Semi-Supervised Learning of Monocular 3D Hand Pose Estimation from Multi-View Images

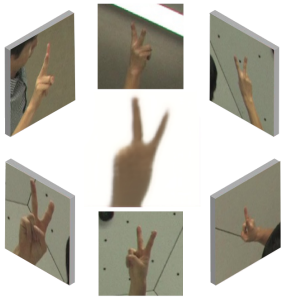
Markus Müller, Georg Poier, Horst Possegger, Horst Bischof,
Institute of Computer Graphics and Vision, Graz University of Technology

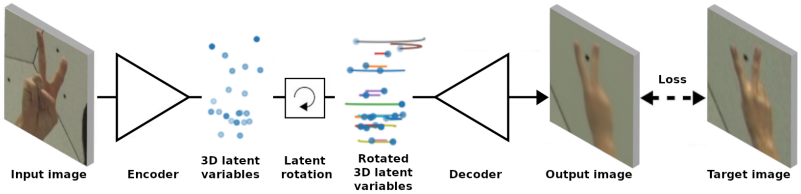
IEEE ICIP 2021

Novel View Synthesis

- Unlabeled multi-view images
- Unsupervised learning
- Network learns to synthesize new views

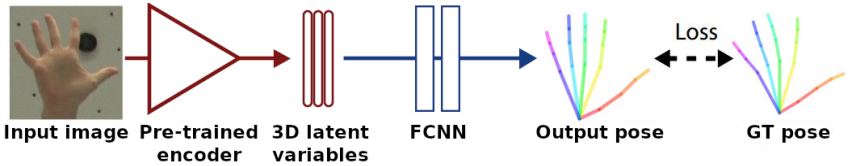
- Encode image to 3D point cloud
- Transform 3D points to novel view





3D Hand Pose Estimation

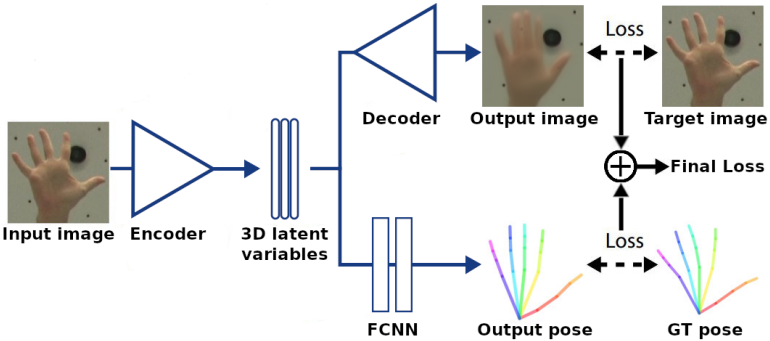
- Rotation matrix connecting both views as additional input
- Learn a mapping from $\mathbf{L}^{3D}$ to the 3D hand pose
- Much less annotations required
- Supervised learning
- Pipeline using a **pre-trained** encoder, inspired by [Rhodin'18]



Rhodin et al. Unsupervised geometry-aware representation learning for 3d human pose estimation. ECCV, 2018

Semi-Supervised Approach

- **Semi-supervised** network
  - Train encoder-decoder and pose network simultaneously



Dataset

- CMU Panoptic Dataset from [Joo'19]
  - Synchronized camera feeds from 31 HD cameras
  - **Training:** 48 683 frames per camera
  - **Testing:** 16 366 frames per camera
- We compare
  - **Pre-trained** network
    - Pre-training with and without augmented images
  - **Semi-supervised** network
  - **Supervised** network
    - Input image gets directly mapped to the 3D pose

Joo et al. Panoptic studio: A massively multiview system for social interaction capture. TPAMI 41(1), 2019

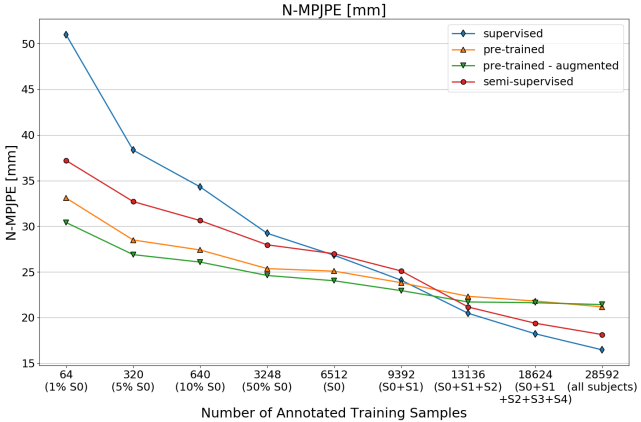
Defined Scenarios

- Provide different levels of supervision

Scenario	# Annotations
Fully supervised training with the 3D annotations of all **9** training subjects	28 592
S0 + S3 + S4 + S5 + S6	18 624
S0 + S3 + S4	13 136
S0 + S3	9 392
S0	6 512
50% of S0	3 248
10% of S0	640
5% of S0	320
1% of S0	64

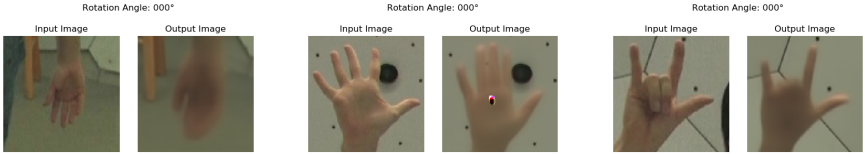
Evaluation

- Normalized Mean Per Joint Position Error (N-MPJPE)

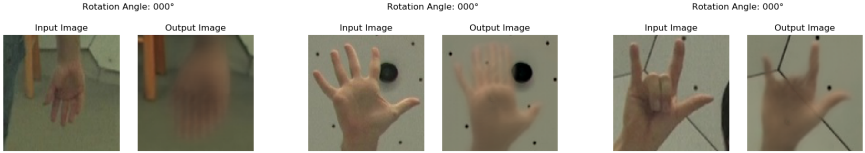


Novel View Synthesis

- Pre-trained encoder

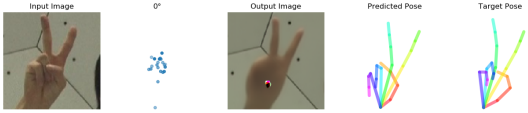


- Pre-trained encoder with augmented unlabeled multi-view images

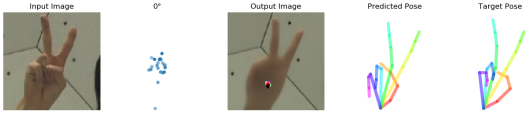


Pre-Trained Network

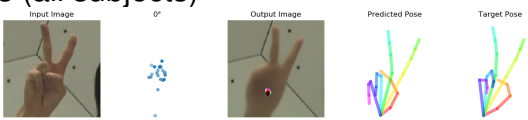
- 320 annotations



- 6 512 annotations

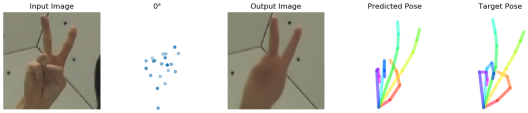


- 28 592 annotations (all subjects)

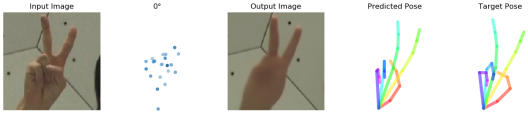


Semi-Supervised Network

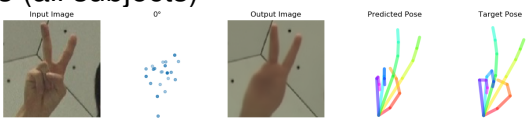
- 320 annotations



- 6 512 annotations

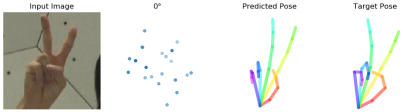


- 28 592 annotations (all subjects)

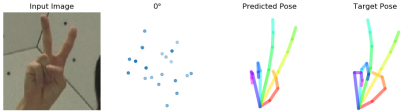


Supervised Network

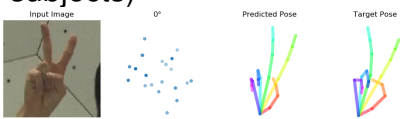
- 320 annotations



- 6 512 annotations



- 28 592 annotations (all subjects)



Conclusion

- Semi-supervised network outperforms fully-supervised methods
- A geometry-aware representation for 3D hand pose estimation performs much better when only few annotated data is available
- Performance increased with augmented multi-view images to pre-train the encoder-decoder network
- Training takes less time for the semi-supervised network than for the pre-trained approach

Thank you for watching!