A Conditional Deep Generative Model of People in Natural Images

Rodrigo de Bem[1,2] Arnab Ghosh[1] Adnane Boukhayma[1] T. Ajanthan[1] N. Siddharth[1] Philip Torr[1]

Department of Engineering Science[1]
University of Oxford, UK

Center of Computational Sciences[2]
Federal University of Rio Grande, Brazil

{rodrigo, arnabg, adnane, ajanthan, nsid, phst}@robots.ox.ac.uk

Abstract

We propose a deep generative model of humans in natural images which keeps 2D pose separated from other latent factors of variation, such as background scene and clothing. In contrast to methods that learn generative models of low-dimensional representations, e.g., segmentation masks and 2D skeletons, our single-stage end-to-end conditional-VAEGAN learns directly on the image space. The flexibility of this approach allows the sampling of people with independent variations of pose and appearance. Moreover, it enables the reconstruction of images conditioned on a given posture, allowing, for instance, pose-transfer from one person to another. We validate our method on the Human3.6M dataset and achieve state-of-the-art results on the ChictopiaPlus benchmark. Our model, named Conditional-DGPose, outperforms the closest related work in the literature. It generates more realistic and accurate images regarding both, body posture and image quality, learning the underlying factors of pose and appearance variation.

1. Introduction

The analysis of visual data containing humans is a central problem in computer vision. In this context, the body posture plays a major role in the process of understanding humans in images and videos [12, 18, 41]. Modelling and learning how natural images of people in particular postures are generated is an important and challenging task. It has many applications, such as, generation of synthetic images and reenactment of body movements in videos [1].

Over the last years, great attention has been given to discriminative models of human pose, specially for 2D pose estimation [3, 4, 27, 35, 40]. Meanwhile, generative models have attracted considerably less interest. This fact may be partially explained by the relatively higher complexity of the latter, mainly due to the large number of random variables involved in the process, as well as the frequently intractable probability distributions over them.

Lately, however, the general research interest for generative models has increased, greatly driven by deep learning methods, such as Variational Autoncoders (VAEs) [17, 30] and Generative Adversarial Networks (GANs) [7]. These approaches have introduced strategies to tackle inference and stochastic learning in the context of intractable probabilistic computations and large datasets.

Generative models for human pose defined over low-dimensional *pose-spaces* are well-known from the literature [6, 28, 33]. Also in more recent works with deep generative models, in particular with VAEs, low-dimensional spaces are adopted, *e.g.*, pose-spaces [37] or segmentation masks [20]. Once such generative models are learned, a common approach is to use image-to-image translation [11] to map the representations to natural images. These multi-stage training and testing processes often reduce the accuracy and control over image generations w.r.t. appearance and body posture [20]. In other cases, as in [25], it prevents sampling, which is a core characteristic and desirable capability of generative models.

To overcome the aforementioned limitations of related works, we learn a deep generative model of natural images, directly on the high-dimensional image space. We explicitly represent the body pose and other latent factors of variation (or just *visual appearance*) as separated and independent random variables in a unified single-stage end-to-end probabilistic model. In this way, we can generate samples from our generative probability distribution by manipulating either pose or appearance directly and independently. Our conditional-VAEGAN architecture, a principled variational method for approximate Bayesian inference, allows us to have a structured and interpretable latent space, by means of a conditional-VAE framework [32], associated with a discriminator module [19], which takes advantage of the high quality image generations from GANs.

To our knowledge, our approach is the first deep generative model capable of generating realistic natural images of people in a unified probabilistic framework, while keeping the body posture and appearance as explicitly separated and interpretable variables. The advantage of that is threefold, as it allows: i) to change the posture of a person in an image, given a conditioning pose (pose-transfer); ii) the sampling from the generative distribution with independent control over pose and visual appearance; iii) the direct and more accurate control over appearance and pose by means of a unified single-stage end-to-end model.

We have validated our method qualitatively on the Human3.6M dataset [10]. Such experiments demonstrate that our model generates realistic images with direct, accurate and independent manipulation of pose and appearance, successfully performing *image reconstruction*, *pose-transfer* and *sampling* with a single network. Moreover, our approach achieves state-of-the-art results on the ChictopiaPlus benchmark, outperforming the closest related work in the literature, the ClothNet-Body network [20], and showing that our model generates more realistic and accurate images w.r.t. both, body posture and image quality, while it learns the underlying latent factors of pose and appearance variation.

2. Related Work

Deep Generative Models. VAEs [17, 30] and GANs [7] have both received great attention in the last years. Both of these deep learning methods are capable of tackling inference and stochastic learning in the context of intractable probabilistic distribution and large datasets. Due to their popularity, several variants of them have been proposed in the literature [19, 26, 38]. In particular, Larsen *et al.* [19] propose the association of VAEs' capability of explicitly modelling latent probability distributions, with GANs' high-quality image generations. Another useful variant of VAEs for conditional generative models are the conditional-VAEs (CVAEs), proposed and applied to image segmentation by Sohn *et al*. [32]. Additionally, conditional-VAEGANs (CVAEGANs) were employed by Bao *et al*. [2] for image inpainting and data augmentation.

Recently, other authors [5, 16, 31] have shown how VAEs can be used to learn structured disentangled representations in the latent space by enforcing partial supervision for a subset of latent variables. Different from previous works, our method is the first to employ a CVAEGAN with a structured, interpretable and disentangled latent space for generating people in natural images.

Deep Generative Models of Humans. Despite the great interest in human pose estimation in the past years, generative models have been far less investigated compared to discriminative approaches. The closest related approach to ours is [20], which presents a generative model based on CVAEs for clothes of segmented people conditioned on pose. However, their generative model works on low-dimensional segmentation masks and an image-to-image translation network [11] is used to render natural images. The segmentation masks (*sketches*) are generated either with a VAE (*ClothNet-full*) or with a CVAE (*ClothNet-body*). The simultaneous, direct and accurate manipulation of pose and appearance is limited because it is done by two separate networks in a two-stages process. In contrast, we learn the generative model directly on the real images using only pose as a conditioner and without the need of body parts' segmentation.

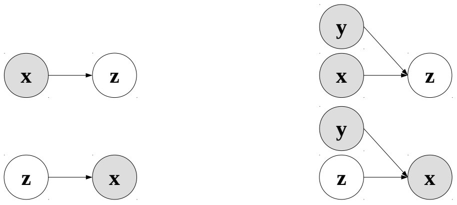
Another related work is the image-to-image translation model proposed by Ma et al. [25]. It uses the U-Net-like model from [29] and no VAE nor CVAE are employed. As in [20], [25] is trained in two stages, as the authors acknowledge that it is difficult for a complete end-to-end framework to cope with both correct poses and appearances simultaneously. However, training is done using pairs of images from the same person in different poses, views and scales, since the approach is designed strictly for pose-transfer. Differently, the Conditional-DGPose accomplishes pose-transfer as a by-product of our formulation. Moreover, it can be performed between images from different people, since it keeps pose and appearance as disentangled random variables. This relevant difference in our CVAEGAN modelling makes our method more general than the ones aforementioned. For instance, it also allows the direct sampling of natural images conditioned on a given pose. In [20], samples of segmentation masks (*sketches*) can be generated and rendered as natural images with the image-to-image model (*portray module*) from [11]. In the image-to-image translation from [25], sampling is not possible at all.

Furthermore, in [37] a hybrid VAEGAN architecture is introduced 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. We use a discriminator in our training to improve the quality of the generated images, following [19]. However, this does not effect our generative probabilistic distribution and neither does it compromise our capability of sampling from it. Finally, considering GAN based generative models, [36] presents a GAN network that can learn motion and content in two separate latent spaces in an unsupervised manner. However, it does not allow an explicit manipulation over the human pose.

3. Deep Variational Autoencoders

In this section, we briefly review variational autoencoders and their relevant variations on which our generative model of people in natural images is based. We refer to [17, 30, 32, 19] for further details.

Variational Autoencoders. VAEs [17, 30] are a class of deep generative models that simultaneously train both a probabilistic *encoder* and a *decoder*, given a training set $\mathcal{D}$ with elements $\mathbf{x} \in \mathcal{D}$. The main idea is that an encoding $\mathbf{z}$ is considered as a latent variable, and the objective is to maximize the likelihood $p(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$. The decoder



(a) VAE

(b) CVAE

Figure 1: Structure of recognition (**top row**) and generative (**bottom row**) models for VAE and CVAE. Here, $\mathbf{z}$ denotes the unobserved latent variables and $\mathbf{y}$ denotes the conditioning variables.

(referred to as the *generative network*) defines the conditional probability $p_\theta(\mathbf{x}|\mathbf{z})$ and the prior over $\mathbf{z}$ is assumed to be the standard normal distribution.

In a high dimensional space, finding the decoder parameters θ that maximize the likelihood is intractable. To tackle this, VAEs use a variational method that approximates the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ using an encoder (referred to as the *inference* or the *recognition network*) $q_\phi(\mathbf{z}|\mathbf{x})$. This approximate posterior is assumed to be a Gaussian whose parameters are the output of a neural network parametrized by ϕ. Under these assumptions, the generative and the inference networks are trained jointly by performing stochastic gradient ascent on the Evidence Lower Bound (ELBO),

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

Here, the first term denotes the expectation over the approximate posterior distribution, which measures the decoder accuracy (or reconstruction error), while the KL-divergence term encourages the approximate posterior to be close to the prior $p(\mathbf{z})$. Computing the expectation in Eq. (1) involves sampling which can be circumvented using a repràmetrization trick [17]. Furthermore, since the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$ are both Gaussian distributions, the KL-divergence can be computed in closed form. The structures of the generative and the recognition models in a standard VAE are shown in Fig. 1a. At test time, only the generative network (decoder) is retained and one can easily generate samples by sampling a latent variable $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and passing it through the decoder. For more details, we refer the reader to [17].

Conditional Variational Autoencoders. CVAEs [32] are a simple extension of the standard VAEs that allow more flexibility in the generative process. In a CVAE, both the input data $\mathbf{x}$ and the latent variable $\mathbf{z}$ are conditioned on $\mathbf{y}$. It means that both, the encoder and the decoder, are now conditioned on $\mathbf{y}$, *i.e.* the corresponding distributions can be written as $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ and $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{y})$. In this case the objective function can be written as,

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

Note that the KL-divergence is between $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ and $p_\theta(\mathbf{z}|\mathbf{y})$, where both the distributions are now conditioned on $\mathbf{y}$. See Fig. 1b for the generative and recognition models of a CVAE. In this work, our CVAE is conditioned on pose, as detailed in the subsequent section.

Conditional-VAEGANs. 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. 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 from the true data distribution $p(\mathbf{x})$. Such objective is defined as

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(D,G) &= \mathbb{E}_{p(\mathbf{x})}\left[\log D(\mathbf{x})\right]\\ &\quad + \mathbb{E}_{p_\theta(\mathbf{z})}\left[\log(1 - D(G(\mathbf{z})))\right]. \qquad (3)\end{aligned}$$

Crucially, learning a customised approximation to the likelihood can result in a much higher quality of generated data, particularly for the visual domain [14]. Thus, bringing together these two different approaches, similarly to [2, 19], our single objective combines both, the CVAE and the GAN objectives, directly as

$$\mathcal{L} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{GAN}}. \qquad (4)$$

4. Our Approach

Considering the previous section, in our Conditional-DGPose model we define $\mathbf{x}$ to be a fully observable random variable correspondent to a RGB image. The random variable $\mathbf{z}$ is unobservable and it corresponds to all latent factors that affect the generation of an image, except for the body pose of the person in it. We may refer to $\mathbf{z}$ as *appearance*, yet it is an over simplified definition. As to the body pose, it is represented by the interpretable and fully observable random variable $\mathbf{y}$. Following, we detail our method.

4.1. Conditional-DGPose

We have extensively tested several variations of deep architectures, culminating with the one shown in Fig. 2. On the latter, all the probability distributions $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$, $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{y})$ and $p_\theta(\mathbf{z}|\mathbf{y})$, as well as the discriminator correspond to deep CNN modules. Implementation details are provided in Tab. 4, Sec. A (Appendix). In the rest of this

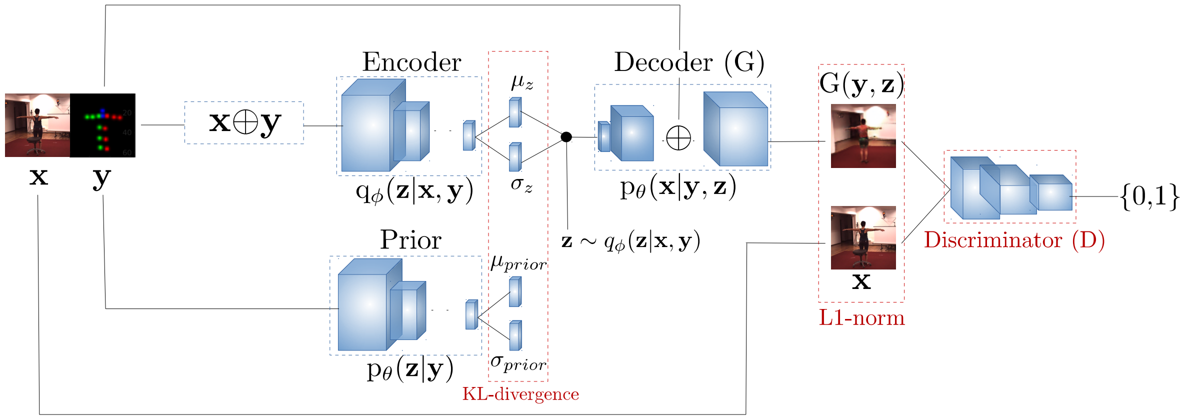

x⊕y
x
y
Encoder
$q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$
μ_z
σ_z
Decoder (G)
$p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z})$
$\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$
Prior
μ_{prior}
σ_{prior}
$p_\theta(\mathbf{z}|\mathbf{y})$
KL-divergence
$G(\mathbf{y},\mathbf{z})$
x
L1-norm
Discriminator (D)
{0,1}

Figure 2: **Conditional-DGPose architecture.** Encoder and Decoder are conditioned on 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$(\mathbf{y},\mathbf{z})$ and the original $\mathbf{x}$ images. G denotes Generator (see Eq.3).

section, we describe training and testing phases. However, we start following detailing the adopted pose representation, as it sets the basis for understanding the other topics in the section.

4.1.1 Pose Representation

It is not evident whether a sparse 2D vector representation would be capable of conveying the spatial information required for reconstructing realistic natural images, taking into account the human body pose. In fact, related work suggest the opposite [20, 25]. On the other hand, the mapping of 2D joints positions to heatmaps has shown to be very effective on several pose estimation approaches [3, 4, 27, 35, 40]. After experimental evaluation (see Sec. 5.1), we have followed [4], such that our heatmap representation consists of P body elements, in a way that $\mathbf{y} \in \mathcal{R}^{P\times H\times W}$, where H and W are the heatmaps' height and width, respectively. As illustrated in Fig. 3, $P = J + R + B$, where J stands for joints, R for rigid parts and B for body. 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, they are modelled 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. The Gaussian representing the rigid part is aligned with the line connecting $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_l$. $\sigma_{p,i}$ is proportional to $|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l|$ and $\sigma_{p,j} = \kappa_p \sigma_{p,i}$, where κ_p is a part-specific ratio, inspired by anthropometric measurements.

Body. The body centre is the mean of the annotated joints. The principal component analysis (PCA) of the joints' positions gives the orientation of the body in the image plane. $\sigma_{p,i}$ and $\sigma_{p,j}$ are proportional to the distance between the extreme projections of the joint onto, respectively, the principal and secondary axes of variation.

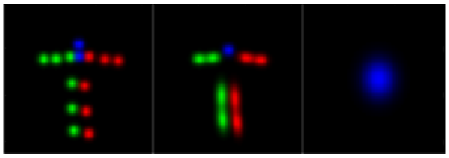


Figure 3: **Pose representation.** 24 heatmaps (14 from the annotated joints, 9 corresponding to rigid parts and 1 corresponding to the whole body). Right, left and central body parts are represented by colours green, blue and red, respectively, in the person-centric representation. Heatmaps of the same kind are gathered and superimposed on a black background to facilitate visualization.

4.1.2 Training

The recognition (encoder) and the generative (decoder) networks conditioned on pose $\mathbf{y}$ are defined by $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ and $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{y})$, respectively. In practice, the Gaussian heatmap labels (conditioning pose) are concatenated to the correspondent input image $\mathbf{x}$ at the encoder's 1st layer; and concatenated to $\mathbf{z}$ at the decoder's 7th layer (see Tab. 4, Sec. A). In both layers, the feature maps match the heatmaps dimensions. This design option was particularly important for the decoder, since the heatmaps effectively *guided the attention* [21] of the network towards the position and area of body parts, improving reconstructions. Additionally, the heatmap labels $\mathbf{y}$ alone are the input for the prior module, which learns the distribution $p_\theta(\mathbf{z}|\mathbf{y})$. Finally, the reconstructed image denoted by $G(\mathbf{y},\mathbf{z})$ and the training image $\mathbf{x}$ are used as input of the discriminator module, which learns to distinguish the real from the reconstructed images.

Following Sec. 3, the loss function $\mathcal{L} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{GAN}}$ (Eq. 4) minimized during training is composed of: i) reconstruction loss L1-norm$(\mathbf{x}, G(\mathbf{y}, \mathbf{z}))$, between the input image $\mathbf{x}$ and the reconstruction $G(\mathbf{y}, \mathbf{z})$; ii) the closed-form KL-divergence KL $[q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}) || p_\theta(\mathbf{z}|\mathbf{y})]$, between the recognition and the prior Gaussian distributions, respectively, $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ and $p_\theta(\mathbf{z}|\mathbf{y})$; and iii) the discriminator cross-entropy loss $\mathcal{L}_{\text{GAN}}$ (Eq. 3).

4.1.3 Testing

Due to the inherent versatility of generative models, our architecture may be employed in different ways, according to the intended task. Thus, the testing stage is divided into *reconstruction*, *pose-transfer* and *sampling*.

Reconstruction and Pose-transfer. For these two procedures, since $\mathbf{x}$ and $\mathbf{y}$ are given, neither the prior module nor the discriminator are employed. For reconstruction, when an image $\mathbf{x}$ and its corresponding pose $\mathbf{y}$ are given as input, the reconstructed image $G(\mathbf{y}, \mathbf{z})$ is obtained as the decoder output (see Fig. 5). On the other hand, for pose-transfer, when $\mathbf{x}$ is used as input along with a different target pose $\mathbf{y}_{target}$, the person in the reconstructed image will keep the appearance of $\mathbf{x}$, along with the body pose defined by $\mathbf{y}_{target}$ (see Fig. 6).

Sampling. Here, neither the encoder module nor the discriminator are employed. Moreover, as expected, no RGB image is given as input. Instead, only a conditioning pose $\mathbf{y}$ is given as the input to the prior module, which defines $p_\theta(\mathbf{z}|\mathbf{y})$. From this prior distribution, the latent appearance $\mathbf{z}$ is sampled and used as input of the decoder network. In this manner, for a given pose, different appearances can be randomly generated from the learned model (see Fig. 7).

5. Experiments and Discussion

We evaluate our Conditional-DGPose architecture on two datasets, the Human3.6M [10] and the ChictopiaPlus [20]. The following parameters were commonly adopted through all experiments: Adam optimizer [15] with learning rate equal to 10^{-4}; weight decay of 5×10^{-4}; network weights initialized randomly for fully-connected layers and with robust initialization [8] for convolutional and transposed-convolutional layers. We crop an image area keeping the person of interest in the central position. Images were normalized to present zero-mean and unit-variance and no other form of data augmentation or preprocessing was employed. Implementation was done using the Caffe framework [13] and the experiments ran in a NVIDIA Titan X. Following we describe in detail the evaluations performed with each one of the datasets.

5.1. Human3.6M

The Human3.6M dataset [10] 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 follow the standard protocol and use sequences of two actors as our test set, while the rest of the data is used for training. A subset of 14 (out of 32) body joints represented by their 2D image coordinates is adopted as ground-truth data, while minor body parts are neglected (*e.g.* fingers). Due to the high frequency of the video acquisition (50Hz), out of images from all 4 cameras, we subsample frames in time, producing subsets for training and test, with 317,989 and 1,280 images, respectively. The images, with original resolution of 1000×1000 pixels, are cropped to 64×64 and grouped in mini-batches of 64 samples.

The benchmark presents images in a controlled environment, thus its is adopted as an initial dataset for a qualitative evaluation. We extensively tested different pose representations and architectures on this dataset. Such experiments supported and guided our design options towards the use of heatmaps instead of 2D pose vectors, since reconstructions were better in the former case, as shown in Fig. 4 and Tab. 1; as well as towards the use of residual blocks in our encoders [23], which have improved our reconstructions. On the other hand, we have not observed benefits of using residual blocks in the decoder.

Finally, as part of our initial qualitative evaluations, we have tested our Conditional-DGPose on the three tasks mentioned in Sec. 4.1.3, namely *reconstruction*, *pose-transfer* and *sampling*. Results are illustrated, respectively, in Figs. 5, 6 and 7.

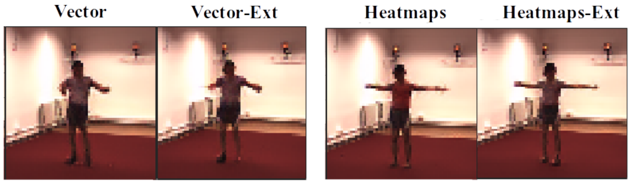


Figure 4: Reconstructed images, obtained with 2D vector versus heatmap representations. Extended versions have rigid parts and body in addition to the joints. We highlight the difficulty for capturing the spatial extend of some body parts, particularly extremities far from the torso, when the vector 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) makes the model more robust to more complex poses.

5.2. ChictopiaPlus

The ChictopiaPlus dataset [20] is an extension of the Chictopia dataset [22]. It augments the original per-pixel annotations for body parts with pose annotation [9], 3D

Pose representation	L1-Norm
Vector (14 joints)	14.52
Vector-Extended (28 joints)	13.91
Heatmaps (14 joints)	13.55
Heatmaps-Extended ⌞ (14 joints + 9 rigid parts + whole body)	**13.41**

Table 1: Average reconstruction errors obtained with our architecture using L1-norm on our Human3.6M test set.

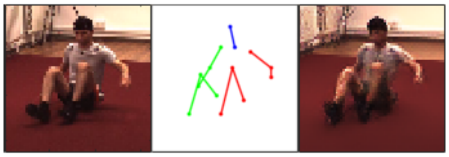


Figure 5: **Reconstruction.** From left to right: original image (64 × 64), original pose and reconstructed image. The pose is illustrated with a skeleton to facilitate visualization, even though we use the heatmap representation.

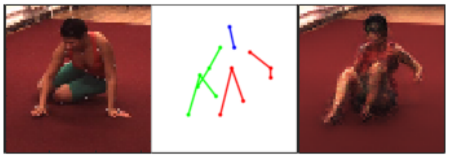


Figure 6: **Pose-transfer.** From left to right: original image (64 × 64), target pose (from Fig. 5) and reconstructed image, in which the target pose has been transfered yet keeping the appearance of the original image. The target pose is used as a conditioner in the generation of the reconstructed image. The pose is illustrated with a skeleton to facilitate visualization, even though we use the heatmap representation.

shape [24] 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.

We employ this benchmark to compare our approach with the closest related work in the literature, the ClothNet-body by Lassner *et al.* [20]. In order to do so, we use the trained models made publicly available by the ClothNet-body authors. We perform quantitative and qualitative comparisons, detailed below, showing that we outperform [20] w.r.t. image quality and body pose reconstructions.

Quantitative Results. 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 model explicitly represent *appearance* and *body pose* as separate variables, we evaluate the two independently with appropriate metrics. Image quality is evaluated using the standard Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) [39] metrics. However, such metrics do not explicitly evaluate the generated poses. Hence, we introduce the use of the Percentage of Correct Keypoints (PCK) metric [42] for this purpose. The PCK computes the percentage of 2D joints correctly located by a pose estimator, given the ground-truth and a normalized distance threshold multiplied by the size of the person's torso.

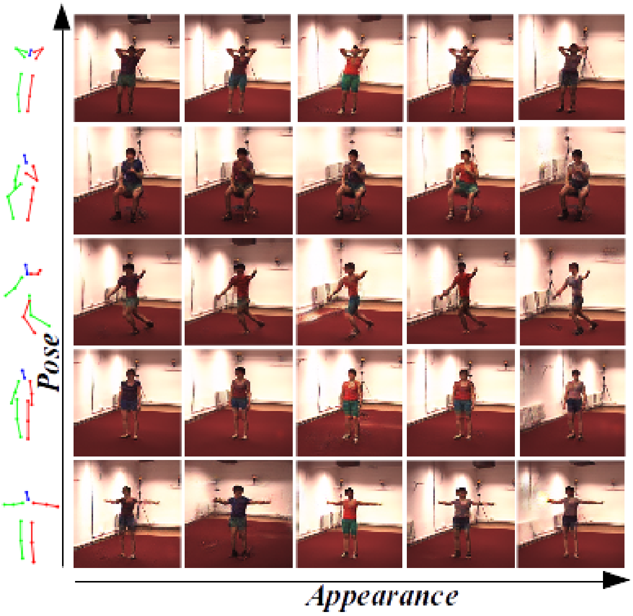


Figure 7: **Sampling.** Random samples of people given pose. As the two axis show, pose and appearance can be independently manipulated. The pose is illustrated with a skeleton to facilitate visualization, even though we use the heatmap representation. Images have 64 × 64 pixels.

To employ the PCK in the evaluation of the reconstructed poses, we use an *off-the-shelf* human pose estimator [27], and initially estimate all poses in the original test set. For our purpose, we assume that these estimations are the ground-truth poses of the test set. Subsequently, we apply the same discriminative estimator over the corresponding reconstructed images, generated by the generative 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 poses directly but only the overall reconstruction quality either by using standard SSIM [25] or the IoU based score, which is specific to the setup of [20], based on reconstruction of segmentation masks. In summary, our PCK metric evaluation measures reconstructions accuracy explicitly considering the generated poses.

Regarding the PCK metric, our model reports 95.14% of accuracy, with PCK score at 0.5, and outperforms [20] by a large margin, which reports 70.89%, as shown is Fig. 8. Moreover, our approach also outperforms [20] w.r.t. the image quality, as can be seen in Tab. 2, which reports the PSNR and the SSIM scores for both methods. The results demonstrate good quality of reconstructions w.r.t. the human pose, suggesting that our Conditional-DGPose model benefits from the single-stage end-to-end approach, in contrast to the multiple stages of training and testing in [20].

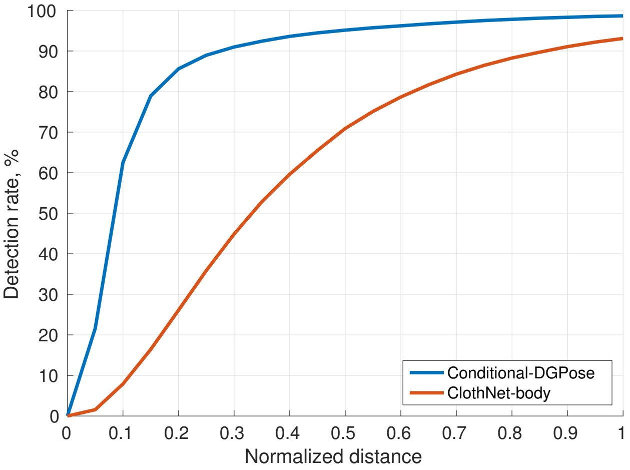
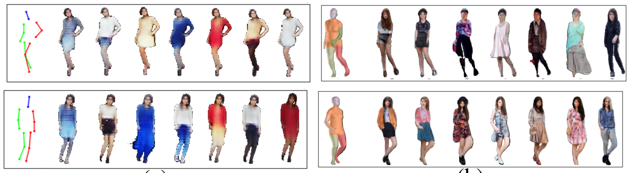


Figure 8: **ChictopiaPlus.** The PCK scores over reconstructed images of our Conditional-DGPose (*blue*) significantly outperforms the ClothNet-body [20] (*red*). Detection rate represents the percentage of joints corrected relocated in the reconstructed images.

	PSNR	SSIM
Conditional-DGPose (ours)	**21.33**	**0.88**
ClothNet-Body [20]	16.89	0.82

Table 2: Quantitative evaluation w.r.t. image quality. Our method outperforms [20] considering both metrics, the PSNR and the SSIM.

Qualitative results. In Fig. 9, we present qualitative sampling results from our model to demonstrate that it generates realistic images with accurate poses. In addition to that, in Fig. 10 (Sec. A), we compare images generated by the Conditional-DGPose to the ones by the ClothNet-body [20]. We note that, while both approaches are capable of generating people in the required poses, our approach performs better in terms of appearances, which are much closer to the original in our case. Even though both methods were able to generate visually good images and poses, the Conditional-DGPose was more accurate in capturing the locations of the body parts, particularly regarding limbs' extremities. This shows that even without the image-to-image translation network our method was able to generate realistic images.



(a) (b)

Figure 9: (a) Random samples from the Conditional-DGPose model for a fixed pose (leftmost image). (b) Random samples from the ClothNet-body [20] for a fixed pose (leftmost image), which are rendered as natural images by an image-to-image translation network [11] over the segmentation masks sampling.

6. Conclusions

In this paper, we have introduced the Conditional-DGPose, a conditional-VAEGAN deep generative model of people in natural images. Our model is conditioned on 2D human pose, allowing the disentangled representation of body posture and other factors of variation in the images. In contrast to other approaches in the literature, we model the problem in the high-dimensional image space. This allows us to generate image samples conditioned on human pose, in opposition to other methods which can only sample in a low-dimensional space (e.g. pose vectors or segmentation masks), relying on image-to-image translations for mappings to the image space. We have evaluated several design options and performed experiments specifically in the context of human pose. In the adopted benchmarks, by the generation of realistic images, our methodology has shown the capability of learning the underlying factors that jointly contribute to the generation of a human body in RGB images. We have successfully validated our model on the Human3.6M dataset and obtained state-of-the-art results in the ChictopiaPlus benchmark, outperforming the closest related method in the literature, the ClothNet-body architecture [20].

A. Appendix

Following, we show reconstructions on ChictopiaPlus (Fig. 10) and provide implementation details (Tabs. 3 and 4). Input heatmaps' channels correspond to: **i)** 14 joints (head top, neck; right shoulder, elbow, wrist, hip, knee and ankle; left shoulder, elbow, wrist, hip, knee and ankle; **ii)** 9 rigid parts (head; right upper arm, lower arm, upper leg and lower leg; left upper arm, lower arm, upper leg, and lower leg; **iii)** 1 body.

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

Table 3: Architecture of the residual block employed in the Conditional-DGPose encoder.

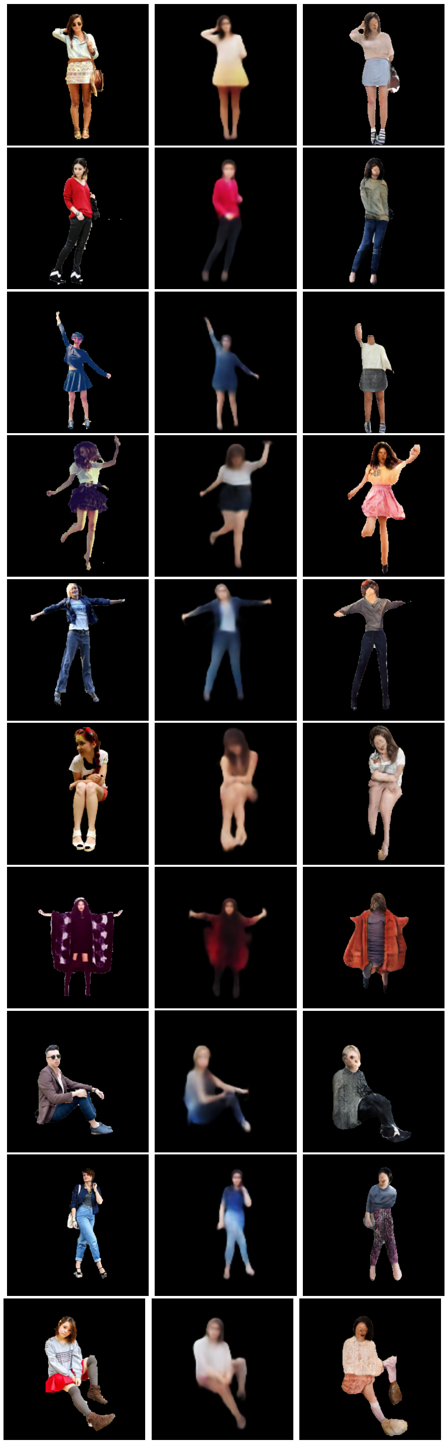


Figure 10: **Reconstructions**. In each row we have, respectively: original image (256 × 256), Conditional-DGPose and ClothNet-body [20] reconstructions. Notice, that the images generated by our model are much closer to the originals in terms of appearance (colours). Moreover, in general, the Conditional-DGPose captures the body parts' locations more accurately, which results in better quantitative results w.r.t. the pose reconstruction, shown in Fig. 8. Limbs' extremities are frequently lost in the ClothNet-body [20] reconstructions. Best viewed if zoomed in digital version.

Layer	Definition
Encoder	
Input: *images*(batch_size=64, channels=3, height=64, width=64) *labels*(batch_size=64, channels=24, height=64, width=64);	
Layer	**Definition**
1	CONCAT(*image*, *labels*)
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
μ	FC-(N100)
σ	FC-(N100)
Prior	
Input: *labels*(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: *sample*(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(*deconv6_output*, *labels*)
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)
$G(\mathbf{y}, \mathbf{z})$	CONV-(N3, K5, S1, P2), TANH
Discriminator	
Input: *decoder_output*(batch_size=64, channels=3, height=64, width=64); *images*(batch_size=64, channels=3, height=64, width=64)	
Layer	**Definition**
1	CONV-(N64, K4, S2, P1), LeakyReLU(0.2)
2	CONV-(N128, K4, S2, P1), BN, LeakyReLU(0.2)
3	CONV-(N256, K4, S2, P1), BN, LeakyReLU(0.2)
4	CONV-(N512, K4, S2, P1), BN, LeakyReLU(0.2)
5	CONV-(N1, K4, S1, P0), SIGMOID

Table 4: Conditional-DGPose architecture for 64 × 64 input images. We use the following abbreviations: N for 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 Tab. 3. The additional layers can be clearly understood. Finally, particular parameters for specific layers are defined between parenthesis after the layers' names.

Acknowledgements

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

References

[1] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018.

[2] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Cvae-gan: Fine-grained image generation through asymmetric training. In *ICCV*, 2017.

[3] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017.

[4] R. de Bem, A. Arnab, S. Golodetz, M. Sapienza, and P. Torr. Deep fully-connected part-based models for human pose estimation. In *ACML*, 2018.

[5] R. de Bem, A. Ghosh, T. Ajanthan, O. Miksik, N. Siddharth, and P. Torr. A semi-supervised deep generative model for human body analysis. In *ECCV 9th Workshop on Human Behavior Understanding*, 2018.

[6] D. J. Fleet. Motion models for people tracking. In *Visual Analysis of Humans*, pages 171–198. Springer, 2011.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*. 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[9] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.

[10] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[12] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACMMM*, 2014.

[14] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[16] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *NIPS*, 2014.

[17] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[18] A. Kleinsmith and N. Bianchi-Berthouze. Recognizing affective dimensions from body posture. In *ACII*, 2007.

[19] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016.

[20] C. Lassner et al. A generative model for people in clothing. In *ICCV*, 2017.

[21] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018.

[22] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *PAMI*, 2015.

[23] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[24] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 2015.

[25] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool. Pose guided person image generation. In *NIPS*, 2017.

[26] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *ICML*, 2017.

[27] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. In *ECCV*, 2016.

[28] G. Pons-Moll and B. Rosenhahn. Model-based pose estimation. In *Visual analysis of humans*, pages 139–170. Springer, 2011.

[29] T. M. Quan, D. G. Hildebrand, and W.-K. Jeong. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics. *arXiv preprint arXiv:1612.05360*, 2016.

[30] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

[31] N. Siddharth, B. Paige, A. Desmaison, J.-W. van de Meent, F. Wood, N. D. Goodman, P. Kohli, and P. Torr. Learning disentangled representations with semi-supervised deep generative models. In *NIPS*, 2017.

[32] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015.

[33] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *CVPR*, 2010.

[34] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *ICLR*, 2016.

[35] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.

[36] S. Tulyakov, M. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018.

[37] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017.

[38] C. Wan, T. Probst, L. J. V. Gool, and A. Yao. Crossing nets: Dual generative models with a shared latent space for hand pose estimation. In *CVPR*, 2017.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[40] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[41] M. Z. Weiyu Zhang and K. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013.

[42] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.