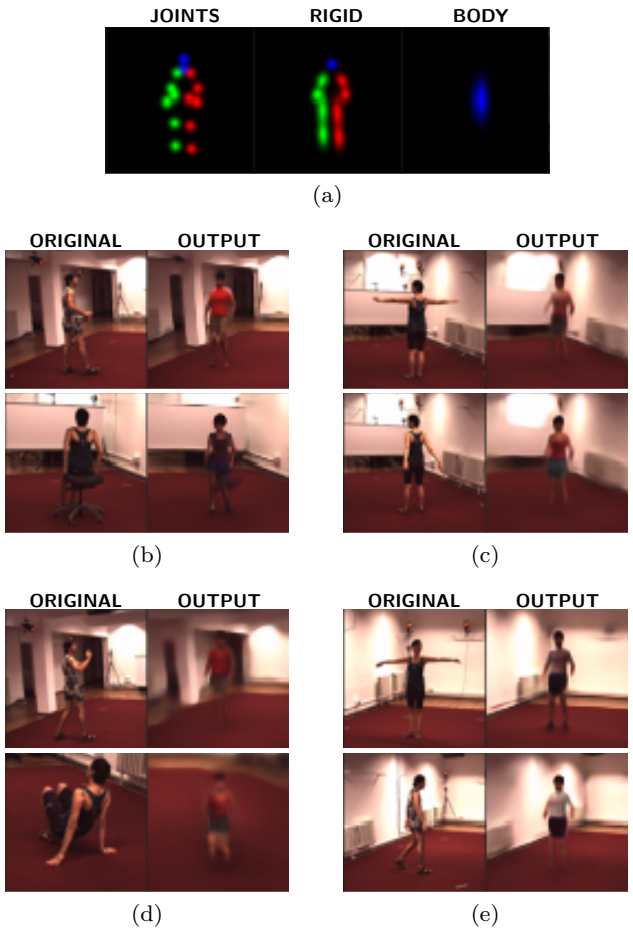


Fig. 31 Indirect Pose-transfer on Human3.6M. Qualitative results with different levels of supervision. (a) Heatmap representation of the target pose (i.e. after being processed by the Mapper module) used for all the subsequent results. Such results show pairs of original images and pose-transfer outputs obtained with the following levels of supervision: (b) 100%, (c) 75%, (d) 50%, and (e) 25%. In the pose-transfer outputs, appearance comes from the original images while the body posture is defined by the target pose.

given observable variable, as in the case of the conditional models or image-to-image translation networks.

Finally, we highlight *indirect pose-transfer* in the DeepFashion dataset, which is a distinctive capability of the Semi-DGPose, in comparison to related methods. In Fig. 34, we compare the indirect pose-transfer results, from our single-stage structured generative model, the Semi-DGPose, with the results from the image-to-image translation baseline, the PG^2 network [56]. It is important to notice that our Semi-DGPose model was not trained specifically for pose-transfer, i.e. it was not trained on pairs of images. On the other hand, the PG^2 architecture is trained on pairs of images of the same person, in different poses, scales or point of views (first two images of each set in Fig. 34). Moreover, in the Semi-DGPose the body pose is estimated by the Encoder network (illustrated in every second image of each

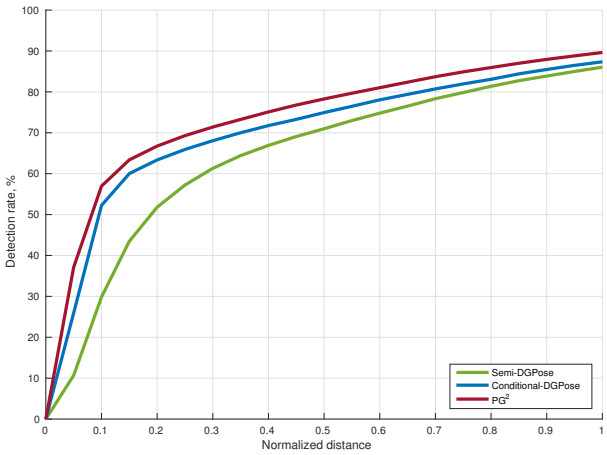


Fig. 32 Accuracy of Poses on DeepFashion. Quantitative evaluation of Semi-DGPose PCK scores over reconstructed poses. The Semi-DGPose (*green*) shows less accurate results, however, in contrast to the Conditional-DGPose (*blue*) and the PG^2 network [56] (*red*), it performs a significantly more complex task, simultaneously executing pose estimation and allowing for semi-supervised learning. Detection rate represents the percentage of joints correctly relocated in the reconstructions.

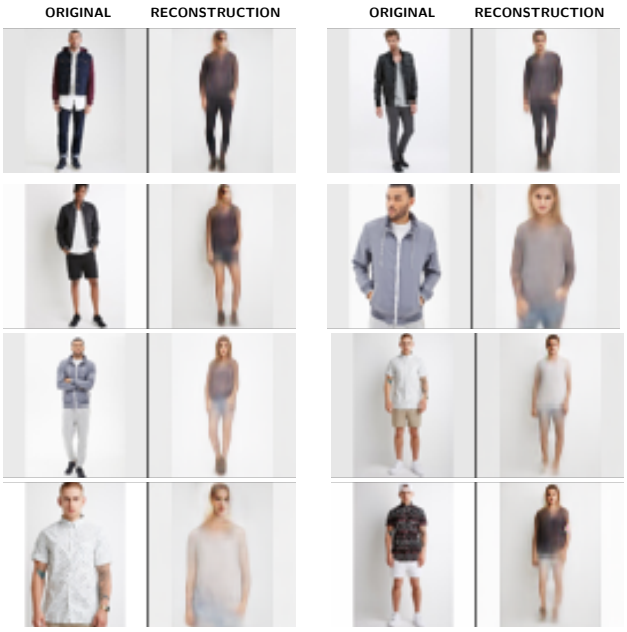


Fig. 33 Reconstructions on DeepFashion. The only input of the Semi-DGPose is the original image. At test time, as pose is estimated in the latent space, discrepancies between the original and reconstructed poses are more frequently observed, in comparison with the Conditional-DGPose. Best viewed if zoomed in digital version.

set in Fig. 34), along with appearance, while in the PG^2 pose is given as an observable variable to the model. Despite such critical competitive disadvantages, we can observe that the Semi-DGPose produce reasonable results in comparison to the ones from PG^2. Lastly, it is crucial to call attention for the capabilities of our Semi-

DGPose approach such as, interpretability of the latent space, pose estimation, sampling and semi-supervised learning, which are not jointly present in the PG^2 or in the related work from the literature. These features justify our approach for learning a deep generative model of people in images and, to our knowledge, significantly differentiate the Semi-DGPose model from prior art.

5.8 Limitations of the Models

Here, we discuss two important limitations common to the Conditional-DGPose and the Semi-DGPose. The first refers to the modelling of appearance in both models. As we mention in Sec. 1, our latent representation of appearance encodes all the visual information in the images (e.g. clothing, skin colours, hairstyles, and background) except for the body pose of the subjects. However, such a strategy does not separate the individual visual characteristics in the latent representation. In Fig. 21 (Sec. 5.6.4), we can observe that as the appearance manifold is traversed, the visual features gradually change altogether. A disentangled representation for appearance itself would be needed for allowing control over specific visual features. Another aspect concerning appearance regards limitations to approximate clothing "seen" few times or "unseen" during training. Interestingly, the extrapolation capabilities shown for unseen poses (see Fig. 18 in Sec. 5.6.2) is not observed for appearance. For example, in the Human3.6M dataset, the low diversity of subjects outfits may eventually prevent the clothing in the reconstructed images to be precisely equal to the ones in the original images, as can be observed in Fig. 30 (Sec. 5.7.1). Other works in the literature refer to this same problem concerning the Human3.6M dataset, e.g. Rhodin et al. [73].

The second relevant limitation refers to our pose representation. Aiming to investigate and explore the capabilities of simple body representations, we have worked only with 2D pose in our models. Such option turns our approaches more general since they are not dependent on 3D information (e.g. 3D models, camera calibration, or multi-view images). It allows, for example, their application on ordinary monocular images. Moreover, this strategy is also less susceptible to body shape variations in comparison to segmentations mask or 3D body meshes, which might not be directly transferable from one person to another. However, such simplicity creates some limitations. An important one concerns the lack of depth information in the body model. Despite the reasonable results obtained with single people in relatively "well-behaved" poses, the models might face difficulties in the presence of stronger self-occlusions associated with particular body postures. In the absence of depth, it is hard to infer, for instance, which one of two overlapping limbs is closer to the camera. Without such explicit information in the body representation, the correct reconstruction might present flaws.

To analyse such issues here, which are present in our both models, we have employed the Conditional-DGPose, trained on the Human3.6M dataset, to perform cross-domain pose-transfer over single images from short video sequences. Employing a sequence of frames allow us to observe how the performance of the model changes according to the concurrent presence of self-occlusion and different poses. In the current experiments, we have used short videos from the JHMDB dataset [39]. Each "in-the-wild" video depicts a single person performing one activity. The dataset provides 2D pose annotations per frame for all videos. Such annotations are used as inputs for the Conditional-DGPose cross-domain pose-transfer. We crop the images maintaining the subjects centralised.

In Fig. 35a, a sequence of frames shows a boy batting a ball while playing baseball (top row) and the correspondent pose-transfer outputs (bottom row). Although the reenacted frames present the gist of the original sequence, already it is possible to observe that overlapping arms and legs appear to be blended in some of the output images (e.g. frames 1 and 5), making evident the problem we have mentioned earlier. Fig. 35b (top row) depicts a football player kicking a ball towards the goal. We call attention for frame 5, in which the self-occluded arm of the original subject turns the upper body of the reconstructed person wider. In frame 9, the concurrent overlapping legs and the unusual pose contribute for an ambiguous posture of the person in the output image, which might be facing forwards or backwards. The particular body pose in frame 25 provokes the misalignment of head, torso and arms in of the body in the output. Finally, even without a task-specific training, we believe that the use of a 3D body representation, which would explicitly encode depth, may be beneficial to mitigate the main issues mentioned above.

6 Conclusions

In this paper, we have presented a comprehensive deep generative model framework for human pose analysis in images. Our models are based on a principled VAEGAN approach and allow the disentanglement of body posture and visual appearance, aiming for the independent manipulation of such factors. With our conditional-VAEGAN model, the Conditional-DGPose, differently from previous art, we have taken such manipulation to extreme cases,

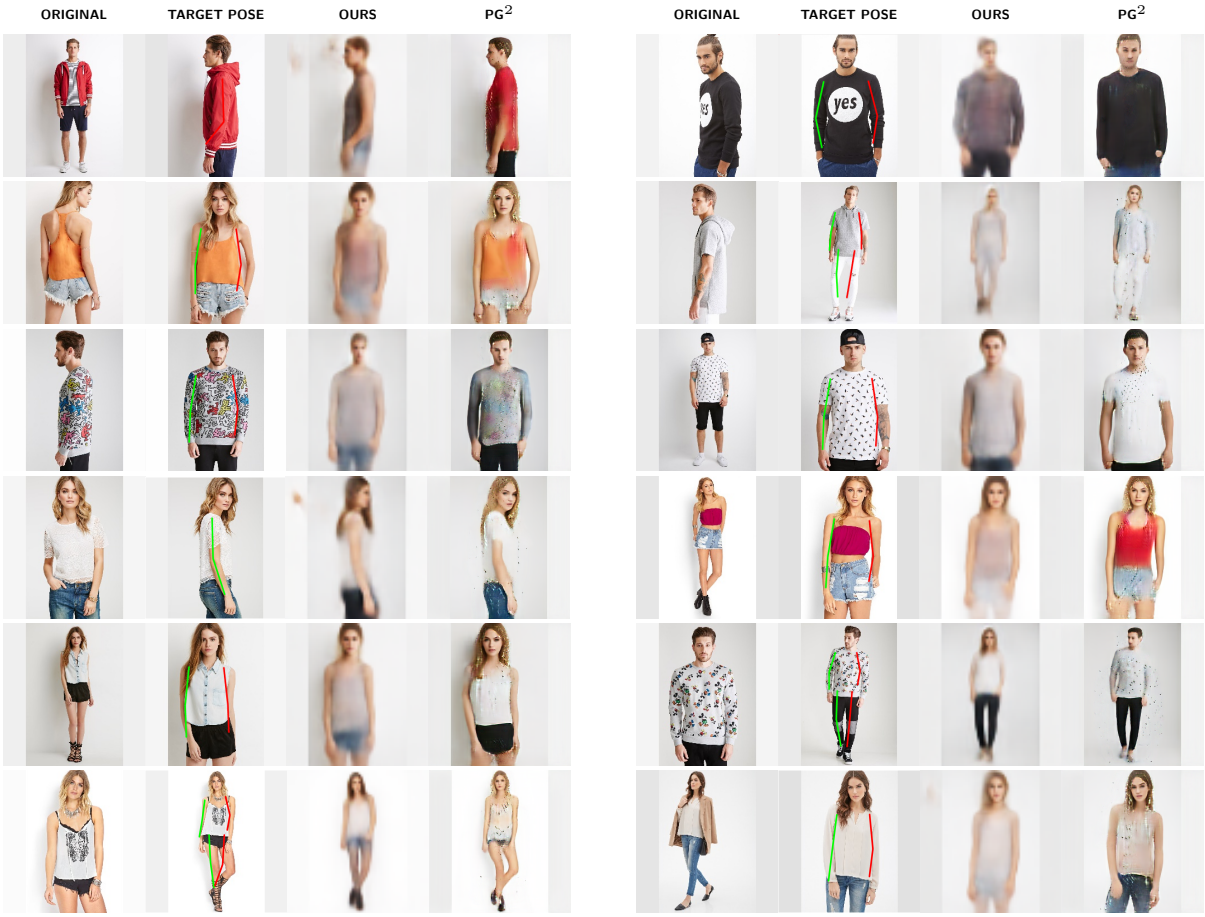


Fig. 34 Indirect pose-transfer in DeepFashion dataset. In each set of images, we have, respectively: the original image, the target image with the superimposed target pose predicted by the Semi-DGPose, the pose-transfer output from the Semi-DGPose and the pose-transfer output from PG^2 [56]. Although tackling a more complex task than [56], which includes the prediction of pose, our results are still reasonable.

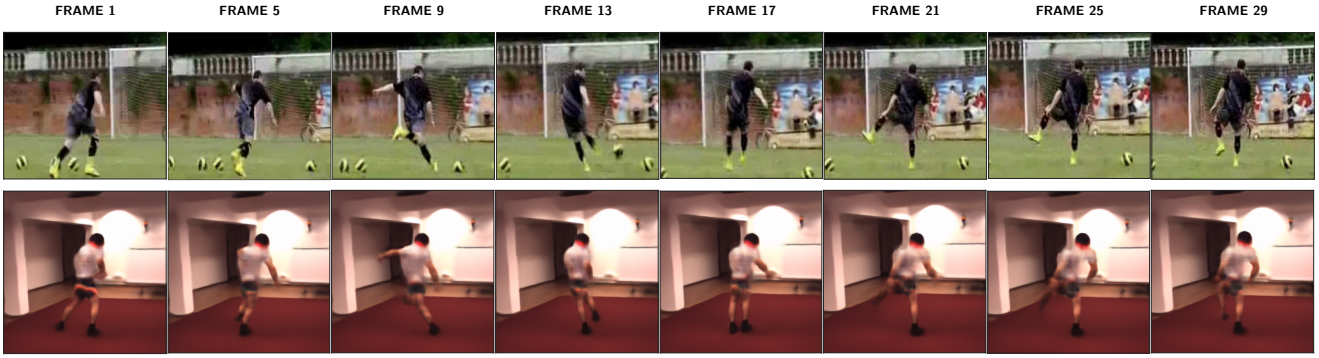
e.g. by performing cross-domain *pose-transfer* and by hallucinating multiple people in a variety of unseen or even unrealistic poses. Moreover, we have achieved state-of-the-art results on image reconstruction conditioned on pose, *outperforming* the closest related comparable baseline [51]. With a single-stage structured semi-supervised VAEGAN architecture, the Semi-DGPose, we pursued the joint *understanding* and *generation* of people in images, not only mapping images to partially interpretable latent representations but also mapping these representations back to the image space. Importantly, such an approach simultaneously allows for reconstruction, direct manipulation, sampling, pose estimation, indirect pose-transfer, and semi-supervised learning. These joint capabilities differentiate the Semi-DGPose from other methods in the literature and demonstrate a real-world application of structured deep generative models with the highly relevant potential of being less dependable of fully-labelled data. We have systematically evaluated our methods on well-known benchmarks, the Human3.6M, the ChictopiaPlus, and the DeepFashion datasets, comparing our results with the closest related baseline methods in the literature [51,56]. Such results and comparisons highlight the novelty and effectiveness of our approaches and its capabilities, despite the significant challenge posed by our aimed goal. We believe that we have shown and reinforced the relevance of employing an interpretable and structured latent space, which allows for semi-supervised learning, as well as the importance of tackling the problem with single-stage end-to-end architectures.

Acknowledgements Rodrigo de Bem is a CAPES Foundation scholarship holder (Process no: 99999.013296/2013-02, Ministry of Education, Brazil).

References

1. 3Lateral: 3Lateral (2018). URL http://www.3lateral.com/
2. Achilles, F., Ichim, A.E., Coskun, H., Tombari, F., Noachtar, S., Navab, N.: Patient mocap: Human pose

(a)



(b)

Fig. 35 Cross-domain pose-transfer over single images from short video sequences from the JHMDB dataset [39]. (a) A sequence of frames shows a boy batting a ball while playing baseball (top row) and the correspondent pose-transfer outputs (bottom row). Mainly due to self-occlusion, some limbs appear blended. (b) A football player is kicking a ball towards the goal (top row) and the correspondent pose-transfer outputs (bottom row). Frames 5, 9, and 25 present important issues due to particular postures and self-occlusion of limbs. Best viewed if zoomed in digital version.

estimation under blanket occlusion for hospital monitoring applications. In: MICCAI (2016)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
4. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: CVPR (2018)
5. de Bem, R., Arnab, A., Sapienza, M., Golodetz, S., Torr, P.: Deep fully-connected part-based models for human pose estimation. In: ACML (2018)
6. de Bem, R., Ghosh, A., Ajanthan, T., Miksik, O., Siddharth, N., Torr, P.: A semi-supervised deep generative model for human body analysis. In: ECCV (HBUGEN) (2018)
7. de Bem, R., Ghosh, A., Ajanthan, T., Siddharth, N., Torr, P.: A conditional deep generative model of people in natural images. In: WACV (2019)
8. Blanz, V., Vetter, T., et al.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH (1999)
9. Boeing: William Fetter's Boeing Man (2018). URL https://secure.boeingimages.com/archive/William-Fetter-s-Boeing-Man-2F3XC5YCZNC.html
10. Bogo, F., Romero, J., Loper, M., Black, M.J.: Faust: Dataset and evaluation for 3d mesh registration. In: CVPR, pp. 3794–3801 (2014)
11. Borshukov, G., Piponi, D., Larsen, O., Lewis, J.P., Tempelaar-lietz, C.: Universal capture - image-based facial animation for "The Matrix Reloaded". In: SIGGRAPH (2005)
12. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: ECCV (2016)
13. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
14. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. arXiv preprint arXiv:1808.07371 (2018)
15. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. CVPR (2017)
16. de La Gorce, M., Fleet, D.J., Paragios, N.: Model-Based 3D Hand Pose Estimation from Monocular Video. TPAMI **33**(9), 1793–1805 (2011)
17. Elgammal, A., Lee, C.S.: Inferring 3d body pose from silhouettes using activity manifold learning. In: CVPR (2004)
18. Enzweiler, M., Gavrila, D.M.: A mixed generative-discriminative framework for pedestrian classification. In: CVPR (2008)
19. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: CVPR (2018)
20. Ezzat, T., Poggio, T.: Facial analysis and synthesis using image-based models. In: FG (1996)
21. Fan, S., Ng, T.T., Koenig, B.L., Herberg, J.S., Jiang, M., Shen, Z., Zhao, Q.: Image Visual Realism: From Human Perception to Machine Computation. TPAMI **40**(9), 2180–2193 (2018)

22. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR, vol. 2 (2005)
23. Fetter, W.A.: A Progression of Human Figures Simulated by Computer Graphics. IEEE Computer Graphics and Applications **2**(9), 9–13 (1982)
24. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. TPAMI **30**(2), 267–282 (2007)
25. Fossati, A., Dimitrijevic, M., Lepetit, V., Fua, P.: Bridging the gap between detection and tracking for 3d monocular video-based motion capture. In: CVPR, pp. 1–8 (2007)
26. Franco, J.S., Boyer, E.: Fusion of multi-view silhouette cues using a space occupancy grid. In: ICCV, pp. 1747–1753 (2005)
27. Geisler, W.S.: Visual perception and the statistical properties of natural scenes. Annu. Rev. Psychol. **59**, 167–192 (2008)
28. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning. MIT press (2016)
29. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
30. Hattori, H., Lee, N., Boddeti, V.N., Beainy, F., Kitani, K.M., Kanade, T.: Synthesizing a scene-specific pedestrian detector and pose estimator for static video surveillance. IJCV **126**(9), 1027–1044 (2018)
31. Hattori, H., Naresh Boddeti, V., Kitani, K.M., Kanade, T.: Learning scene-specific pedestrian detectors without real data. In: CVPR (2015)
32. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV (2015)
33. Hilton, A., Beresford, D., Gentils, T., Smith, R., Sun, W.: Virtual people: capturing human models to populate virtual worlds. In: Proceedings of Computer Animation, pp. 174–185 (1999)
34. Ian Spriggs: (2018). URL http://www.ianspriggs.com/
35. Ichim, A.E., Bouaziz, S., Pauly, M.: Dynamic 3d avatar creation from hand-held video input. TOG **34**(4), 45 (2015)
36. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV (2016)
37. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI (2014)
38. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 (2016)
39. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV (2013)
40. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
41. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC (2010)
42. Kanade, T., Rander, P., Narayanan, P.J.: Virtualized reality: Constructing virtual worlds from real scenes. IEEE Multimedia **4**(1), 34–47 (1997)
43. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018)
44. Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L.: Identifying natural images from human brain activity. Nature **452**(7185), 352 (2008)
45. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015)
46. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: NIPS (2014)
47. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
48. Krizhevsky, A., Hinton, G., et al.: Factored 3-way restricted boltzmann machines for modeling natural images. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (2010)
49. Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.: Deep convolutional inverse graphics network. In: NIPS (2015)
50. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: ICML (2016)
51. Lassner, C., Pons-Moll, G., Gehler, P.V.: A generative model for people in clothing. In: ICCV (2017)
52. Lee, C.S., Elgammal, A.: Facial expression analysis using nonlinear decomposable generative models. In: International Workshop on Analysis and Modeling of Faces and Gestures (2005)
53. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. TPAMI (2015)
54. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016)
55. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM TOG (2015)
56. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. In: NIPS (2017)
57. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation (2018)
58. MacDorman, K.F., Green, R.D., Ho, C.C., Koch, C.T.: Too real for comfort? uncanny responses to computer generated faces. Computers in human behavior **25**(3), 695–710 (2009)
59. Magnenat-Thalmann, N., Thalmann, D.: Virtual humans: Thirty years of research, what next? Visual Computer **21**(12), 997–1015 (2005)
60. Magnenat-Thalmann, N., Thalmann, D.: Handbook of Virtual Humans. John Wiley & Sons (2006)
61. MakeHuman: (2018). URL http://www.makehumancommunity.org/
62. Massiceti, D., Siddharth, N., Dokania, P., Torr, P.H.: FlipDial: A generative model for two-way visual dialogue. In: CVPR (2018)
63. Massive Software: (2018). URL http://www.massivesoftware.com/
64. Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L.: Visual analysis of humans. Springer (2011)
65. Müller, M., Casser, V., Lahoud, J., Smith, N., Ghanem, B.: Sim4cv: A photo-realistic simulator for computer vision applications. International Journal of Computer Vision **126**(9), 902–919 (2018)

66. NASA: Space flight human-system standard volume 1. Tech. Rep. NASA-STD-3001, National Aeronautics and Space Administration - NASA (1995)
67. Neverova, N., Alp Guler, R., Kokkinos, I.: Dense pose transfer. In: ECCV (2018)
68. Newell, A., Yang, K., Deng, J.: Stacked Hourglass Networks for Human Pose Estimation. In: ECCV (2016)
69. Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D.H.: Synthesizing realistic facial expressions from photographs. In: SIGGRAPH, p. 19. ACM (2006)
70. Poser: (2018). URL https://www.posersoftware.com/
71. Pumarola, A., Agudo, A., Sanfeliu, A., Moreno-Noguer, F.: Unsupervised person image synthesis in arbitrary poses. In: CVPR (2018)
72. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. ICML (2014)
73. Rhodin, H., Salzmann, M., Fua, P.: Unsupervised geometry-aware representation for 3d human pose estimation. In: ECCV (2018)
74. Rogez, G., Schmid, C.: Mocap-guided data augmentation for 3d pose estimation in the wild. In: NIPS (2016)
75. Rogez, G., Schmid, C.: Image-based synthesis for deep 3d human pose estimation. International Journal of Computer Vision **126**(9), 993–1008 (2018)
76. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. TOG **36**(6), 245 (2017)
77. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
78. Rosales, R., Athitsos, V., Sigal, L., Sclaroff, S.: 3d hand pose reconstruction using specialized mappings. In: ICCV, vol. 1, pp. 378–385. IEEE (2001)
79. Schulman, J., Heess, N., Weber, T., Abbeel, P.: Gradient estimation using stochastic computation graphs. In: Advances in Neural Information Processing Systems, pp. 3528–3536 (2015)
80. Seemann, E., Nickel, K., Stiefelhagen, R.: Head pose estimation using stereo vision for human-robot interaction. In: FG (2004)
81. Shan, Q., Adams, R., Curless, B., Furukawa, Y., Seitz, S.M.: The Visual Turing Test for Scene Reconstruction. In: 2013 International Conference on 3D Vision, pp. 25–32 (2013)
82. Shotton, J., Fitzgibbon, A.W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011)
83. Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable gans for pose-based human image generation (2018)
84. Siddharth, N., Paige, B., Desmaison, A., van de Meent, J.W., Wood, F., Goodman, N.D., Kohli, P., Torr, P.H.: Learning disentangled representations with semi-supervised deep generative models. In: NIPS (2017)
85. Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annual review of neuroscience **24**(1), 1193–1216 (2001)
86. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: NIPS (2015)
87. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE Computer Graphics and Applications **27**(3), 21–31 (2007)
88. Starck, J., Miller, G., Hilton, A.: Video-based character animation. In: SIGGRAPH, pp. 49–58. ACM (2005)
89. Theis, L., van den Oord, A., Bethge, M.: A note on the evaluation of generative models. In: ICLR (2016)
90. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR (2016)
91. 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)
92. Trumble, M., Gilbert, A., Hilton, A., Collomosse, J.: Deep autoencoder for combined human pose estimation and body model upscaling. In: ECCV18 (2018)
93. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: CVPR, pp. 1526–1535 (2018)
94. Unreal Engine: (2018). URL https://www.unrealengine.com
95. Valenza, E., Simion, F., Cassia, V.M., Umiltà, C.: Face preference at birth. Journal of experimental psychology. Human perception and performance **22**(4), 892–903 (1996)
96. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 109–117 (2017)
97. von Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. Eurographics (2017)
98. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: ICCV (2017)
99. Wang, T.C., Liu, M.Y., Zhu, J.Y., Yakovenko, N., Tao, A., Kautz, J., Catanzaro, B.: Video-to-Video Synthesis. In: Advances in Neural Information Processing Systems 31 (NIPS), pp. 1152–1164 (2018)
100. Wang, Y., Huang, X., Lee, C.S., Zhang, S., Li, Z., Samaras, D., Metaxas, D., Elgammal, A., Huang, P.: High resolution acquisition, learning and transfer of dynamic 3-d facial expressions. In: Computer Graphics Forum, vol. 23, pp. 677–686. Wiley Online Library (2004)
101. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
102. Wang, Z., Merel, J.S., Reed, S.E., de Freitas, N., Wayne, G., Heess, N.: Robust imitation of diverse behaviors. In: NIPS (2017)
103. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)
104. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: ICCV (2017)
105. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)
106. Yuille, A., Kersten, D.: Vision as bayesian inference: analysis by synthesis? Trends in cognitive sciences **10**(7), 301–308 (2006)
107. Zanfir, M., Popa, A.I., Zanfir, A., Sminchisescu, C.: Human appearance transfer. In: CVPR (2018)
108. Zhang, Y., Guo, Y., Jin, Y., Luo, Y., He, Z., Lee, H.: Unsupervised discovery of object landmarks as structural representations. In: CVPR (2018)

A DGPose Architectures Details

Here, we provide implementation details of our both architectures considering the following inputs: images $\mathbf{x}$ (batch_size=64, channels=3, height=64, width=64) and heatmaps $\mathbf{y}_h$ (batch_size=64, channels=24, height=64, width=64). Regarding the heatmap labels, the channels correspond to: **i)** 14 joints (head top, neck, right shoulder, right elbow, right wrist, right hip, right knee, right ankle, left shoulder, left elbow, left wrist, left hip, left knee, left ankle); **ii)** 9 rigid parts (head, right upper arm, right lower arm, right upper leg, right lower leg, left upper arm, left lower arm, left upper leg, left lower leg); **iii)** 1 whole body. Finally, in Tabs. A2 and A3, we show the full definition of both, the Conditional-DGPose and the Semi-DGPose, respectively.

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(CONV-2, *previous_layer_output*)

Table A1 Architecture of the residual block employed in the DGPose encoder.

Encoder	
Input: $\mathbf{x}$(batch_size=64, channels=3, height=64, width=64); $\mathbf{y}_h$(batch_size=64, channels=24, height=64, width=64)	
Layer	**Definition**
1	CONCAT($\mathbf{x}$, $\mathbf{y}_h$)
2	CONV-(N64, K7, S2, P1), LeakyReLU(0.01)
3	CONV-(N128, K3, S2, P1), BN, ReLU
4	CONV-(N256, K3, S2, P1), BN, ReLU
5	CONV-(N512, K3, S2, P1), BN, ReLU
6	CONV-(N512, K3, S2, P1), BN, ReLU
7	CONV-(N512, K3, S2, P1), BN, ReLU
8	RESIDUAL-(N512, K3, S1, P1)
9	RESIDUAL-(N512, K3, S1, P1)
10	RESIDUAL-(N512, K3, S1, P1)
11	RESIDUAL-(N512, K3, S1, P1), SIGMOID
$\mu_{\mathbf{z}}$	FC-(N100)
$\sigma_{\mathbf{z}}$	FC-(N100)
Prior	
Input: $\mathbf{y}_h$(batch_size=64, channels=24, height=64, width=64)	
Layer	**Definition**
1	CONV-(N128, K4, S2, P1), LeakyReLU(0.2)
2	CONV-(N256, K4, S2, P1), BN, LeakyReLU(0.2)
3	CONV-(N512, K4, S2, P1), BN, LeakyReLU(0.2)
4	CONV-(N1024, K4, S2, P1), BN, LeakyReLU(0.2)
5	CONV-(N100, K4, S1, P0), SIGMOID
μ_{prior}	FC-(N100)
σ_{prior}	FC-(N100)
Decoder	
Input: $\mathbf{z}$(batch_size=64, channels=100)	
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(DECONV-6, $\mathbf{y}_h$)
8	CONV-(N512, K5, S1, P2), BN, LeakyReLU(0.2)
9	CONV-(N256, K5, S1, P2), BN, LeakyReLU(0.2)
10	CONV-(N128, K5, S1, P2), BN, LeakyReLU(0.2)
11	CONV-(N128, K5, S1, P2), BN, LeakyReLU(0.2)
G($\mathbf{y}_h$, $\mathbf{z}$)	CONV-(N3, K5, S1, P2), TANH
Discriminator	
Input: G($\mathbf{y}_h$, $\mathbf{z}$)(batch_size=64, channels=3, height=64, width=64); $\mathbf{x}$(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 A2 Conditional-DGPose architecture for 64×64 input images. We use the following abbreviations: N for the number of kernels/neurons, K for kernel size, S for stride and P for zero padding. Concerning the layers, 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, detailed at Table A1. The additional layers can be clearly understood. Finally, particular parameters for specific layers are defined between parenthesis after the layers' names.

Encoder	
Input: $\mathbf{x}$(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	CONV-(N512, K3, S2, P1), BN, ReLU
5	CONV-(N512, K3, S2, P1), BN, ReLU
6	CONV-(N512, K3, S2, P1), BN, ReLU
7	RESIDUAL-(N512, K3, S1, P1)
8	RESIDUAL-(N512, K3, S1, P1)
9	RESIDUAL-(N512, K3, S1, P1)
10	RESIDUAL-(N512, K3, S1, P1), SIGMOID
$\mu_{\mathbf{z}}$	FC-(N100)
$\sigma_{\mathbf{z}}$	FC-(N100)
$\mu_{\mathbf{y}_v}$	FC-(N48)
$\sigma_{\mathbf{y}_v}$	FC-(N48)
Mapper	
Input: $\mathbf{y}_v$(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)
$\mathbf{y}_h$	DECONV-(N24, K4, S2, P1), SIGMOID
Decoder	
Input: $\mathbf{z}$(batch_size=64, channels=100); $\mathbf{y}_v$(batch_size=64, channels=48); $\mathbf{y}_h$(batch_size=64, channels=24, height=64, width=64)	
Layer	**Definition**
1	CONCAT($\mathbf{z}$, $\mathbf{y}_v$)
2	RESHAPE(batch_size=64, channels=148, height=1, width=1)
3	DECONV-(N512, K4, S1, P0), BN, LeakyReLU(0.2)
4	DECONV-(N256, K4, S2, P1), BN, LeakyReLU(0.2)
5	DECONV-(N128, K4, S2, P1), BN, LeakyReLU(0.2)
6	DECONV-(N64, K4, S2, P1), BN, LeakyReLU(0.2)
7	DECONV-(N128, K4, S2, P1), BN, LeakyReLU(0.2)
8	CONCAT(DECONV-6, $\mathbf{y}_h$)
9	CONV-(N512, K5, S1, P2), BN, LeakyReLU(0.2)
10	CONV-(N256, K5, S1, P2), BN, LeakyReLU(0.2)
11	CONV-(N128, K5, S1, P2), BN, LeakyReLU(0.2)
12	CONV-(N128, K5, S1, P2), BN, LeakyReLU(0.2)
G($\mathbf{y}_v$, $\mathbf{z}$)	CONV-(N3, K5, S1, P2), TANH
Discriminator	
Input: G($\mathbf{y}_v$, $\mathbf{z}$)(batch_size=64, channels=3, height=64, width=64); $\mathbf{x}$(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 A3 Semi-DGPose architecture for 64×64 input images. We use the following abbreviations: N for the number of kernels/neurons, K for kernel size, S for stride and P for zero padding. Concerning the layers, 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, detailed at Table A1. The additional layers can be clearly understood. Finally, particular parameters for specific layers are defined between parenthesis after the layers' names.