**Bachelor Thesis**

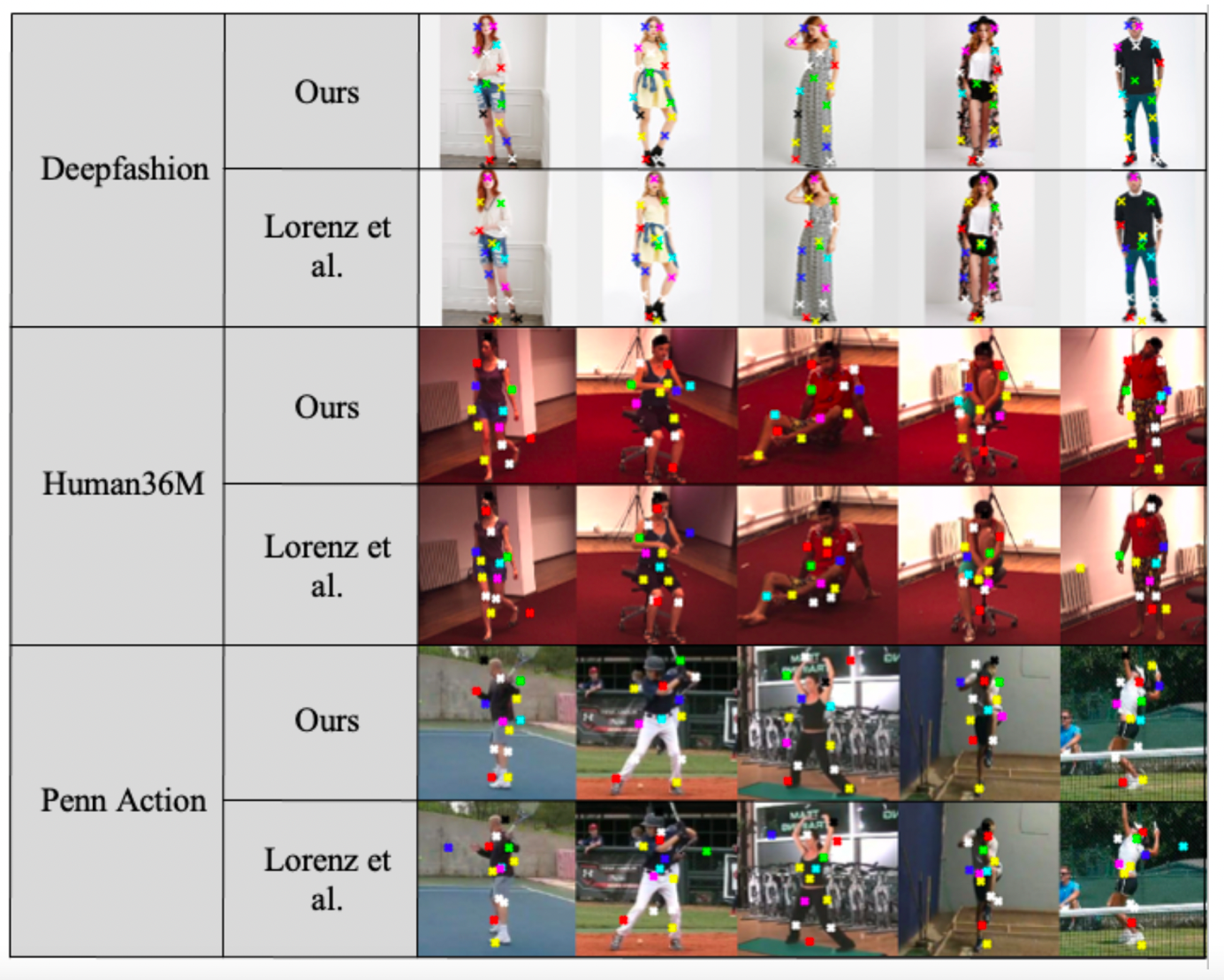
*“Unsupervised Human Landmark Detection with Vision Transformers”*

Initial Challenges:

* Difficulties to distinguish background from foreground
* Instability (numerical and performance-wise)

Idea: Global dependencies important to explain human pose (pose of the hands – feet)

🡪 **Replace convolutional modules** in baseline model (Lorenz et al.) with **ViTs**.



**Application on KIA Project**

Challenging Dataset: 🡪 Multiple persons in some scenes

🡪 Large occlusions

* Not predictable with our model architecture
* Instead: **Preprocessing** to get single-person images and usage of **pretrained model** (on PennAction) with subsequent **finetuning**



Ein Bild, das Text, Screenshot enthält.

Automatisch generierte Beschreibung

**Master Thesis**

*“Unsupervised Representation Learning for Video”*

Current Challenges:

* Not clear, how models interpret videos (relevance of texture / motion?)
* Previous study (“*ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*”) has identified a bias towards texture for images:

Ein Bild, das Text, Säugetier enthält.

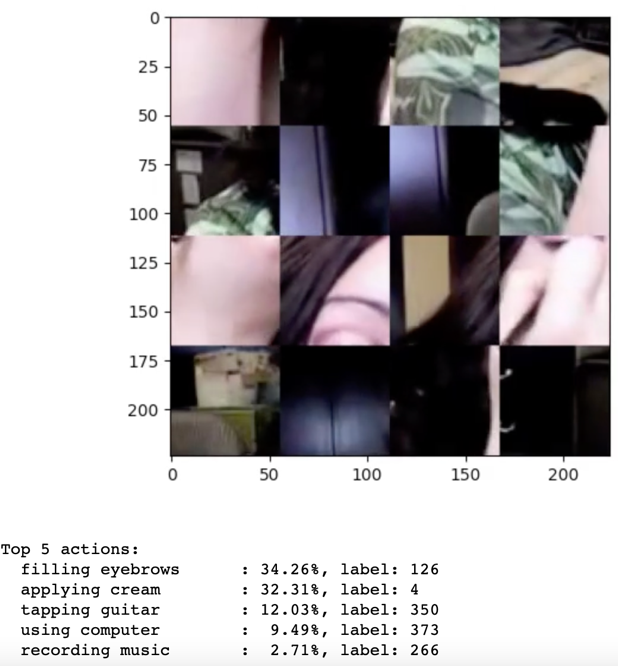
Automatisch generierte Beschreibung

* Study of student in our group suggests, that this might also be the case for videos
* Both FVD and OT score sensitive against appearance augmentations (Gaussian Blur, Gaussian Noise, Brightness, Saturation etc.)

Goal:

* Intensify study on how appearance and shape augmentations are reflected in performance and FVD / OT score
* Video models should be sensitive to motion – is this currently the case?

🡪 collect further statistics for more complex augmentations (patch-shuffling, style-transfer etc.)



Video still recognized despite loss of shape

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Are there differences between models trained supervised / unsupervised?
* Potentially develop new metric which better reflects robustness / performance