1 Finetuning

As we now have pretrained our models, we need to finetune the models, such that they are specialized to yield optimial results on the ClimbAlong dataset. The following section describes the finetuning of these models. This includes the the various experiments we perform, the preprocessing of the data, the configuration details we use, as well as the obtained results.

In the finetuning stage we will be using the already developed pose-estiamtor to train our temporal-inclusive models. However, we will be freezing the pose-estimator, such that the weights of the model will not change during the training and we will thus only train our temporal-inclusive models. We do this for the following three reasons: (1) the training of the models will be quicker, as we just need to train the tempoeral-inclusive models and not the already developed pose-estimator, (2) we get an greater understanding of the effects of our models when combined with the pose-estimator, as we can clearly see how big of a difference it makes by adding our tempoeral-inclusive models, and (3) we lower the probability of overfitting, as we have less tuneable parameters.

1.1 Data Preprocessing

For the ClimbAlong dataset we perform only minor preprocessing. First, the preprocessing of each video is done by having the already developed pose-estimator process the video, such that we have the output heatmaps of the pose-estimator, containing all of the pose-estimations of each video. Next, we preprocess the heatmaps by setting all negative values to 0 and normalizing each heatmap, such that each heatmap sums up to the fixed value $c = 255$ that we used when preprocessing the BRACE and Penn Action datasets, essentially making the heatmaps more similar to the preprocessed heatmaps of BRACE and Penn Action. These heatmaps will then be used as the input for our models.

For the groundtruth heatmaps we create twenty five heatmaps of each frame, similarly to how we did it for the BRACE and Penn Action datasets, however, in this case we use the predicted bounding-box of the pose-estimator as our bounding-box. In cases where the groundtruth keypoint is placed outside of the bounding-box, we place the groundtruth keypoint at the closest border of the bounding-box.

1.2 Training Details

Data Configuration Generally, we follow a similar approach to how we did in the pretraining stage. We again use a window-size of $k = 5$ frames, resulting in a total of $9,419$ windows. Also here are we using $c = 255$ as a representation of the placement of each keypoint. We also split the dataset into a non-overlapping and non-repeating training, validation and test set, consisting of 60%, 20% and 20% of the data, respectively. However, we note that one incorrect frame can have a huge impact on the evaluation results, as this frame is used five times during evaluation, due to the small dataset size. For that reason, for the validation and test set we make sure that none of the windows of the same set.

Setups As the finetuning dataset is so small, the fitting of the models is very quick, making us fit all of the developed models from the pretraining stage. For each model we pick the epoch from the pretraining stage, that yielded the highest validation accuracy and use that for finetuning.

Training Configuration The optimization parameters are very similar to the ones from the pretraining stage. We again use the ADAM optimzer with a batch size of 16 and the MSE

loss-function. During training, we again keep track of the lowest reached validation loss of an epoch and use learning-rate reduction and early-stopping in a similar manner to how we did in the pretraining stage. However, unlike the pretraining stage, we here use a smaller initial learning rate of 10^{-6}, as the weights only need to be fineadjusted, making us believe that greater learning rate would skew the weights too much.

1.3 Training and Validation Results

- Vi skal kun bruge én epoch - nok pga. pretræning

1.4 Test Results

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Mean threshold distance*	0.80			1.60			3.21		
Setup	1.1	1.2	1.3	1.1	1.2	1.3	1.1	1.2	1.3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
Conv3D	49.7	52.3	53.1	**95.7**	**95.7**	**95.8**	99.2	99.3	99.3
DeciWatch	**76.7**	**76.7**	**68.1**	94.4	94.4	87.3	99.2	99.2	96.3
bi-ConvLSTM - sum.	37.8	34.9	39.0	91.8	92.1	92.2	99.4	**99.7**	99.2
bi-ConvLSTM - concat.	35.9	39.0	38.5	93.1	93.6	92.6	**99.8**	**99.7**	**99.7**

Table 1: Testing accuracies of the various developed models for shifting-scalar $k = 1$. All the accuracies are in percentage. *: The mean maximum distance between the predicted keypoint and corresponding groundtruth keypoint for the prediction to count as being correct, using the units of the heatmap coordinates.

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Mean threshold distance*	0.80			1.60			3.21		
Setup	2.1	2.2	2.3	2.1	2.2	2.3	2.1	2.2	2.3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
Conv3D	**46.5**	51.6	**47.3**	**95.5**	**95.5**	**95.8**	99.2	99.3	99.2
DeciWatch	39.4	**63.3**	28.5	68.2	91.0	69.2	93.1	99.1	93.7
bi-ConvLSTM - sum.	38.8	37.4	35.9	92.7	92.1	91.2	99.4	**99.5**	99.3
bi-ConvLSTM - concat.	39.2	39.5	37.1	92.5	92.9	92.6	**99.6**	99.3	**99.6**

Table 2: Testing accuracies of the various developed models for shifting-scalar $k = 2$. All the accuracies are in percentage. *: The mean maximum distance between the predicted keypoint and corresponding groundtruth keypoint for the prediction to count as being correct, using the units of the heatmap coordinates.

- DeciWatch tager meget stor effekt af shifting-scalar
- Shifting-scalar har ikke den store effect på unipose1 og unipose2
- Frame skipping har en stor effekt på deciwatch eller uniposes
- Unipose2 er lidt bedre end unipose1
- Uniposes tager stortset ingen effekt af den ekstra noise
- Uniposes performer lidt bedre i 1.2 end i 1.1

- Baseline performer lidt bedre i 1.2 end i 1.1
- Sammelign med andre modeller fra andre datasæt?
- Frame skipping har ingen effekt på baseline (det har næsten en positiv effekt)
- Noise har lidt effekt på PCK@0.05, men ikke på de andre

	Conv3D			DeciWatch			bi-ConvLSTM sum.			bi-ConvLSTM concat.			Total
Setup	1.1	1.2	1.3	1.1	1.2	1.3	1.1	1.2	1.3	1.1	1.2	1.3	
Nose	100	100	100	99.8	99.8	97.5	100	100	99.9	100	99.7	99.9	99.7
Ear	100	100	100	99.8	99.8	97.7	99.8	99.9	100	100	100	99.9	99.7
Shoulder	99.9	100	99.9	99.8	99.8	97.5	100	100	99.9	100	100	100	99.7
Elbow	99.9	99.9	99.9	99.4	99.5	96.9	100	100	100	100	99.9	100	99.6
Wrist	99.8	99.9	99.9	99.1	99.2	96.2	100	99.9	99.8	100	99.9	100	99.5
Pinky	93.4	93.1	94.4	98.3	98.4	94.4	97.2	98.8	97.0	98.0	99.0	98.6	96.7
Index finger	99.0	98.8	98.8	98.2	98.3	94.0	99.5	98.7	97.0	99.6	99.4	99.4	98.4
Thumb	98.9	98.8	98.9	98.2	98.3	95.0	96.8	99.6	97.8	99.7	98.6	99.6	98.3
Hip	99.9	100	100	99.8	99.8	97.6	100	100	100	100	100	100	99.8
Knee	100	100	99.9	99.7	99.7	97.3	100	100	100	100	100	100	99.7
Ankle	100	100	100	99.5	99.5	96.9	100	100	99.9	100	100	99.9	99.6
Heel	100	100	100	99.3	99.3	96.3	99.3	99.9	99.9	99.9	100	99.8	99.5
Foot	99.9	100	100	99.0	99.1	95.4	99.6	99.8	99.4	99.8	100	99.8	99.3

Table 3: Keypoint-specific testing PCK@0.2-accuracies of the various models for shiting-scalar $k = 1$.All the accuracies are in percentage.

	Conv3D			DeciWatch			bi-ConvLSTM sum.			bi-ConvLSTM concat.			Total
Setup	1.1	1.2	1.3	1.1	1.2	1.3	1.1	1.2	1.3	1.1	1.2	1.3	
Nose	100	100	100	89.2	99.8	94.3	100	99.9	99.7	99.7	99.9	100	98.5
Ear	100	99.8	100	96.7	99.7	93.9	99.8	99.8	100	99.9	99.9	99.9	99.1
Shoulder	99.9	99.7	99.9	93.8	99.2	92.0	99.9	99.8	100	99.9	100	100	98.7
Elbow	99.8	99.9	99.9	90.2	99.4	89.8	100	99.5	100	100	100	100	98.2
Wrist	99.8	100	99.9	84.9	99.1	94.3	99.8	99.7	99.8	99.8	100	99.7	98.1
Pinky	93.1	93.7	93.9	98.4	98.4	93.4	97.7	98.0	97.9	99.1	96.0	98.2	96.5
Index finger	98.9	99.0	98.8	98.2	98.3	93.5	99.4	99.1	99.2	99.6	98.0	97.5	98.3
Thumb	98.6	98.6	98.6	98.3	98.4	94.3	95.7	96.6	98.6	97.5	98.3	99.6	97.8
Hip	100	99.9	100	93.0	99.1	93.9	99.9	99.8	99.9	99.8	100	99.9	98.8
Knee	100	100	99.9	79.5	99.5	93.8	100	99.9	99.9	99.9	99.9	100	97.7
Ankle	100	99.9	100	87.1	99.4	96.3	100	100	100	100	100	100	98.6
Heel	100	100	100	99.1	99.1	95.7	99.9	99.9	99.8	99.9	99.6	99.9	99.4
Foot	99.9	100	100	99.0	98.9	94.9	99.9	99.4	98.4	99.8	99.6	99.8	99.1

Table 4: Keypoint-specific testing PCK@0.2-accuracies of the various models for shiting-scalar $k = 2$.All the accuracies are in percentage.

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Mean threshold distance*	0.87			1.77			3.55		
Setup	1.1	1.2	1.3	1.1	1.2	1.3	1.1	1.2	1.3
Identity function	21.2	21.2	21.2	65.5	65.5	65.5	84.7	84.7	84.7
Conv3D	58.4	61.4	61.7	98.7	98.9	99.0	99.6	99.8	99.7
DeciWatch	82.4	82.3	75.3	95.6	95.7	92.8	99.1	99.1	97.7
bi-ConvLSTM - sum.	45.7	45.0	47.6	97.3	96.9	97.0	99.6	99.6	99.1
bi-ConvLSTM - concat.	44.5	46.1	48.5	97.4	97.9	97.9	99.6	99.5	99.6

Table 5: Testing accuracies of the various developed models for shifting-scalar $k = 1$ on the additional test video. All the accuracies are in percentage. *: The mean maximum distance between the predicted keypoint and corresponding groundtruth keypoint for the prediction to count as being correct, using the units of the heatmap coordinates.

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Mean threshold distance*	0.87			1.77			3.55		
Setup	2.1	2.2	2.3	2.1	2.2	2.3	2.1	2.2	2.3
Identity function	21.2	21.2	21.2	65.5	65.5	65.5	84.7	84.7	84.7
Conv3D	56.2	60.0	56.6	98.9	98.8	98.8	99.7	99.7	99.7
DeciWatch	42.8	69.8	39.8	72.7	93.5	81.0	96.0	99.0	96.8
bi-ConvLSTM - sum.	44.8	46.2	45.0	96.9	95.9	97.1	99.5	99.6	99.5
bi-ConvLSTM - concat.	45.9	47.9	46.7	96.7	97.1	98.1	99.6	99.4	99.6

Table 6: Testing accuracies of the various developed models for shifting-scalar $k = 2$ on the additional test video. All the accuracies are in percentage. *: The mean maximum distance between the predicted keypoint and corresponding groundtruth keypoint for the prediction to count as being correct, using the units of the heatmap coordinates.

2 Further Test Results

3 Technical Details