Motivation

- **Goal**: Accurate 3D hand pose estimation from single depth image
- **Straightforward approach**: Predictor for 3D pose
  - Bottom-up / Feed-forward network
  - Many neurons in the visual cortex provide feedback
- **Adding top-down information with a synthesizer**
  - Synthesizes depth image from 3D pose
  - Difficult to estimate pose by minimizing difference between synthesized and input image!
- **Our approach**: Train an updater to predict updates for the initial 3D pose of a predictor using a learned model of the hand
- **All components are implemented as Convolutional Neural Networks and trained from data.**

Method

- **Training a predictor CNN** to predict initial 3D pose $p$ from depth image $D$
  \[
  \arg \min_p \sum_{(p',\mathcal{D})} \|\text{pred}_4(\mathcal{D}) - p\|^2 + \gamma \|\Phi\|^2
  \]
- **Training a synthesizer CNN** to generate depth image $\text{synth}(p)$ from 3D pose $p$
  \[
  \arg \min_p \sum_{(p',D)} \frac{1}{D} \|\text{synth}_a(p) - D\|^2
  \]
- **Training an updater CNN** to predict 3D pose update from two depth images
  \[
  \arg \min_{\mathcal{D}} \sum_{(p',\mathcal{D})} \max(0, \|p' + \text{updater}_g(D, \text{synth}(p')) - p\|^2 - \lambda \|p' - p\|^2)
  \]
- **Effectively augment the training data $\mathcal{T}_D$**
  - Add Gaussian noise to 3D poses
  - Predictions on training data
  - Samples from error distribution of predicted updates
  - More details about the training can be found in the paper.

Results

- **Synthesizing accurate depth images from 3D pose**
  - Observed
  - Synthesized
  - Difference
- **Predicted pose updates**
  - 1st iteration
  - 2nd iteration
  - 3rd iteration
- **Update mechanism is robust to image noise**
  - Minimize image difference:
    \[
    \tilde{p} = \arg \min_p \|D - \text{synth}(p)\|^2
    \]
  - Our approach:
  - **Quantitative evaluation of predicted 3D pose**
    - Comparison with SOTA on NYU dataset
    - Different initialization