Multi-Instance Pose Networks: Rethinking Top-Down Pose Estimation

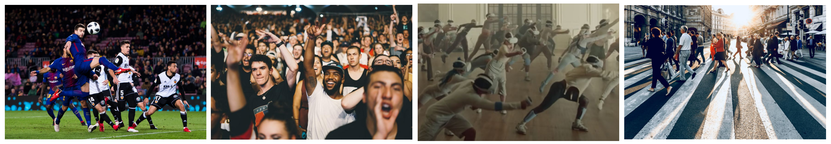
Rawal Khirodkar[1], Visesh Chari[2], Amit Agrawal[3], Ambrish Tyagi[4]
[1]Carnegie Mellon University, [2]Waymo, [3]Amazon, [4]Cruise

https://rawalkhirodkar.github.io/mipnet/

2021 ICCV OCTOBER 11-17 VIRTUAL

1. Introduction

Goal: Multi-Person 2D Pose Estimation under severe Occlusion



Sports | Social Events | Virtual Reality | Self-Driving

Contributions:

- We address the limiting *single person assumption* in top-down methods.
- A novel approach to modify any backbone to predict multiple pose instances using parameter efficient feature modulation.
- MIPNet is state-of-the-art on occlusion datasets, OCHuman & CrowdPose.

2. Key Challenges

Top-Down Methods:

✓ Accurate due to scale invariance.

✗ Fails under occlusion and crowding.

Sun et. al. **Deep High-Resolution Representation Learning for Human Pose Estimation**, CVPR 2019.
Xiao et. al. **Simple Baselines for Human Pose Estimation and Tracking**, ECCV 2018.
Newell et. al. **Stacked Hourglass Networks for Human Pose Estimation**, ECCV 2016.

Bottom-Up Methods:

✓ Robust to occlusion.

✗ Lacks precise keypoint localization.

Cao et. al. **OpenPose: Realtime Multi-Person Pose Estimation using Part Affinity Fields**, TPAMI, 2019.

Occluded Pose Estimation:

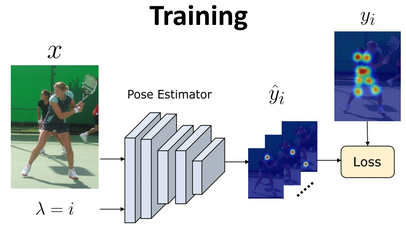
✓ Multi-instance predictions.

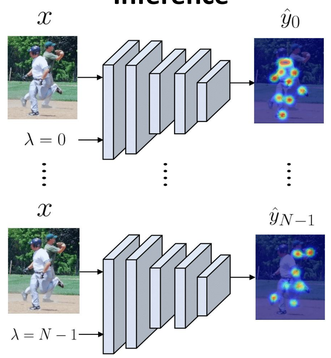
✗ Parameter inefficient/Three stages.

Li et. al. **CrowdPose Crowded Scenes Pose Estimation and a New Benchmark**, CVPR, 2019.

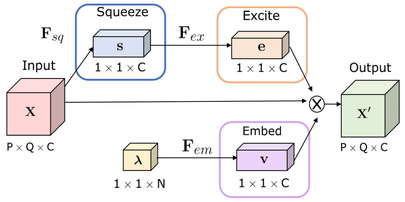
3. Multi-Instance Pose Prediction

MIPNet takes two inputs to predict pose - the rgb bounding box image, **x** and a scalar instance selector, **λ**. **λ** ∈ [0, N-1] where N is a hyper-parameter.





Multi-Instance Modulation Block



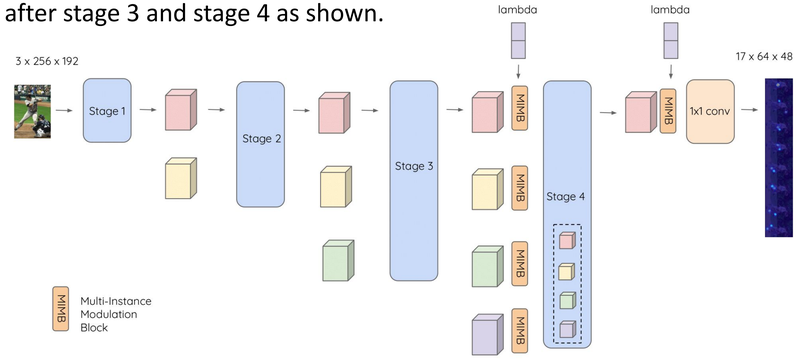
Losses

n = ground-truth poses in image and *MSE(..)* is the weighted heatmap pose loss.

$$\mathcal{L}_i = \begin{cases} \text{MSE}(y_i, P(x, \lambda = i)), i < \min(n, N) \\ \text{MSE}(y_0, P(x, \lambda = i)), \text{otherwise} \end{cases}$$

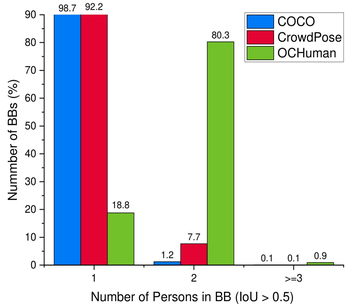
4. Architecture

Any top-down architecture can be converted to MIPNet by inserting MIMB blocks across the backbone. We insert 5 blocks in HRNet backbone after stage 3 and stage 4 as shown.



Comparison to State-of-the-Art

Method	COCO (AP)	CrowdPose (AP)	OCHuman (AP)
MaskRCNN	64.8	57.2	20.2
AlphaPose	70.1	61.0	21.1
OPEC-Net	73.9	**70.6**	29.1
SBL	73.7	60.8	24.1
HRNet	75.5	69.3	37.2
MIPNet	**75.7 (+0.2)**	70.0 (+0.7)	**42.5 (+5.3)**
HRNet (gt bb)	78.1	72.8	65.0
MIPNet (gt bb)	**78.8 (+0.7)**	**73.7 (+0.9)**	**74.1 (+9.1)**



Qualitative Results