Automatic 3D Annotations applied to 3D Hand+Object Pose Estimation

Vincent Lepetit
ENPC ParisTech

Shreyas Hampali

HOnnotate: A Method for 3D Annotation of Hand and Object Poses. Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

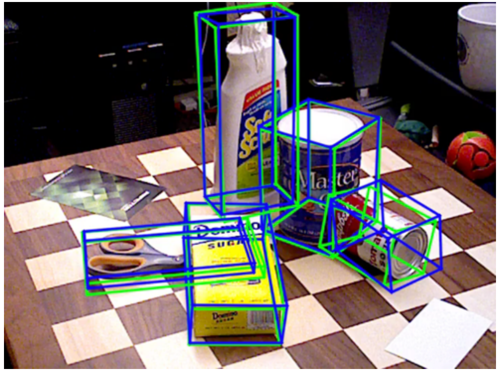
École des Ponts ParisTech

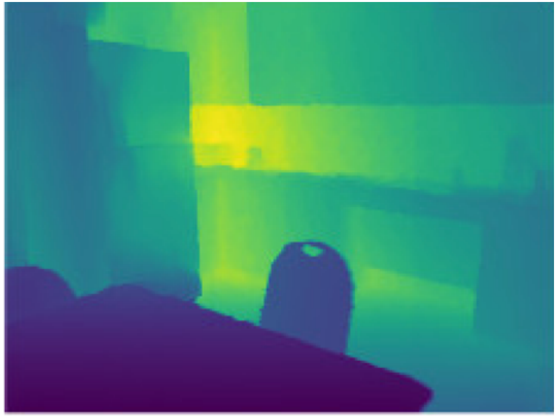
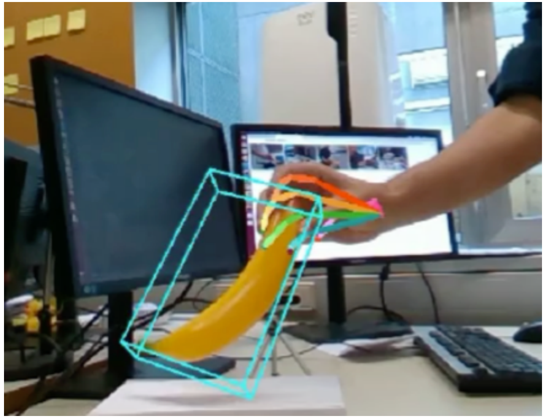
TU Graz

Qualcomm

Christian Doppler Forschungsgesellschaft

Deep Learning is great for 3D Computer Vision



How Can We Train a Deep Network for 3D Computer Vision?

1. On annotated real images;

annotation is difficult, time consuming, ..

Pix3D dataset

2. On synthetic images;

domain gap, content creation, ..

Structured3D dataset

3. Using self-learning;

cool

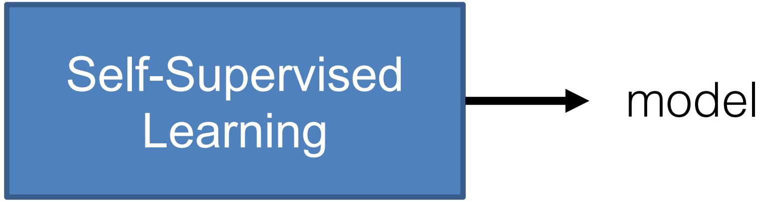
How Can We *Evaluate* a Deep Network for 3D Computer Vision?

1. On *accurately* annotated real images;

Proposed Approach

A method for automatically creating a dataset of 3D annotations of real images that we can evaluate;

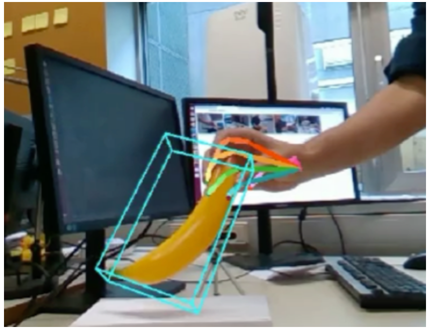
Different from self-learning: We can validate the created training set; We can use the dataset to validate a method.



Creating a Dataset for 3D Hand[+Object] Pose Estimation

NYU hand dataset

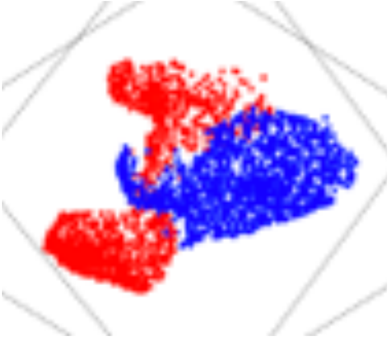
FREIhand dataset



GANerated hand dataset

First-Person Hand Action dataset

Automated Annotations



- 1 or more RGB-D cameras;
- temporal constraints.

Object 3D model from YCB [Xiang et al, 2018]

MANO model [Romero et al, 2017]

Bayesian Formulation

$$\max_{\{(\mathbf{p}_t^H, \mathbf{p}_t^O)\}_t} \prod_t \prod_c p((I_t^c, D_t^c) \mid \mathbf{p}_t^H, \mathbf{p}_t^O)\; p(\mathbf{p}_{t+1}^H, \mathbf{p}_{t+1}^O \mid \mathbf{p}_t^H, \mathbf{p}_t^O)\; p(\mathbf{p}_t^H, \mathbf{p}_t^O)$$

Bayesian Formulation

$$\max_{\{(\mathbf{p}_t^H,\mathbf{p}_t^O)\}_t} \prod_t \prod_c \overbrace{p((I_t^c, D_t^c) \mid \mathbf{p}_t^H, \mathbf{p}_t^O)}^{\text{RGB-D likelihoods}} \underbrace{p(\mathbf{p}_{t+1}^H, \mathbf{p}_{t+1}^O \mid \mathbf{p}_t^H, \mathbf{p}_t^O)}_{\text{temporal constraints}} \overbrace{p(\mathbf{p}_t^H, \mathbf{p}_t^O)}^{\text{physical constraints}}$$

hand pose at time t

object pose at time t

depth map from camera $\mathbf{c}$ at time t

color image from camera c at time t

RGBD Likelihood

$$\max_{\{(\mathbf{p}_t^H,\mathbf{p}_t^O)\}_t} \prod_t \prod_c \underbrace{p((I_t^c, D_t^c) \mid \mathbf{p}_t^H, \mathbf{p}_t^O)}_{\text{RGBD likelihoods}}\; p(\mathbf{p}_{t+1}^H, \mathbf{p}_{t+1}^O \mid \mathbf{p}_t^H, \mathbf{p}_t^O)\; p(\mathbf{p}_t^H, \mathbf{p}_t^O)$$

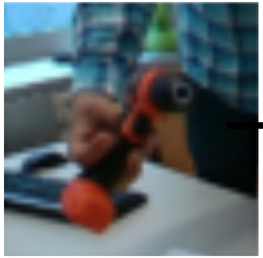
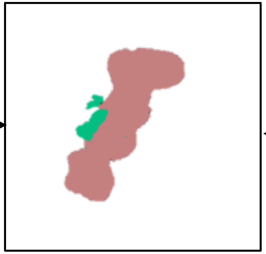
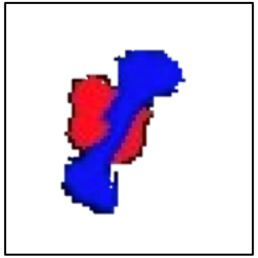


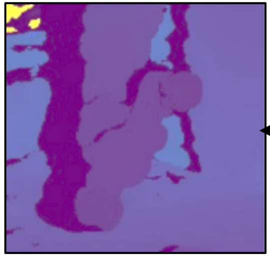
image I_t^c



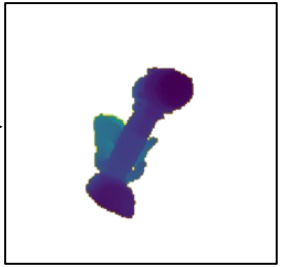
observed masks from I_t^c



predicted masks for $\mathbf{p}_t^H, \mathbf{p}_t^O$



observed depth D_t^c



predicted depth for $\mathbf{p}_t^H, \mathbf{p}_t^O$

- Efficient way to deal with occlusions hand/object;
- Gradient computed with differential renderer.

Physical Constraints

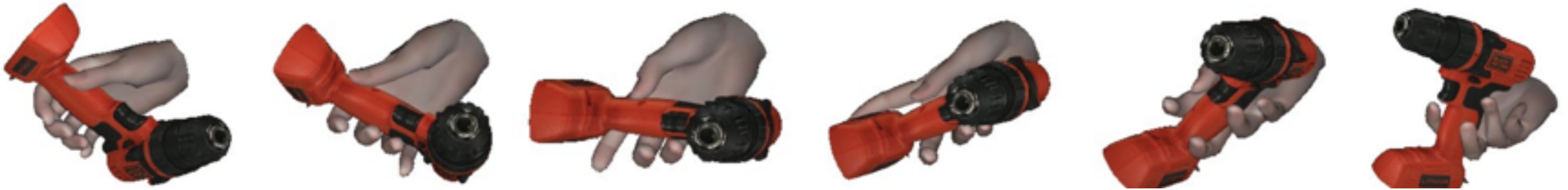
$$\max_{\{(\mathbf{p}_t^H,\mathbf{p}_t^O)\}_t} \prod_t \prod_c p((I_t^c, D_t^c) \mid \mathbf{p}_t^H, \mathbf{p}_t^O)\, p(\mathbf{p}_{t+1}^H, \mathbf{p}_{t+1}^O \mid \mathbf{p}_t^H, \mathbf{p}_t^O)\, \overbrace{p(\mathbf{p}_t^H, \mathbf{p}_t^O)}^{\text{physical constraints}}$$

joint angle constraints

no intersection constraint

Temporal Constraints

$$\max_{\{(\mathbf{p}_t^H,\mathbf{p}_t^O)\}_t} \prod_t \prod_c p((I_t^c, D_t^c) \mid \mathbf{p}_t^H, \mathbf{p}_t^O) \underbrace{p(\mathbf{p}_{t+1}^H, \mathbf{p}_{t+1}^O \mid \mathbf{p}_t^H, \mathbf{p}_t^O)}_{\text{temporal constraints}} p(\mathbf{p}_t^H, \mathbf{p}_t^O)$$



simple 0-order motion model

Optimization

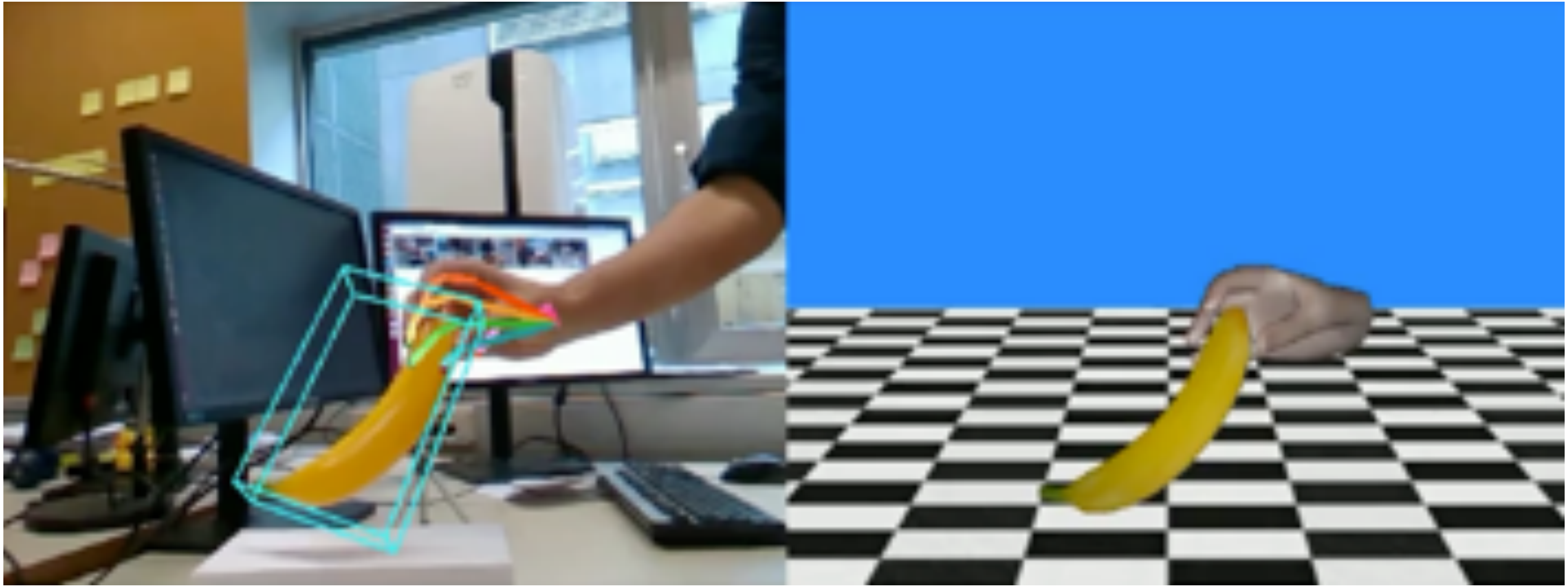
$$\max_{\{(\mathbf{p}_t^H, \mathbf{p}_t^O)\}_t} \prod_t \prod_c p((I_t^c, D_t^c) \mid \mathbf{p}_t^H, \mathbf{p}_t^O)\; p(\mathbf{p}_{t+1}^H, \mathbf{p}_{t+1}^O \mid \mathbf{p}_t^H, \mathbf{p}_t^O)\; p(\mathbf{p}_t^H, \mathbf{p}_t^O)$$

Negative log:

$$\begin{aligned}\min_{\{(\mathbf{p}_t^H, \mathbf{p}_t^O)\}_t} \sum_t \quad \sum_c \; & \alpha \| S_t^c - S(\mathbf{p}_t^H, \mathbf{p}_t^O) \|^2 + \beta \| D_t^c - D(\mathbf{p}_t^H, \mathbf{p}_t^O) \|^2 + \\ & \gamma E_{\text{joints}}(\mathbf{p}_t^H) + \delta E_{\text{inters}}(\mathbf{p}_t^H, \mathbf{p}_t^O) + \\ & \epsilon E_{\text{temp}}(\mathbf{p}_t^H, \mathbf{p}_t^O, \mathbf{p}_{t-1}^H, \mathbf{p}_{t-1}^O, \mathbf{p}_{t+1}^H, \mathbf{p}_{t+1}^O) + \\ & \eta E_{3D}(\{D_t^c\}_c, \mathbf{p}_t^H, \mathbf{p}_t^O)\end{aligned}$$

Optimized using Adam.

Automated 3D Annotations



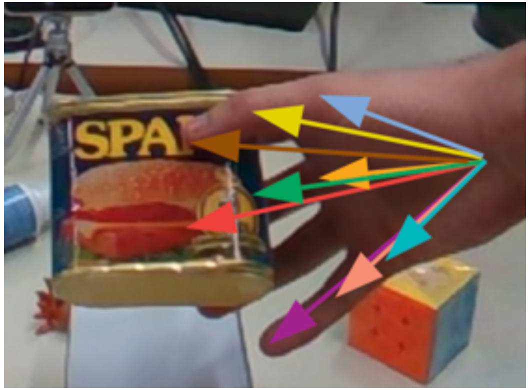
Validating our Annotations

Manual annotations of 3D joints on 100 randomly selected time steps;

Done directly on the point cloud created from 4 cameras;

→ Mean error is 8mm with 4 cameras;
→ Mean error is 10mm with 1 camera.

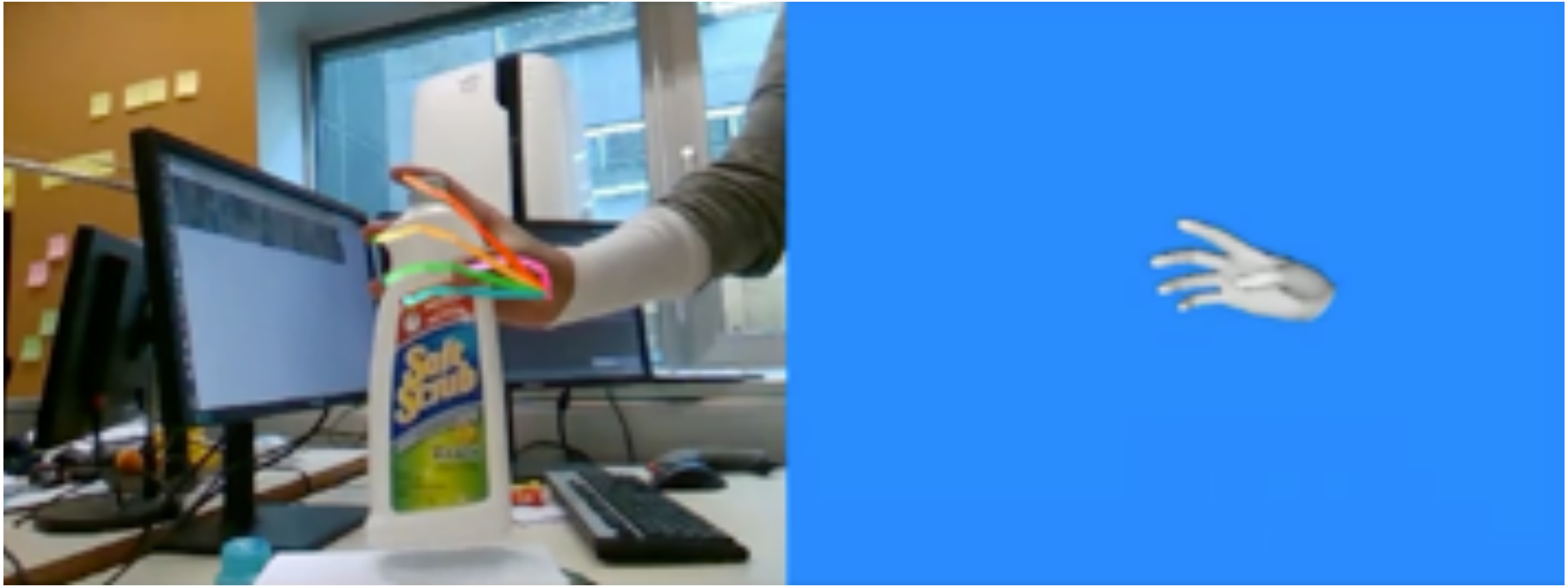
Using our 3D Annotations for Single RGB Frame Prediction



We train a network to predict:
– 2D keypoint locations;
– Root relative joint directions.

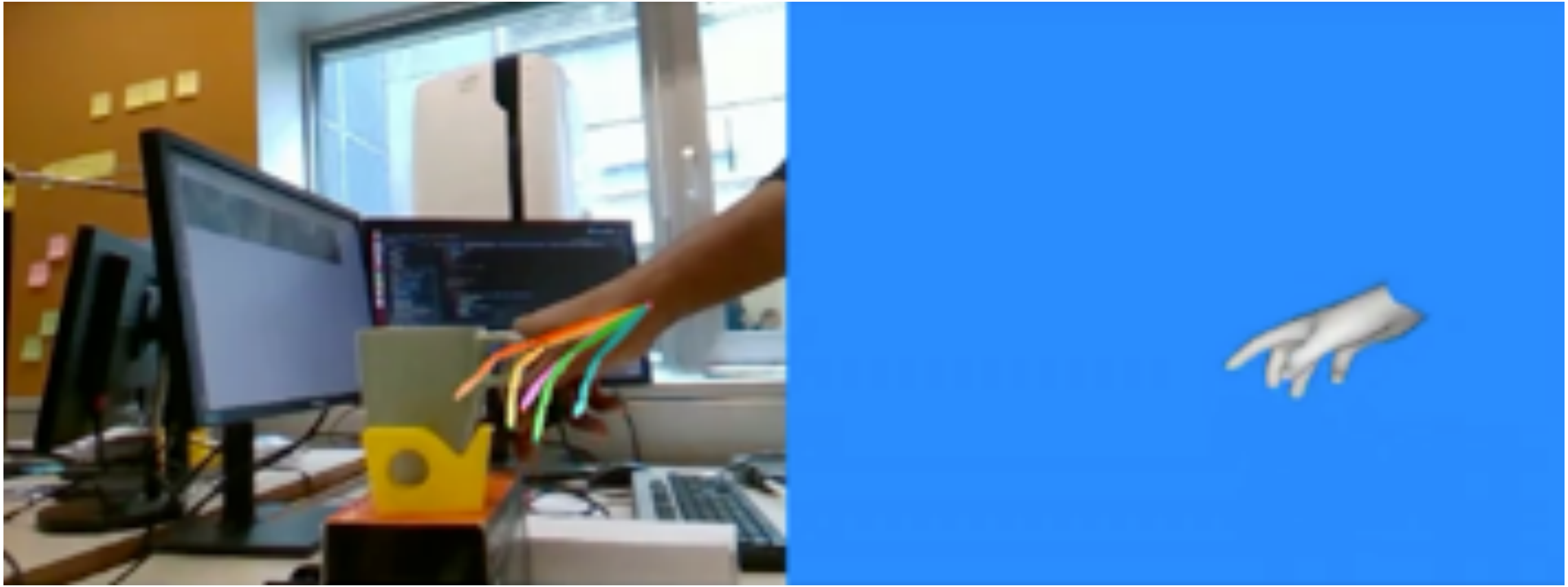
+ MANO model fitted to these predictions

Using the Annotations for Single RGB Frame Prediction



(objects are unknown)

Using the Annotations for Single RGB Frame Prediction



(objects are unknown)

Thanks for listening!

Questions?

Shreyas Hampali

HOnnotate: A Method for 3D Annotation of Hand and Object Poses. Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

Dataset and code: https://www.tugraz.at/index.php?id=40231

École des Ponts ParisTech

TU Graz

Qualcomm

Christian Doppler Forschungsgesellschaft