Supplementary Material

Additional t-SNE Visualizations

We provide further t-SNE visualizations of cls token encoded with our model and generated skeleton sequences. The dimension of the cls token is 512. For the t-SNE we choose the perplexity 50, the number of iterations 1000 and the learning rate 1000 as well.

We show the t-SNE visualizations of our cls tokens trained on NW-UCLA, NTU 60 and NTU 120 dataset. We evaluate the models on NW-UCLA and NTU 120 dataset partitioned in NTU 60 and NTU 61-120. We evaluate the NTU 60 pretrained model on NTU 61-120. For NTU 60 pretrained model evaluated on NTU 60 dataset refer to Figures 6 and 5. The t-SNE plots of the NTU 120 pretrained model give a hint on the performance of the network. The t-SNE visualization shows that on NTU 60 the clusters are better separated than on NTU 61-120, which is in line with our results in Table 4.

Additional Skeleton Sequences

For the visualization of normalized skeleton sequences, we select 5 equidistant frames out of the original skeleton sequence. Where one frame means one set of joints. For the NTU dataset the number of joints per frame is 25. For a better visibility we add an offset of $n \cdot 0.6$ to the x-axis [8] values of frame n. All skeleton sequences are extracted from the NTU 60 dataset.

In the Figures 15 - 17 we show conditioned data generation with our transformer autoencoder. The skeleton sequences with the same encoder input (skeleton sequences a)-d) and e)-h)) share some significant and important movements, such as grab the foot for action "wear a shoe" (see Figure 15) or leaning forward for action "vomiting condition" (see Figure 16). Due to the random noise input for the decoder, the skeleton sequences b)-d) and f)-h) differ in joint position such as feet, hand or head position.

For skeleton sequences with little per joint variance for the majority of joints (see Figure 17 a)-d)) we observe smaller variance in the generated sequences compared to sequences with larger variance (see Figure 17 e)-h)).

LEP with fewer labels.

We evaluate the generalization of our model within a dataset using the LEP with only a fraction of the full dataset. To this end we train our model unsupervised on the full train dataset. Thereafter, we train a linear layer according to the LEP with a randomly selected fraction of the full train dataset (1%, 10%, 50%). We choose the same hyperparamters as in the LEP (see subsection 4.3).

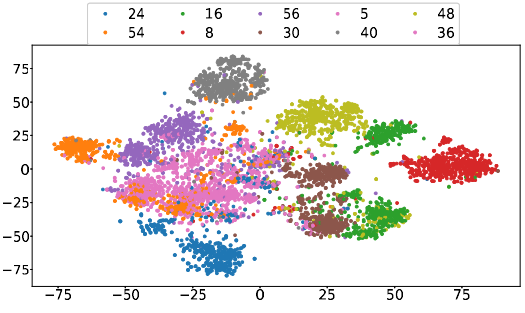

Figure 6. The t-SNE visualization of feature embeddings of the NTU 60 training data. 10 random categories are sampled. The activites are as follow: 24: "kicking something", 16: "wear a shoe", 56: "giving something to other person", 5: "drop", 48: "nausea or vomiting condition", 54: "point finger at the other person", 8: "sitting down", 30: "typing on a keyboard", 40: "cross hands in front (say stop)", 36: "shake head".

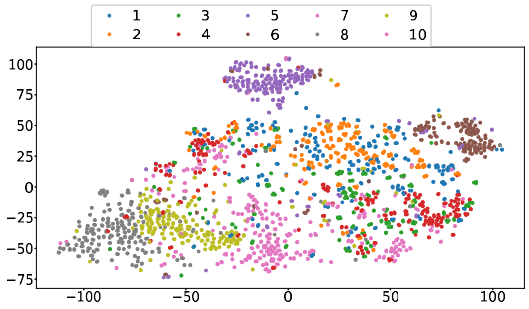

Figure 7. The t-SNE visualization of cls token outputs of NW-UCLA train dataset with the model trained on NW-UCLA. The activites are as follow: 1: "pick up with one hand", 2: "pick up with two hands", 3: "drop trash", 4: "walk around", 5: "sit down", 6: "stand up", 7: "donning", 8: "doffing", 9: "throw", 10: "carry".

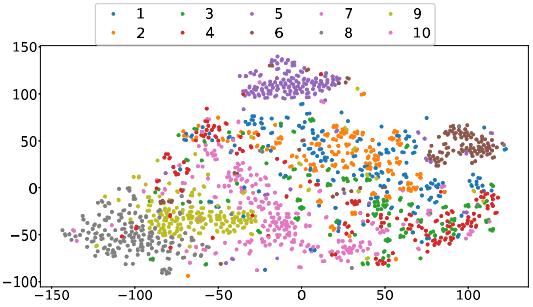

Figure 8. The t-SNE visualization of cls token outputs of NW-UCLA validation dataset with the model trained on NW-UCLA. The activites are as follow: 1: "pick up with one hand", 2: "pick up with two hands", 3: "drop trash", 4: "walk around", 5: "sit down", 6: "stand up", 7: "donning", 8: "doffing", 9: "throw", 10: "carry".

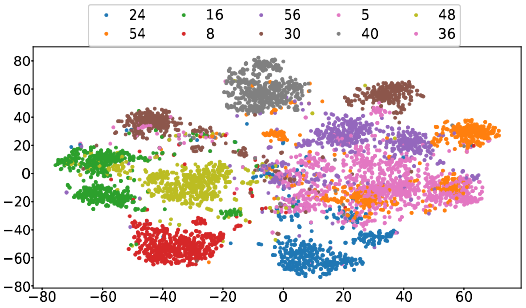

Figure 9. The t-SNE visualization of cls token outputs of NTU 60 train dataset with the model trained on NTU 120. The activites are as follow: 24: ”kicking something”, 16: ”wear a shoe”, 56: ”giving something to other person”, 5: ”drop”, 48: ”nausea or vomiting condition”, 54: ”point finger at the other person”, 8: ”sitting down”, 30: ”typing on a keyboard”, 40: ”cross hands in front (say stop)”, 36: ”shake head”.

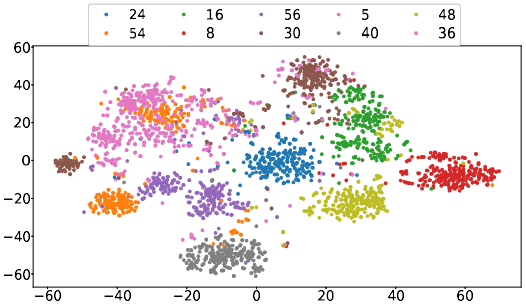

Figure 10. The t-SNE visualization of cls token outputs of NTU 60 validation dataset with the model trained on NTU 120. The activites are as follow: 24: ”kicking something”, 16: ”wear a shoe”, 56: ”giving something to other person”, 5: ”drop”, 48: ”nausea or vomiting condition”, 54: ”point finger at the other person”, 8: ”sitting down”, 30: ”typing on a keyboard”, 40: ”cross hands in front (say stop)”, 36: ”shake head”.

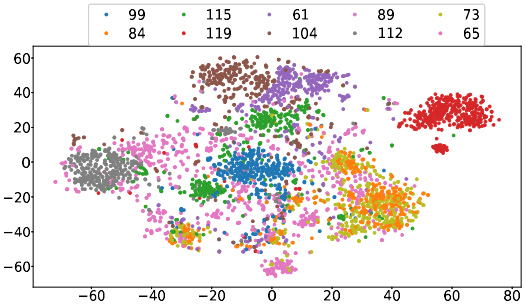

Figure 11. The t-SNE visualization of cls token outputs of NTU 60-120 train dataset with the model trained on NTU 60. The activities are as follow: 99: ”running on the spot”, 115: ”take a photo of other person”, 61: ”put on headphone”, 89: ”put something into a bag”, 73: ”staple book”, 84: ”play magic cube”, 119: ”support somebody with hand”, 104: ”stretch oneself”, 112: ”high-five”, 65: ”tennis bat swing”.

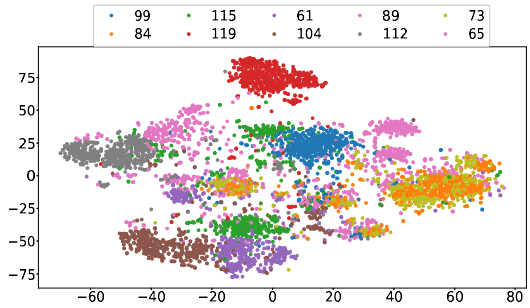

Figure 12. The t-SNE visualization of cls token outputs of NTU 60-120 validation dataset with the model trained on NTU 60. The activities are as follow: 99: ”running on the spot”, 115: ”take a photo of other person”, 61: ”put on headphone”, 89: ”put something into a bag”, 73: ”staple book”, 84: ”play magic cube”, 119: ”support somebody with hand”, 104: ”stretch oneself”, 112: ”high-five”, 65: ”tennis bat swing”.

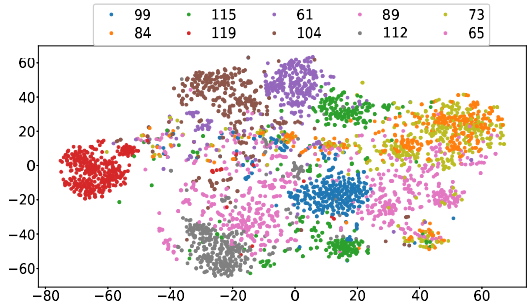

Figure 13. The t-SNE visualization of cls token outputs of NTU 60-120 train dataset with the model trained on NTU 120. The activities are as follow: 99: ”running on the spot”, 115: ”take a photo of other person”, 61: ”put on headphone”, 89: ”put something into a bag”, 73: ”staple book”, 84: ”play magic cube”, 119: ”support somebody with hand”, 104: ”stretch oneself”, 112: ”high-five”, 65: ”tennis bat swing”.

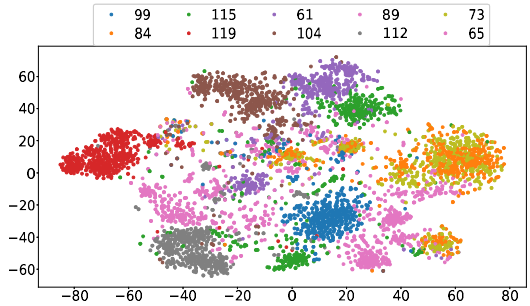

Figure 14. The t-SNE visualization of cls token outputs of NTU 60-120 validation dataset with the model trained on NTU 120. The activities are as follow: 99: ”running on the spot”, 115: ”take a photo of other person”, 61: ”put on headphone”, 89: ”put something into a bag”, 73: ”staple book”, 84: ”play magic cube”, 119: ”support somebody with hand”, 104: ”stretch oneself”, 112: ”high-five”, 65: ”tennis bat swing”.

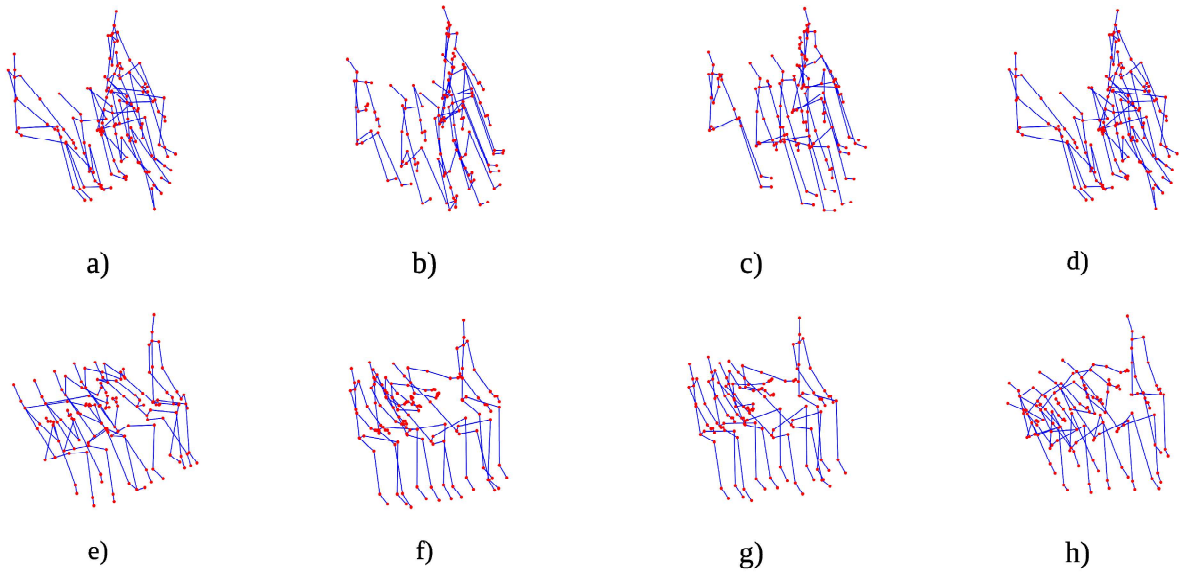

Figure 15. Input and output sequences of the transformer autoencoder. We use our transformer autoencoder to condition the skeleton sequence generation. In a) the encoder input with activity A016: ”wear a shoe” from the train dataset is shown. The decoder input is random noise. In b)-d) the decoder outputs for a) as encoder input are shown. In e) the encoder input with activity A016: ”wear a shoe” from the validation dataset is shown. The decoder input is random noise. In f)-h) the decoder outputs are shown for e) as encoder input.

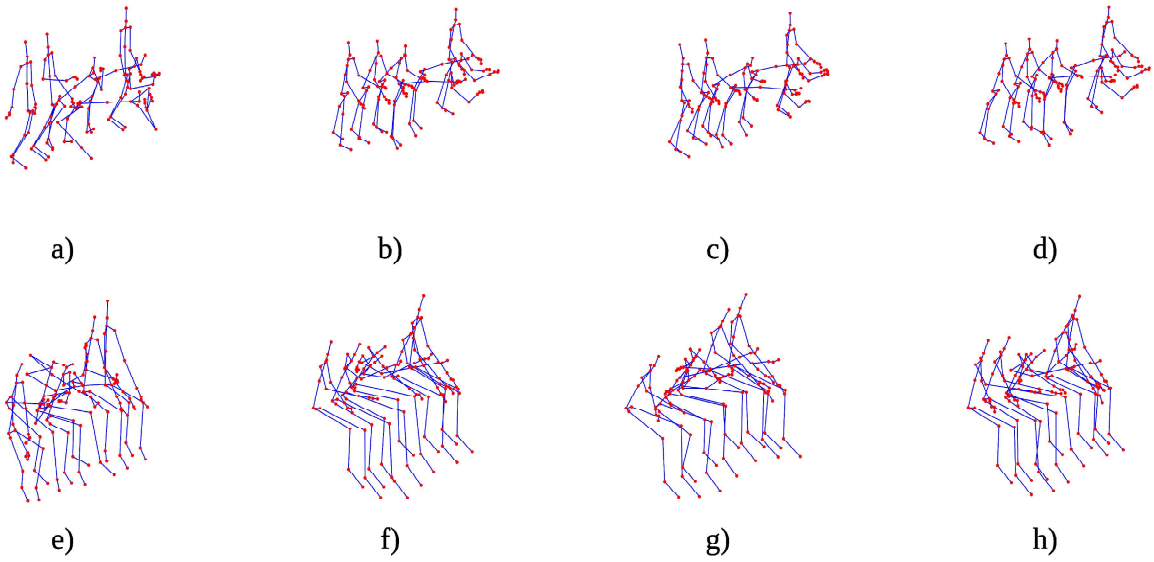

Figure 16. Input and output sequences of the transformer autoencoder. We use our transformer autoencoder to condition the skeleton sequence generation. In a) the encoder input with activity A048: ”nausea or vomiting condition” from the train dataset is shown. The decoder input is random noise. In b)-d) the decoder outputs for a) as encoder input are shwon. In e) the encoder input with activity A048: ”nausea or vomiting condition” from the validation dataset is shown. The decoder input is random noise. In f)-h) the decoder outputs for e) as encoder input are shown.

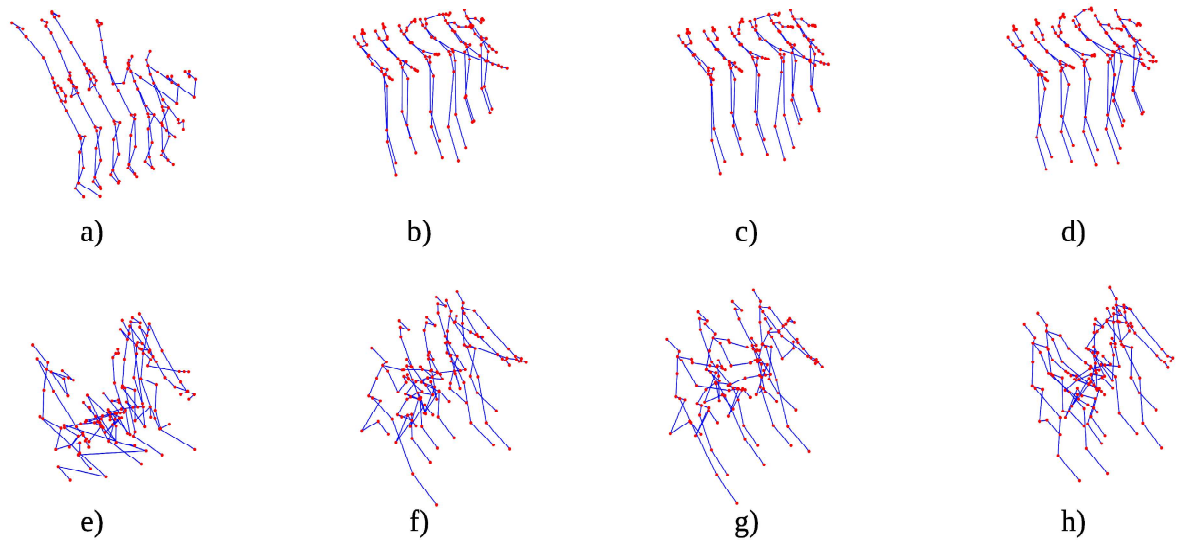

Figure 17. Input and output sequences of the transformer autoencoder. We use our transformer autoencoder to condition the skeleton sequence generation. In a) the encoder input with activity A104: ”stretch oneself” from the train dataset is shown. The decoder input is random noise. In b)-d) the decoder outputs for a) as encoder input are shown. In e) the encoder input with activity A104: ”stretch oneself” from the validation dataset is shown. The decoder input is random noise. In f)-h) the decoder outputs for e) as decoder input are shown.

Models	NTU 60 xview			NTU 60 xsub			NTU 120 xset			NTU 120 xsub		
	1%	10%	50%	1%	10%	50%	1%	10%	50%	1%	10%	50%
LongT GAN [38]	-	-	-	35.2	**62.0**	-	-	-	-	-	-	-
AS-CAL [18]	**53.5**	57.3	67.3	**47.2**	52.2	61.0	**38.3**	43.0	53.0	**36.0**	42.3	52.6
CR-AE [15]	26.5	-	**76.0**	32.3	-	65.0	18.9	-	53.7	16.1	-	50.7
TAHAR (Ours)	36.4	**61.2**	74.8	35.5	58.1	**66.4**	25.3	**49.1**	**60.6**	21.9	**44.9**	**58.2**

Table 6. Comparison of unsupervised trained autoencoders evaluated on action recognition datasets. The values are classification accuracies in [%] with LEP. The linear layer is trained with a fraction of the dataset.