One Style is All You Need to Generate a Video

Sandeep Manandhar and Auguste Genovesio
IBENS, Ecole Normale Supérieure
75005 Paris, France
sandeep.manandhar@bio.ens.psl.eu, auguste.genovesio@ens.psl.eu

1. Wave-like nature of trajectories in StyleGAN2's latent space

One can iteratively invert a sequence of video frames into a sequence of latent(style) vectors in StyleGAN2's network [2]. Optimizing for the i^{th} frame leads to a vector w_i, which then can be used as a starting point for the optimization for w_{i+1} corresponding to the next video frame. With this iterative strategy one can obtain a sequence of latent vectors $w_0, w_1, w_2, ..., w_n$. Upon principal component analysis of such sequence of latent vectors, waves like structures are revealed as shown in Figure 1. We leverage upon this observation to propose our network that utilizes sinusoidal bases to form a latent space for temporal styles. During the training time, we embed the style vectors of subsequent video frames in a close vicinity guided by the topology formed by the sinusoidal bases. More formally,

$$\begin{aligned} m^t &= F_t(z_m), \\ w_m^t &= m^t * v(t), \\ w_m^{t+1} &= m^t * v(t+1), \end{aligned} \tag{1}$$

where F_t is the 4-layered multi layer perceptron network (please refer to Figure 2 of the main paper) that maps a random vector z_m to a vector m^t. Here, $v(t)$ is derived from the *time2vec* embedding at time t, and w_m^t, w_m^{t+1} are two successive temporal style vectors.

2. Datasets

2.1. MEAD

Our MEAD dataset consists of all 8 emotions of 30 individuals (15 males and 15 females). We used the same ROI to crop the region containing faces in all videos. The video frames are then rescaled to 256^2 image dimension. The faces are not aligned. During the training, we excluded entire set of videos corresponding to randomly selected individuals expressing certain emotions. Therefore, the training set contains sequences of some individuals that never display certain emotions. We then generated videos of those individuals displaying the missing emotion (from the training set) for the evaluation. The generated sequences can be viewed in the accompanying videos.

2.2. UTDMHAD

This dataset contains 754 videos of 8 individuals performing 27 different actions. For our training and test set, we cropped the video frames around a region containing the actors and rescaled them to 256^2 image dimension. However, we train the network on 128×128 image dimension. In our training set, we removed entire set of each actions from randomly selected actors. During the evaluation, we generated only the sequences with actors with corresponding missing actions. Because the test set was small, we computed FVD score using only 108 generated sequences for all the methods. The generated sequences can be viewed in the accompanying videos.

3. Human Evaluation

We performed a survey to carry out human evaluation on the generated videos for the MEAD dataset. We created a set of 10 videos containing 6 sequences. We randomly selected a real video and 5 generated videos (ours and baselines), and

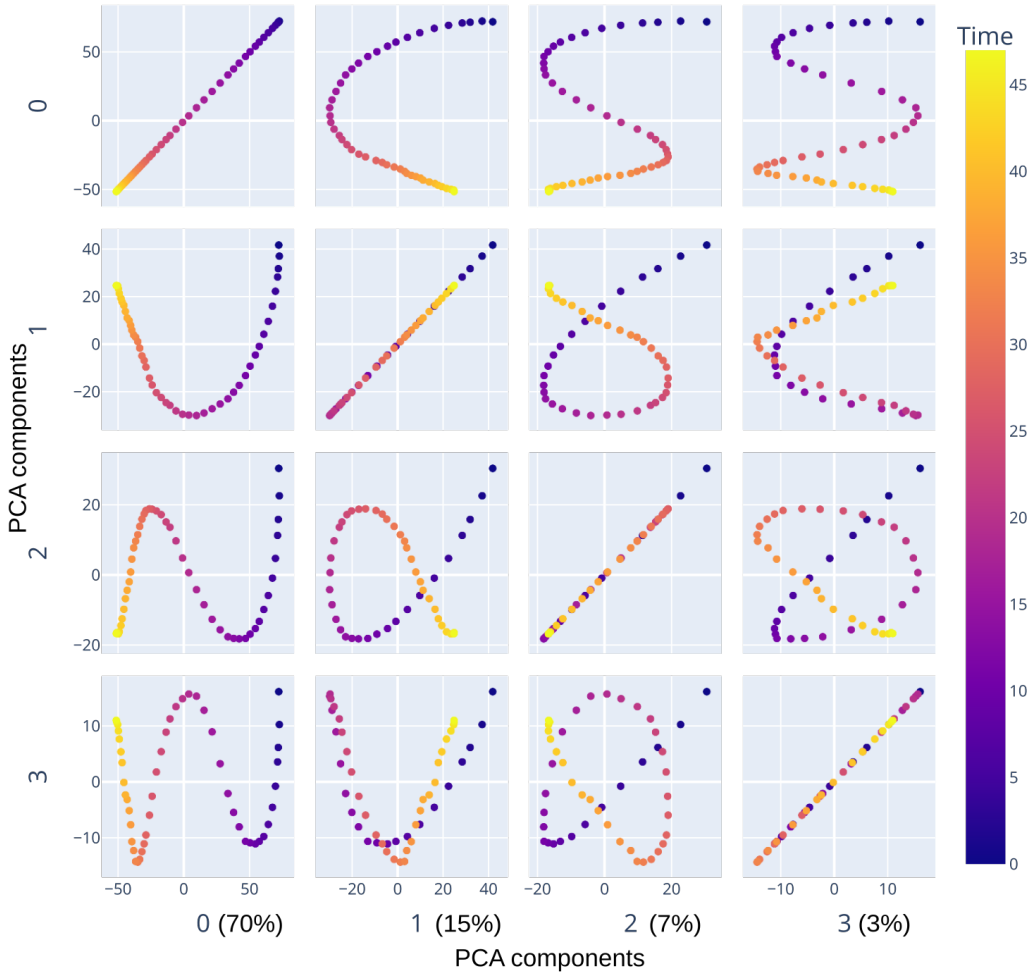

Figure 1. Plotting 4 principal components of the sequence of style vectors obtained via iterative GAN-inversion in StyleGAN2's latent space (for frames in Figure 8). The numbers in parenthesis in the labels along the x-axis denote the percentage of variance contained in the corresponding principal component.

presented them side-by-side. The sequences are 32 frames long as ImaGINator is only capable of generating 32 frames. Furthermore, MOCOGAN-HD cannot generate consistent video frames for longer duration. Each participants were given 3 sets of videos and asked to rank 6 sequences in each videos in descending order of how realistic they found the sequences to be. The videos were viewed by the participants in their own personal devices.

4. Action classification

We used an implementation of [1] to train their network on UTD-MHAD dataset. The method uses sequences of skeletal data as input to classify the action present in a given video. We trained the network using the skeletal data extracted from the same training set videos that was used to train our conditional method. We then tested the classifier on the test set containing the skeletal data of 4 real sequences for each action. Here, the classifier was able to obtain 77% $top-1$ accuracy. The confusion matrix for this case is presented in Figure 2. We also classified the actions from the sequences that were generated by ImaGINator and our method(s). Figure 3 shows that the videos generated by ImaGINator were not accurately classified. As we can see in the accompanying videos, the generated videos by ImaGINator do not have good clarity. Figures 4 and 5

show the confusion matrices of the classification of the videos generated by our model after training it using single time-point and three time-points in D_t. Figure 6 shows the confusion matrix of the classification of the videos generated by our model trained without D_t. For the sake of brevity, we have abbreviated the words "right, left, hand, arms, feet, counter" in the official UTDMHAD label names as "R,L,H,A,F,C" respectively, in the figures depicting the confusion matrices. When the model is trained using three time-points, the generated sequences were being classified with accuracy of 68.5%, which is not bad compared to the real sequences.

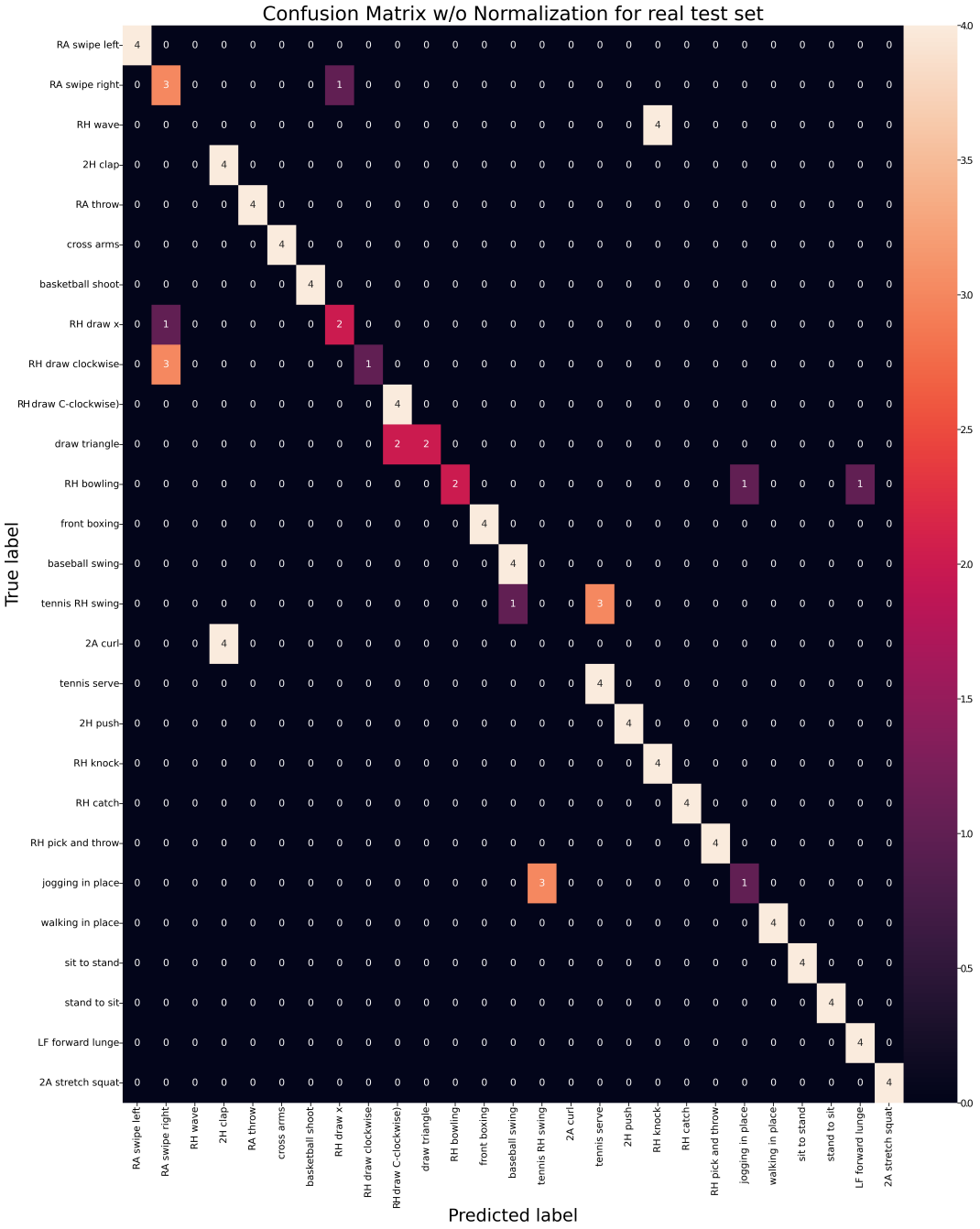

Figure 2. Confusion matrix for the classification of 27 actions in UTD-MHAD for real test sequences using [1] with 77% *top-1* accuracy.

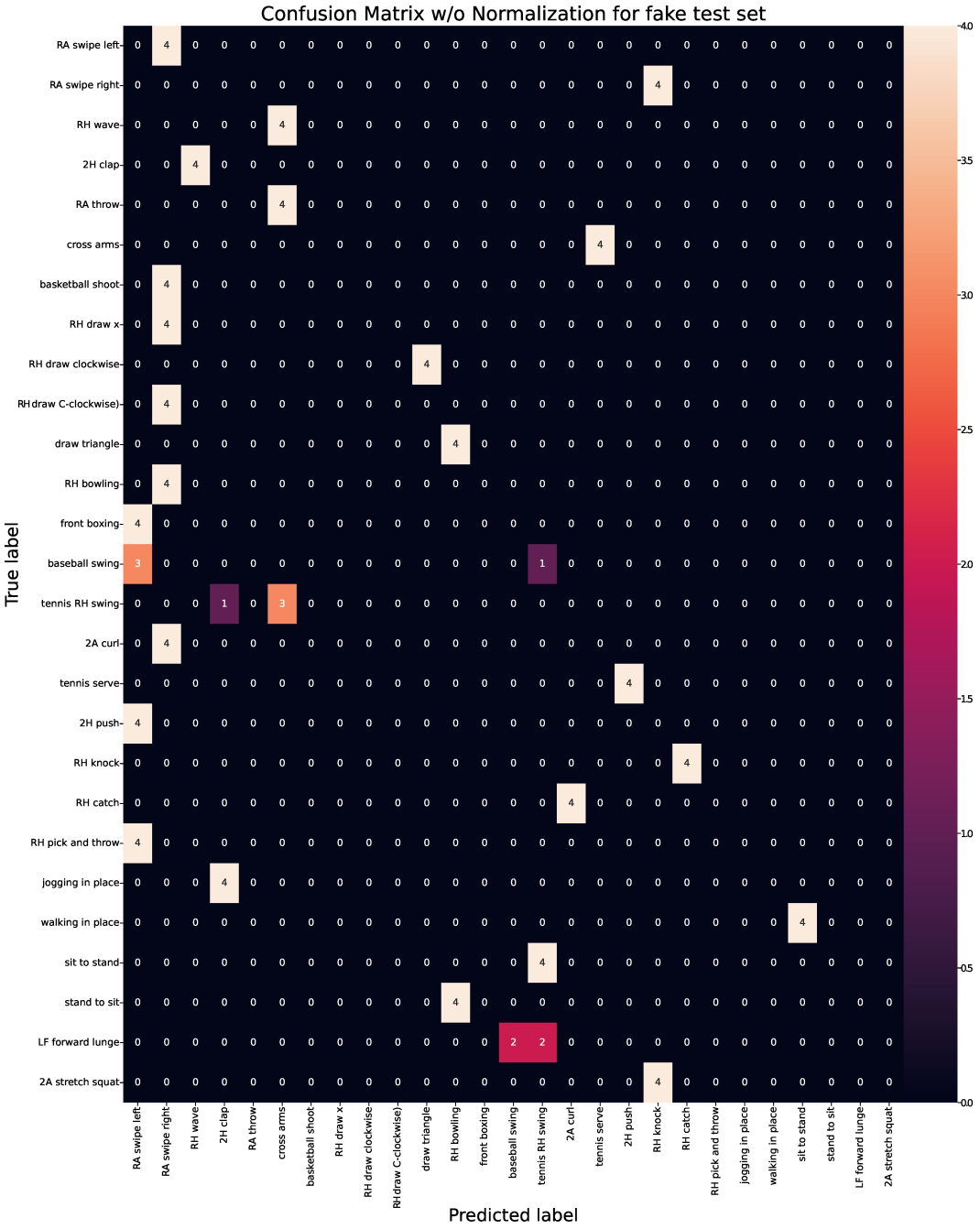

Figure 3. Confusion matrix for the classification of 27 actions in UTD-MHAD for generated sequences by ImaGINator using [1]

5. GAN-inversion for temporal styles

We preform the GAN-inversion for temporal styles of our model by optimizing the following loss function with respect to the motion code m (as computed in Eq. (1))

$$\mathcal{L} = \sum_{i=0}^{N} (\lambda \mathcal{L}_2^i + \mathcal{L}_{LPIPS}^i) \tag{2}$$

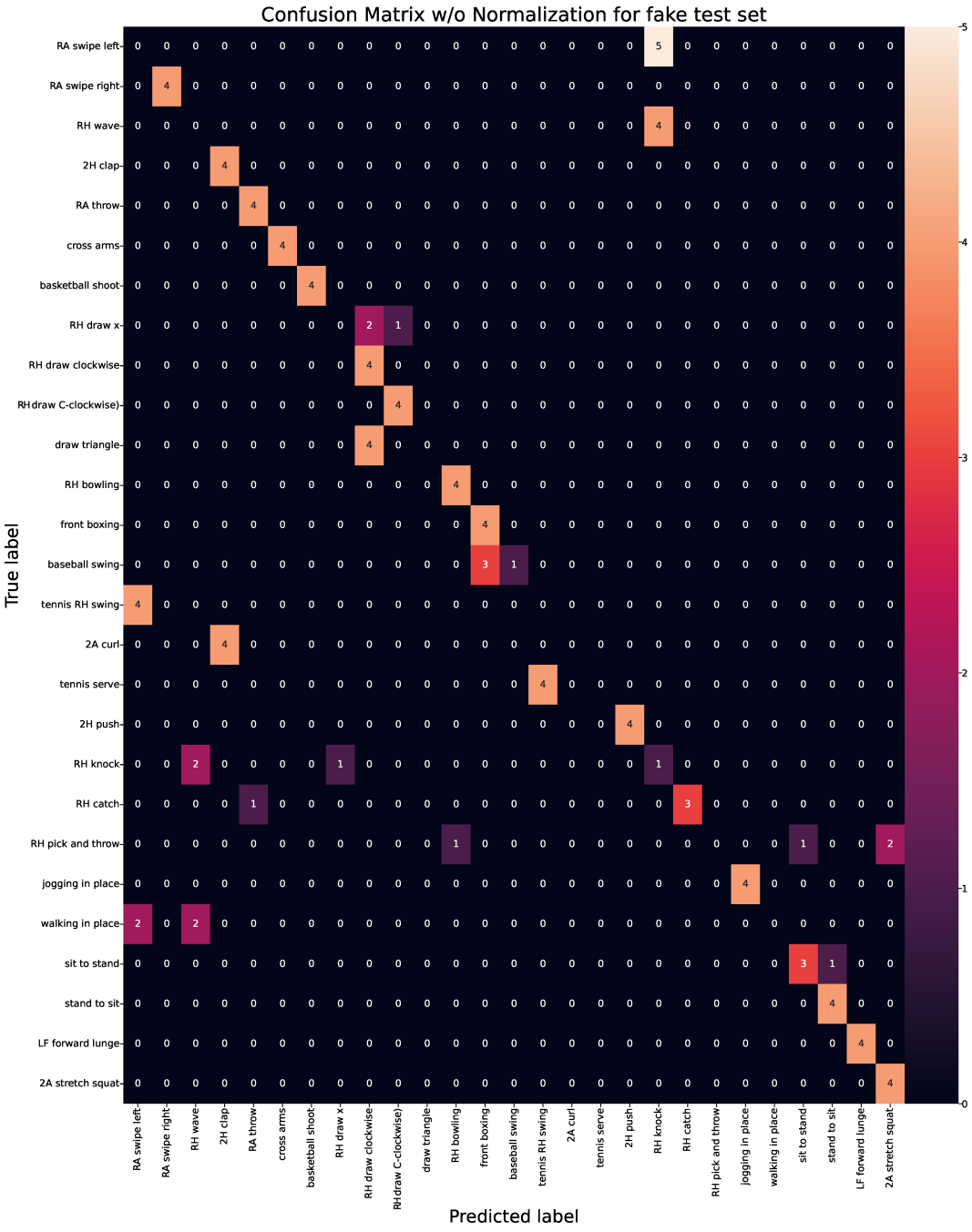

Figure 4. Confusion matrix for the classification of 27 actions in UTD-MHAD for generated sequences by our method using single time-point in D_t using [1] with 57% *top-1* accuracy.

where $\mathcal{L}_2^i$ is L_2 distance between the pixels of i^{th} video frame of a real video x_i and fake video frame $G\left([w_c, mv(i)]\right)$, and $\mathcal{L}_{LPIPS}^i$ is the perceptual loss [3] between the features extracted from the i^{th} real and fake frames. Here, $v(i)$ is the temporal embedding obtained from the *time2vec* module of the generator and m is the temporal code we would like to optimize for. We do not need to optimize for $v(i)$ as it is generated by the *time2vec* module with the corresponding time-point of the video frame as input. In our experiments, we inverted the real videos where the actor-id and action-id are known. Thus, we fixed w_c by appropriately. We run the optimization for 300 steps.

We show the principal components of the optimized temporal codes in Figure 7. The wave-like structures are clearly

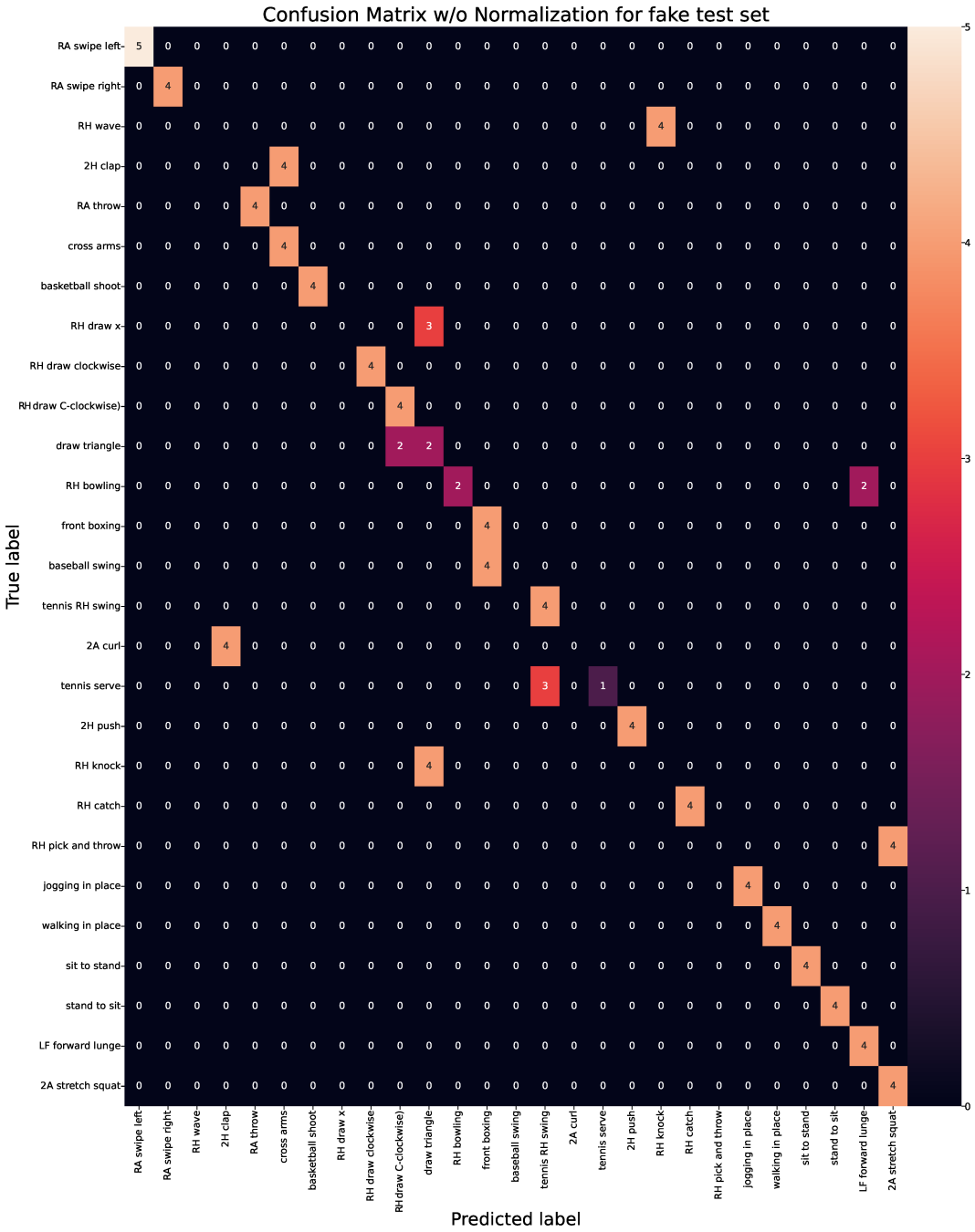

Figure 5. Confusion matrix for the classification of 27 actions in UTD-MHAD for generated sequences by our method using 3 time-point in D_t using [1] with 68.5% *top-1* accuracy.

visible thanks to the sinusoidal bases in *time2vec*. The bottom row of Figure 8 presents few frames of the videos generated by using the temporal style obtained after the GAN-inversion. The obtained temporal style can be reused in generating video of other actors as shown in the accompanying videos.

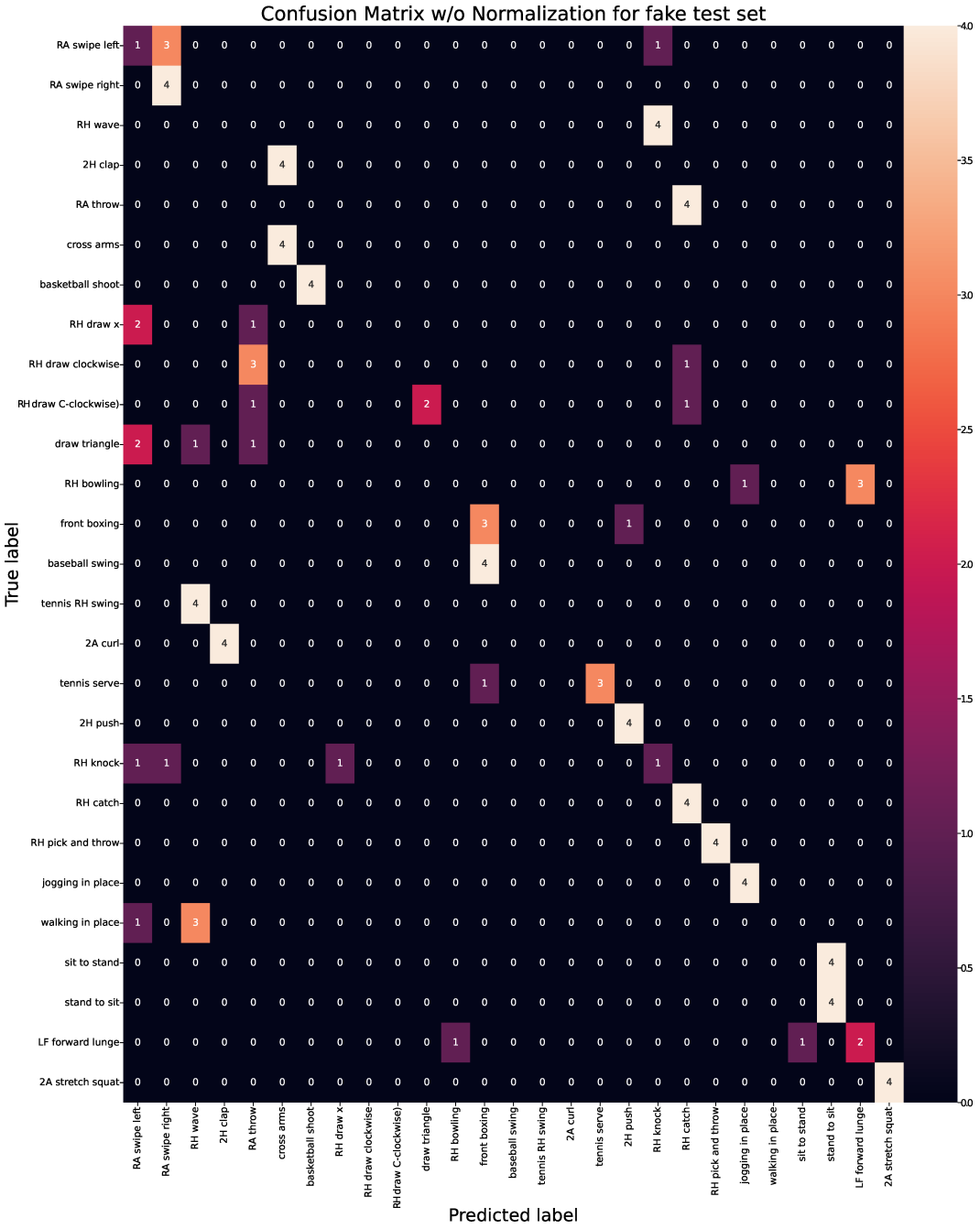

Figure 6. Confusion matrix for the classification of 27 actions in UTD-MHAD for generated sequences by our method without D_t using [1] with 42.6% *top-1* accuracy.

6. List of emotion-specific sentences for GAN-inversion test

Here we list out the 39 emotion-specific sentences excluding the neutral sequences used to evaluate our GAN-inversion. The tests were conducted for ”level 2” expressions. After cleaning up the dataset by removing missing and duplicate sentences, we test over the following 39 sentences. The 3 digit number represents the file-id provided in the dataset.

Angry

006: The cat’s meow always hurts my ears

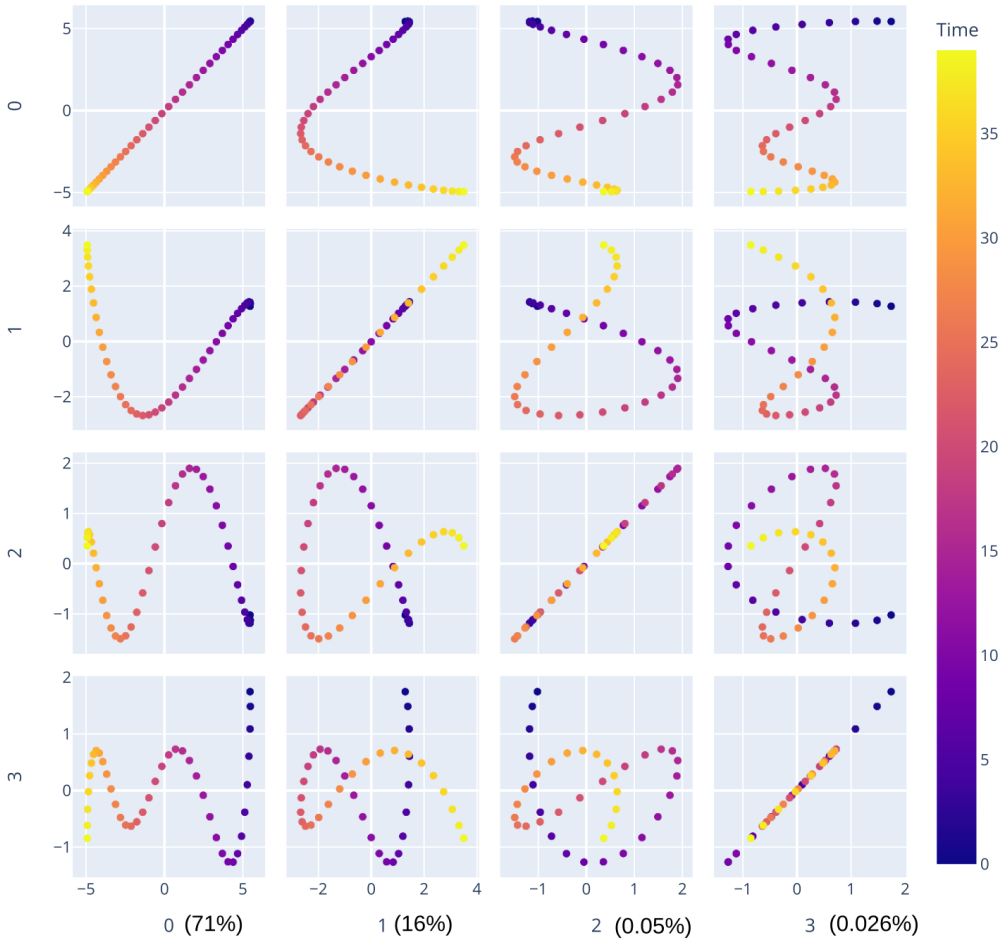

Figure 7. Plotting 4 principal components of the sequence of style vectors obtained after GAN-inversion of temporal style of our conditional network (for frames in Figure 8). The numbers in parenthesis in the labels along the x-axis denote the percentage of variance contained in the corresponding principal component.

007: Why else would Danny allow others to go

008: Why do we need bigger and better bombs

009: Nuclear rockets can destroy airfields with ease

010: You're so preoccupied that you've let your faith grow dim

011: Cory and Trish played tag with beach balls for hours

Contempt

006: Only lawyers love millionaires

007: It's illegal to postdate a check

008: He stole a dime from a beggar

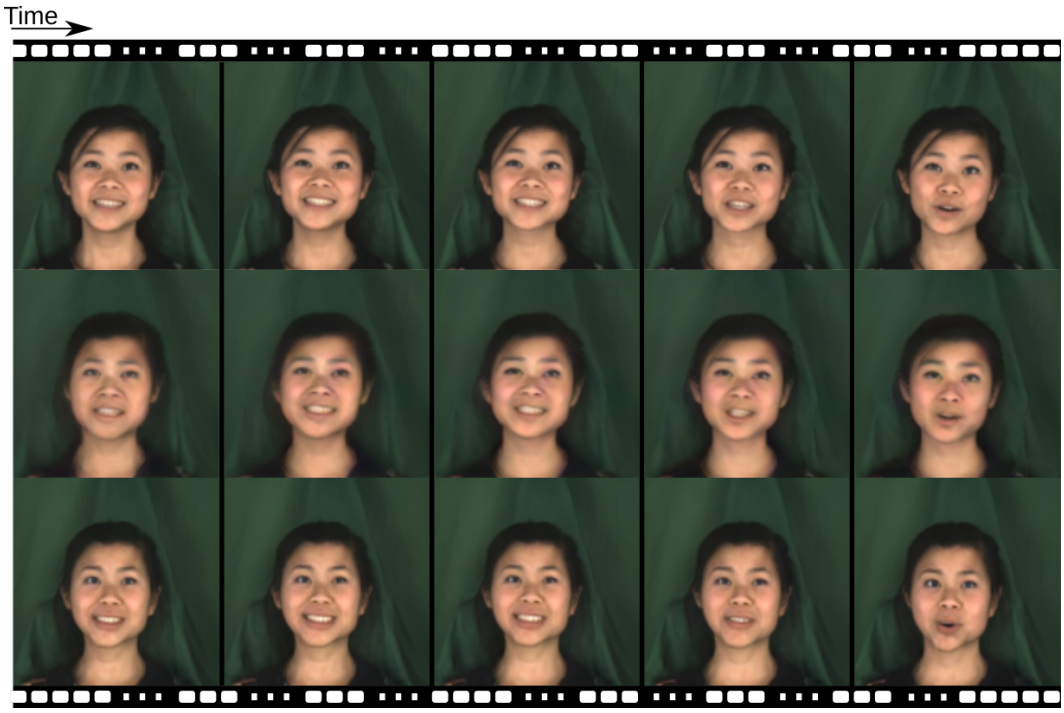

Figure 8. **Top row**: Few frames of real video. **Mid row**: frames obtained after iterative GAN-inversion of latent vectors in StyleGAN2 latent space. **Bottom row**: frames obtained after GAN-inversion of temporal styles of our conditional network. Please refer to the accompanying videos.

009: His failure to open the store by eight cost him his job

010: Let us differentiate a few of these ideas

011: The big dog loved to chew on the old rag doll

Disgust

006: Young children should avoid exposure to contagious diseases

007: Military personnel are expected to obey government orders

008: Basketball can be an entertaining sport

009: How good is your endurance

010: Barb burned paper and leaves in a big bonfire

011: December and January are nice months to spend in Miami

Fear

007: Would you allow acts of violence

008: The high security prison was surrounded by barbed wire

010: The fish began to leap frantically on the surface of the small lake

011: Straw hats are out of fashion this year

Happy

006: Tim takes Sheila to see movies twice a week

007: They used an aggressive policeman to flag thoughtless motorists

008: When you're less fatigued, things just naturally look brighter

009: By that time perhaps something better can be done

011: Project development was proceeding too slowly

Sad

006: We can die too, we can die like real people. People never live forever

007: He didn't figure her at all, and if he found out a woman, it'd be bad

008: There would still be plenty of moments of regret and sadness and guilty relief

009: She drank greedily and murmured thank you as he lowered her head

010: There's no chance now of all of us getting away

011: Before Thursday's exam review every formula

Surprise

006: The patient and the surgeon are both recuperating from the lengthy operation

007: He ate four extra eggs for breakfast

008: While waiting for Chipper, she crisscrossed the square many times

009: I just saw Jim near the new archaeological museum

010: I took her word for it, but is she really going with you

011: The viewpoint overlooked the ocean

7. Limitations and Future work

While our study has achieved convincing and promising results in the realm of style-based conditional video generation and video GAN inversion, several limitations and avenues for future research warrant consideration. First, it is important to note that our experiments are primarily concentrated on scenarios involving single actors executing simple actions. The current method could encounter challenges when attempting to generate video scenes featuring multiple actors with intricate interactions. The empirical choice of k, i.e. the number of Fourier bases in our experiments may not be optimal to capture complex dynamics. A possible solution could consist of adopting a multi-resolution approach, whereby lower-frequency bases are introduced during coarser stages, progressively incorporating higher-frequency elements in finer stages. Furthermore, our current video GAN-inversion succeeds in a conditional setting. Without providing the actor-id, the optimization methods fail so far. This model would benefit a robust optimization method that could disentangle the actor from the action during the inversion process.

8. Acknowledgement

This work has received support under the program "Investissements d'Avenir" launched by the French Government and implemented by the ANR, with the references: ANR-10-LABX-54 MEMO LIFE ANR-11-IDEX-0001-02 PSL*. S.M. was funded by Inserm ITMO Cancer - TOTEM. This work was granted access to the HPC resources of IDRIS under the allocation 2020-AD011011495 made by GENCI.

References

[1] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, 2022. 2, 3, 4, 5, 6, 7

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1

[3] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5