Learning Pose Specific Representations by Predicting Different Views

Georg Poier David Schinagl Horst Bischof

Graz University of Technology

Project page: poier.github.io/PreView/

CVPR 2018 SALT LAKE CITY • JUNE 18-22

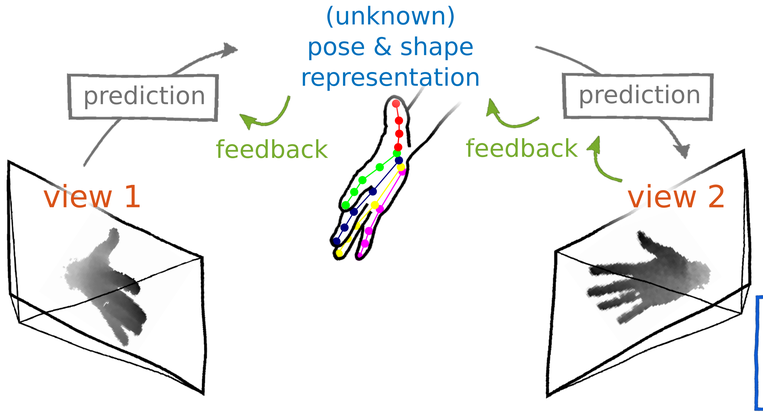
Motive

- Learning accurate models needs large amount of labeled data
- Accurate labeling important
- Synthetic data not enough

➜ **High Labeling effort**

➜ **Want to learn without labels**

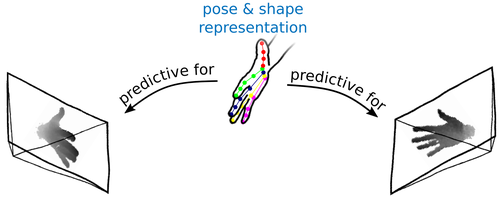
!?



But... what did we do? erm... learn?

Idea

- **Observation**
  If you know the pose
  ➜ *you can imagine the appearance from any view*



- We don't know the pose
- But we can easily capture another view
- Reformulate task
  as learning to predict different view
  ➜ should reveal pose representation

Let's do it!

Implementation

According to our observation, there exists a function g_i that maps a pose representation θ to any view i:

$$\mathbf{x}^{(i)} = g_i(\theta)$$

What we search for is the mapping from input to pose:

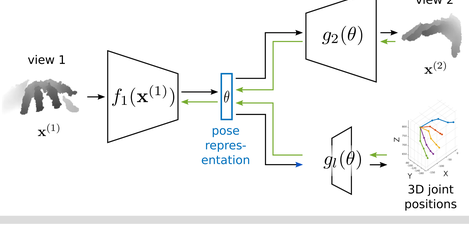
$$\theta = f_i(\mathbf{x}^{(i)})$$

Substituting θ, i.e, subsequently applying f and g, we can map from one view to another:

$$\mathbf{x}^{(j)} = g_j(f_i(\mathbf{x}^{(i)}))$$

➜ Learning the full mapping should reveal an implicit pose representation

- Trained using reconstruction objective
- Assumption/Constraint: Lowest dimensional representation from which different view can be reconstructed will be pose specific
- Semi supervised learning via additional joint position loss term



Works!

Analysis

- Nearest neighbors in latent space


Autoencoder

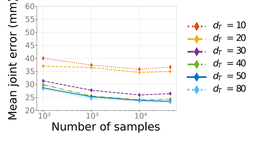

PreView (ours)

- Predictability of pose from latent representation
  - test via linear mapping

n	Autoencoder	PreView (Ours)	
100	48.0 ± 0.76	**33.4** ± 1.18	−30.4%
1,000	47.2 ± 0.29	**29.6** ± 0.32	−37.3%
10,000	47.3 ± 0.08	**29.0** ± 0.14	−38.7%
43,640	47.1 ± 0.08	**29.0** ± 0.09	−38.4%

Mean joint error on NYU-CS

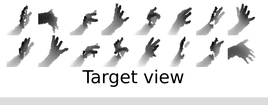
Representation size



- View Predictions


Input view


Input reconstruction (Autoenc.)


Target view

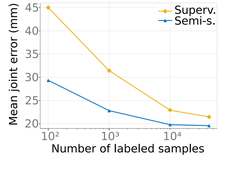

Prediction for different view (Ours)

Pose Estimation Results

- Semi-supervised learning for pose estimation
  - NYU-CS (Mean joint error (ME))

Number of labeled samples	100	1,000	10,000	43,640
DeepPrior [CVWW 2015]	44.99	36.99	30.31	27.97
Crossing Nets [CVPR 2017]	67.65	36.35	28.97	25.57
DeepPrior++ [ICCVw 2017]	38.07	31.01	24.14	20.87
Semi-superv. Autoenc.	31.58	24.05	21.32	20.74
Semi-superv. PreView (Ours)	**29.35**	**22.83**	**19.81**	**19.60**

Ablation



  - MV-hands (**published novel dataset**)

	ME	FS80	JS80
DeepPrior++ [ICCVw 2017]	34.17	0.22	0.57
Supervised (Baseline)	26.35	0.36	0.67
Semi-superv. Autoencoder	25.20	0.38	0.68
Semi-superv. PreView (Ours)	**24.14**	**0.39**	**0.69**