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

Markus Müller[1,2] Georg Poier[1,3] Horst Possegger[1] Horst Bischof[1]

[1]Graz University of Technology, [2]Naked Labs, [3]Reactive Reality

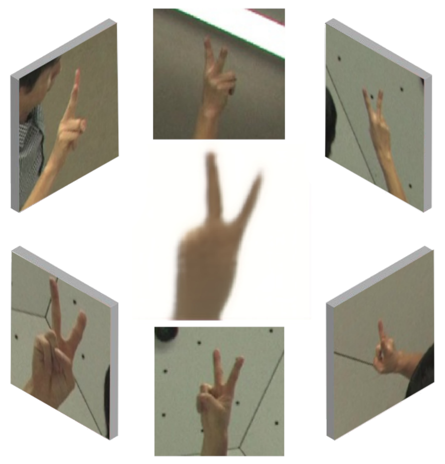
1462

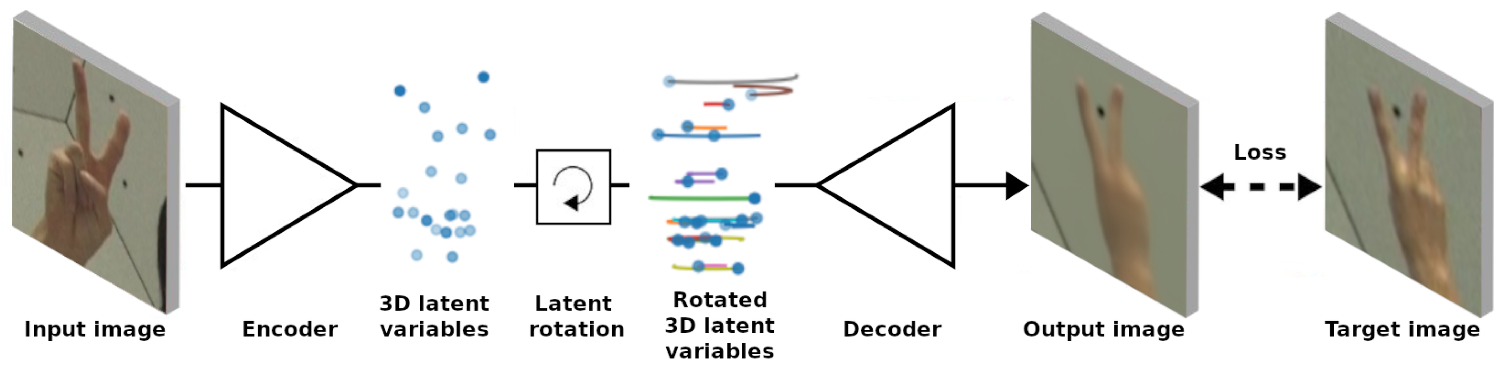
Motivation

- We address the long standing problem of data **annotation**, which is especially **expensive** for 3D hand pose estimation
- We use **unlabeled data** to reduce the required amount of annotated data
- We propose a **semi-supervised** method, that learns to estimate the 3D pose of a monocular hand image
- We build upon the work of Rhodin *et al.* [2] on 3D human pose estimation which we adapt and optimize to work for hands and to **jointly** handle **labeled and unlabeled data** in an end-to-end manner

Idea

- We learn a **geometry-aware representation** of the human hand from multi-view images without annotations
- The **encoder-decoder** is trained to predict an image seen from one view given an image captured from a different view

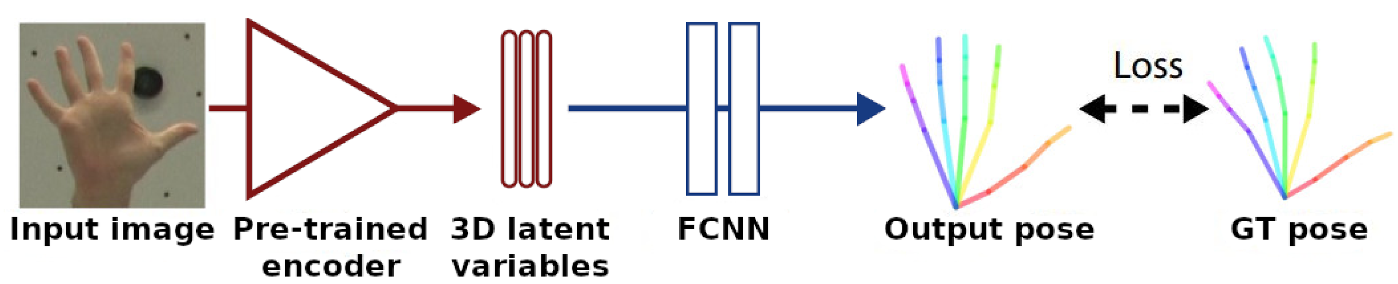




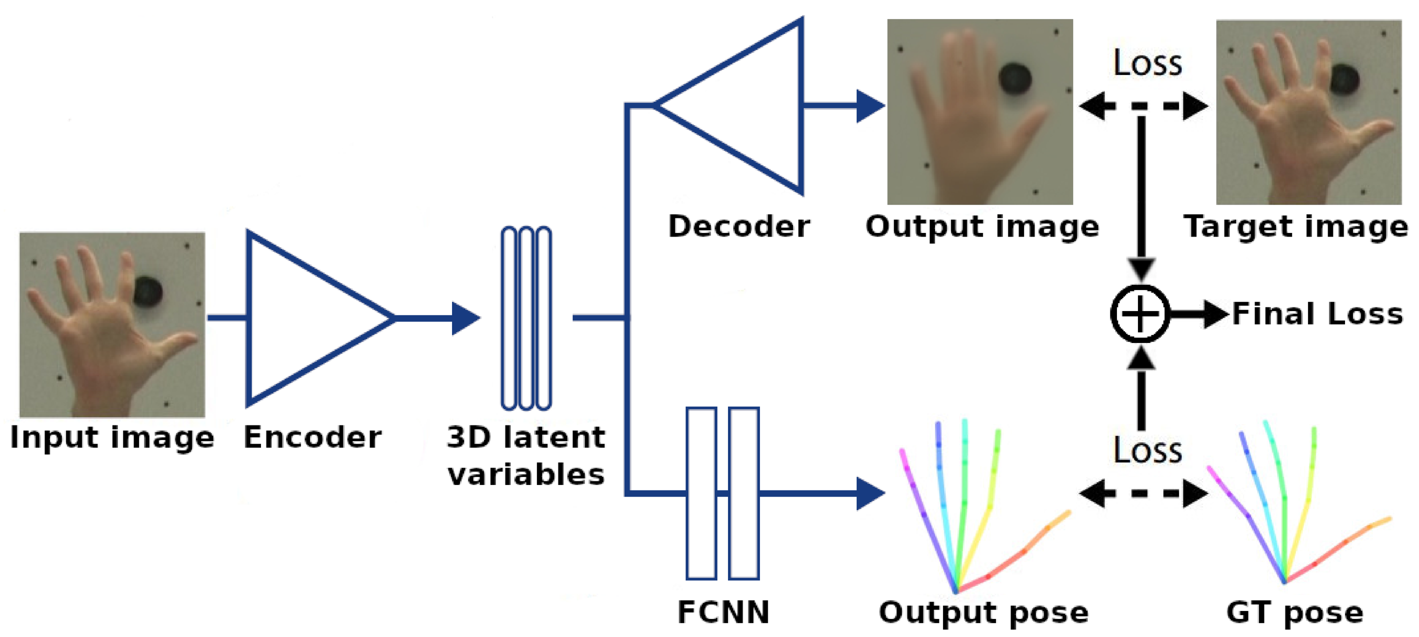
- We train an encoder-decoder network to learn an **unsupervised geometry-aware representation** for 3D hand pose estimation
- We use the **latent representation** to learn a mapping to the 3D pose in a supervised manner
- The latent representation already captures the 3D geometry, therefore
  - → the mapping is much simpler and
  - → considerably fewer examples are required to learn the mapping
- We use sequences of RGB images acquired from multiple synchronized and calibrated cameras
- We feed the rotation matrix $\mathbf{R}^{i\to j}$ connecting $\mathbf{I}_t^i$ and $\mathbf{I}_t^j$ as an additional input to the encoder and decoder, and train them to encode $\mathbf{I}_t^i$ and resynthesize $\mathbf{I}_t^j$

Geometry-Aware 3D Hand Pose Estimation

- Novel views of the corresponding hand pose can be rendered by manipulating the rotation parameter $\mathbf{R}^{i\to j}$
- We split the background and the appearance of the hand from the latent representation to only encode the 3D geometry
- We model the latent representation $\mathbf{L}^{3\mathrm{D}} \in \mathbb{R}^{3\times N}$ of the input image as a set of N points in 3D space
- $\mathbf{L}^{3\mathrm{D}}$ has the semantic meaning of a skeleton with $K = 21$ hand joints encoded as a vector $\mathbf{P} \in \mathbb{R}^{3K}$
- We learn a mapping $\mathcal{F} : \mathbf{L}^{3\mathrm{D}} \to \mathbb{R}^{3K}$
- $\mathcal{F}$ is modeled as a Fully Connected Neural Network (FCNN)
- This mapping requires only a small amount of annotated data
- **Pre-trained** network as proposed by Rhodin *et al.* [2] combines the encoder-decoder network (unsupervised training) and the FCNN (supervised training):



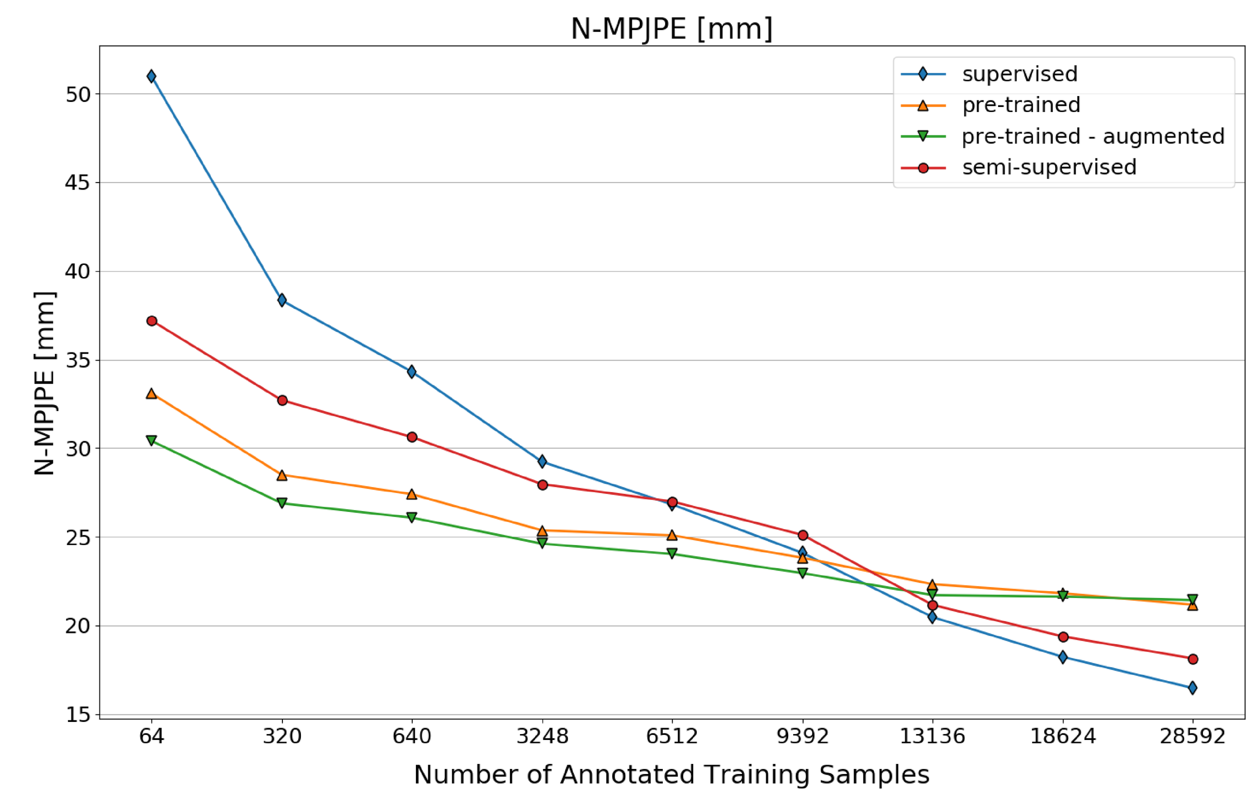
- Our **semi-supervised** network simultaneously optimizes the weights of the encoder, decoder and FCNN:



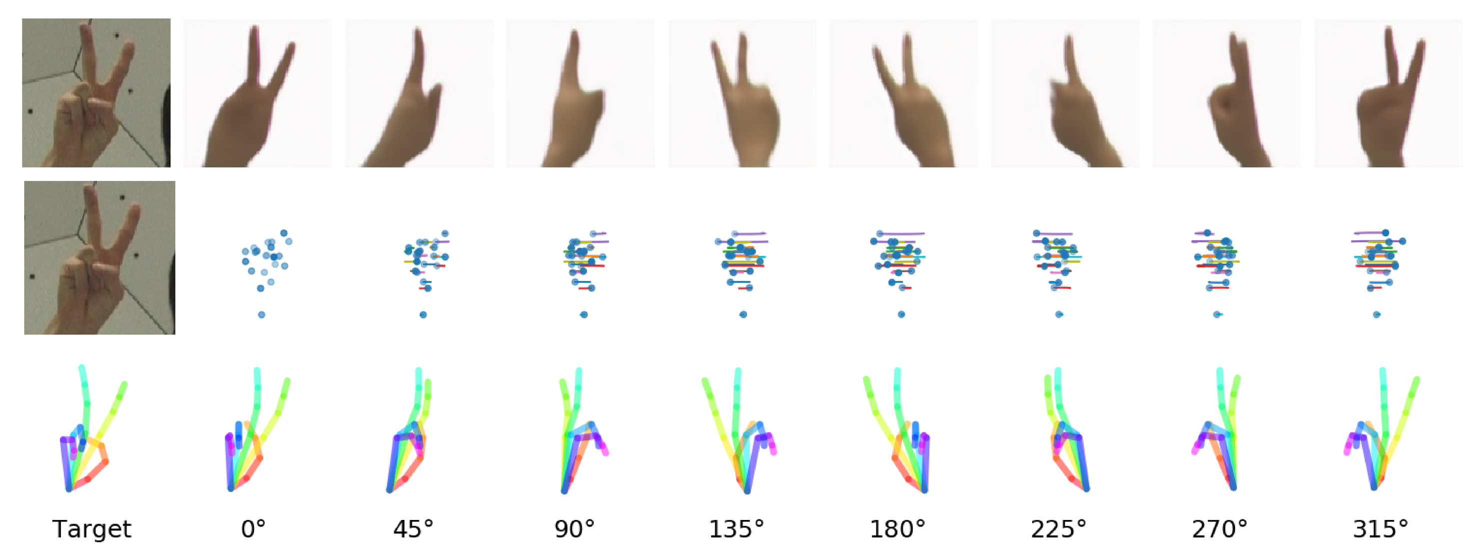
- **Semi-supervised** setup almost **halves the training time**
- We use a **random minority oversampling** method to compensate for the imbalances in labeled and unlabeled data

Findings

- **Supervised**: The network directly maps an input image to the 3D pose, without pre-training the encoder with unlabeled images
- **Pre-trained**: This is the hand pose estimation network using a pre-trained encoder
- **Semi-supervised**: The encoder-decoder network and the pose network are trained simultaneously
- Define different levels of supervision to train the FCNN
- Normalized Mean Per Joint Position Error on CMU Panoptic dataset [1]



- Qualitative results for Novel View Synthesis, 3D latent variables and pose predictions of the **semi-supervised** network:



References

[1] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. C. Nabbe, I. A. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(1):190–204, 2019.

[2] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation learning for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.