DGPose: Disentangled Semi-supervised Deep Generative Models for Human Body Analysis

Rodrigo de Bem, Arnab Ghosh, Thalaiyasingam Ajanthan,
Ondrej Miksik⋆, N. Siddharth, and Philip H.S. Torr

University of Oxford
{rodrigo, arnabg, ajanthan, omiksik, nsid}@robots.ox.ac.uk
philip.torr@eng.ox.ac.uk

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

Keywords: semi-supervised learning, deep generative models, variational autoencoders, human pose estimation
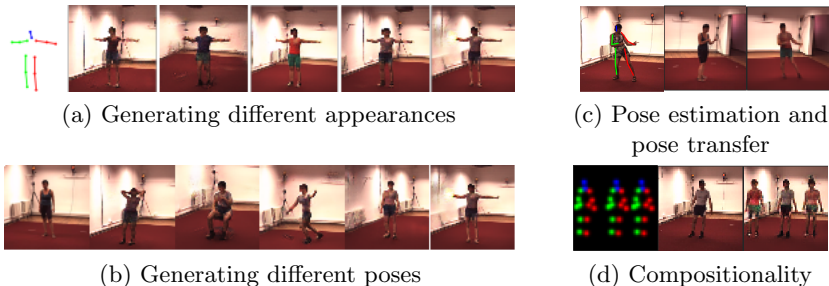
1 Introduction

Robust human-body analysis has been a long-standing goal in computer vision, with many applications in gaming, human-computer interaction, shopping or health-care [1,2,3,4]. Typically, most approaches to this problem have focused on discriminative models [1,5,6,7,8,9,10,11,12,13,14,15], where the given visual input (images or videos) is transformed to a suitable abstract form (*e.g.* human pose) by performing supervised learning. While these approaches do exceptionally well on their prescribed task, as evidenced by performance on pose estimation benchmarks, they fall short on two particular criteria of interest: a) reliance on fully-labelled data, and b) the ability to (conditionally) generate novel data from the abstractions.

The former is a fairly onerous requirement, particularly when dealing with real-world visual data, needing many hours of human-annotator time and effort

---

⋆ Ondrej Miksik has done this work at the University of Oxford. He is currently with Emotech Labs.

(a) Generating different appearances

(c) Pose estimation and pose transfer

(b) Generating different poses

(d) Compositionality

Fig. 1. Applications of our deep generative models for natural images of people. (a) For a given pose (first image), it generates different appearances. (b) For a given appearance (first image), it generates different poses. (c) For an estimated pose (first image) and a given appearance (second image), it combines both. (d) For manipulated poses (first image) and a given appearance (second image), it *hallucinates* people in the scene.

to collect. Thus, being able to relax the reliance on labelled data is a highly desirable goal. The latter addresses the ability to manipulate the abstractions directly, with a view to *generating* visual data with semantically consistent behaviour; *e.g.* moving the pose of an arm results in generation of images or videos where that arm is correspondingly displaced. Such *generative* modelling, in contrast to discriminative modelling, enables an *analysis-by-synthesis* approach to human-body analysis, where one can generate images of humans in combinations of poses and clothing unseen during training. This has many potential applications. For instance, it can be used for performance capture and reenactment of RGB videos, as already possible for faces [16], and still incipient for whole human bodies. It can also be used to generate images in a user specified pose to enhance datasets with minimal annotation effort. Such an approach is typically tackled using deep generative models (DGMs) [17,18,19] – an extension of standard generative models that incorporate neural networks as flexible function approximators. Such models are particularly effective in complex perceptual domains such as computer vision [20], language [21], and robotics [22], effectively delegating bottom-up feature learning to neural networks, while simultaneously incorporating top-down probabilistic semantics into the model. They solve both the deficiencies of discriminative approach discussed above by a) employing unsupervised learning, thereby removing the need for labels, and b) embracing a fully generative approach.

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

In this work, in order to ameliorate the aforementioned issue, while still eschewing reliance on fully-labelled data, we adopt the *structured semi-supervised* approach [23,24] from the Variational Autoencoder (VAE) [17,18] literature. 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 (*e.g.* appearance) is left to be non-interpretable. Weak (semi) supervision acts as a mean 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 [25], and c) interpretable subset of latent variables defined through model structure. In this work, we further extend these semi-supervised models with a discriminator-based [19] loss function, yielding a semi-supervised VAEGAN [26]. Our approach, with its main applications illustrated in Fig. 1, is formulated in a principled, unified probabilistic framework and allows end-to-end training.

In summary, our main contributions are:

i) a large-scale real-world application of structured semi-supervised deep generative models for natural images, separating pose from appearance in the analysis of the human body,

ii) a quantitative and qualitative evaluation of the generative capabilities of such models, and

iii) a demonstration of their utility in performing pose-transfer, without being explicitly trained for such a task.

2   Preliminaries

Deep generative models (DGMs) come in two broad flavours – Variational Autoencoders (VAEs) [17,18], and Generative Adversarial Networks (GANs) [19]. 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

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

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

VAEs also admit an extension to *conditional* generative models [27], simply by incorporating a conditioning variable $\mathbf{c}$, to derive

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

For 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 [23,24]. For learning in this framework, the objective can be expressed as the combination of supervised and unsupervised objectives. Let $\mathcal{D}_u$ and $\mathcal{D}_s$ denote the unlabelled and labelled subset of the dataset $\mathcal{D}$, and let the joint recognition network factorise as $q_\phi(\mathbf{y}, \mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{y}|\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$. Then, the combined objective summed over the entire dataset corresponds to

$$\mathcal{L}_{\text{SS}}(\theta, \phi; \mathcal{D}) = \sum_{\mathbf{x}^n \in \mathcal{D}_u} \mathcal{L}_u(\theta, \phi; \mathbf{x}^n) + \gamma \sum_{(\mathbf{x}^m, \mathbf{y}^m) \in \mathcal{D}_s} \mathcal{L}_s(\theta, \phi; \mathbf{x}^m, \mathbf{y}^m) \qquad (3)$$

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

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

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

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

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

Crucially, learning a customised approximation to the likelihood can result in a much higher quality of generated data, particularly for the visual domain [28].

A more recent family of DGMs, VAEGANs [26], bring together these two different approaches into a single objective that combines both the VAE and GAN objectives directly as

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{GAN}}. \qquad (7)$$

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

3   Our Approach

As set out in the preliminaries in Section 2, we use the VAEGAN [26] as the basis framework for our generative models. Note that, in incorporating semi-supervised learning, the semi-supervised VAEGAN includes two distinct tasks.
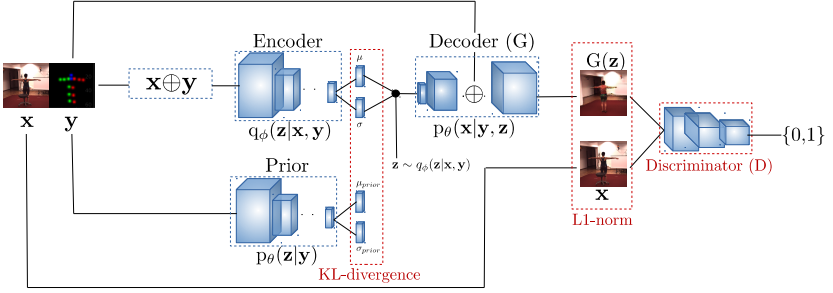
Fig. 2. Conditional-DGPose architecture. Encoder and Decoder are conditioned to the pose $\mathbf{y}$. The Prior module learns the Gaussian distribution $p_\theta(\mathbf{z}|\mathbf{y})$, which is used to regularize the Gaussian distribution $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ by the KL-divergence loss. The sampling of the *appearance* $\mathbf{z}$, which is the Decoder input, is done using the reparametrization trick [17]. The L1-norm and the Discriminator losses are computed over the reconstructed $G(\mathbf{z})$ and the original $\mathbf{x}$ images. G denotes Generator (see Eq.6). More details in the supplementary.

First, it involves learning a recognition network that can predict pose (interpretable) and *appearance* (non-interpretable) for any given data. Second, it involves learning a generative network that combines a given pose with an appearance to generate visual data 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 (texture, colour, shapes, *etc.*) about the given input.

Given a generative model that takes both appearance $\mathbf{z}$ and pose $\mathbf{y}$ as inputs to produce data $\mathbf{x}$, a reasonable first step can be to just 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** (Fig. 2), with reference to the fact that we simply employ a Conditional VAE model [27]. Its lower bound is similar to Eq. 2, given by

$$\log p_\theta(\mathbf{x}|\mathbf{y}) \geq \mathcal{L}_{\text{CVAE}}(\phi, \theta; \mathbf{x}|\mathbf{y}) \; = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z}|\mathbf{y})}{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})} \right], \qquad (8)$$

and its final objective function is defined as $\mathcal{L} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{GAN}}$, in contrast to the standard VAEGAN objective (Eq. 7). Here, all data is "labelled" with pose, but the primary goal is to verify if a low-dimensional conditioning variable has an effect in the conditional generative model. This approach does also lend itself to evaluating the accuracy of the reconstructed images w.r.t. the human body poses and the image quality.

Once we have verified that the conditional approach works, we can then proceed to evaluating our semi-supervised VAEGAN, referred to as **Semi-DGPose** (Fig. 3), as all this changes from the previous setup is that the encoding distribution is no longer conditioned on the pose, but instead predicts it as per

Fig. 3. Semi-DGPose architecture. The Encoder receives **x** as input. The KL-divergence loss between the Gaussian distribution $q_\phi(\mathbf{y}, \mathbf{z}|\mathbf{x})$ and the weak Gaussian priors $p(\mathbf{y})$ and $p(\mathbf{z})$ works as an regulariser for unsupervised training samples (see Eq. 4). For the supervised samples, a regression loss is used (see Eq. 5). The sampling of *appearance* and pose is done using the reparametrization trick [17] and propagated to the Decoder. The low-dimensional pose vector **y** is mapped to a heatmap representation by the Mapper module. The L1-norm and the Discriminator losses are computed over the reconstructed G(**z**) and the original **x** images. G denotes Generator (see Eq.6). More details in the supplementary.

Eq. 3–6. We again adapt the standard VAEGAN objective (Eq. 7), but now to use the semi-supervised VAE objective instead of the standard unsupervised one, yielding $\mathcal{L} = \mathcal{L}_{SS} + \mathcal{L}_{GAN}$ as the final objective function of this model.

4   Experiments

In this section, we extensively evaluate our models, namely, Conditional-DGPose and Semi-DGPose. First, we present details about datasets, metrics, training, architectures and pose representation, which are common to both the models and then provide quantitative and qualitative results specific to each model.

Datasets.   We use the Human3.6M [29], and the ChictopiaPlus [30] datasets. The Human3.6M is a widely used dataset 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 two actors as our validation set, while the rest of the data was used for training. We use a subset of 14 (out of 32) body joints represented by their (x, y) 2D image coordinates as our ground-truth data, neglecting minor body parts (*e.g.* fingers). Due to the high frequency of the video acquisition (50Hz) in the Human3.6M, 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 validation, with $317, 989$ and $1, 280$ images, respectively. All the images have resolution of 1000×1000 pixels.

   The ChictopiaPlus dataset [30] is an extension of the Chictopia dataset [31]. It augments the original per-pixel annotations for body parts with pose annotation [32], 3D shape [33] and facial segmentation. In contrast to the Human3.6M

dataset, in which each actor wears always the same outfit, it contains $23,011$ training, $2,913$ validation and $2,873$ test images of segmented people (without background) dressed in a great variety of clothes. All the images have resolution of 286×286 pixels.

Metrics.   Quantitative evaluation of generative models is inherently difficult [34] and usually a great deal of emphasis is placed on qualitative evaluation of reconstructed (generated) samples. Since our models explicitly represent *appearance* and *body pose* as separate variables, we evaluate the two independently with appropriate metrics. Image quality is evaluated using a standard Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [35] metrics. However, such metrics do not explicitly evaluate the generated poses. Hence, we introduce the use of the Percentage of Correct Keypoints (PCK) metric [36], which computes the percentage of 2D joints correctly located by a pose estimator, given the ground-truth and a normalized distance threshold. To employ the PCK in the evaluation of the reconstructed poses, we use an established (discriminative) human pose estimator [37], and initially estimate all poses in the original validation set. For our purpose, we assume that such estimations are the ground-truth poses of the validation set. Subsequently, we apply the same discriminative estimator over the reconstructed validation images, generated by our trained models. 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 in the reconstructed images, at the same locations, up to the distance threshold. Related works do not evaluate the accuracy of the generated pose directly but only the overall reconstruction quality either by using standard SSIM [38] or the IoU based score which is specific to the setup of [30], based on reconstruction of segmentation masks. In short, our evaluation based on the PCK metric measures reconstructions accuracy by also considering the accuracy of the generated poses.

Training.   All experiments were trained with mini-batch consisting of 64 images. We used the Adam optimizer [39] with initial learning rate set to 10^{-4}. The weight decay regulariser was set to 5×10^{-4}. Network weights were initialized randomly for fully-connected layers and with robust initialization [40] for convolutional and transposed-convolutional layers. Except when stated differently, for all images and all models, we used a 64×64 pixels crop centering 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 [41] and all experiments ran on an NVIDIA Titan X GPU.

Encoder/Decoder Architectures.   In both of our Conditional-DGPose and Semi-DGPose models, encoders, decoders and discriminators are all CNNs. We have performed extensive experiments to evaluate different alternatives for such CNNs architectures, which have culminated in our best performing models.

Pose Representation.   In representing pose as a latent variable, we have a couple of choices. We can encode just the (x, y) positions of the joints them-

selves in vector form, or alternately, we can construct a heatmap representation of the joints, as adopted successfully in many discriminative pose estimation works [14,37,12,42]. In our experiments, the heatmap-based pose representation produced higher quality results compared to the standard vector-based representation. Hence, we use the heatmap pose representation in all our final models.
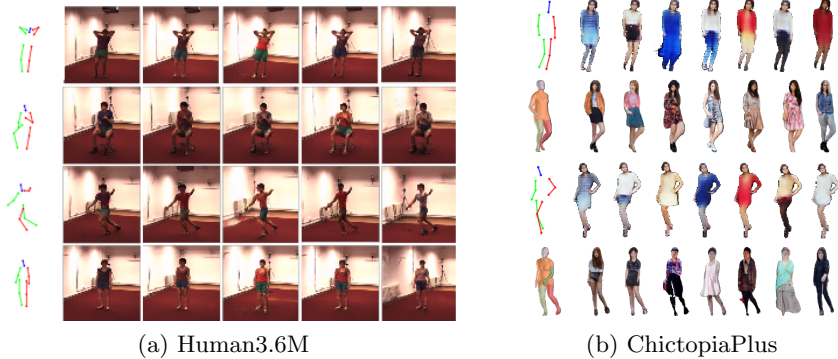
4.1 Conditional-DGPose

In this part, we evaluate the Conditional-DGPose model on the Human3.6M [29] and on the ChictopiaPlus [30] datasets. For the latter one, we also compare with the approach of Lassner *et al.* [30] which is closely related to our model as it has the capability of generating images conditioned on pose represented by segmentation masks. In addition to quantitative and qualitative results, we also demonstrate the flexibility of our model on potential applications.

Quantitative Results. In order to provide quantitative evaluation of pose reconstructions, we employed the PCK metric, as described previously in this section (see **Metrics**). On the Human3.6M dataset, our model achieves 89.22% of accuracy w.r.t. pose reconstruction, using the PCK metric normalized at 0.5 (see Fig. 4a for the overall PCK curve). The ChictopiaPlus dataset allows us to compare with Lassner *et al.* [30]. In this dataset, our model reports 95.14% of accuracy, with PCK score at 0.5, and outperforms [30] by a large margin, which reports 70.89% (see Fig. 4b and Fig. 6). Our results demonstrate good quality of our reconstructions w.r.t. the human pose on both datasets and hence benefits of single stage end-to-end Conditional-DGPose model, in contrast to the multiple stages of training and testing in [30].

(a) Human3.6M validation set   (b) ChictopiaPlus test set
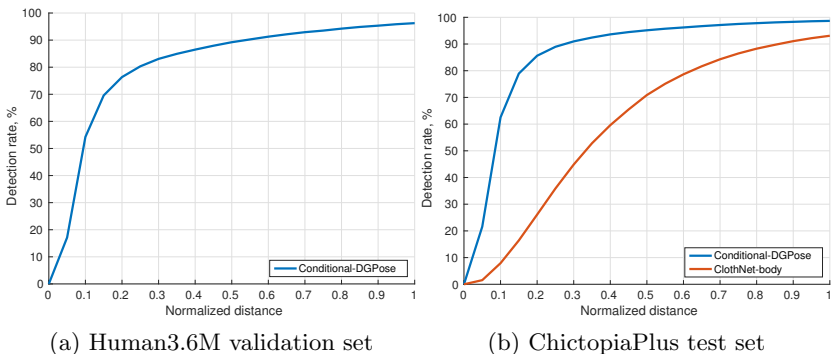
Fig. 4. (a) **Human3.6M:** PCK scores over reconstructed images of our Conditional-DGPose. (b) **ChictopiaPlus:** the PCK scores over reconstructed images of our Conditional-DGPose (*blue*) significantly outperforms the ClothNet-body [30] (*red*). Detection rate represents the percentage of joints corrected relocated in the reconstructed images.

(a) Human3.6M   (b) ChictopiaPlus

Fig. 5. Random sampling. Results on Human3.6M (**a**) and ChictopiaPlus (**b**) by changing pose and appearance independently. In (**b**), we compare 128×128 samples from our Conditional-DGPose method (odd rows), with 256×256 samples from Lassner *et al.* [30] (even rows). We also compare the different pose representations in the first column. We derive our heatmaps from joint positions, while segmentation masks are derived from 3D meshes and joints in [30].

Qualitative results. In Fig. 5 we present qualitative results of our model to demonstrate that it generates realistic images with accurate poses. In addition to that, we compare images generated by our model and Lassner *et al.* [30] in Fig. 6. Notice that, while both approaches are capable of generating people in required poses, our approach significantly outperforms [30] in terms of generated appearance which is much closer to the original in our case. Even though both the methods were able to generate visually good images and poses, Conditional-DGPose was more accurate in capturing the locations of the body parts. This shows that even without he image-to-image translation network our method was able to generate realistic images. Next, we show qualitative results on the pose-transfer and compositionality applications.

Pose transfer. In this task, 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 appearance of the input image, however, the generated person is "moved" into the required target pose. The target pose may be composed manually, extracted from other image by means of any other (even discriminative) pose estimator or provided interactively by an user. This is illustrated in Fig. 7, where we employed target poses from LSP dataset [43], that have completely different poses in a drastically different environment compared to our training set. The quality of the generations show that our generative model could disentangle pose and appearance and generate images with poses that do not exist in the training data.

Compositionality. Next we show in Fig. 8, 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,
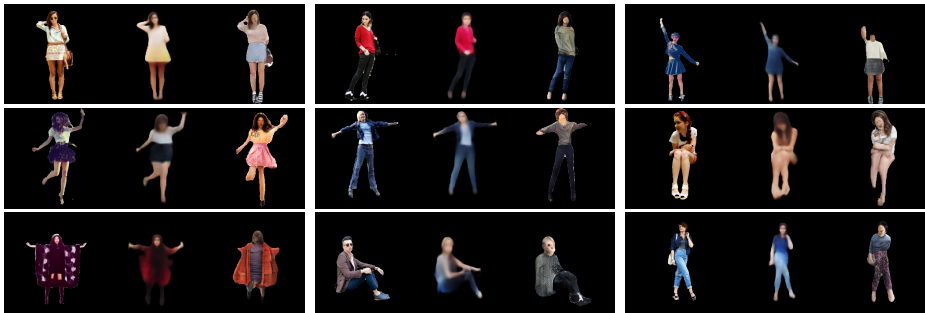
Fig. 6. Reconstructions. In each trio of images we have, respectively: original image (256 × 256), Conditional-DGPose (128 × 128) and ClothNet-body (256 × 256) [30] reconstructions. Notice, that the images generated by our model are much closer to the originals in terms of appearance (colours). Moreover, in general, the Conditional-DGPose captures the body parts' locations more accurately, which results in better quantitative results w.r.t. the pose reconstruction, shown in Fig. 4b. Best viewed if zoomed in digital version.
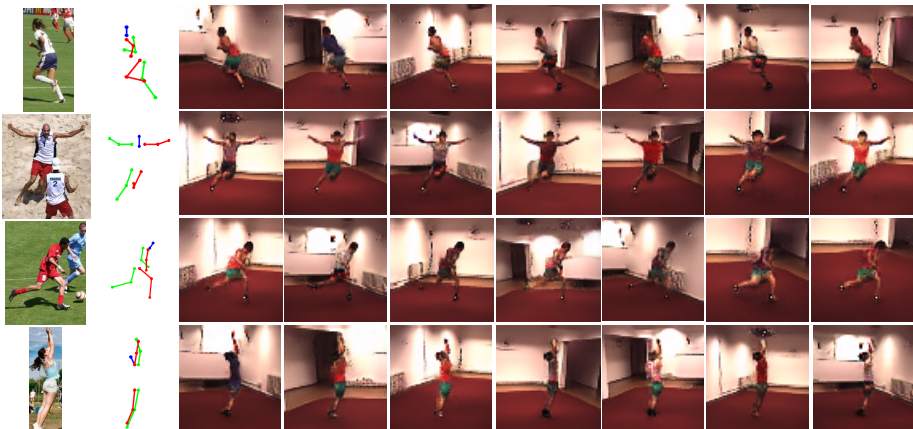
Fig. 7. Pose-transfer. Here we illustrate the *pose-transfer* capability of our Conditional-DGPose. On the leftmost column we show test images from the LSP dataset [43], along with their corresponding ground-truth 2D pose annotations, composed by 14 joints. These are taken as conditioners (*target-poses*) on our model for the generation of the reconstructions, shown from the thrid to the rightmost column. As can be observed, the *target-poses* are transfered to the validation 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.

and generate an image where all persons are in the same pose, but one of them is *e.g. shorter* and another *thinner*, as shown in Fig. 9a. In an extreme case, we can even generate "unreal" images containing only certain body parts (*i.e.* heads) or disconnecting them from the rest of the body, as in Figs. 9b and 9c, respectively. Note that the training dataset is composed of only single person images, thus the model has never seen any image with multiple people or only some separate body parts. This clearly demonstrates that the learned latent space of our model
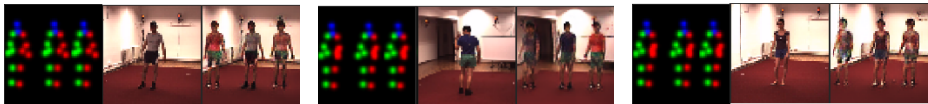
Fig. 8. Composing multiple people. The Conditional-DGPose model was only trained with images containing only one person. The middle figure in each set is the source image and the right image is the output, generated conditioned to the heatmap pose representation in the left image.

| (a) | (b) | (c) |
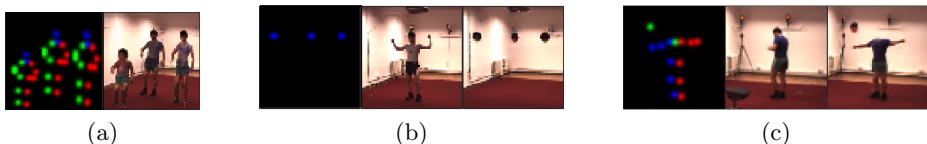
Fig. 9. Composing "unreal" images. We illustrate the versatility of the model extrapolating the generation of images to unseen scenes: (**a**) sampled image in which the pose representation in the center was manually translated and scaled, producing two additional bodies: one shorter and chunkier (*left*) 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.

is indeed disentangled. To the best of our knowledge, this capability has not been demonstrated with any other existing models [30,38].
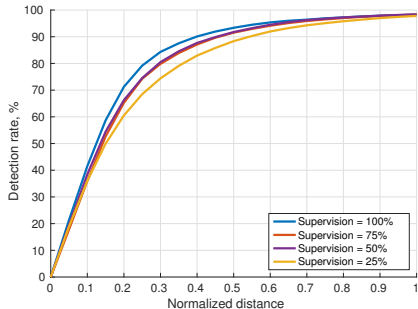
4.2   Semi-DGPose

Here we evaluate our Semi-DGPose model on the Human3.6M [29] dataset. We show quantitative and qualitative results, focusing particularly on the pose estimation and on the *indirect* pose transfer capabilities, described later in this section.

Quantitative results. In order to evaluate the efficacy of our semi-supervised model, we perform "relative" comparison on the Human3.6M dataset. 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-supervised part to see how the gradual decrease in the level of supervision affects the final performance of the method on the same validation set.

We first cross-validated the hyper-parameter α which weights the regression loss (see Eq. 5, in Sec. 2) and found that $\alpha = 100$ yields the best results. We keep $\gamma = 1$ in all experiments (see Eq. 3, in Sec. 2). Qualitatively, the reconstructions of the fully-supervised model are comparable with the ones obtained using the Conditional-DGPose. We evaluated it across different levels of supervision, with the PSNR and SSIM metrics and show results in Fig. 10a. We also evaluated the pose estimation accuracy of the Semi-DGPose model. It achieves 93.85%
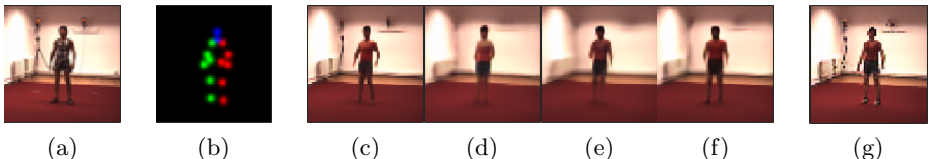
Level of supervision	PSNR	SSIM
100%	22.27	0.89
75%	21.36	0.86
50%	21.49	0.87
25%	20.06	0.83

(a)

(b)

Fig. 10. Quantitative evaluations of Semi-DGPose: (**a**) PSNR and SSIM measures for different levels of supervision, (**b**) PCK scores for different levels of supervision. Note that, even with 25% supervision, our Semi-DGPose obtains 88.35% PCK score, normalized at 0.5.


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

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

PCK score, normalized at 0.5, in the fully-supervised setup (100% of supervision over the training data). This pose estimation accuracy is on par with the state-of-the-art pose estimators on unconstrained images [44]. However, since the Human3.6M was captured in a controlled environment, a standard (discriminative) pose estimator can be expected to perform better. The overall PCK curves corresponding to each percentage of supervision in the training set is shown in Fig. 10b. Note that, even with 25% supervision, our model obtains 88.35% PCK score, normalized at 0.5.

Qualitative Results. Here we show qualitative results of the Semi-DGPose model on the Human3.6M dataset. Initially, in Fig. 11, we show reconstructed images obtained with different levels of supervision. It allows us to observe how image quality is affected when we gradually reduce the availability of labels. Following that we evaluate results on pose estimation and on *indirect* pose transfer.

Pose Estimation. Here we complement the previous quantitative evaluation of pose estimation. We highlight this distinctive capability of our Semi-DGPose generative model, which is not present in related works in the literature [30,38]. Again, we aim to analyse how the gradual decrease of supervision in the training set affects the quality of pose estimation on the validation images. Results are shown in Fig. 12.
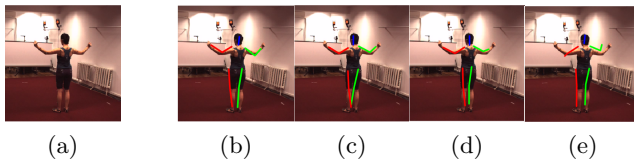
Fig. 12. Pose estimation. Original image (**a**), followed by estimations, over the original image, with: (**b**) 100%, (**c**) 75%, (**d**) 50% and (**e**) 25% of supervision.
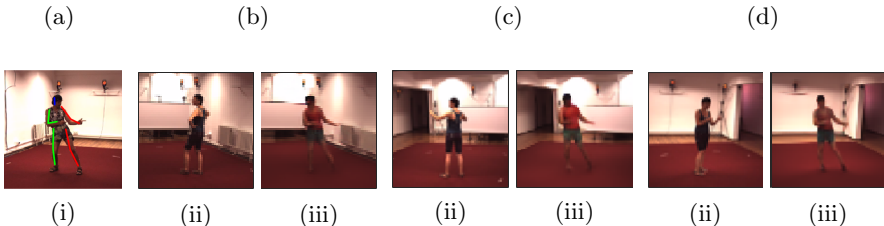
Fig. 13. *Indirect* **pose transfer:** (**i**) the latent target pose representation **y** is estimated (encoder). The pairs (**b**), (**c**) and (**d**), show (**ii**) the image from which the latent *appearance* **z** is estimated (encoder); (**iii**) the output image generated as a combination of **y** and **z** (decoder). The person's outfit in the output images (**iii**) is approximated to the ones in images (**ii**), however restricted to the low diversity of outfits observed in training data. Backgrounds of images (**ii**) are reproduced in the output images (**iii**) and all them differ from the one in image (**a**).

Indirect **Pose Transfer.** Another important and distinctive capability of our Semi-DGPose model is what we called *indirect* pose transfer. As both latent variables, corresponding to pose **y** and *appearance* **z**, can be inferred by the model's encoder (recognition network) at test time, latent variables extracted from different images can be combined in a subsequent step, and employed together as inputs for the decoder (generative network). The result of that is a generated image combining appearance and body pose, extracted from two different images. The process is done in three phases, as illustrated in Fig. 13: i) the latent pose representation **y** is estimated from the first input image through the encoder; ii) the latent *appearance* representation **z** is estimated from a second image, also through the encoder, iii) **y** and **z** are propagated through the decoder, and a new image is generated, combining body pose and appearance, respectively, from the first and second *encoded* images.

5   Related Work

Generative modelling for human body analysis has a long history in computer vision [45,46,47,48]. However, despite the great interest in human body analysis in the past years, deep generative models have been far less investigated compared to their discriminative counterparts [1,5,6,7,8,9,10,11,12,13,14,15]. Recently, Lassner *et al.* [30] presented a deep generative model based on a CVAE conditioned on human pose which allowed generating images of segmented people and their clothing. However, this model does not encode pose using raw

image data but only low dimensional (binary) segmentation masks and "image-to-image" transfer network [49] is used to generate realistic images. In contrast, we learn the generative model directly on the raw image data using only pose as a conditioner and without the need of body parts segmentation. Difficulty of generating both correct poses and detailed appearance simultaneously in an end-to-end fashion is admitted also by Ma *et al.* [38]. In order to tackle this issue, they proposed a two stage model, where the first part focuses on generating a global structure of the human and the second stage fills in appearance details, however the generative process is conditioned on image rather than pose. Hence, generating images conditioned on pose is not trivial which is again in contrast to our approach.

In a concurrent work to ours, Siarohin *et al.* [50] improves approach of [38] by making it single-stage and trainable end-to-end. While this approach is relatively similar to ours, the key difference is that the human body joints (keypoints) are given to the algorithm (detected by another off-the-shelf discriminative method) while our method learns to encode them directly from the raw image data. Hence, our model allows sampling of different poses independent of appearance. Finally, Ma *et al.* [51] proposed a model for learning image embeddings of foreground, background and pose variables encoded as interpretable variables. However, this model has to rely on an off-the-shelf pose estimator to perform pose-transfer but our model can perform pose estimation even in a semi-supervised setting in addition to image generation. To sum it up, the existing approaches do not have the flexibility to manipulate pose independently of appearance and they have to be explicitly trained to allow pose transfer. This is in sharp contrast to our approach, where we only learn for pose *prediction* and pose transfer is a by-product.

Apart from this, Walker *et al.* [52] proposed a hybrid VAEGAN architecture for forecasting future poses in a video. Here, a low-dimensional pose representation is learned using a VAE and once the future poses are predicted, they are mapped to images using a GAN generator. Following [26], we use a discriminator in our training to improve the quality of the generated images, however, in contrast to [26], the latent space of our approach is interpretable which enables us to sample different poses and appearance. Considering GAN based generative models, Tulyakov *et al.* [53] presents a GAN network that learns motion and content in two separate latent spaces in an unsupervised manner. However it does not allow an explicit manipulation over the human pose.

6   Conclusions

In this paper we have presented a deep generative model for human pose analysis in natural images. To this end, we have adapted the structured semi-supervised variational auto-encoder approach. Our model allows independent manipulation of pose and appearance and hence enables applications such as pose-transfer without being explicitly trained for such a task. In addition to that, the semi-supervised setting relaxes the need for labelled data. We have systematically evaluated our model on the Human3.6M and ChictopiaPlus datasets and showed that it enables applications such as pose transfer and outperforms related work.

References

1. 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)
2. von Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. Eurographics (2017)
3. Achilles, F., Ichim, A.E., Coskun, H., Tombari, F., Noachtar, S., Navab, N.: Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications. In: MICCAI. (2016)
4. Seemann, E., Nickel, K., Stiefelhagen, R.: Head pose estimation using stereo vision for human-robot interaction. In: FG. (2004)
5. Gkioxari, G., Toshev, A., Jaitly, N.: Chained predictions using convolutional neural networks. In: ECCV. (2016)
6. U. Rafi, I.Kostrikov, J.G., Leibe, B.: An efficient convolutional network for human pose estimation. In: BMVC. (2016)
7. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: ICCV. (2015)
8. Chen, X., Yuille, A.L.: Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations. In: NIPS. (2014)
9. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: CVPR. (2016)
10. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV. (2016)
11. Belagiannis, V., Zisserman, A.: Recurrent human pose estimation. FG (2016)
12. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4724–4732
13. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: ECCV. (2016)
14. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. CVPR (2017)
15. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. (2017)
16. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR. (2016)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. ICML (2014)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014)
20. Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.: Deep convolutional inverse graphics network. In: NIPS. (2015)
21. Massiceti, D., Siddharth, N., Dokania, P., Torr, P.H.: FlipDial: A generative model for two-way visual dialogue. In: CVPR. (2018)
22. Wang, Z., Merel, J.S., Reed, S.E., de Freitas, N., Wayne, G., Heess, N.: Robust imitation of diverse behaviors. In: NIPS. (2017)

23. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: NIPS. (2014)
24. 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)
25. Schulman, J., Heess, N., Weber, T., Abbeel, P.: Gradient estimation using stochastic computation graphs. In: Advances in Neural Information Processing Systems. (2015) 3528–3536
26. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: ICML. (2016)
27. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: NIPS. (2015)
28. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR. (2018)
29. 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)
30. Lassner, C., Pons-Moll, G., Gehler, P.V.: A generative model for people in clothing. In: ICCV. (2017)
31. 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)
32. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV. (2016)
33. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM TOG (2015)
34. Theis, L., van den Oord, A., Bethge, M.: A note on the evaluation of generative models. In: ICLR. (2016)
35. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
36. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011)
37. Newell, A., Yang, K., Deng, J.: Stacked Hourglass Networks for Human Pose Estimation. In: ECCV. (2016)
38. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. In: NIPS. (2017)
39. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015)
40. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV. (2015)
41. 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)
42. 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)
43. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC. (2010)
44. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: ICCV. (2017)
45. Jaeggli, T., Koller-Meier, E., Van Gool, L.: Learning generative models for monocular body pose estimation. In: ACCV. (2007)

46. Rauschert, I., Collins, R.T.: A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In: ECCV. (2012)
47. Sigal, L., Balan, A., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NIPS. (2008)
48. Rosales, R., Sclaroff, S.: Combining generative and discriminative models in a framework for articulated pose estimation. IJCV (2006)
49. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 (2016)
50. Siarohin, A., Sangineto, E., Lathuiliere, S., Sebe, N.: Deformable gans for pose-based human image generation. arXiv preprint arXiv:1801.00055 (2017)
51. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. arXiv preprint arXiv:1712.02621 (2017)
52. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. arXiv preprint arXiv:1705.00053 (2017)
53. Tulyakov, S., Liu, M., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. arXiv preprint arXiv:1707.04993