Can we automatically refine a generic ConvNet for a specific task?

- Multi-person pose estimation
- 3D human pose estimation
- 2D human pose estimation

Limitation 1: 3D Pictorial Structures
- Human3.6M & HumanEva
- MPII & FLIC

Limitation 2: Humans cannot annotate metric 3D information.
- MoCap systems for capturing ground truth

Some tasks live in the small-data regime.
- "Personalizing" 2D human pose
- 2D pose detectors are still not perfect out-of-the-box.

Motivation
- ConvNets have achieved impressive results on large scale human pose estimation benchmarks.

But, ground truth data is not always readily available!
- Some tasks live in the small-data regime.

Human3.6M & HumanEva

Multi-person pose estimation
- MPII Multi-Person & CoC Keypoints
- Humans3.6M & HumanEva

Multi-view Setup
- The input is a set of videos from a calibrated multi-view setup.

Generic 2D pose
- ConvNet

Heatmaps
- The input is a set of videos from a calibrated multi-view setup.

3D Pictorial Structures
- Generic 2D pose estimates in the form of heatmaps for each view.

3D Annotations
- A 3D Pictorial Structures model estimates the globally optimal 3D pose.

Annotations Leveraging
- The harvested 3D pose estimates can be used as high quality annotations for human pose estimation tasks.

We train a ConvNet that takes a single color image as input, and predicts the 3D pose.

Results
- Multi-view pose estimation using only generic 2D pose detector - no retraining.
- Single-view 3D human pose estimation
- On par with the state-of-the-art without using 3D ground truth for training.

We refine a generic ConvNet for 2D pose by using the automatic 3D annotations projected to the 2D image.

Testing Code
- tinyurl.com/PoseHarvesting

Automatic refinement of generic 2D pose detector
- Consistent benefit over all body parts