Appendices

This Appendix is organized as follows. We firstly present details on the architecture of our proposed G^{3}AN in Appendix A. Secondly in Appendix B, we provide the details of our proposed Factorised spatio-temporal Self-Attention. Finally, we disclose in Appendix C numerous example frames of generated video sequences.

A. Architecture Details of Generator

We append here details of our model. We recall that our Generator consists of five G^3-modules, designed to generate videos containing 16 frames with a spatial scale of 64×64 (see Figure 2 for details). We note that we implement *ConvTranspose1d* and *ConvTranspose2d* convolutions using *ConvTranspose3d* convolution by setting spatial dimension 1×1 and temporal dimension 1 respectively. Output dimensions in G_V after each G^3 module is shown in Table 1.

	C	$H \times W \times T$
G^3_0	1024	$4 \times 4 \times 2$
G^3_1	512	$8 \times 8 \times 4$
G^3_2	256	$16 \times 16 \times 8$
G^3_3	128	$32 \times 32 \times 16$
G^3_4	64	$64 \times 64 \times 16$
output	3	$64 \times 64 \times 16$

Table 1: **Output dimensions in G_V after each G^3 module.**

B. Factorized spatio-temporal Self-Attention

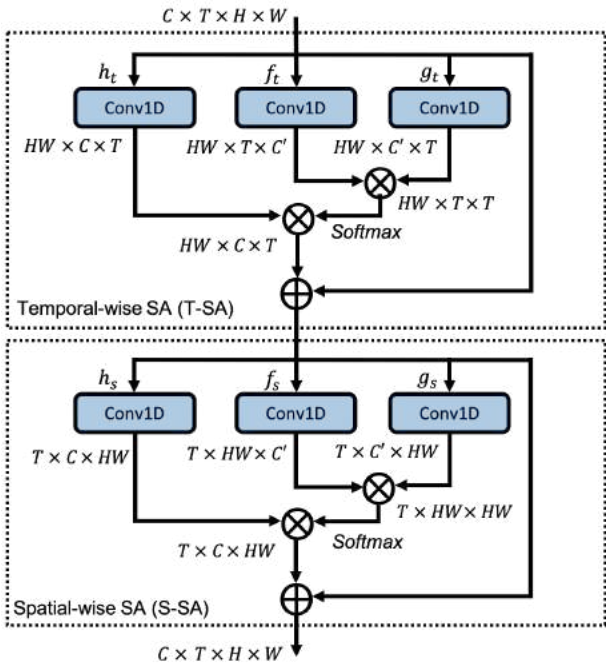
We provide details of proposed Factorized spatio-temporal Self-Attention (F-SA) in this section.

Our F-SA contains a Temporal-wise Self-Attention (T-SA) followed by a Spatial-wise Self-Attention (S-SA) (see Figure 1a). Given spatio-temporal feature maps in the G_V stream, $F_{V_n} = x \in \mathbb{R}^{C \times T \times H \times W}$, where T and $H \times W$ denote temporal and spatial size, respectively. We firstly perform T-SA on $C \times T$ dimensions of x, where attention is only calculated along T for each position in x (see Figure 1b). Then, S-SA is performed on $C \times H \times W$ dimensions and attention maps are obtained for all spatial position at each time step (see Figure 1c).

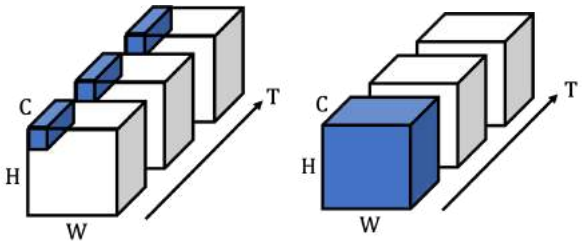
While we apply T-SA, x is firstly transformed into two feature spaces f_t and g_t, in order to compute temporal self-attention

$$a^t_{s,ji} = \frac{\exp(t_{s,ij})}{\Sigma^T_{i=1} \exp(t_{s,ij})}, \; where \; t_{s,ij} = f_t(x_{s,i})^T g_t(x_{s,j}) \tag{1}$$

where a^t_{ji} indicates the correlation between j^{th} and i^{th} time instances for each position s in x. Then we apply attention



(a) F-SA



(b) Temporal-wise SA (T-SA) (c) Spatial-wise SA (S-SA)

Figure 1: Factorized spatio-temporal Self-Attention (F-SA) module.

maps on $h_t(x)$, which is the transformed feature map of x in h_t feature space. Finally we multiply the output of the attention layer by a scalar parameter γ_t and we add back the input feature map in order to obtain the final output of T-SA y^t.

$$y^t_{s,j} = \gamma_t \sum_{i=0}^{T} a^t_{s,ji} h_t(x_{s,i}) + x_{s,j}, \; h_t(x_{s,i}) = W_{h_t} x_{s,i} \tag{2}$$

Similar to T-SA, S-SA uses f_s, g_s and h_s to project y^t into three different feature spaces. γ_s is a learnable scalar parameter multiplied with the output after attention layer. S-TA is computed as following for each time step t.

$$a^s_{t,ji} = \frac{\exp(s_{t,ij})}{\Sigma^N_{i=1} \exp(s_{t,ij})}, \; where \; s_{t,ij} = f_s(x_{t,i})^T g_s(x_{t,j}) \tag{3}$$

$$y^s_{t,j} = \gamma_s \sum_{i=0}^{N} a^s_{t,ji} h_s(x_{t,i}) + x_{t,j},\ h_s(x_{t,i}) = W_{h_s} x_{t,i} \quad (4)$$

In the above formulation, f_t, g_t, h_t, f_s, g_s and h_s are implemented as $1 \times 1 \times 1$ convolutions. For memory efficiency, we reduce channel numbers to $C^{'} = C/k$, where $k = 8$ for f_t, g_t, f_s and g_s in all our experiments.

C. Generated samples

Due to page limitation in the main paper, we here provide additional generated samples. We firstly show unconditional generated samples on UvA-Nemo in Sec C.1 and conditional generated samples pertaining to MUG and Weizmann in Sec C.2. Then we provide results of manipulating both appearance and motion latent representations in Sec C.3. Finally, in Sec C.4 we show results of our model on motion transfer task.

C.1. Unconditional Generation

Here we sample one z_a with three different z_m. Results show that, given different motion representations, our model can generate videos of same appearance with diverse motion (see Figure 3).

C.2. Conditional Generation

We show generated results on MUG and Weizmann datasets in Figure 4-7. For each dataset, results are sampled from two different z_a. We combine each z_a with a one-hot motion category label and for each motion category we sample two z_m. Results show that proposed model can provide diverse *intra-class* samples.

C.3. Latent representations Analysis

In order to understand latent representation of appearance and motion, we manipulate each dimension in z_a and z_m.

Appearance. For appearance, we increase the value in each dimension in order to observe the changes, videos are represented as rows in each figure. From top to bottom, appearance values are increased. We illustrate the manipulated results in two dimensions of each dataset (see Figure 8-13).

Motion. Similar to appearance, we also manipulate each dimension in motion representations. We observe, different dimensions can control different factors, *e.g.*, starting position and motion intensity. Here we show one example for each dataset (see Figure 14-15). From top to bottom in each figure, motion values are increased.

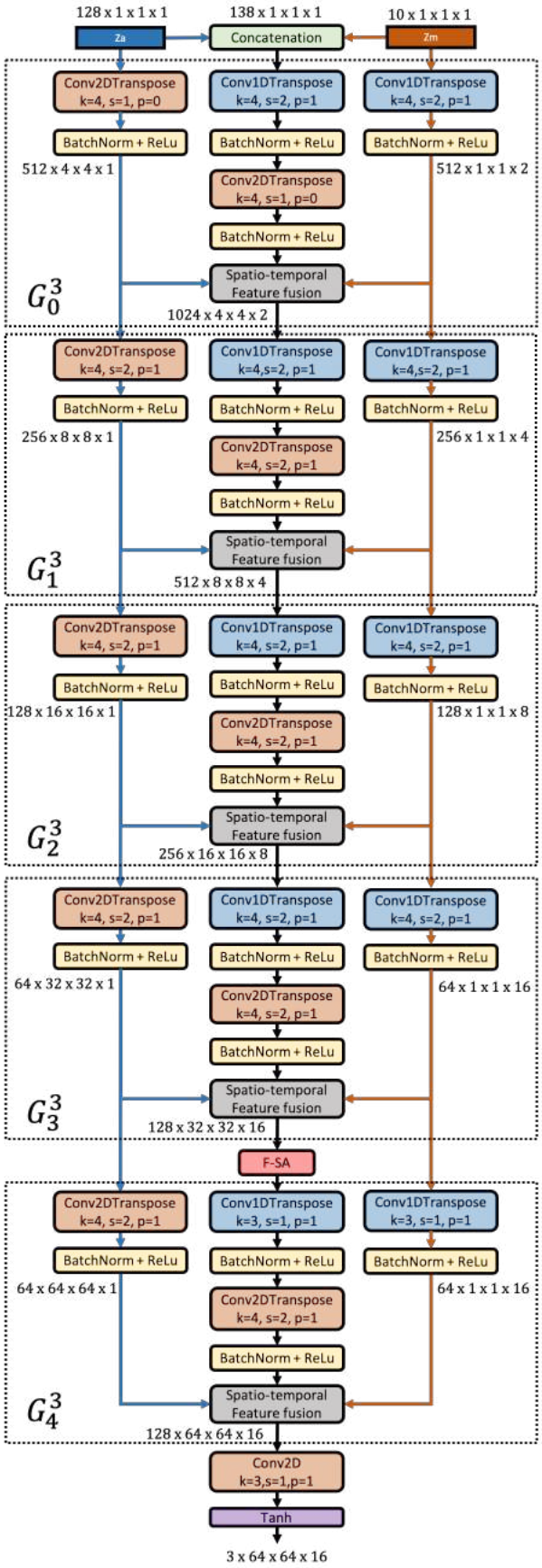


Figure 2: **Generator Architecture.**

C.4. Motion Switch

In this experiments, we firstly sample two sets of latent representations, (z_{a_0}, z_{m_0}, c_0) and (z_{a_1}, z_{m_1}, c_1) to get two videos, where c_0 and c_1 indicate two motion categories. Then we switch motion, obtaining two new sets (z_{a_0}, z_{m_1}, c_1) and (z_{a_1}, z_{m_0}, c_0), we observe that the obtained new videos preserve the appearance and switch the motion. As shown in Figure 16, in each sub-figure, top two rows represent videos obtained from the original two sets while bottom rows represent the switched results.

C.1 Unconditional Generation

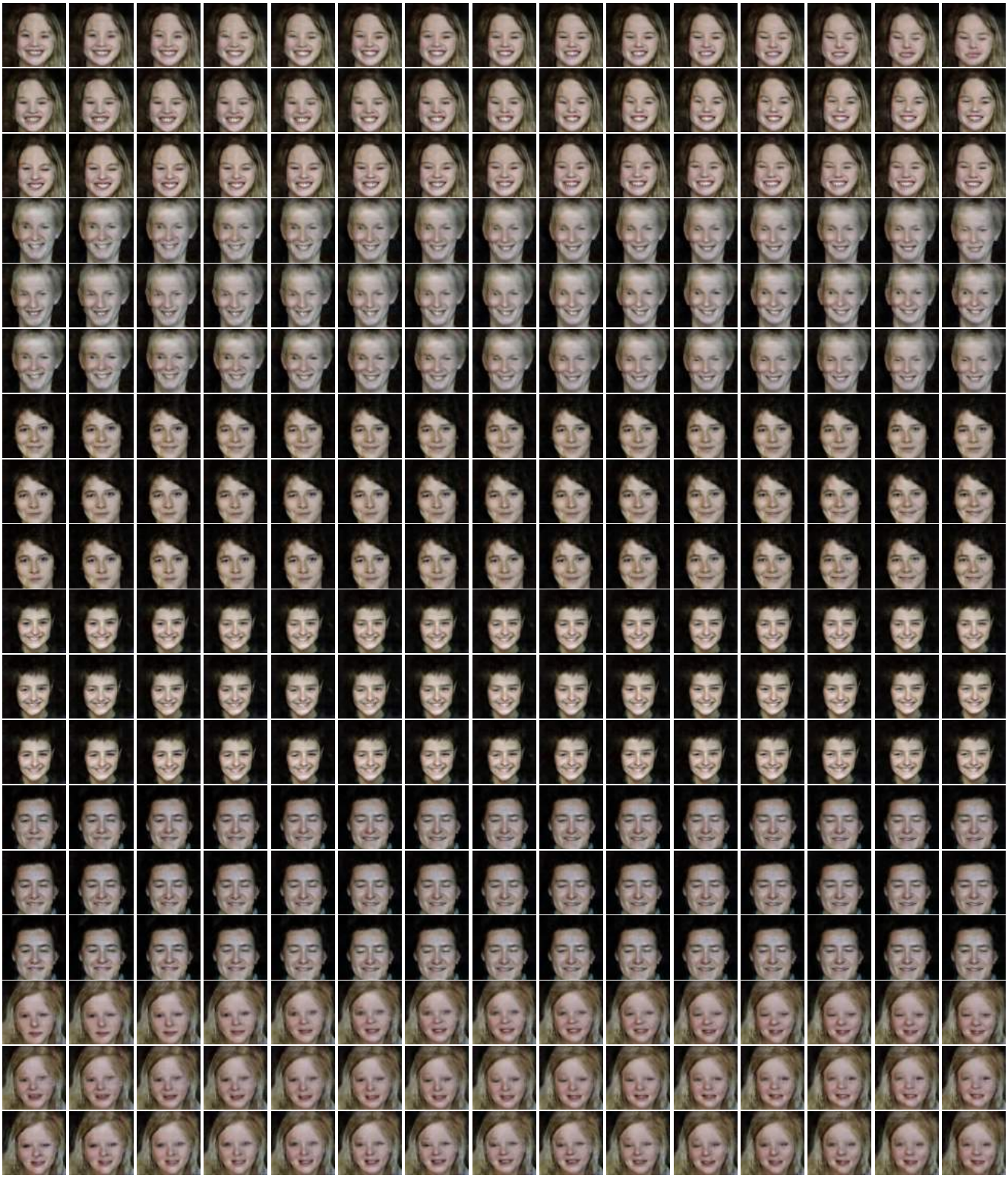


Figure 3: Unconditionally generated samples from G^{3}AN on UvA-NEMO. We combine each z_a with three different z_m, obtaining three different videos for the same appearance. Each row represents a video sequence.

C.2 Conditional Generation

a: subject 1, *sad*

b: subject 1, *anger*

c: subject 1, *surprise*

d: subject 1, *disgust*

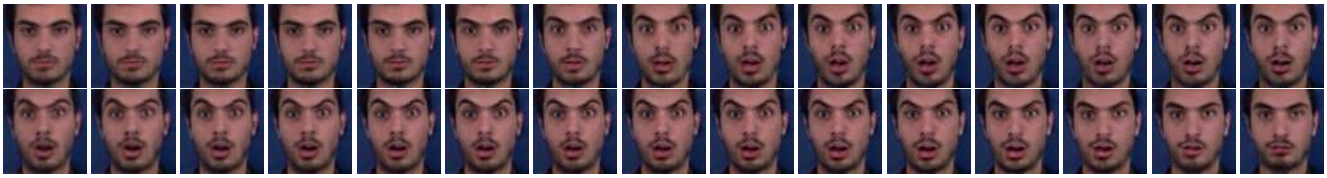
e: subject 1, *happy*

f: subject 1, *fear*

Figure 4: Conditionally generated samples from G^{3}AN on MUG dataset. Each row represents the result generated by combining a one-hot category label with the same z_a and randomly sampled z_m as input.

a: subject 2, *sad*

b: subject 2, *anger*



c: subject 2, *surprise*

d: subject 2, *disgust*

e: subject 2, *happy*

f: subject 2, *fear*

Figure 5: Conditionally generated samples from G^{3}AN on MUG dataset. Each row represents the result generated by combining a one-hot category label with the same z_a and randomly sampled z_m as input.

a: subject 1, *one-hand waving*

b: subject 1, *two-hands waving*

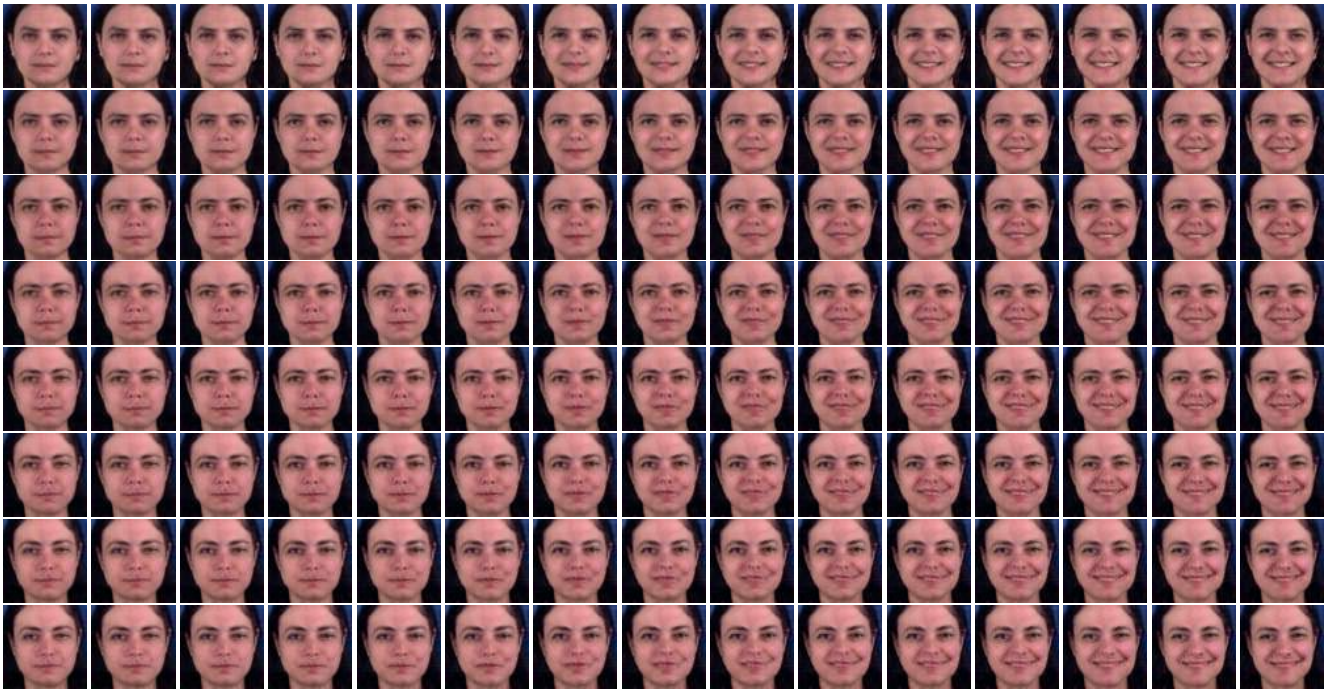
Figure 6: Conditionally generated samples from G^{3}AN on Weizmann dataset. Each row represents the result generated by combining a one-hot category label with the same z_a and randomly sampled z_m as input.

a: subject 2, *one-hand waving*

b: subject 2, *two-hands waving*

Figure 7: Conditionally generated samples from G^{3}AN on Weizmann dataset. Each row represents the result generated by combining a one-hot category label with the same z_a and randomly sampled z_m as input.

C.3. Appearance Manipulation



a: subject 1

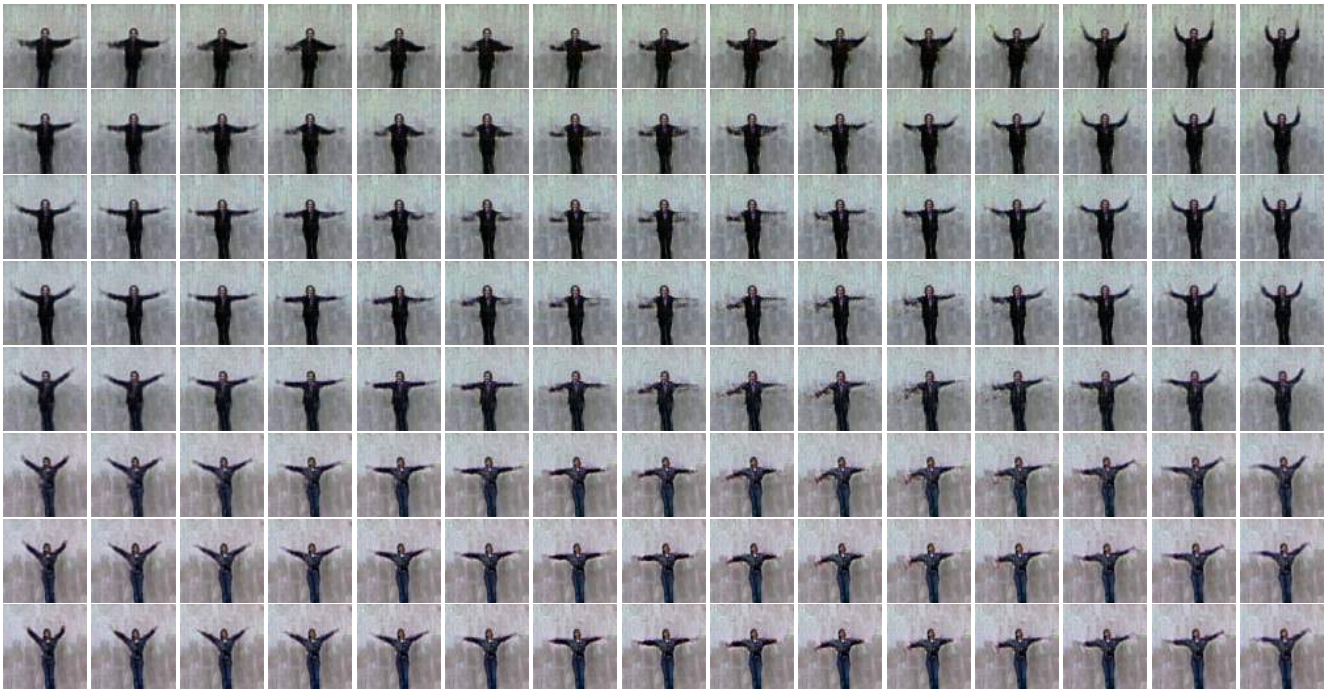
b: subject 2

Figure 8: Results of manipulating *first dimension* in appearance representation on MUG dataset. *a* and *b* are results from two randomly sampled z_a. From top to bottom in each sub-figure, values of *first dimension* are increased.

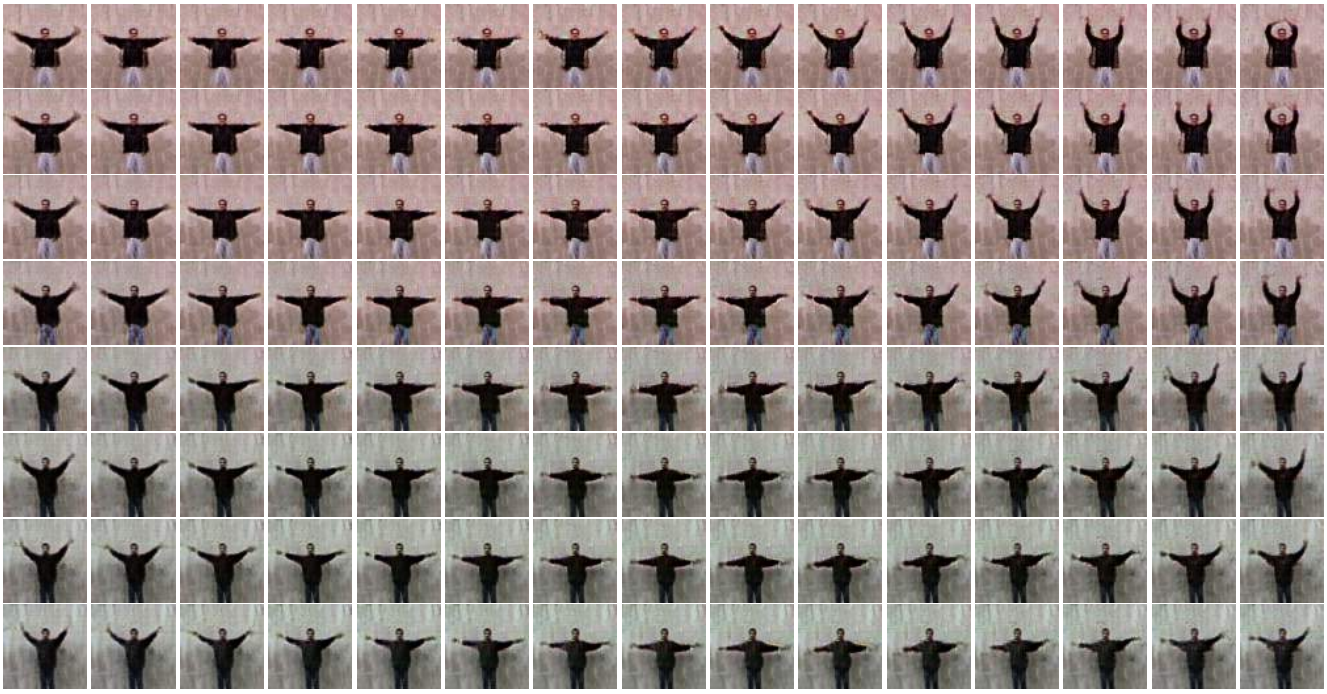
a: subject 1

b: subject 2

Figure 9: Results of manipulating *second dimension* in appearance representation on MUG dataset. *a* and *b* are from two randomly sampled z_a. From top to bottom in each sub-figure, values of *second dimension* are increased.

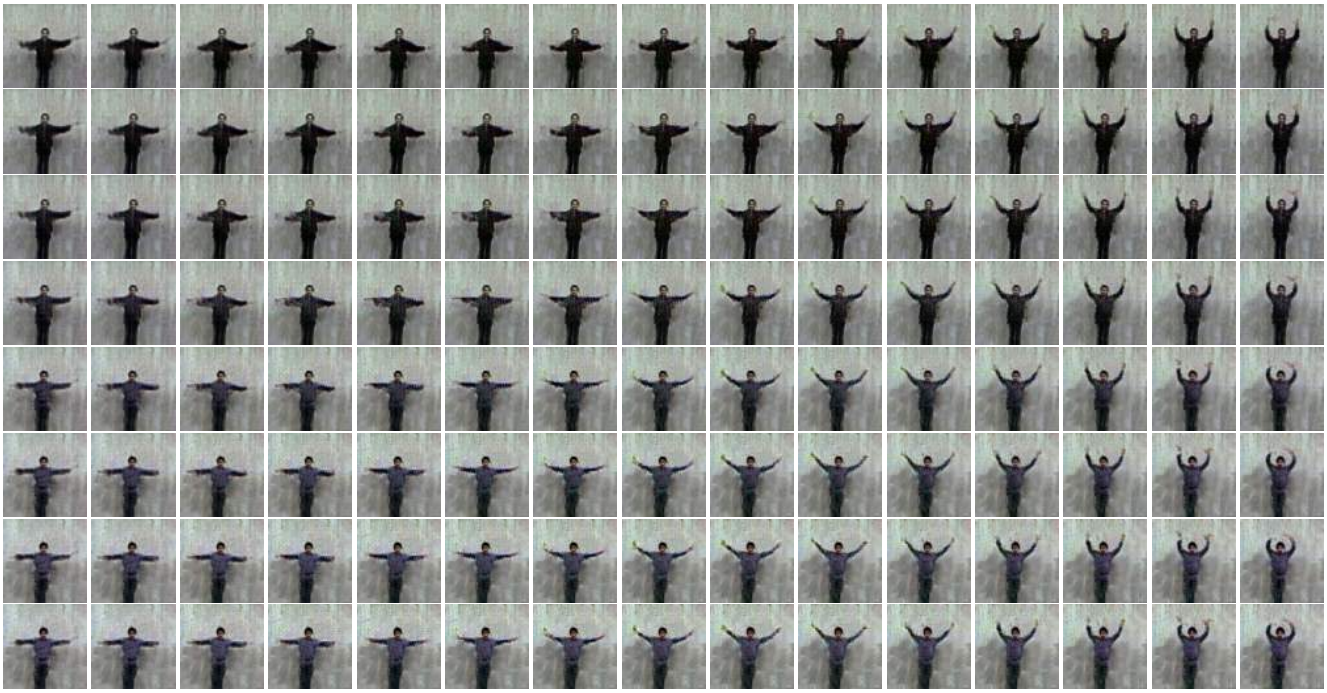


a: subject 1

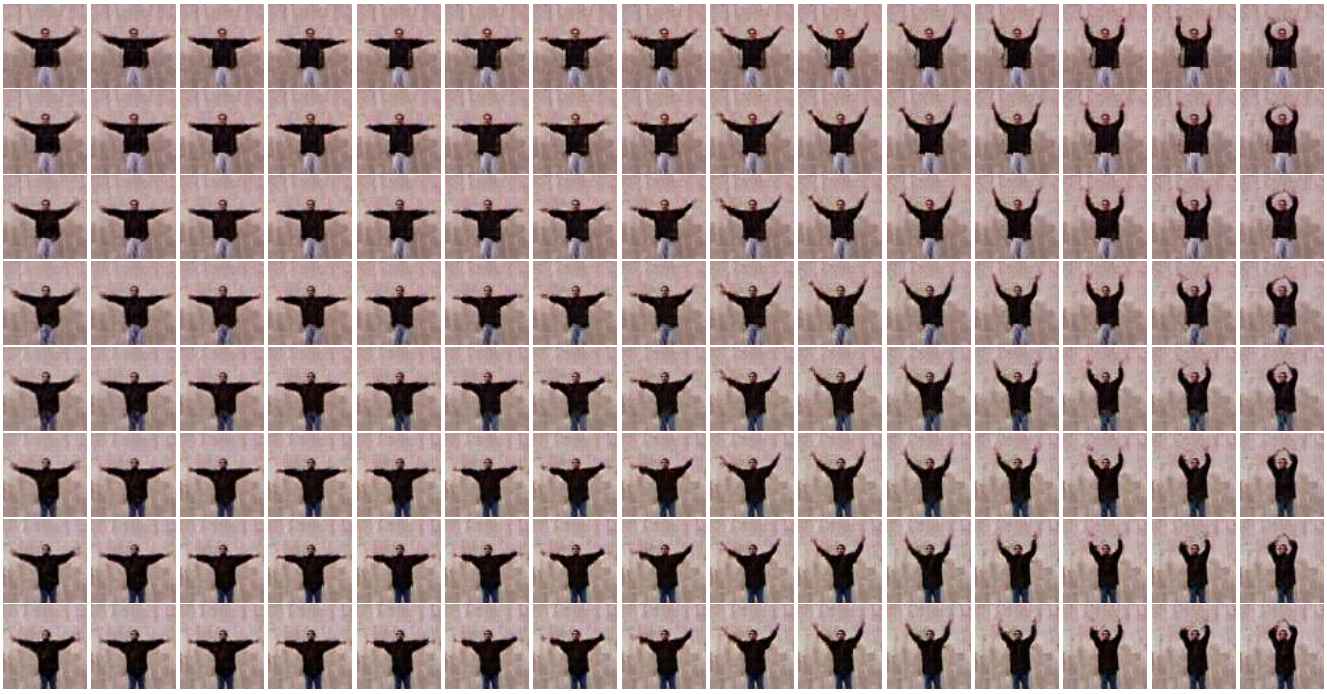


b: subject 2

Figure 10: Results of manipulating *first dimension* in appearance representation on Weizmann dataset. *a* and *b* are from two randomly sampled z_a. From top to bottom in each sub-figure, values of *first dimension* are increased.



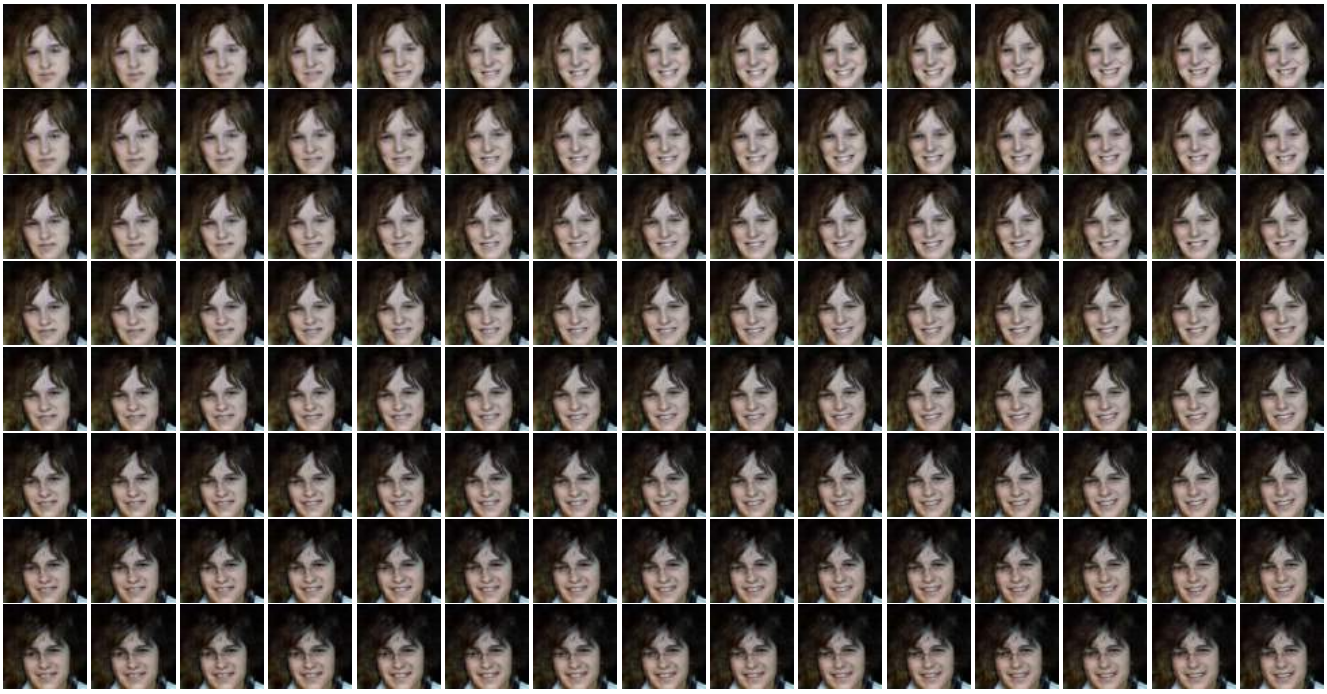
a: subject 1



b: subject 2

Figure 11: Results of manipulating *second dimension* in appearance representation on Weizmann dataset. *a* and *b* are from two randomly sampled z_a. From top to bottom in each sub-figure, values of *second dimension* are increased.

a: subject 1



b: subject 2

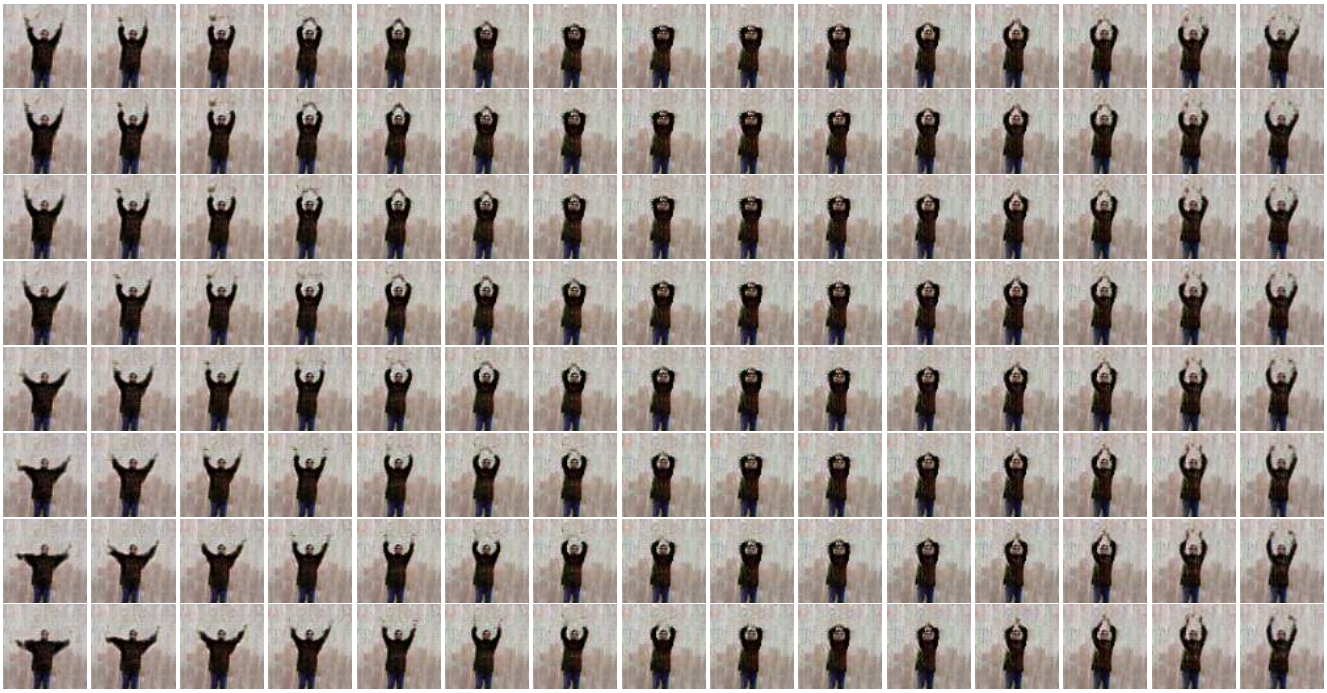
Figure 12: Results of manipulating *first dimension* in appearance representation on UvA-NEMO dataset. *a* and *b* are from two randomly sampled z_a. From top to bottom in each sub-figure, values of *first dimension* are increased.

C.3. Motion Manipulation

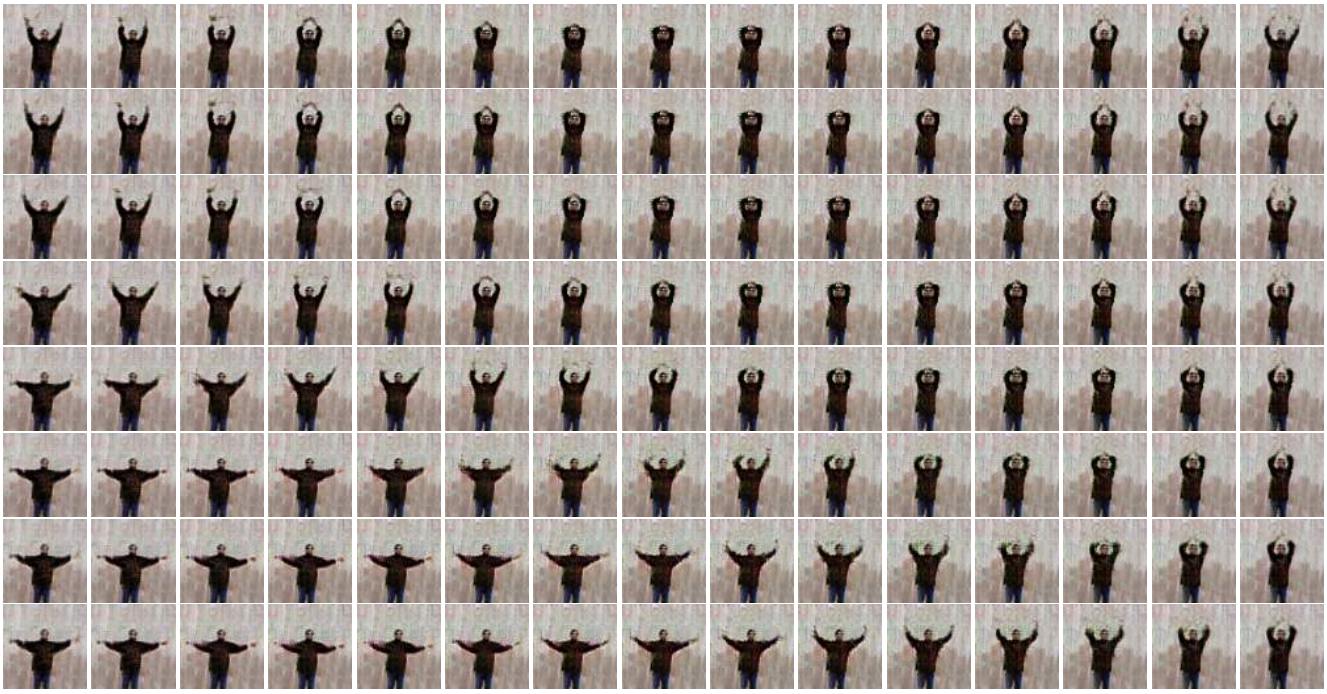
a: second dimension

b: *sixth dimension*

Figure 14: Results of manipulating motion representation on UvA-NEMO dataset. *a* and *b* are results of manipulating *first* and *sixth* dimensions. From top to bottom in each sub-figure, values are increased.



a: *first dimension*



b: *second dimension*

Figure 15: Results of manipulating motion representation on Weizmann dataset. *a* and *b* are results of manipulating *first* and *second* dimensions. From top to bottom in each sub-figure, values are increased.