Training a Feedback Loop for Hand Pose Estimation

Markus Oberweger, Paul Wohlhart and Vincent Lepetit

Institute for Computer Vision and Graphics

Graz University of Technology

Motivation

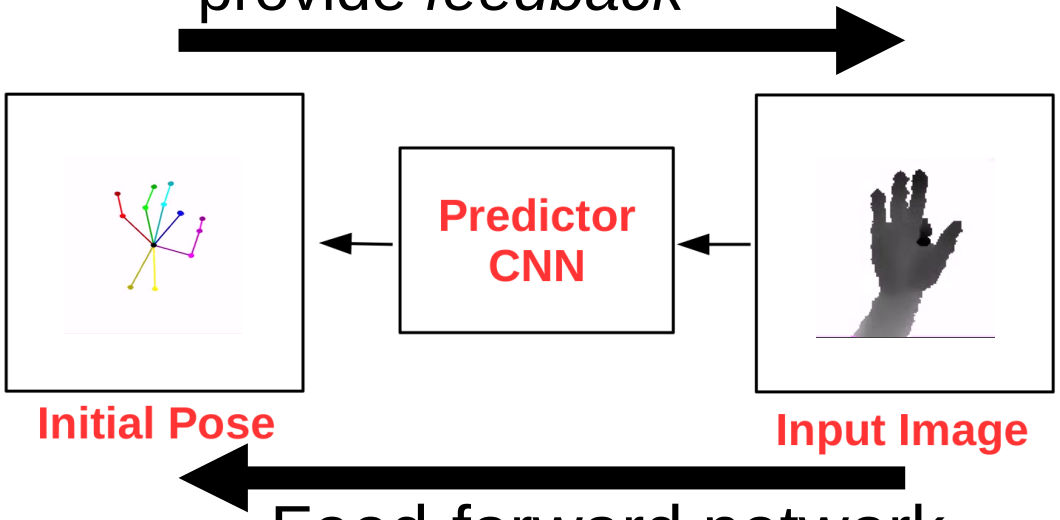
- Important for Human Computer Interaction, Augmented Reality
- Hot topic: [Tang et al., ICCV'15], [Tzionas&Gall, ICCV'15], [Li et al., ICCV'15], [Supančič et al., ICCV'15], [Choi et al., ICCV'15], [Rogez et al., ICCV'15], [Xu&Cheng, ICCV'13], [Tompson et al., ToG'14], [Tang et al., ICCV'13], [Sridhar et al., ICCV'13], [Sharp et al., CHI'15], [Qian et al., CVPR'14], [Oikonomidis et al., ICCV'11], [Melax et al., GIC'13], [Kuznetsova et al., ICCV'13], [Keskin et al., ICCV'11], [de La Gorce et al., PAMI'11], [Ballan et al., ECCV'12], Leap Motion
- Our goal: Accurate 3D pose from single depth image

Challenges

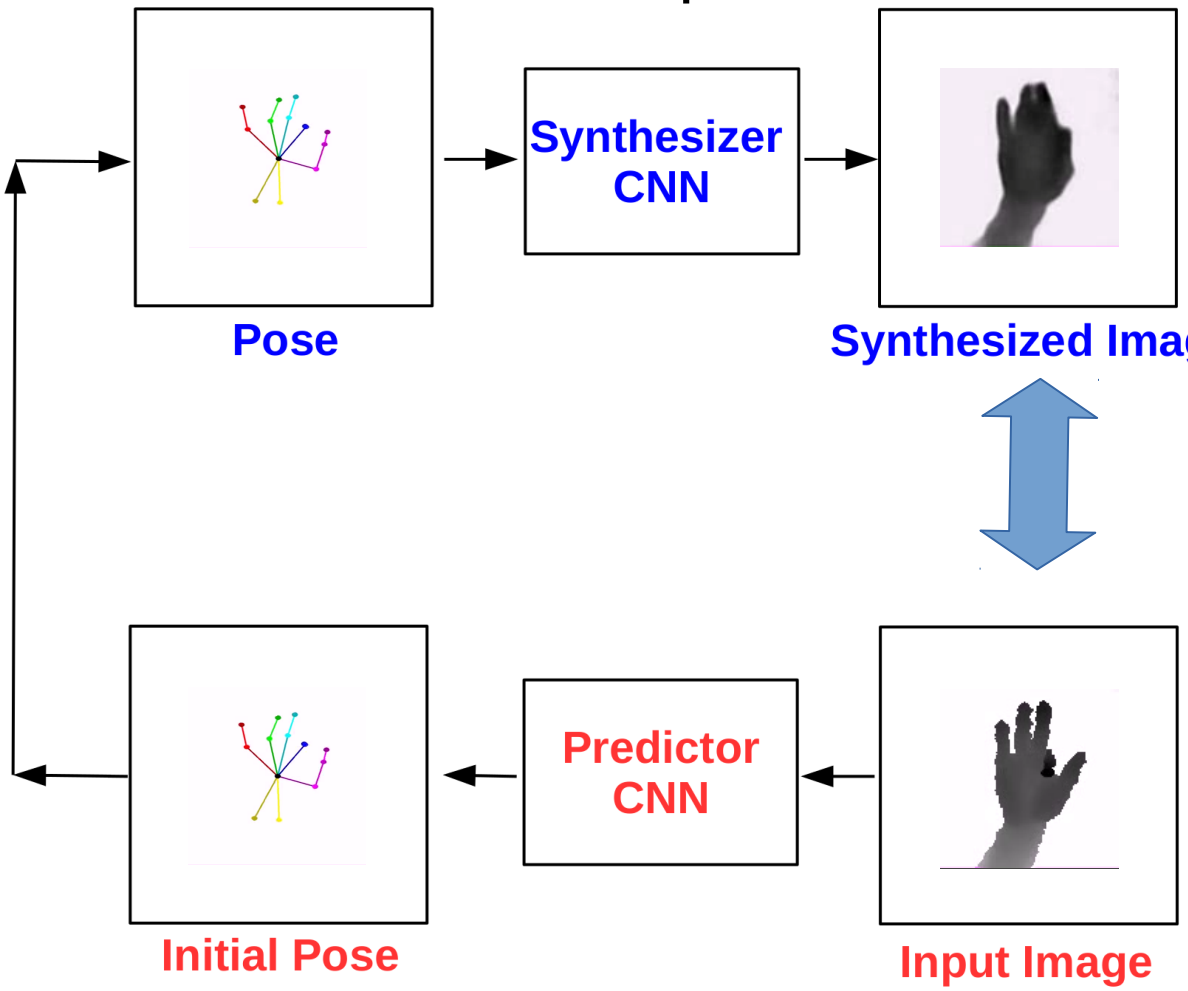
- Self-occlusions
- Self-similarity
- Many degrees-of-freedom
- Noisy input

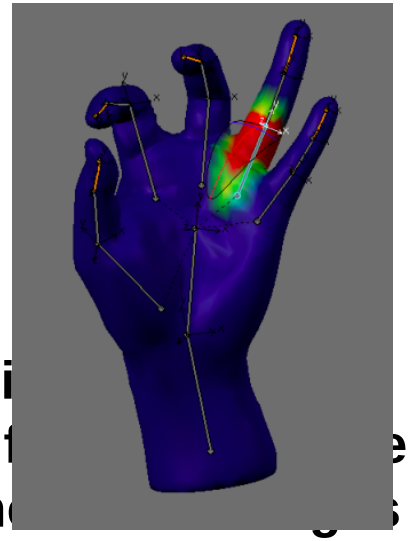
A Predictor

Many neurons in the visual cortex provide *feedback*



Starting a Feedback Loop:
Training a CNN to synthesize depth images from pose





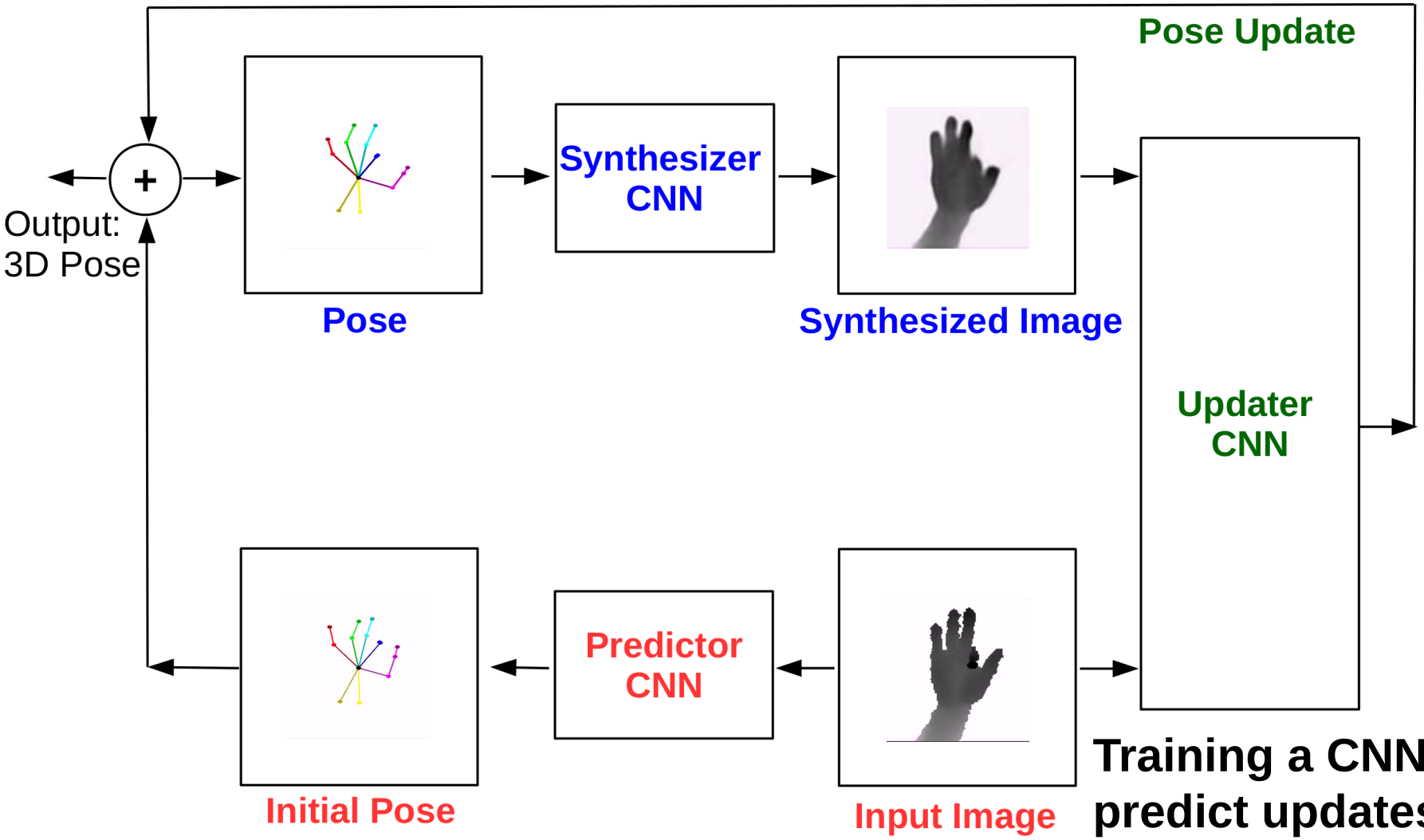
[Sharp et al.]

Mi… dif… th… to estimate the pose?

No need for a 3D model

Does not work well in practice!

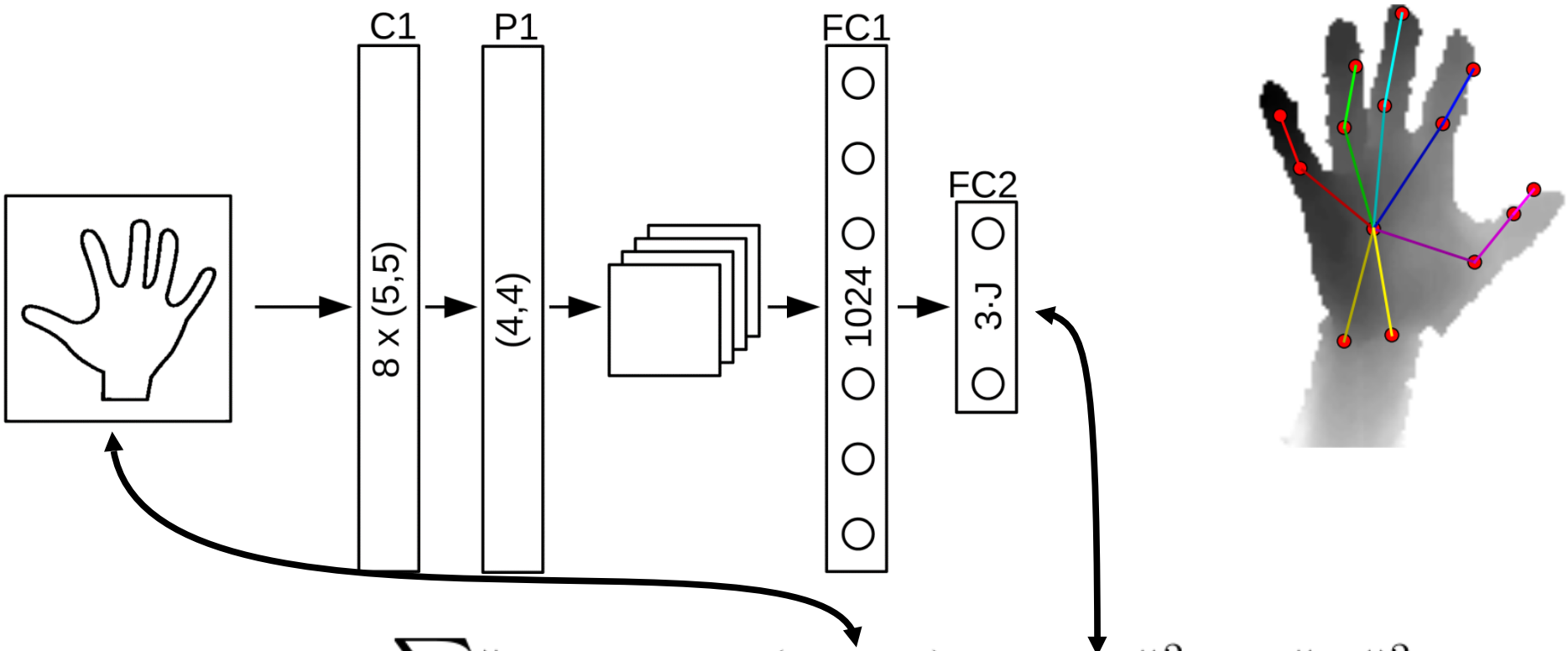
Iterating the Feedback Loop



Training a CNN to predict updates for the pose

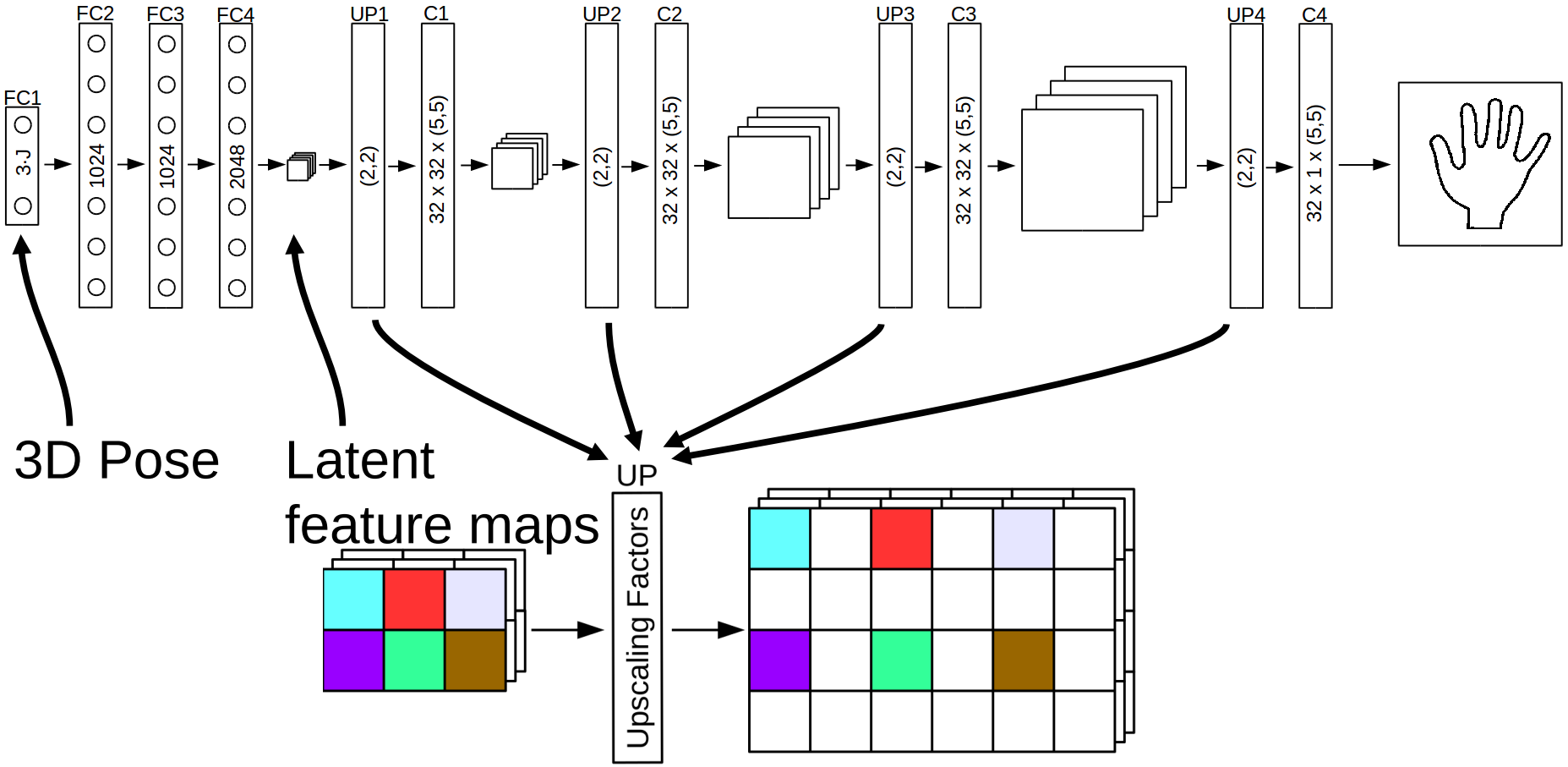
The Predictor

- Simple and very fast CNN for pose initialization



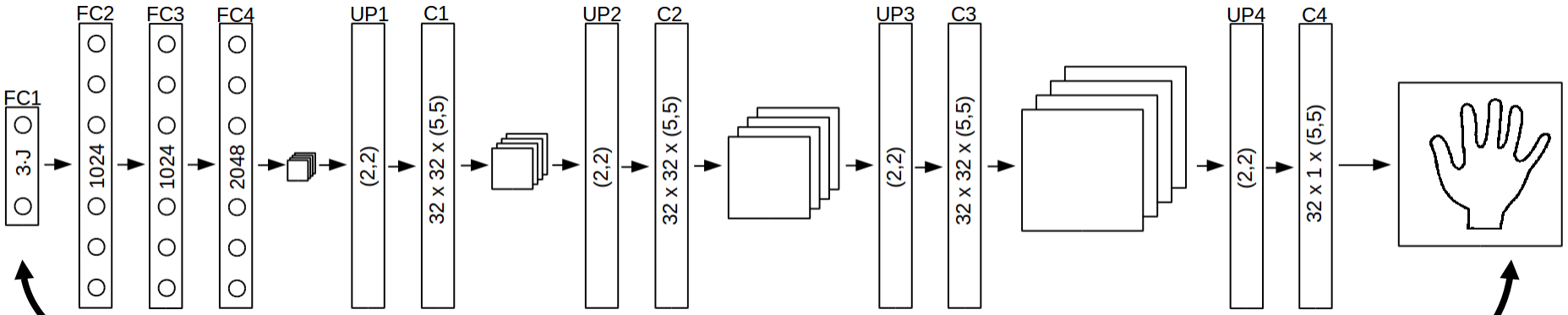
$$\arg\min_{\Phi} \sum_{\tau} \|\text{predictor}_{\Phi}(\text{depth}) - \text{pose}\|_2^2 + \gamma\|\Phi\|_2^2$$

The Synthesizer



[Zeiler et al., ICCV'11;
Dosovitskiy et al., CVPR'15]

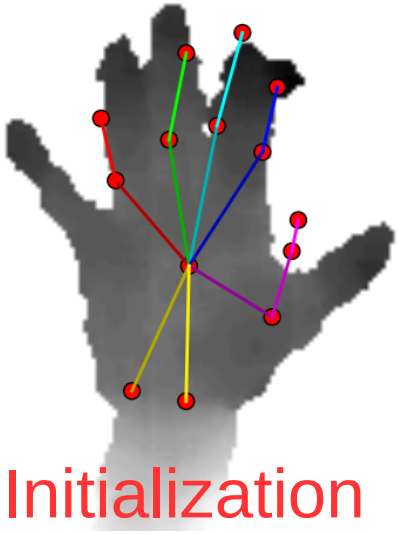
The Synthesizer



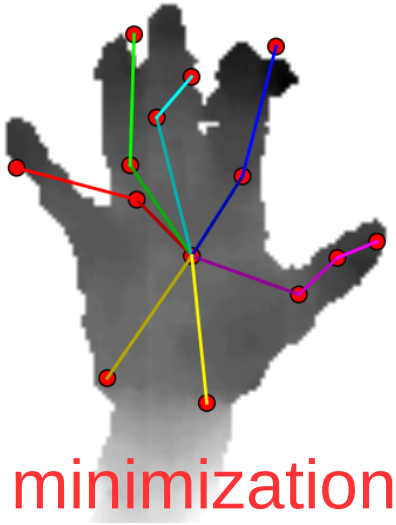
$$\arg\min_{\Theta} \sum_{\mathcal{T}} \|\text{synthesizer}_{\Theta}(\text{pose}) - \text{depth}\|_2^2$$

Directly using the Synthesizer for Optimization?

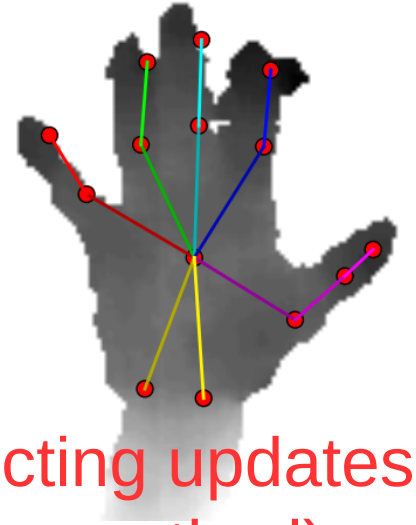
$$\widehat{\text{pose}} = \arg\min_{\text{pose}} \|\text{depth} - \text{synthesizer}(\text{pose})\|_2^2$$



Initialization

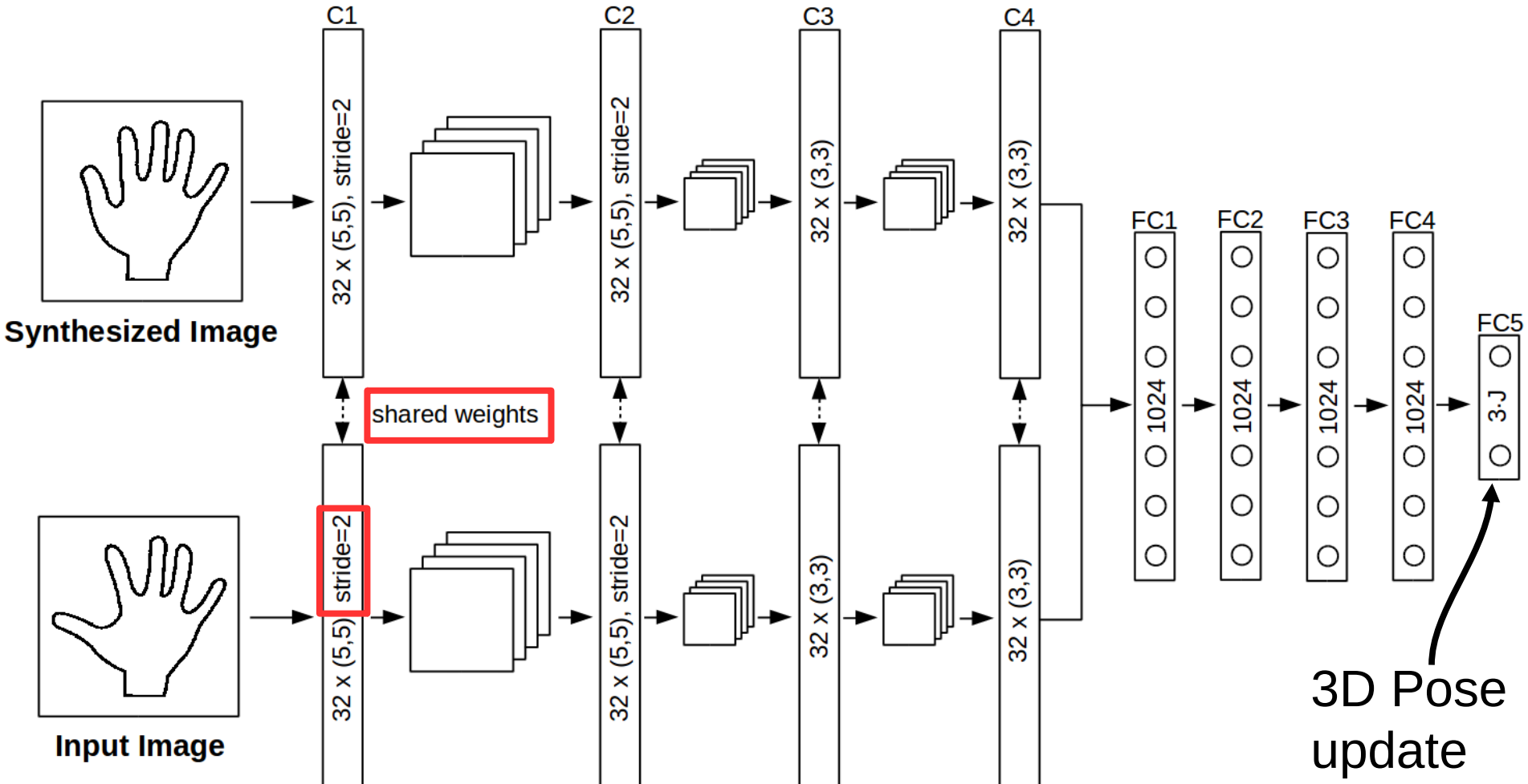


Direct minimization



Predicting updates (our method)

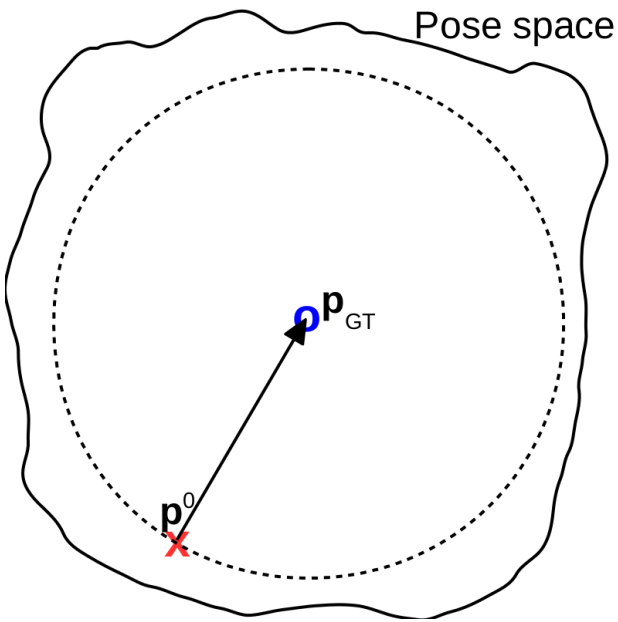
The Updater


C1
C2
C3
C4
32 x (5,5), stride=2
32 x (5,5), stride=2
32 x (3,3)
32 x (3,3)
Synthesized Image
shared weights
stride=2
32 x (5,5)
32 x (5,5), stride=2
32 x (3,3)
32 x (3,3)
Input Image
FC1
FC2
FC3
FC4
FC5
1024
1024
1024
1024
3·J
3D Pose update

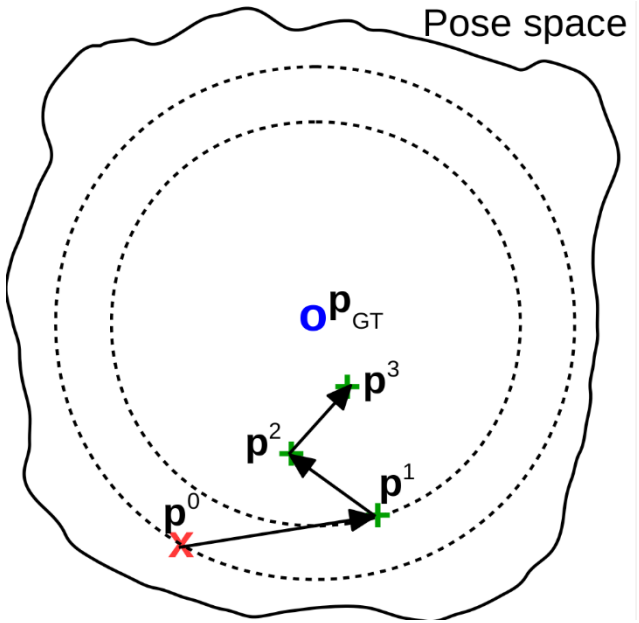
The Updater

- Predict updates to increase accuracy of pose

Difficult!



Feasible!



- Get closer

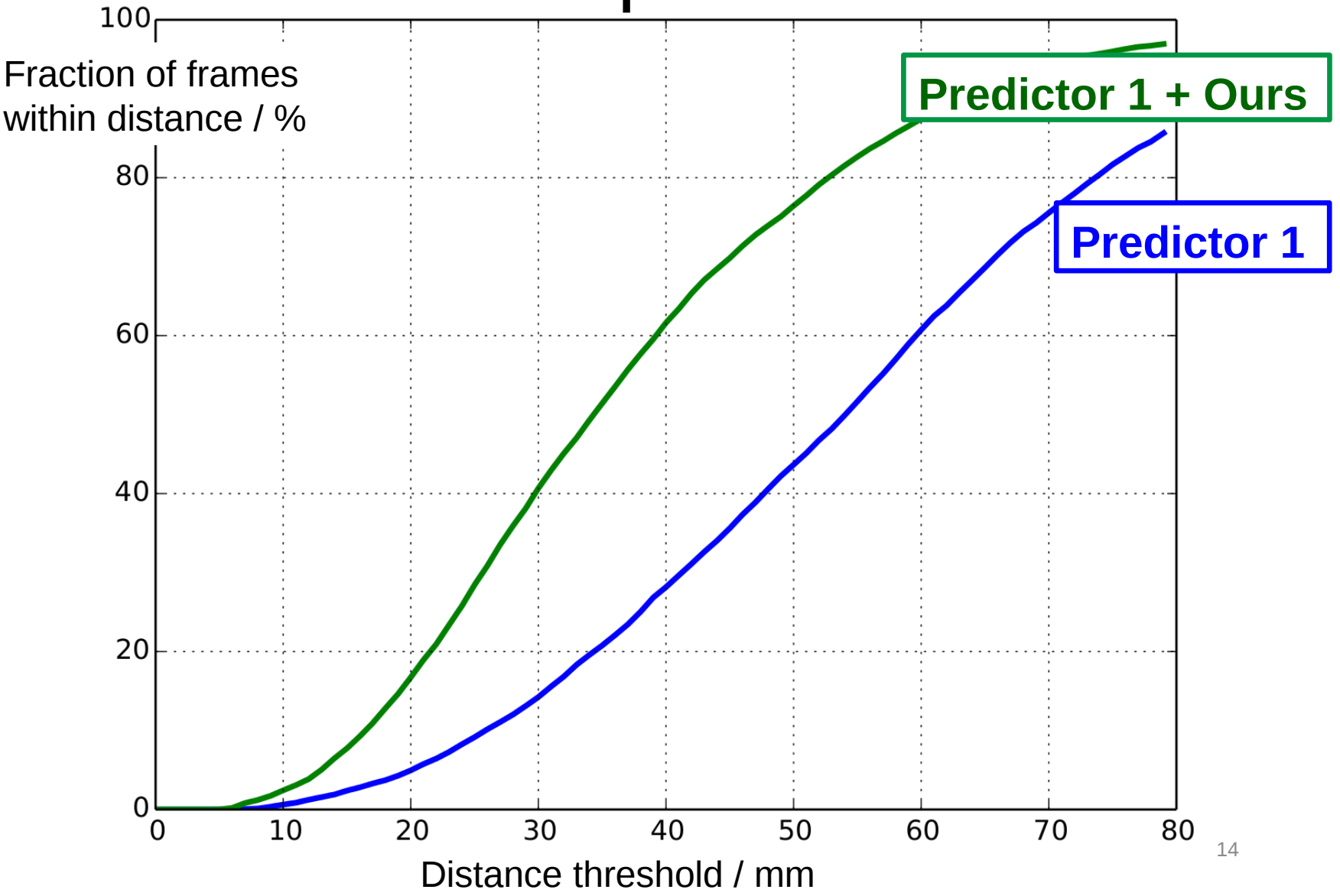
$$\|\mathbf{p} + \text{updater}(\text{depth}, \text{synthesizer}(\mathbf{p})) - \mathbf{p}_{\text{GT}}\|_2 < \lambda \|\mathbf{p} - \mathbf{p}_{\text{GT}}\|_2$$

The Updater

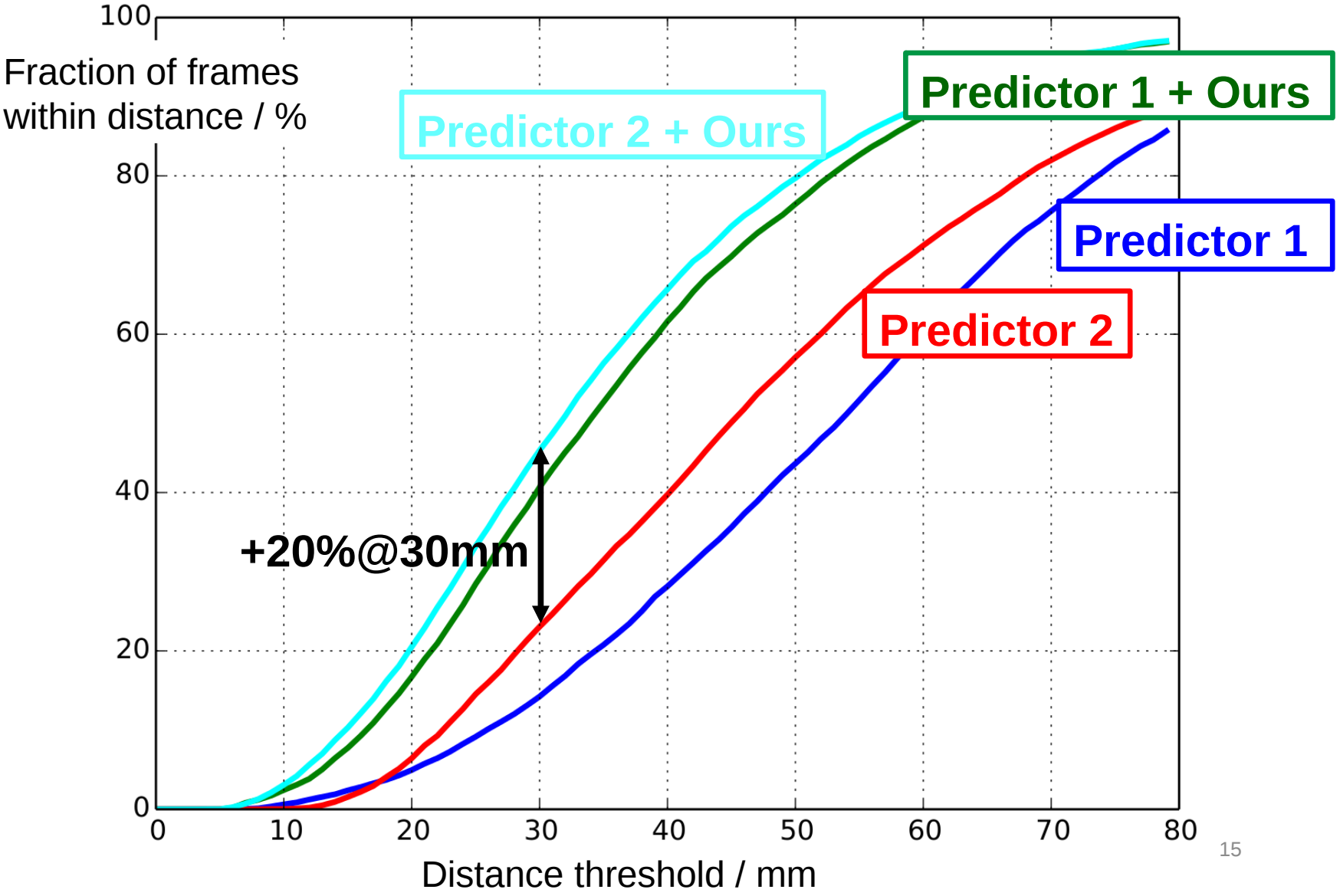
- Training is actually more complex...

$$\arg\min_{\Omega} \sum_{(\mathcal{D},\mathbf{p})\in\mathcal{T}} \sum_{\mathbf{p}'\in\mathcal{T}_{\mathcal{D}}} \max(0, \|\mathbf{p}' + \text{updater}_{\Omega}(\mathcal{D}, \text{synth}(\mathbf{p}')) - \mathbf{p}\|_2 - \lambda\|\mathbf{p}' - \mathbf{p}\|_2)$$

Results for Simple Initialization

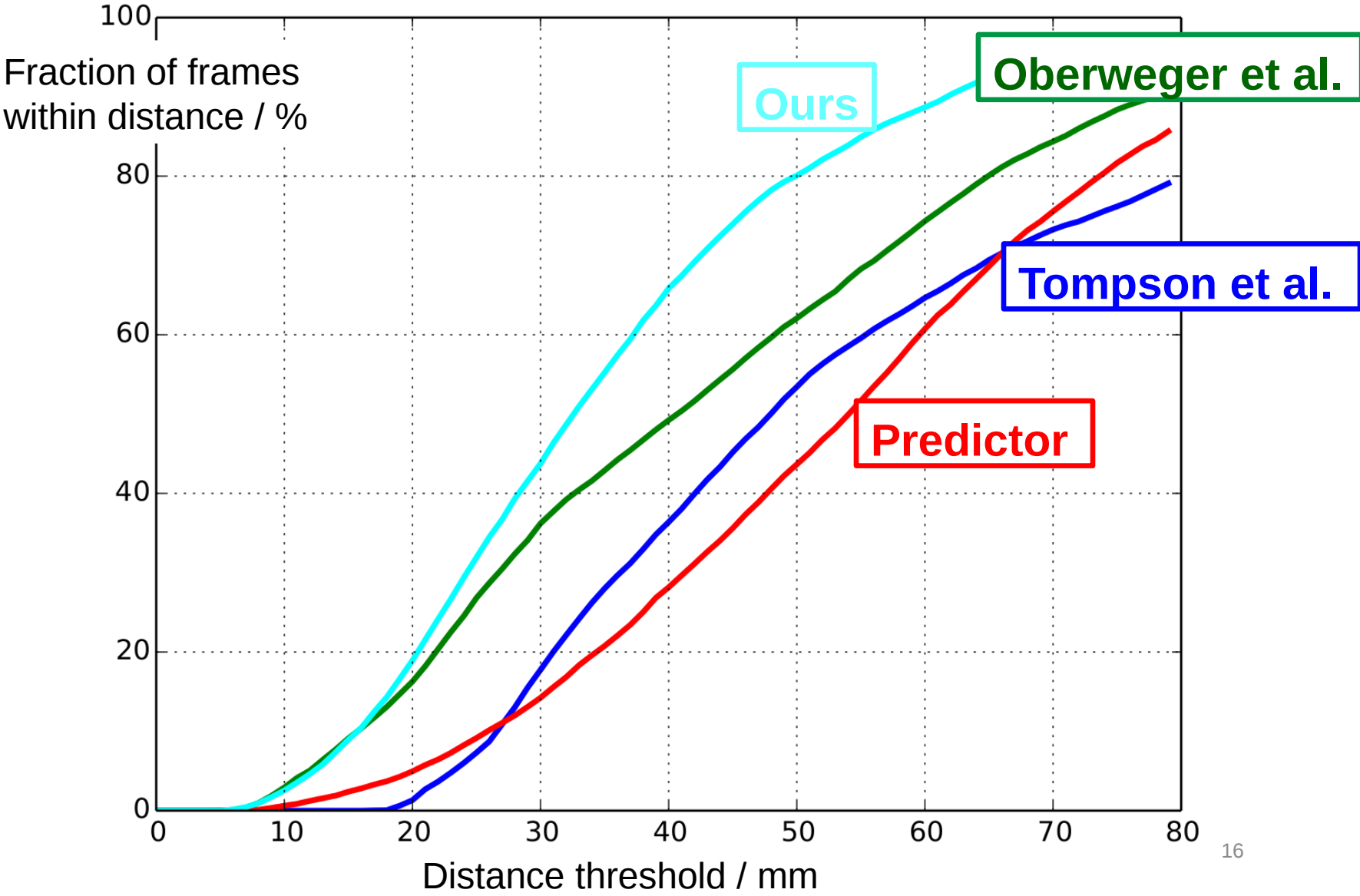


Better Initialization: Better Results

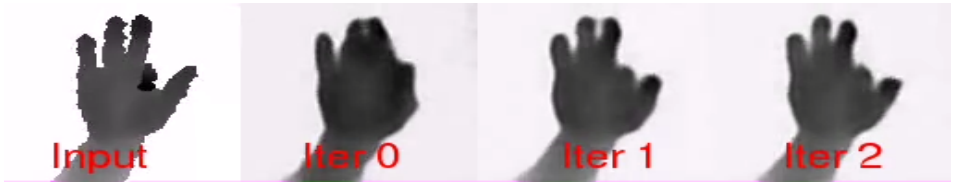


New SOTA on NYU Dataset

[Tompson et al., ToG'1



Our Results on NYU Dataset



Conclusion

- Everything is learned: No hand-crafted 3D model and similarity measure needed!
- Trained feedback loop provides updates for pose
- Predicting updates much more robust than minimizing image difference
- Efficient implementation possible (~400fps on GPU)
- General approach, applicable to other pose estimation problems

Thanks! Questions?

{oberweger,wohlhart,lepetit}@icg.tugraz.at