Supplementary Material
3D Shape Variational Autoencoder Latent Disentanglement via Mini-Batch Feature Swapping for Bodies and Faces

Simone Foti Bongjin Koo Danail Stoyanov Matthew J. Clarkson
University College London
s.foti@cs.ucl.ac.uk

1. Mesh Operators

Since traditional neural network operators are not well suited for the non-Euclidean nature of meshes, we rely on spiral++ convolutions [16] and on the sampling operators defined in [35].

The creation of spiral sequences is at the core of the adopted convolution. Spirals are a simple yet effective approach to aggregate neighbouring mesh vertices into ordered sequences. Given a vertex, the spiral sequence is obtained by arbitrarily selecting one neighbour and following a clockwise spiral until the spiral length is reached. The receptive field of these convolutional operators can be expanded dilating the spirals (i.e. not selecting certain vertices along the sequence). Denoting by $\mathcal{S}(n, l)$ the spiral centred at vertex n with length l, the convolution at layer k is defined as:

$$\mathbf{x}_n^{(k)} = \text{MLP}^{(k)}\Big(\mathop{\Big\|}_{j \in \mathcal{S}(n,l)} \mathbf{x}_j^{(k-1)}\Big)$$

where $\|$ is the concatenation operation over the vertices in the spiral $\mathcal{S}(n, l)$, $\mathbf{x}_n^{(k)}$ are the vertex features at layer k, and MLP is a multilayer perceptron. Note that spirals are fixed during training because they are pre-computed only once for all vertices.

Pooling and un-pooling operators are matrix multiplications between the vertex features of a given layer and a sparse matrix. The sparse matrices are both pre-computed during a mesh simplification procedure that iteratively contracts the two vertices with the smallest quadric error. In particular, the pooling matrix $Q_d \in \{0, 1\}^{N_{k+1} \times N_k}$ is a sparse matrix where $Q_d(p, q) = 1$ if vertex q has been preserved during quadric sampling and $Q_d(p, q) = 0$ otherwise. The un-pooling matrix $Q_u \in \mathbb{R}^{N_k \times N_{k+1}}$ leaves the preserved vertices unchanged by setting $Q_u(q, p) = 1$. Contracted vertices are expressed in barycentric coordinates with respect to the closest preserved triangle, and then their corresponding elements in Q_u are set to the barycentric weights. This allows to restore the contracted vertices.

2. Latent Space Interpolation

We performed two latent interpolation experiments. Fig. 7 shows the effect of interpolating $\mathbf{z}$ between the latent representation of two different shapes. Fig. 8 shows the effects of changing each $\mathbf{z}^\omega$ of one shape with the corresponding $\mathbf{z}^\omega$ of the other shape, which is equivalent to progressively replacing features of the initial mesh with those of the target mesh. The interpolation experiment of Fig. 8 is better represented in the *supplementary video* [1]. The video also shows the interpolation between each pair of $\mathbf{z}^\omega$. The two experiments prove that our method creates a smooth latent space where per-feature modifications are possible.

The *supplementary video* also shows per-variable latent interpolation experiments for all different methods. Interestingly, while intermediate faces generated with our method are a plausible interpolations between the initial and target shape, intermediate faces generated with other methods often belong to substantially different identities.

3. Random Generation and Latent Disentanglement

For each method we report a more comprehensive set of randomly generated samples (Fig. 9) than those already depicted in Fig. 3. Then, we show the full latent disentanglement experiments detailed in Sec. 4. In particular, Fig. 10 and the *supplementary video* extend Fig. 3 by showing for each $\mathbf{z}^\omega$ the effects caused by traversing its latent variables. Similarly, when our method is trained on bodies, Fig. 11 extends Fig. 5D. Finally, Fig. 12 shows the effects of traversing each latent variable of VAE ($\beta = 1e^{-2}$), VAE ($\beta = 1e^{-4}$), DIP-VAE-I, DIP-VAE-II, and Factor VAE. Since these methods do not have a structured latent representation, it is not possible to distinguish different $\mathbf{z}^\omega$ like in Fig. 10.

[1] The supplementary video is available at the following link https://youtu.be/w9WF0mZe1ig

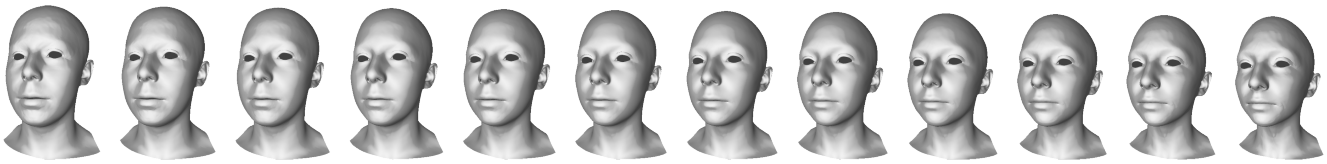

Figure 7. Latent interpolation experiment. An initial and a target shape are selected from the test set. Then, their latent representation $\mathbf{z}$ is computed by feeding the shapes in the encoder network E. 10 intermediate latent vectors are thus computed by linearly interpolating all the latent variables. The shapes generated from these latent vectors smoothly transition from the initial (leftmost shape) to the final (rightmost shape) shape

.

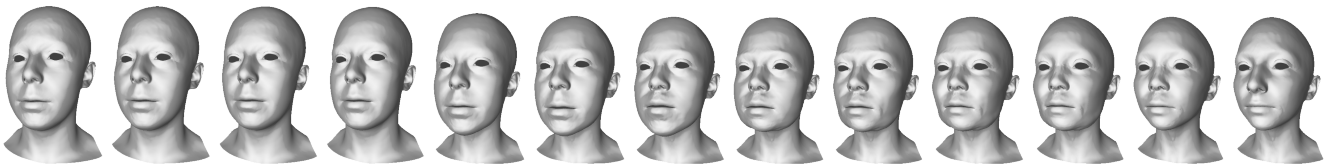

Figure 8. Per-feature latent interpolation experiment. Given the same initial and target latent vectors used in Fig. 7, subsets of the latent representation corresponding to different features ($\mathbf{z}^\omega$) are progressively replaced. In fact, the first face is the initial face, the second face is obtained by replacing the values of the $\mathbf{z}^\omega$ controlling the eye region of the initial face with those controlling the eye region in the target shape. The third face is obtained from the second face by replacing the $\mathbf{z}^\omega$ controlling the ears. Then, the subsequent shapes are obtained replacing the $\mathbf{z}^\omega$ controlling: temporal, neck, back, mouth, chin, cheeks, cheekbones, forehead, jaw, and nose. Each shape is obtained starting from the one on its left, and therefore the last one also corresponds to the target shape.

4. Ablation Study

One of the strengths of our method is its intuitiveness and the small number of changes required to convert a VAE into our method. The mini-batch feature swapping and the latent consistency loss are indeed the only changes required, but we decide to analyse also two important components that characterise our implementation of the VAE: the Laplacian regulariser used in the loss function and the instance normalisation. The ablation study, whose results are depicted in Fig. 13, is performed re-training the proposed model with the necessary modifications. When κ is set to 0, $\mathcal{L}_c$ is ignored (see Eq. 3). Despite this appears to be equivalent to the VAE, note that in this case the mini-batch feature swapping is performed. Observing the latent perturbations in Fig. 13 (No z Cons) we see the importance of the latent consistency loss. As expected, curating only the mini-batching does not allow to obtain a structured and disentangled latent space. Observing the random samples depicted in Fig. 13 (No Lapl) and obtained from the proposed method re-trained with $\alpha = 0$, we notice that the contributions of $\mathcal{L}_L$ are more subtle. Nevertheless when this term is removed we notice a more irregular surface as well as some surface discontinuity (e.g. top part of the head in the first sample or neck of the second sample). Finally, from Fig. 13 (No Norm) we observe the importance of the normalisation, which helps the generation of realistic faces.

5. Generalisation Capabilities

We evaluate the ability of our model to generate meshes outside the training data distribution by fitting all the CoMA subjects in their neutral expressions. Starting from the mean latent representation, we iteratively generate new meshes. For the first 80 iterations we optimise $\mathbf{z}$ with a mean squared error over 24 manually selected facial landmarks, for the remaining 170 iterations we use a Chamfer distance between the vertices generated with our model and those of the target mesh from CoMA. Also for this experiment we use the ADAM optimiser setting the learning rate to $lr = 5e^{-3}$. Errors are computed as per-vertex distances between each generated vertex and the closest vertex of the target mesh. To evaluate the robustness to noise we repeat the experiment perturbing the target vertices of meshes from CoMA with different amounts of random noise. When noise is applied, errors are computed with respect to the target without noise. As we could expect, in Fig. 14, we show that errors increase linearly with the amount of noise applied for all methods. Nevertheless, errors remain low, thus proving good generalisation capabilities. In addition, our method is the one with the lowest errors despite all methods were trained on the same dataset.

Figure 9. Random sample generation for VAE ($\beta = 1e^{-2}$), VAE ($\beta = 1e^{-4}$), DIP-VAE-I, DIP-VAE-II, and Factor VAE.

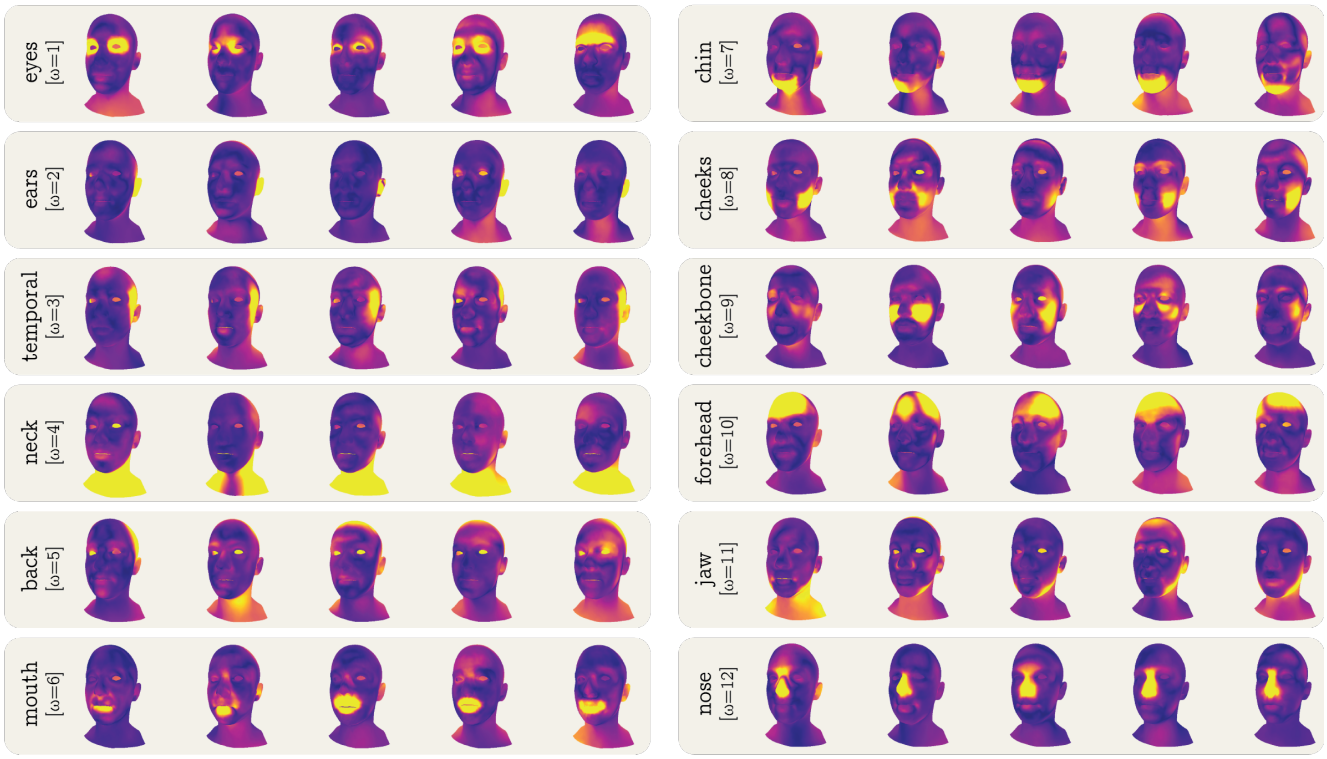

Figure 10. Complete latent traversals of proposed method trained on faces.

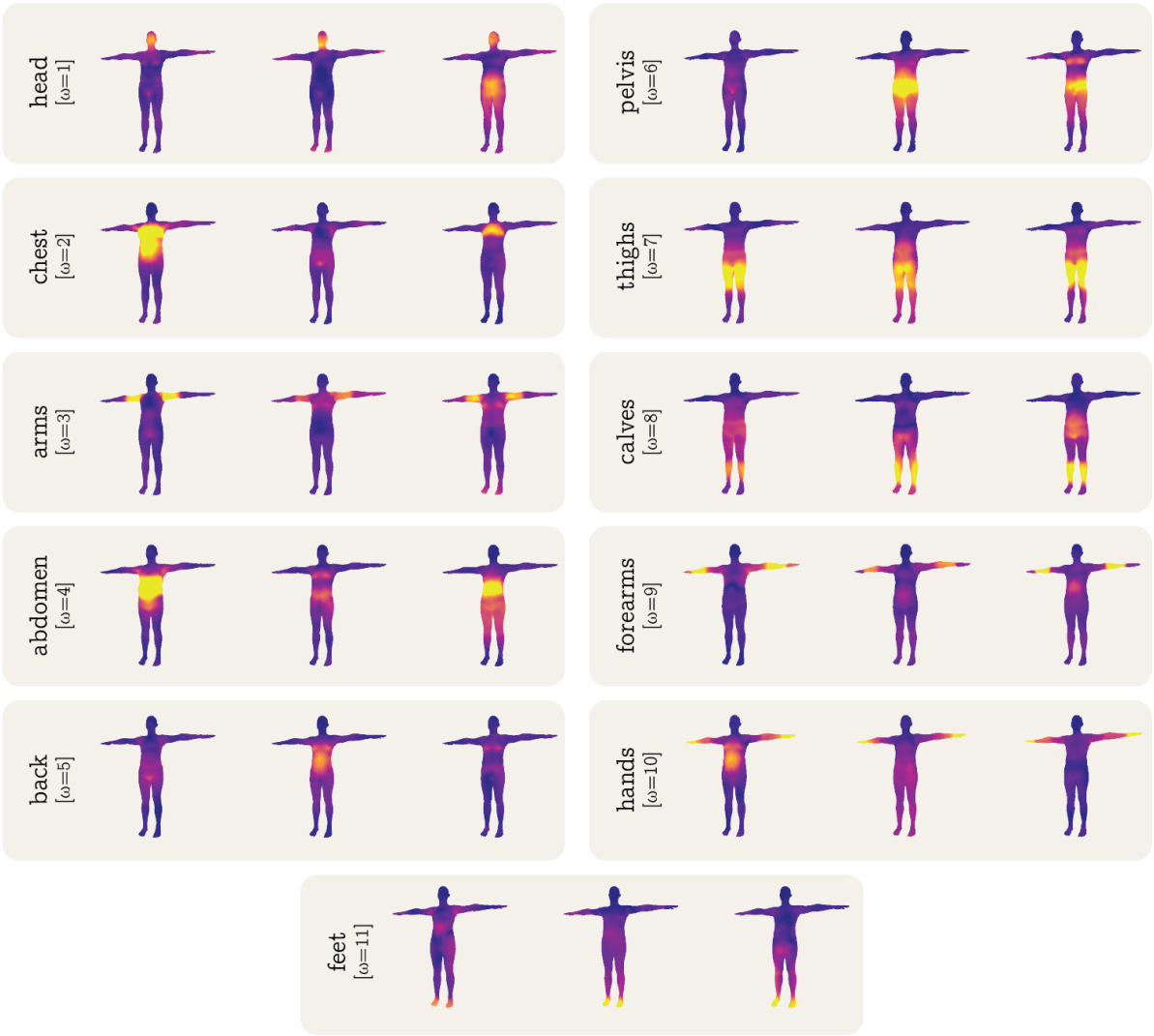

Figure 11. Complete latent traversals of proposed method trained on bodies.

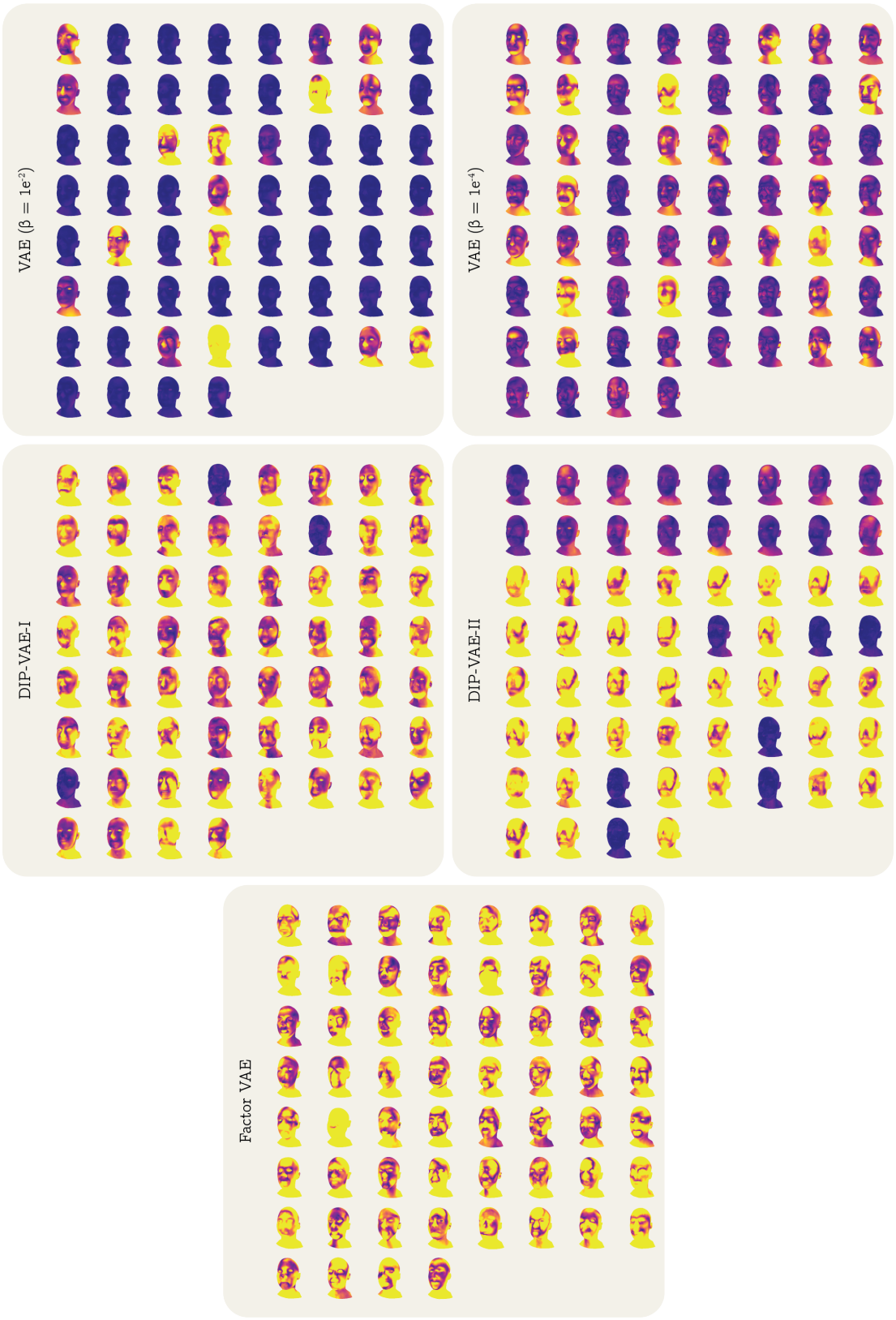

Figure 12. Complete latent traversals for VAE ($\beta = 1e^{-2}$), VAE ($\beta = 1e^{-4}$), DIP-VAE-I, DIP-VAE-II, and Factor VAE.

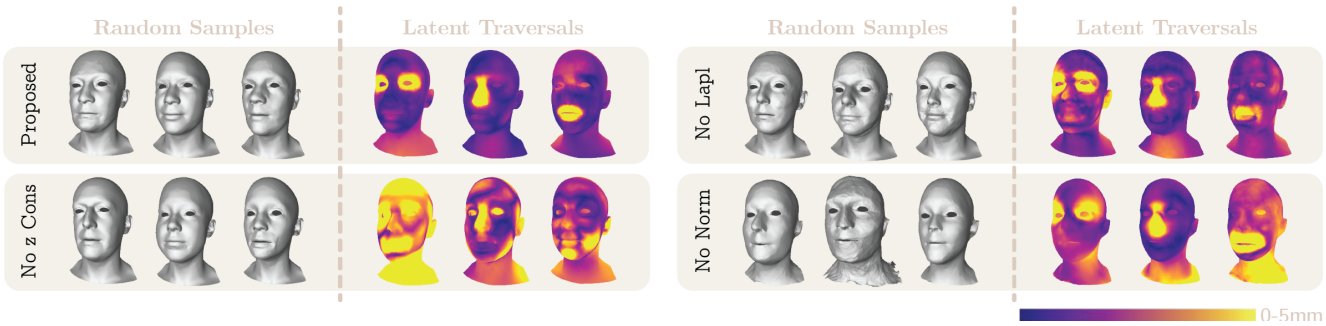

Figure 13. Ablation Study. The proposed method is ablated to examine the effects of the Laplacian regulariser, of the latent consistency loss, and of the instance normalisation. To observe how each of them contributes to the definition of the proposed method we show random samples and vertex-wise distances representing the effects of traversing three randomly selected latent variables.

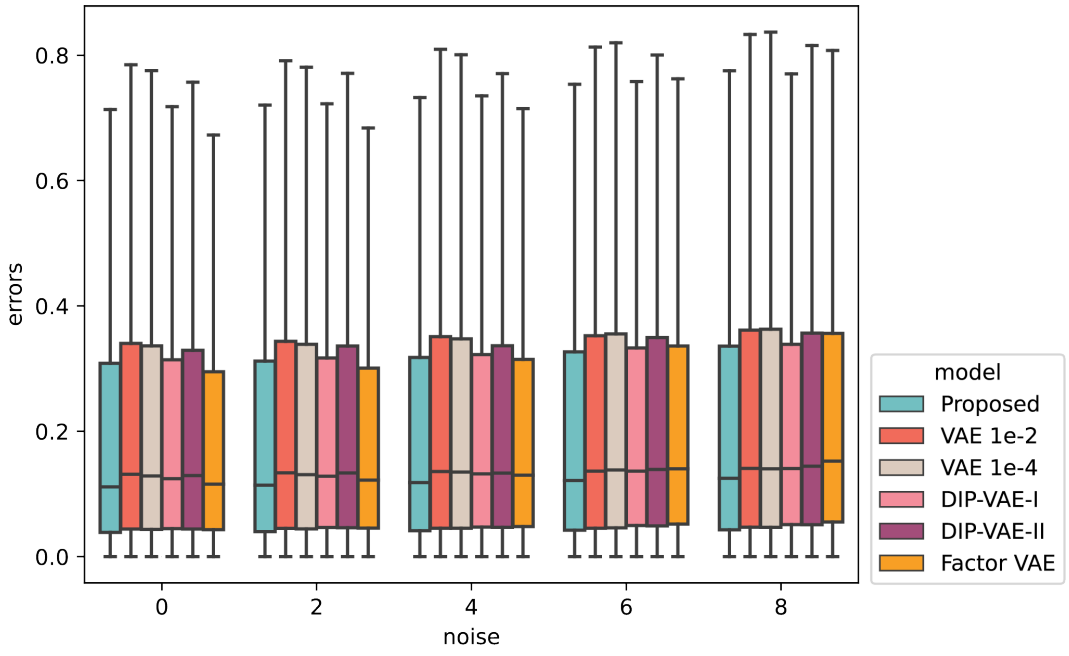

Figure 14. Generalisation capabilities evaluated by fitting CoMA subjects in neutral expressions with the proposed method as well as VAE ($\beta = 1e^{-2}$), VAE ($\beta = 1e^{-4}$), DIP-VAE-I, DIP-VAE-II, and Factor VAE.

6. Societal Impact

Our work focuses on the generation of 3D shapes of bodies and faces, but shapes without textures and materials are far from being realistic. For this reason, we believe that our work does not raise disinformation or immediate security concerns. Nevertheless, it still involves sensitive human data and solutions to disentanglement could potentially find future applications to face image manipulation.

Finally, considering that the limited size of our model does not require long trainings (see Implementation Details in Sec. 4), the proposed method does not cause significant environmental impacts.