Adversarial Parametric Pose Prior
Supplementary Material

A. Derivation of SLERP in d-dim case

Let us denote two vectors of unit length:

$$\begin{cases} \mathbf{x}, \mathbf{y} \in \mathbb{S}^d \\ \|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1 \\ < \mathbf{x} \cdot \mathbf{y} >= \cos\theta, \quad \theta \in (0, \pi) \end{cases} \tag{1}$$

where $< \mathbf{x} \cdot \mathbf{y} >$ is the inner product between vector $\mathbf{x}$ and $\mathbf{y}$, with θ as the angle between them on a d-dimensional unit sphere $\mathbb{S}^d$. With slight abuse of notations, we use $\mathbf{x}$ to refer to both a point on a sphere, when talking about the sampled d-dimensional point, and also a vector from the origin of this sphere, when talking about trigonometric operations.

Let us call the point $\mathbf{z} \in \mathbb{S}^d$ lying on the interpolation path from $\mathbf{x}$ to $\mathbf{y}$ if:

$$\begin{cases} \|\mathbf{z}\|_2 = 1 \\ < \mathbf{z} \cdot \mathbf{x} >= \cos\phi, \quad \phi \in [0, \theta] \\ < \mathbf{z} \cdot \mathbf{y} >= \cos(\theta - \phi) \end{cases} \tag{2}$$

Then $\mathbf{z}$ can be found via the Spherical Linear Interpolation (SLERP) formula:

$$\mathbf{z} = \frac{\sin(\theta - \phi)}{\sin\theta}\mathbf{x} + \frac{\sin\phi}{\sin\theta}\mathbf{y} \tag{3}$$

It can be shown through rather simple but rigorous trigonometry that Eq. 3 does indeed satisfy all the properties in 2.

Initially, SLERP arose in the task of 3D rotations for solid objects [8] and Eq. 3 can be naturally derived from the 4D rotation intuition via quaternions. However, Eq. 3 is still valid for d-dimensional vectors, even though quaternion intuition is not applicable anymore. Next, we prove interpretation used in this equation.

Derivation

In our derivation we will only base on properties of Eq. 2. One can always search for $\mathbf{z}$ as the linear combination of $\mathbf{z}_{\parallel}$, the part that lies in the span of ($\mathbf{x}$, $\mathbf{y}$) (dimensionality 2), and $\mathbf{z}_{\perp}$, the part that is orthogonal to it (dimensionality $d-2$):

$$\mathbf{z} = \mathbf{z}_{\parallel} + \mathbf{z}_{\perp} = \alpha\mathbf{x} + \beta\mathbf{y} + \mathbf{z}_{\perp} \tag{4}$$

Using Eq. 2, one can find the connection between α and β coefficients:

$$\begin{aligned} \cos\phi &= < \mathbf{z} \cdot \mathbf{x} > \\ &= < (\alpha\mathbf{x} + \beta\mathbf{y} + \mathbf{z}_{\perp}) \cdot \mathbf{x} > \\ &= \alpha < \mathbf{x} \cdot \mathbf{x} > + \beta < \mathbf{y} \cdot \mathbf{x} > + < \mathbf{z}_{\perp} \cdot \mathbf{x} > \\ &= \alpha + \beta\cos\theta + 0 \\ \cos(\theta - \phi) &= < \mathbf{z} \cdot \mathbf{y} > \\ &= < (\alpha\mathbf{x} + \beta\mathbf{y} + \mathbf{z}_{\perp}) \cdot \mathbf{y} > \\ &= \alpha < \mathbf{x} \cdot \mathbf{y} > + \beta < \mathbf{y} \cdot \mathbf{y} > + < \mathbf{z}_{\perp} \cdot \mathbf{y} > \\ &= \alpha\cos\theta + \beta + 0 \end{aligned} \tag{5}$$

The equations in the system 5 are independent and linear with respect to α and β, hence, we can solve it and obtain the desired coefficients:

$$\begin{aligned} \alpha &= \frac{\sin(\theta - \phi)}{\sin\theta} \\ \beta &= \frac{\sin\phi}{\sin\theta} \end{aligned} \tag{6}$$

Now, let us look at the norm of the vector $\mathbf{z}$. As the components $\mathbf{z}_{\perp}$ and $\mathbf{z}_{\parallel}$ are orthogonal to each other, the squared norm of $\mathbf{z}$ is the sum of the two:

$$\|\mathbf{z}\|^2 = \|\mathbf{z}_{\perp}\|^2 + \|\mathbf{z}_{\parallel}\|^2 = 1 \tag{7}$$

The norm of the $\mathbf{z}_{\parallel}$ can be shown to be equal to 1 (explicitly using computed coefficients of 6):

$$\|\mathbf{z}_{\parallel}\|^2 = \|\alpha\mathbf{x} + \beta\mathbf{y}\|^2 = \alpha^2 + \beta^2 + 2\alpha\beta\cos\theta = ... = 1 \tag{8}$$

From 7 and 8, one can conclude that the norm of $\mathbf{z}_{\perp}$ is 0. Hence, the interpolant point $\mathbf{z}$ does indeed lie in the two-dimensional hyperspace of $\mathbf{x}$ and $\mathbf{y}$.

One can reparameterize the angle ϕ as $\frac{t}{T}\theta$ where t goes from 0 to T (as in the formula in the main paper). Then the unit vector $\mathbf{z}$ can be seen as smoothly "rotating" from $\mathbf{x}$ to $\mathbf{y}$ in d-dimensional space.

B. Ablation on *Recall* experiments

Sampling rarer poses from the data. Sampling uniformly across the ground-truth data can cause a bias towards mean poses, hence, hindering recall metrics. To overcome this issue and apply the same sampling for all prior distributions, one can sample from the t-SNE embedding space, as

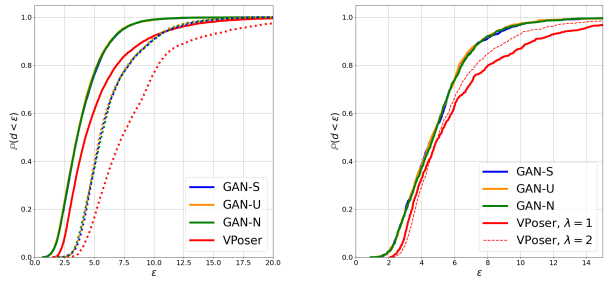


Figure 1. Various *Recall* strategies. Higher means better in all charts. *Left*: Results of the main paper, experiments with data from the *Train* set are drawn with solid lines, and from the *Test* set with dashed lines. Datapoints are sampled uniformly. *Right:* 1) Sampling from t-SNE embedding space instead of uniform sampling from the (*Train*) dataset does not affect the order between the models (solid lines). 2) Tuning covariance coefficient while sampling from the latent space of unbounded models indeed improves Recall for VPoser (at $\lambda = 2$, dashed line), although does not bring any noticeable change for GAN-N model.

it would diminish the clustering effect and let rare poses be sampled more often. We repeat the *Recall* experiment with the aforementioned technique. We search for the nearest neighbor for each uniformly sampled point in t-SNE space. The results are shown in Fig. 1 with solid lines. Compared to results from the main paper, we see that models have lower recall but the order between the models remains the same.

Sampling rarer poses from the latent space. In the main paper, for VPoser we use the standard sampling from a normal distribution (with the covariance matrix λI, where $\lambda = 1$), as also done in the original paper [6]. However, the sampling strategy can affect the diversity of the obtained samples. For example, sampling with the covariance $\lambda > 1$ would give rarer poses more often and vice versa. We experimented with sampling for unbounded models from $\mathcal{N}(0, \lambda I)$ with λ values in the range $[0.5, 100]$. We plot the results for the best λ value in Fig. 1. Sampling with higher λ picks unusual poses more often, increasing Recall up to a point. We found that $\lambda = 2$ — denoted by a dashed line in the figure — yields the best Recall in terms of mean and median statistics for the VPoser model. Yet our GAN model still dominates. Increasing λ for the GAN-N model does not bring any noticeable improvement.

C. Interpolation sequences

In Fig. 2 we show more interpolation sequences for different generative models: GAN-based GAN-S, GAN-U, GAN-N and VAE-based VPoser [6]. The procedure of sampling and interpolation is the same as in the manuscript. Also, in Fig. 3 we provide the transition distances that correspond to the interpolations in Fig. 2. For GAN-based models transition distances for the sampled pairs lie in the range $[10^{-3}, 10^{-2}]$ mm (per-vertex distance), which is in much smaller range than VPoser. Note that the average result for all samples is provided in the main paper (Fig.5). It is clear that VAE-based VPoser [6] applies most of the transition either at the beginning or at the end of interpolations, while GAN-based models, especially GAN-S and GAN-U, provide smooth interpolations.

SLERP for unbounded latent spaces. In the main paper, while experimenting with interpolation in unbounded models, VPoser and GAN-N, we perform linear interpolation. With non-zero probability, the linear interpolation between the two vectors passes close to 0. While in the reality during training, it is highly unlikely to sample the vectors close to zero vector. Even more than that, the norm of the Gaussian vector is distributed according to the χ^2_d distribution, which has a non-zero mean. It implies that such high-dimensional vectors must concentrate around thin spherical shells (with radii close to the mean) and have low probability of living around zero point. It might be then concluded that for the unbounded models used in the present paper, VPoser and GAN-N, the linear interpolation (e.g., through close-to-zero space) is less sensible than the spherical interpolation (through denser regions along spherical shells).

We carry out the corresponding experiment, comparing interpolation smoothness for unbounded models with spherical and with linear interpolation. Fig. 4 provides the comparison between linear and spherical interpolations. Contrary to the chi-squared intuition, the linear interpolation results for unbounded models are in fact smoother than the SLERP ones.

It is worth noting that it is not trivial to explore the structure of high-dimensional spaces and explain why linear interpolation works better in this case. We also observe interpolations in VAE tend to stick to one pose and then jump to another. Our guess is that the continuous and smooth behaviour of neural network functions might fill the gaps between feasible points to create poses in all cases, however, the generated poses might not be transitioning linearly from one pose to another. We leave a provable clarification of this phenomenon for future research.

Why GAN interpolation is smoother than VAE. The results of this experiment both quantitatively (Table in the main paper) and qualitatively (Fig. 2) indicate that VAE's interpolations are much more abrupt than the ones of proposed GAN models (especially, GAN-S). VAEs behave like a step-function when interpolating between poses as the generated samples can jump suddenly from one pose to another. We believe this is due to the training mechanism of the decoders in VAEs, which maps multiple random samples, through noise injection, to the same exemplar. By contrast, in GANs the generated pose does not have a fixed tar-

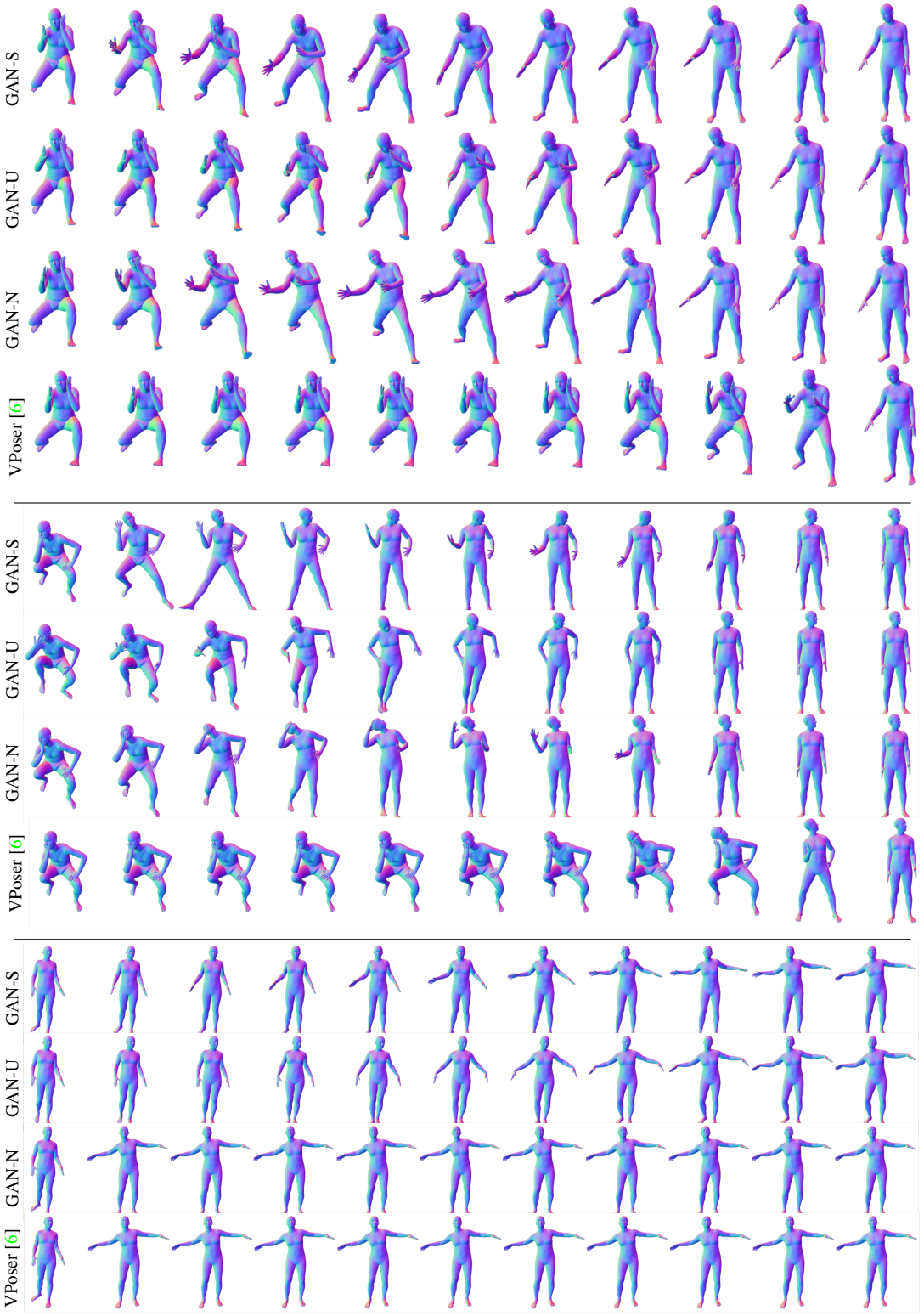
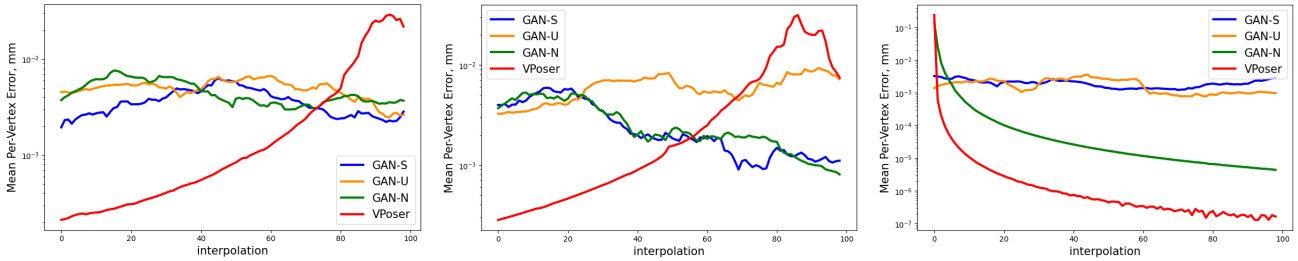


Figure 2. Examples of interpolations for different generative models, 10 intermediate interpolants each. Corresponding transition distances (100-step interpolations) are provided in Fig. 3.



(a) Transition distances for the sequence in Fig. 2 *top*. Smooth for GANs, abrupt for VAE.

(b) Transition distances for the sequence in Fig. 2 *middle*. All GANs provide hardly smooth yet physically plausible interpolation, while VPoser [6] smoothly moves the *head* of the body and then "jumps" to the final pose.

(c) Transition distances for the sequence in Fig. 2 *bottom*. GAN-N model demonstrates less smooth interpolations than bounded GAN models, closer to VPoser [6] and locally stucks at one pose (transiton distances $\sim 10^{-5}$ mm per vertex).

Figure 3. Mean per-vertex transition distances of different generative models for the samples presented in Fig. 2. Each interpolation path takes 100 steps from start to end, while the corresponding images in Fig. 2 show only each 10-th step. It is clear that VAE-based VPoser [6] provides abrupt transitions, while GAN-based models provide smoother transitions in the output space.

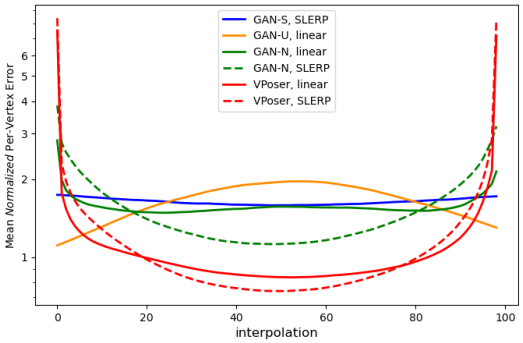


Figure 4. Average normalized mesh interpolations. Applying linear interpolation to samples from Gaussian distribution (for VPoser and GAN-N models) is superior over the spherical sequencing. Solid lines are taken from the main paper.

get as the model is free to map the latent to any meaningful sample from the pose distribution, validated by the discriminator. This yields a better latent organization. In GAN-S, this is further enforced by the structure of the latent space, which is uniform and isotropic.

D. Image-to-Mesh regression

We demonstrate qualitative results of injecting our module GAN-S into the pretrained HMR architecture and finetuning it on recent in-the-wild pseudo ground-truth dataset COCO_{EFT} [2] (see Sections 3.3 and 4.4 in the main paper for details). We show the worst predictions of our model on H3.6M [1] validation subset, chosen according to *Protocol 2* (corresponding to the results in Table 4 of the main paper).

E. GANs for shape parameters

The SMPL model [4] is parametrized by two sets of parameters: pose Θ and shape β. In the main paper, we explore the properties of the former for a fair comparison with VPoser [6] that only explores the pose prior, while the latter stays out of context. Despite having relatively constrained shape parameters in SMPL, by using a data-driven PCA model, the unbounded nature of PCA does not restrict the SMPL output to always be plausible.

Luckily, the GAN-based approach that we used for poses can easily be applied for shapes as well. The only difference is the choice of architecture for the shape discriminator.

To provide a joint model for pose and shape, we explore two GAN variants (using spherical input space prior):

- Disentangled "shape-only" ("GAN-β") and "pose-only" ("GAN-Θ") models.
- Entangled "shape+pose" ($\mathbf{z} \mapsto \beta, \Theta$), denoted as "GAN-$\beta\Theta$".

In the first case, we train an independent shape model GAN-β and use it together with GAN-Θ, while in the second case, we train both jointly, using a shared $\mathbf{z}$ space. Note that in the main paper all GAN-based models are of a kind "GAN-Θ", as they map input points to the SMPL-pose space.

It is important to note that for training GAN-{S,U,N} models we used AMASS [5] dataset, which contains an abundant number of body poses. However, the number of subjects (different body shapes) is very few (346 in total). To overcome this issue, for training shape-oriented models we choose another dataset, SURREAL [9], which is composed of synthetic SMPL bodies. In SURREAL, the shapes β are sampled from CAESAR dataset [7], which contains about $4k$ different shapes. This number is still very limited, compared to tens of millions of Θ-poses in AMASS, however, we are still able to train generative models as a proof of concept.

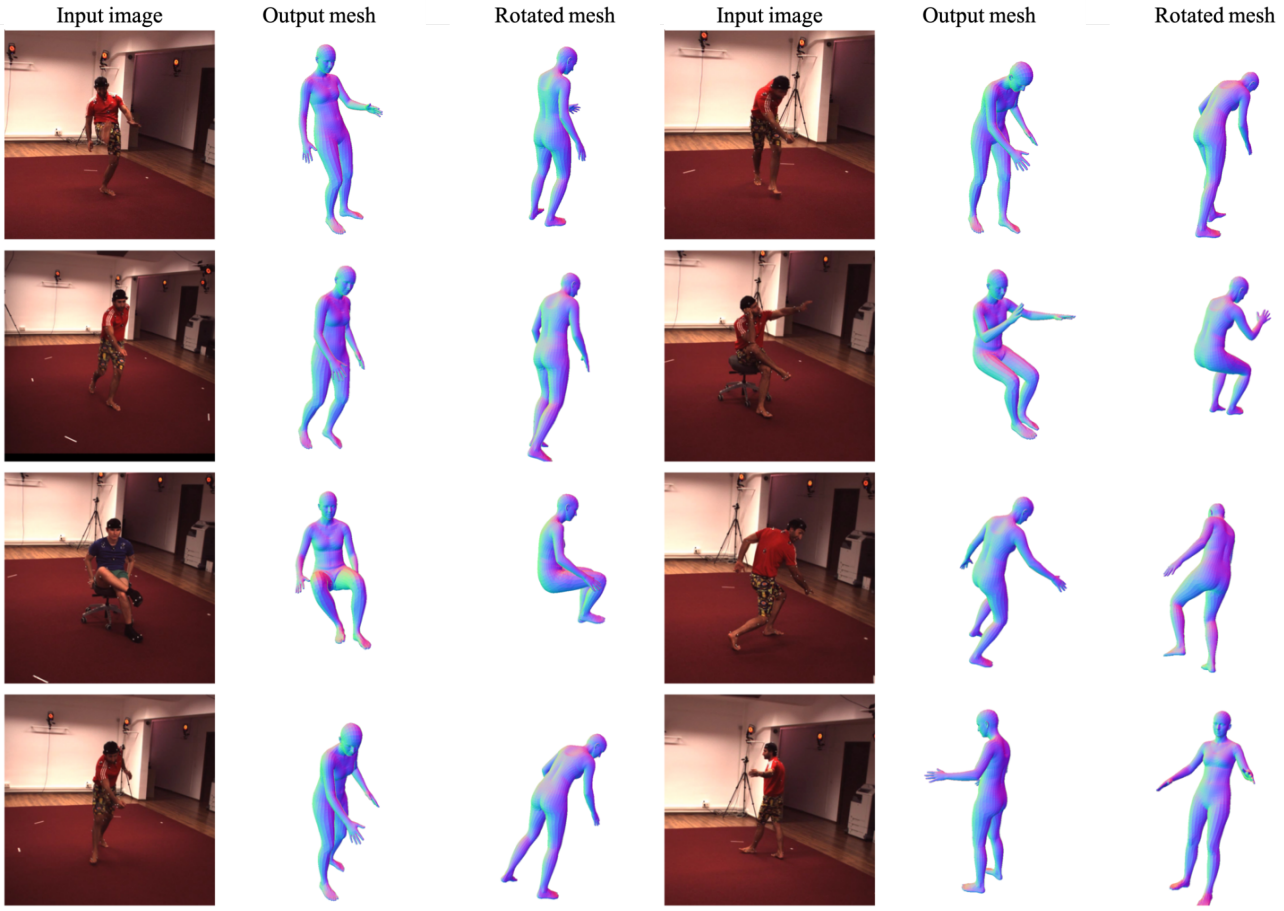


Figure 5. Mesh predictions in Section 4.4 experiments of the main paper with the highest ("worst") P-MPJPE on the subset of H3.6M [1] validation set (according to H3.6M *Protocol 2*). Examples are split in triplets: *left* - the input image, *middle* - the predicted mesh (with predicted rotation), *right* - the same mesh rotated by 90° along vertical axis. The error decreases *from left to right* and *top to bottom* (*top-left* sample obtains the highest P-MPJPE value).

E.1. Disentangled GAN-β and GAN-Θ

Trained GAN-β ($\mathbf{z}_1 \mapsto \beta$) coupled with trained GAN-Θ model ($\mathbf{z}_2 \mapsto \Theta$, from the main paper) might serve as a full prior of the original SMPL [4] model. This approach learns disentangled priors for pose and shape, in the same spirit as SMPL that represents pose and shape parameters independently. However, the realistic pose and shape are *not* completely disentangled. It means that pose and shape independently may account for plausible humans but combined together give a body with self-interpenetrations, as illustrated in Fig. 7.

As for the architecture of the shape discriminator, we follow HMR [3] and use a 2-layer MLP.

E.2. Entangled GAN-$\beta\Theta$

To train a GAN for pose and shape together, we use the discriminators of GAN-Θ (GAN-S in the main paper) and GAN-β together, and train a generator with a shared input space $\mathbf{z}$. In this model, we also use a "$\beta + \Theta$" discriminator that penalizes the full generated SMPL vector. In total we have $K + 1(\text{pose}) + 1(\text{shape}) + 1(\text{pose+shape})$ discriminators.

The examples of random interpolations in the input space of $\mathbf{z}$ can be found in Fig. 6. As pose and shape are entangled together, it prevents utilizing such a mixed model in applications where one of these characteristics needs to remain fixed.

At the same time, not every sample corresponds to a plausible body (even with self-interpenetrations allowed). Our experiments show that generated samples might not look as humans at all. We demonstrate it in Fig. 8 with corresponding t-SNE visualizations for each GAN model.

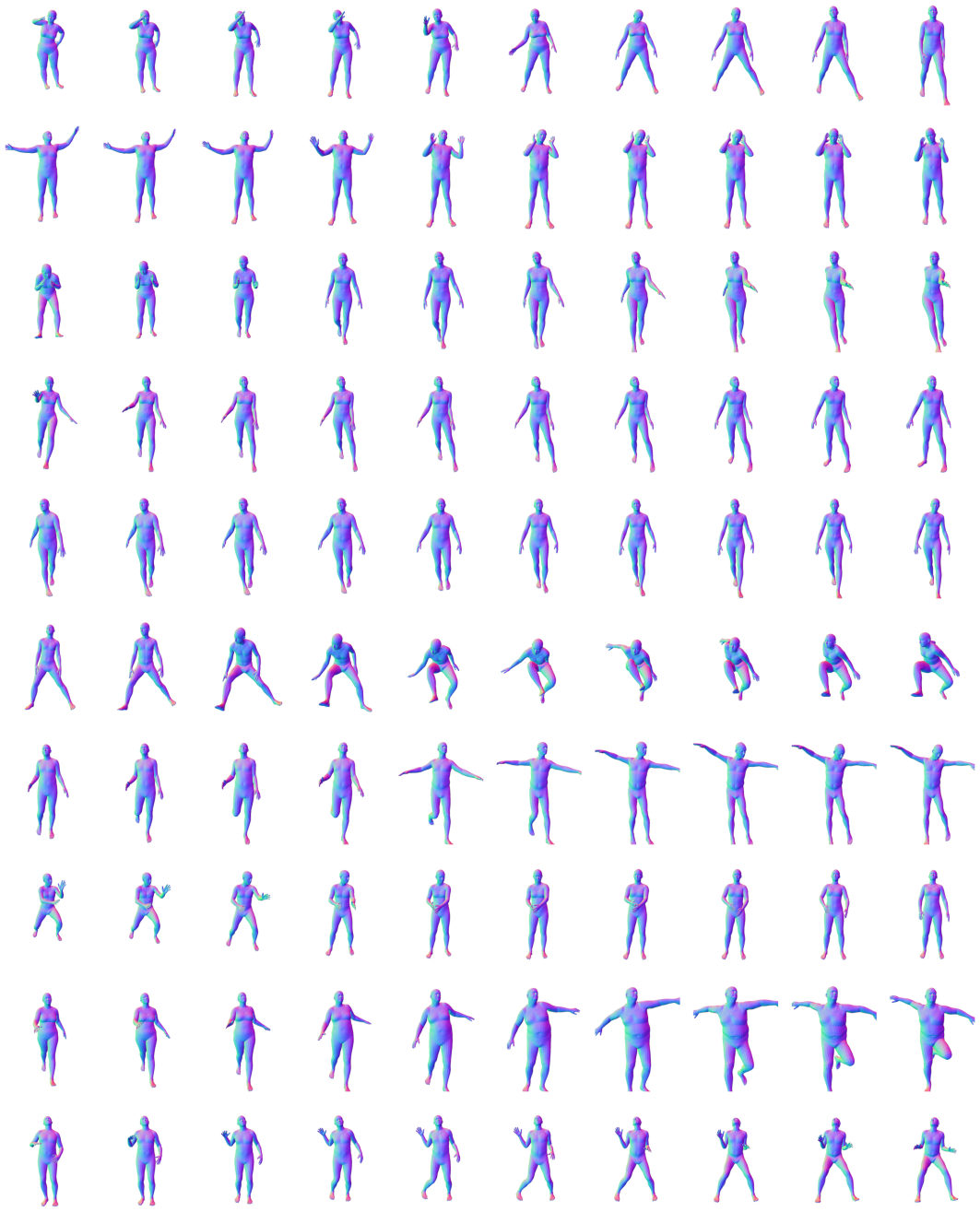


Figure 6. Examples of interpolating between random points in the latent space of GAN-$\beta\Theta$ mixed model. Pose and shape are entangled with each other, which complexifies the usage of such a mixed model in applications.

F. Future work

In most situations, independently sampling pose and shape parameters will result in realistic bodies. However, this is not always the case. Consequently, the task of generating plausible bodies when mixing shape and pose together needs to be further investigated. It might be resolved, for example, by using the formulation of a conditional GAN, where generating pose depends on some shape features or vice versa. We explore these aspects in our further work.

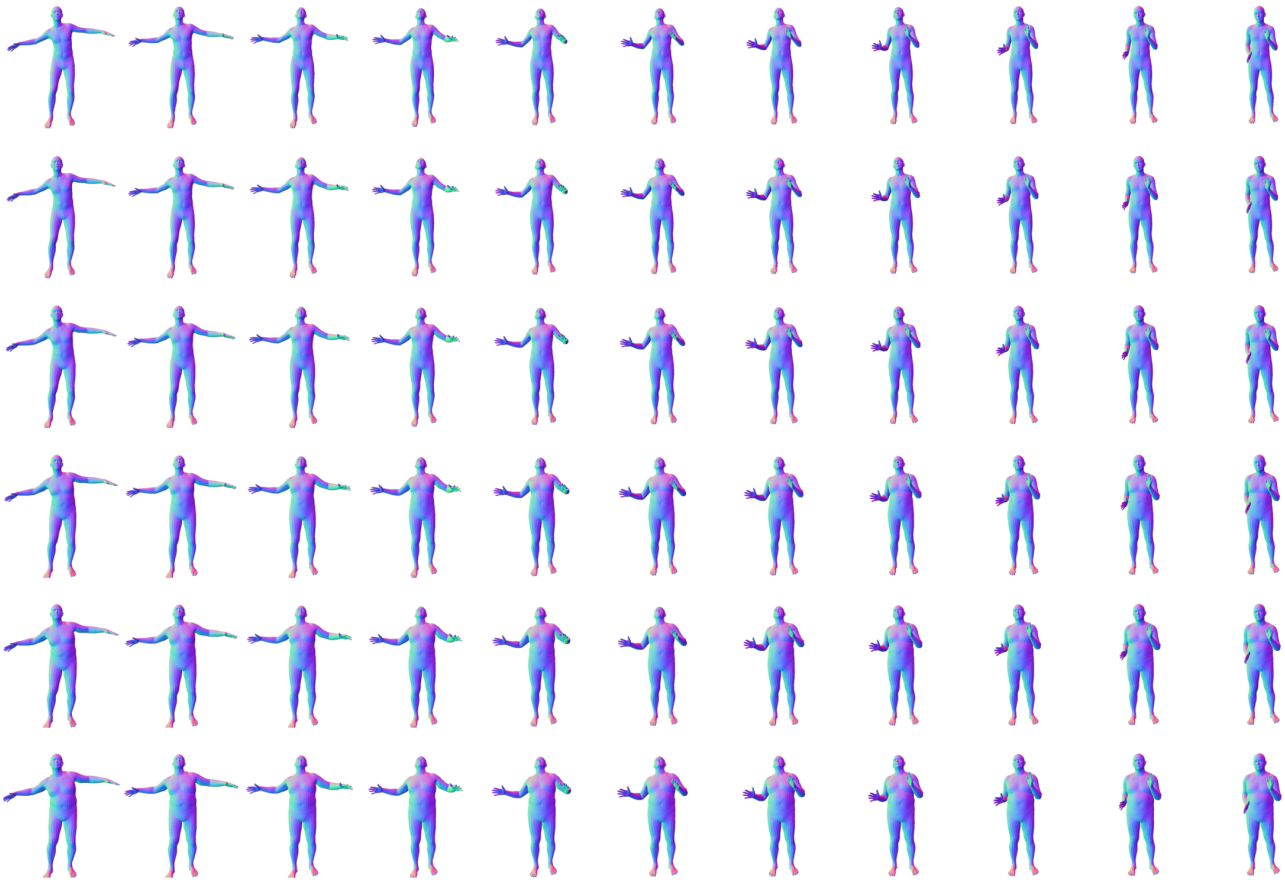
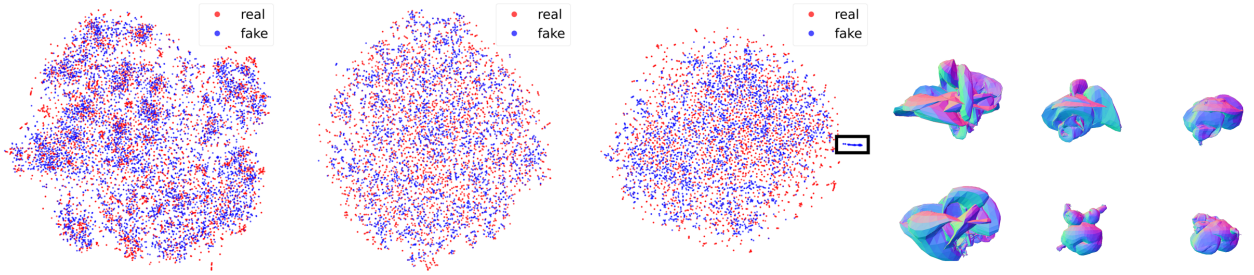


Figure 7. Example of generating bodies via two independent models, GAN-Θ and GAN-β. *Pose* varies along axis x, *Shape* varies along axis y. Having indepedent latent models allows to disentangle shape and pose for generating body. However, as pose and shape are in fact dependent, it might lead to generating implausible bodies (see *last column, bottom rows*).



(a) t-SNE on GAN-Θ (pose only, main paper).

(b) t-SNE on GAN-β (shape only).

(c) t-SNE on GAN-$\beta\Theta$ (pose and shape). The cluster of outliers (*black rectangle*) corresponds to implausible bodies.

(d) Unrealistic samples of GAN-$\beta\Theta$ (from the *black rectangle* in (c)).

Figure 8. t-SNE of samples from the entangled GAN-$\beta\Theta$ (c) identifies a cluster of generated points that are not consistent with real samples. This cluster consists of outliers, shown in (d), that correspond to unrealistic bodies. t-SNE plots for independent models, GAN-Θ (a) and GAN-β (b), demonstrate consistent data coverage.

References

[1] C. Ionescu, I. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 4, 5

[2] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020. 4

[3] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-To-End Recovery of Human Shape and Pose. In *Conference on Computer Vision and Pattern Recognition*, 2018. 5

[4] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M.J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM SIGGRAPH Asia*, 34(6), 2015. 4, 5

[5] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 4

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 3, 4

[7] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 2017. 4

[8] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985. 1

[9] G. Varol, J. Romero, X. Martin, N. Mahmood, M. Black, I. Laptev, and C. Schmid. Learning from Synthetic Humans. In *Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017. 4