# Yang, Ouyang, Wang, Ren, Li, Wang

3D Human Pose Estimation in the Wild by Adversarial Learning

## 1 Summary

The authors introduce a GAN framework for 3D human pose estimation in the wild. The generator is the 3D human pose estimator, and produces a single 3D pose from a monocular RGB image that the discriminator would hopefully think is legitimate. The discriminator tries to determine whether the (*input image*, *output pose*, *geometric descriptor of output pose*) tuple could have come from the real (annotated) distribution of such tuples, creating a loss that pushes the space of generated 3D poses toward the space of ground-truth 3D poses. Notably, the discriminator is multi-source, meaning it takes not only standard information relevant to image-pose correspondence, but also a novel *geometric descriptor* which contains information about human body articulation.

The point: there are no in-the-wild datasets with 3D annotations. There are only constrained-environment datasets with 3D annotations and in-the-wild datasets with 2D annotations. Here, we leverage both of these datasets as a replacement for that which we don't have, distinguishing the 3D pose output for both datasets' images from the ground-truth 3D poses that we *do* have.

The result: domain shift. We are able to shift the domain of *predicted, in-the-wild* 3D poses to the domain of *ground-truth* 3D poses (which is supposed to represent the domain of *all* valid 3D poses).

## 2 One-Sentence Summary

A pose estimator tries to generate 3D poses from all kinds of images so as to fool a discriminator, which is itself trained to disambiguate generated poses from available ground-truth poses using multiple information sources (image, 2D/3D predictions, *geometric descriptor* encoding relative distances between body parts).

## 3 Generator: 3D Pose Estimator

- The generator is two-stage (predicts 2D, then predicts depth).
- The 2D module uses the state-of-the-art stacked hourglass architecture.
  - input: $256 \times 256$ image
  - output: one $64 \times 64$ heatmap for each joint
- The depth regression module consists of several residual modules.
  - input: one $64 \times 64$ heatmap for each joint, intermediate image features
  - output: a vector containing the estimated depth for each joint

## 4 Discriminator

- The discriminator says "real" or "fake" (i.e. "ground-truth" or "not ground-truth").
- All predicted poses (for both in-lab and in-the-wild images) are used as "fake" examples.
- Ultimately, the generator learns to produce 3D poses that are indistinguishable from the ground-truth.

- Architecture:
  - input:
    - original image
    - *geometric descriptor*: pairwise relative locations and distances between joints
    - heatmaps and depth maps for joints
  - each input passes through its own CNN
  - the outputs of the three CNNs are concatenated and passed to fully connected layers
  - output: classification as "real" or "fake"

# 5 Training Process

- Pre-train pose estimator $G$ on 3D human pose dataset.
- Alternately optimize the generator $G$ and the discriminator $D$.

# 6 Testing Process

- Run the pose estimator $G$ on a new image.

# References

[1] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, Xiaogang Wang. 3D Human Pose Estimation in the Wild by Adversarial Learning. *arXiv preprint arXiv:1803.09722 (2018).*