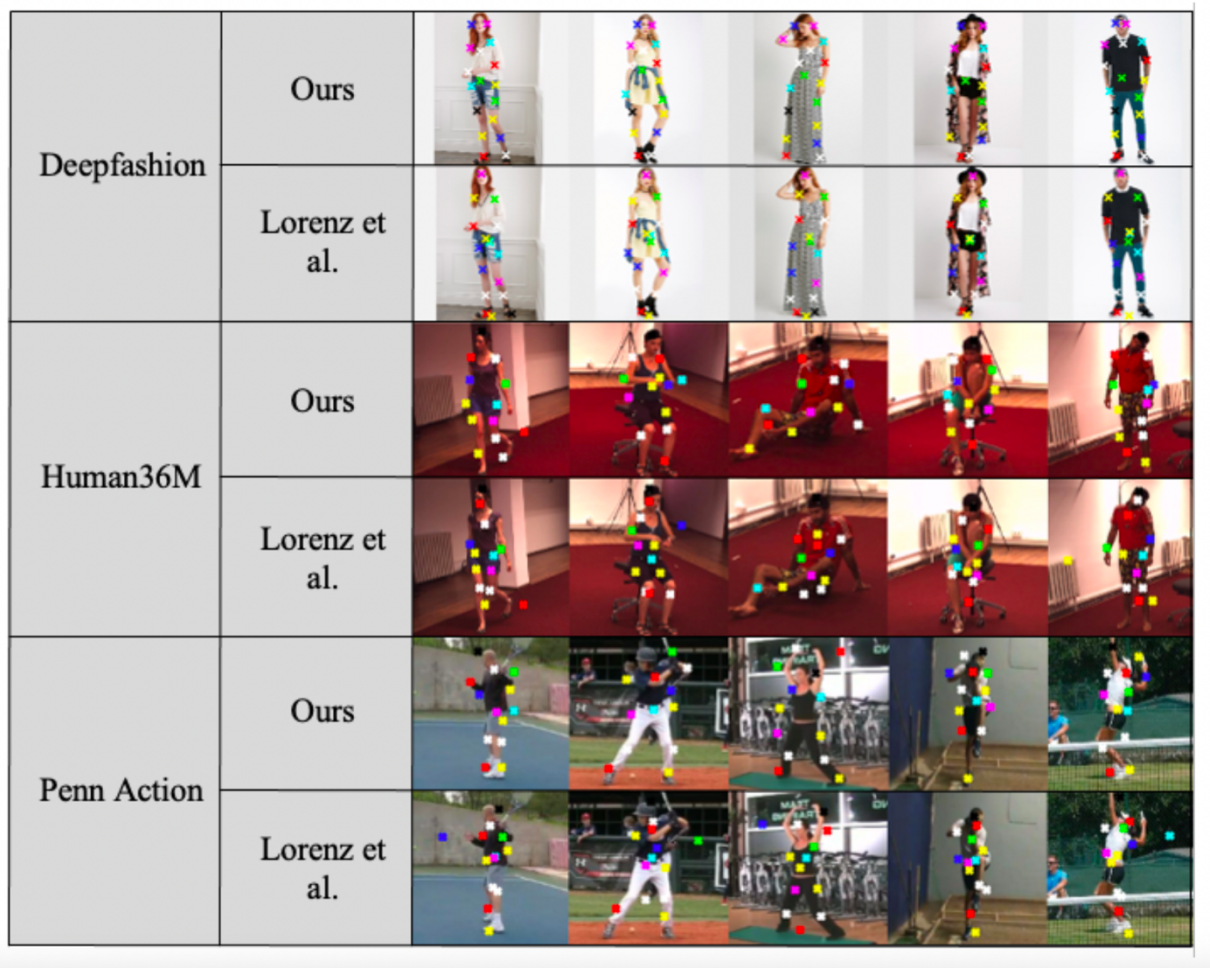
Bachelor Thesis

"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**

Landmark predictions with model pre-trained on PennAction (no fine-tuning)

Landmark predictions with model pre-trained on PennAction (with fine-tuning)

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:

(a) Texture image

81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image

71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

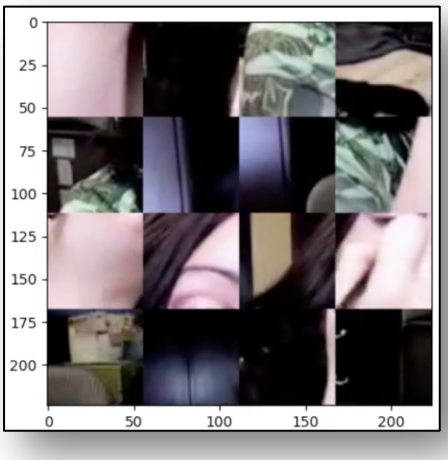
(c) Texture-shape cue conflict

63.9% **Indian elephant**
26.4% indri
9.6% black swan

- 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

```
Top 5 actions:
  filling eyebrows      : 34.26%, label: 126
  applying cream        : 32.31%, label: 4
  tapping guitar        : 12.03%, label: 350
  using computer        :  9.49%, label: 373
  recording music       :  2.71%, label: 266
```

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