3D Human Pose Estimation

공대현
Vision and Display System Lab.
Sogang University

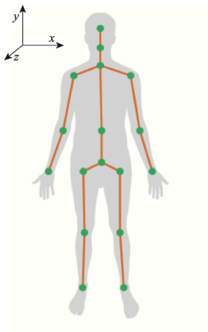
Outline

- 3D Human Pose Estimation
  - Background

- 3DMPPE - *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*
  - Abstract
  - Model (3 Phase)
    - *DetectNet : Mask R CNN (ICCV 2017)*
    - *RootNet :* ***RootNet***
    - *PoseNet : Integral Human Pose Regression (ECCV 2018)*
  - Performance

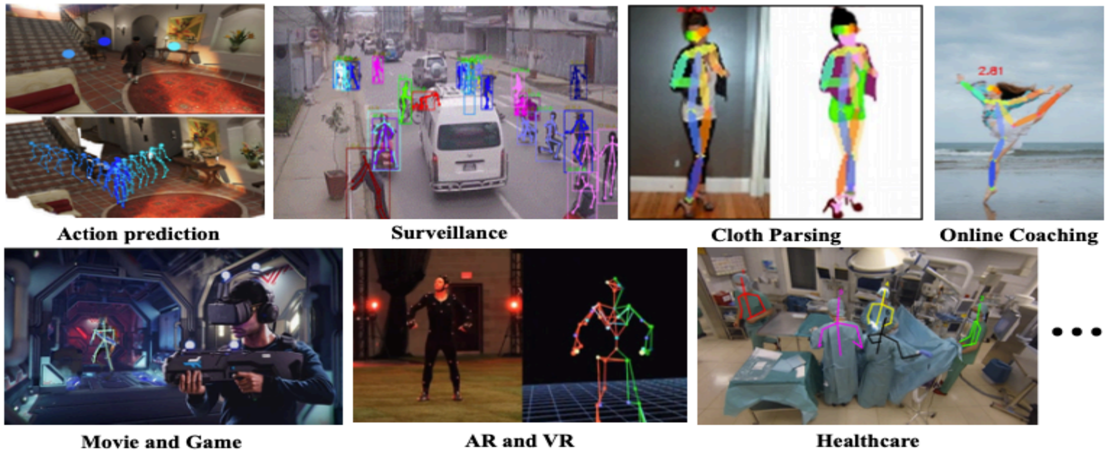
3D Human Pose Estimation

- **Background**



- Work: 15~17개의 Joints(x, y, z) 추정

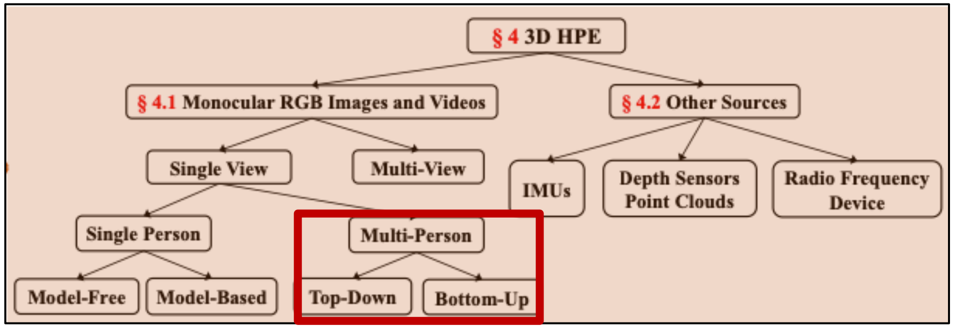
- Application:



3D Human Pose Estimation

- **Background**

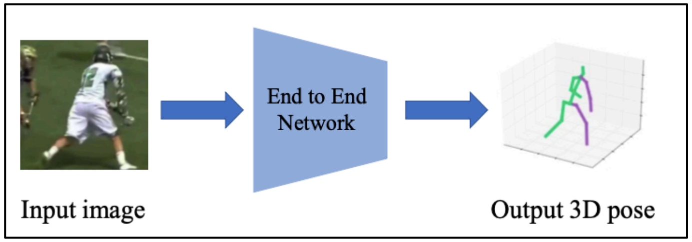
- 분류

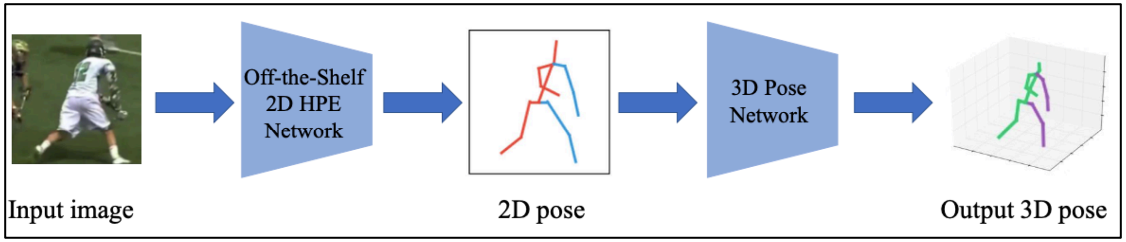


- Single Person HPE Methods

1) Direct Estimation

2) 2D to 3D Lifting



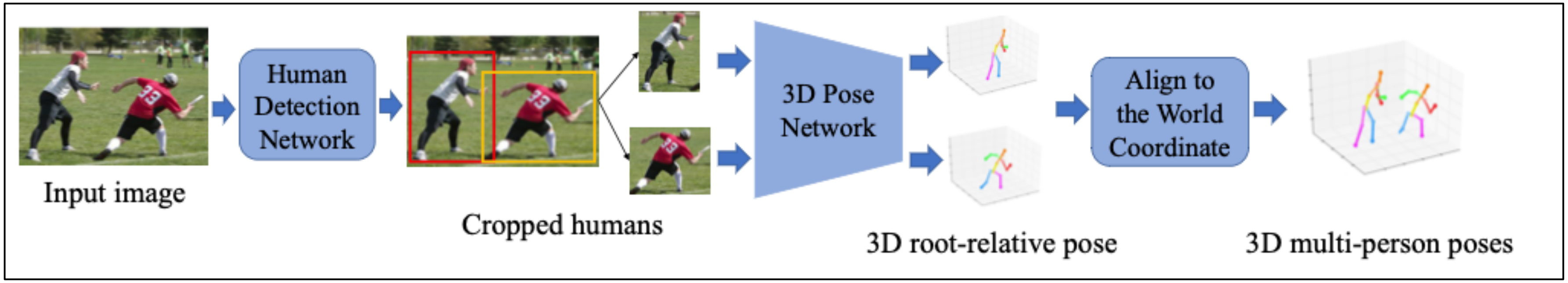


3D Human Pose Estimation

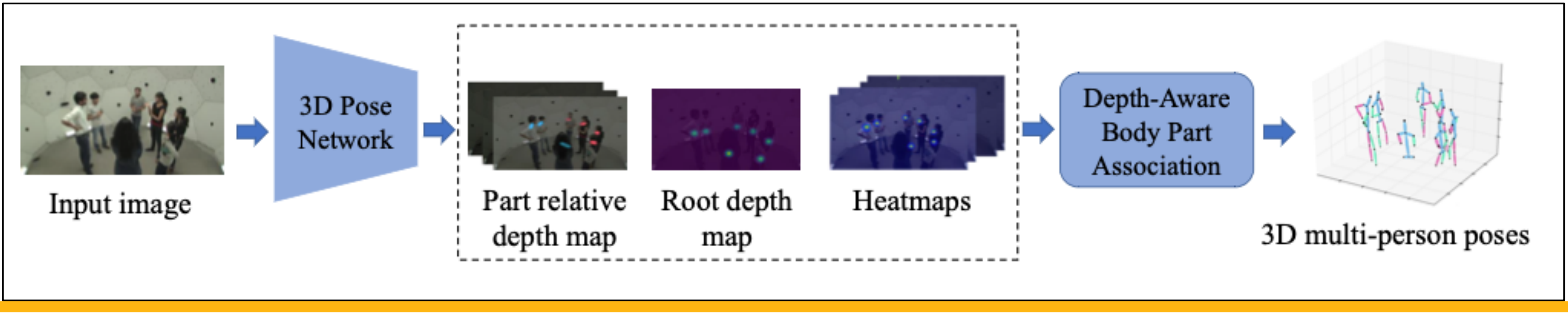
- **Background**

  - Multi Person HPE Methods

    1) Top-Down Approach (장점: Performance, 단점: Slow, Global Information 유실)



    2) Bottom-Up Approach (장점: Fast, 단점: Low Performance,)



3D Human Pose Estimation

- **Background**

  - Evaluation Metrics

  1) MPJPE (Mean Per Joint Position Error)

$$\boxed{MPJPE = \frac{1}{N}\sum_{i=1}^{N} \|J_i - J_i^*\|_2}$$

  2) 3DPCK (3D Percentage of Correct Keypoints)

     ➢ Threshold: 150mm, 그 이하면 Correct

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Abstract**

- Monocular View, Single Camera , Multi Person , Top-Down

- Contribution : Pinhole Camera 원리를 이용한 방법으로 Depth Estimation 성능을 높임

+ $3DPCK_{abs}$ First Report

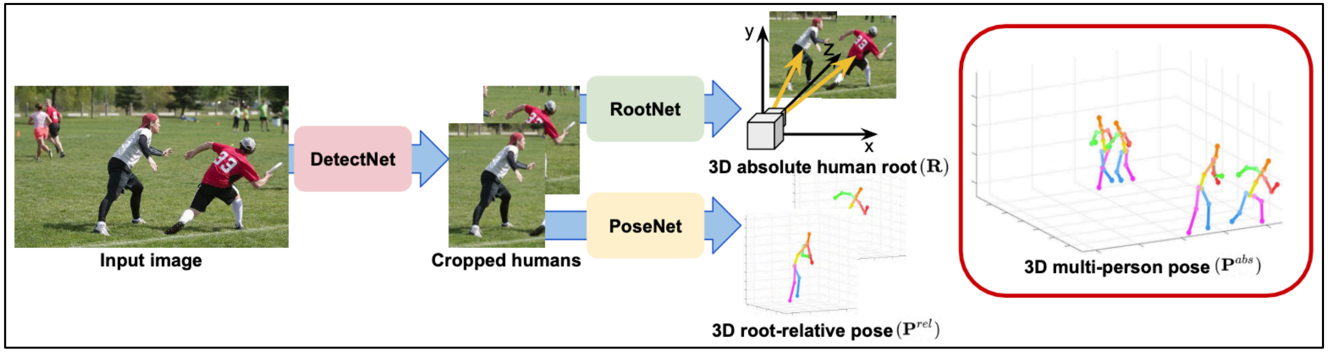
- Overview of Model

DetectNet, RootNet, PoseNet 3단계의 Phase로 구성

1) DetectNet : Human Bounding Box(Object Detection)
2) RootNet : Absolute Depth
3) PoseNet : 3D Relative KeyPoints

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

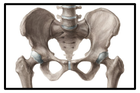
- **Model**
  - Pipeline



  - DetectNet
    - Human Bounding Box를 Detect하는 Phase
    - 이 논문에선 Mask R-CNN 을 사용
    - Mask R-CNN (DetectNet Framework)
      - High Object Detection Performance
      - 여기선 Human or not Human만 판단하여 Bounding Box Estimation

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

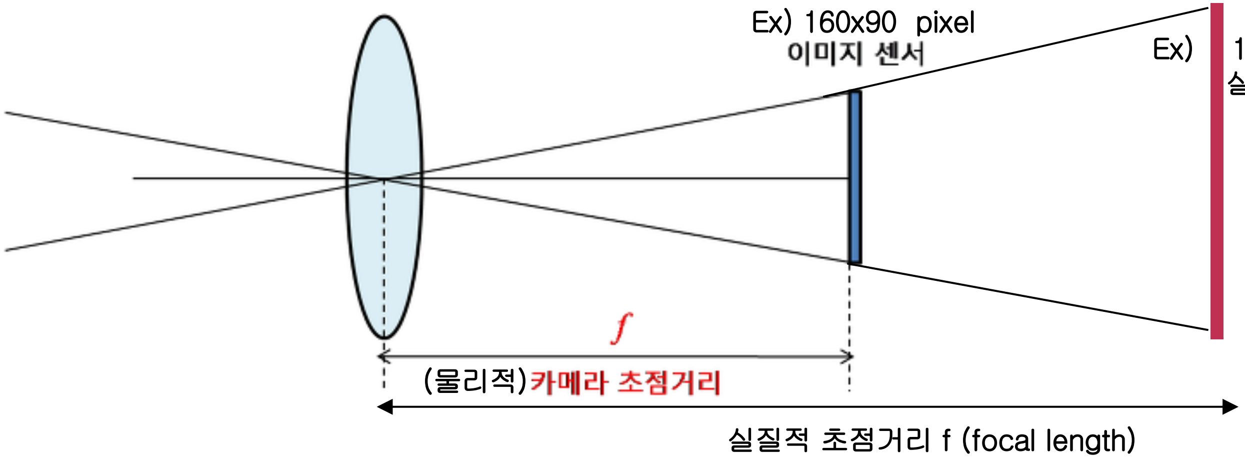
- **Model**
  - RootNet
    - Estimate Camera-Centered coordinates of the **human Root R** = (x_R, y_R, Z_R)
    - Input: Human Bounding Box, Image
    - Output: Root of Each human
    - Depth 추정에 PinHole Camera Projection 원리 이용



Pelvis Bone

(x_R, y_R, Z_R)

- PinHole Camera Model에서 Focal Length를 사용해서 **R**을 추정

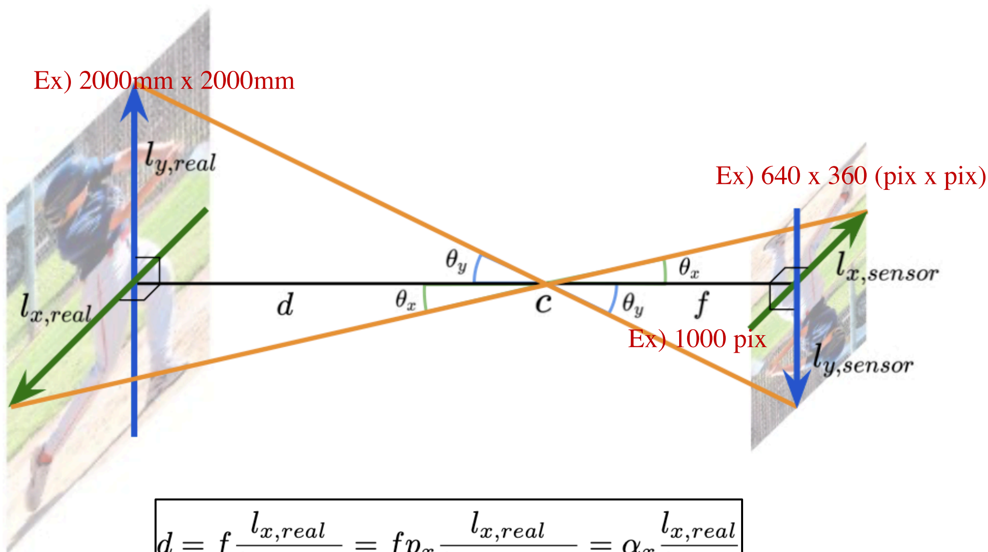


3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Model**

  - RootNet

    - Pinhole Camera Model

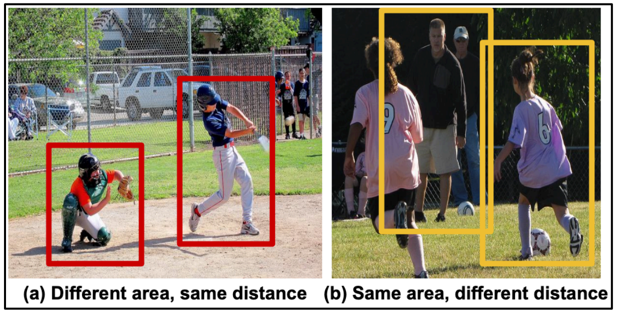


$$\boxed{d = f\frac{l_{x,real}}{l_{x,sensor}} = fp_x\frac{l_{x,real}}{l_{x,sensor}p_x} = \alpha_x\frac{l_{x,real}}{l_{x,img}}}$$

- $l_{x,real}$, $l_{y,real}$: 실제 사람 크기(mm)
- d : 구하려는 depth (mm)
- f : focal length (pix)
- $l_{x,sensor}$, $l_{y,sensor}$: 해상도(pix)
- p_x , p_y: pix 단위를 mm 단위로 바꿔주는 factor (pix/mm)
- α_x , α_y: focal length(mm)
- 결론 : l_{real} (2000mm로 fix), f (focal length) , l_{sensor} 만 알면 d(root depth)를 추정할 수 있음

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

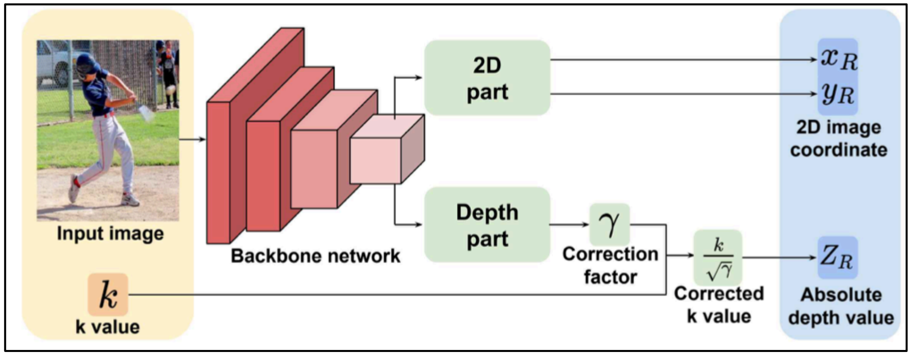
- **Model**
  - RootNet
    - Pinhole Camera Model



문제점 :

1) 사람의 크기인 l_{real} 을 2000mm X 2000mm 고정?
2) Bounding Box 크기와 Focal Length만으로 정확한 Depth를 추정할 수 있을까? – 그림 (a) , (b)

해결책: Input Image로부터 추출된 Feature로 Correction factor ***γ*** 를 학습시킴



3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Model**
  - RootNet
    - Pinhole Camera Model

Loss Function

$$\boxed{L_{root} = \|\mathbf{R} - \mathbf{R}^*\|_1}$$

장점

1) γ 값은 Input image에만 의존: 서로 다른 카메라 내부 파리미터(Focal length)를 갖는 Dataset들을 Flexible하게 학습가능

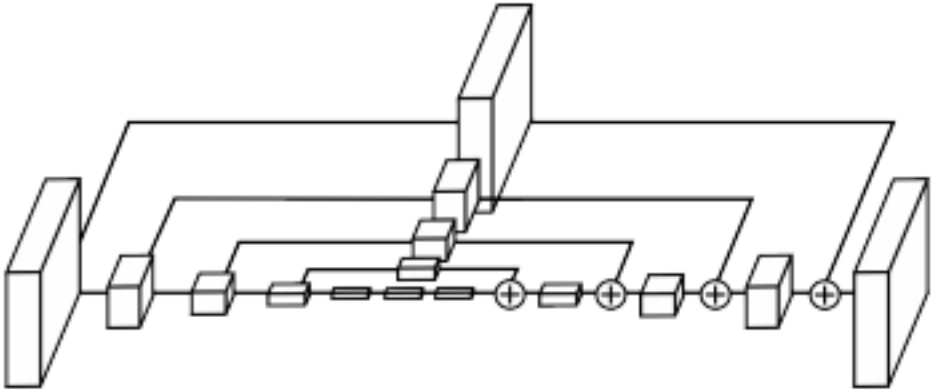
2) 다른 카메라로 찍은 Dataset Test시에도 카메라 내부 Parameter만 알면 Inference 가능

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Model**
  - PoseNet
    - *Integral Human Pose Regression (ECCV 2018)*
    - 3D Pose $P_j^{rel} = (x_j, y_j, Z_j^{rel})$를 추론(Single Person)
    - Contribution: Joint Point를 학습시키는 다른방식인 **Heatmap Representation** 과 **Joint Regression** 을 합쳐서 각각의 단점을 해소시킴

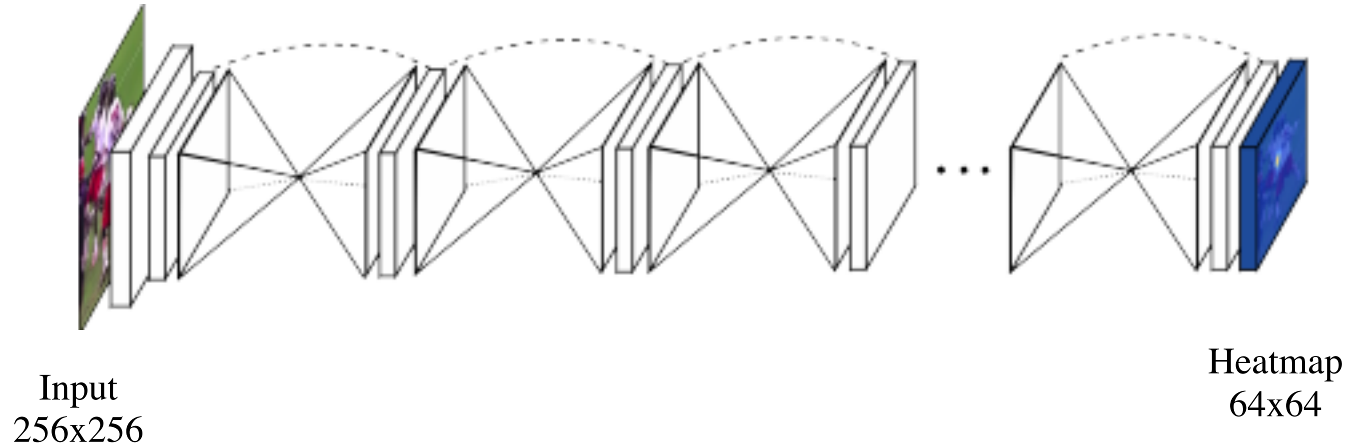
3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Model**
  - PoseNet
    - *Integral Human Pose Regression (ECCV 2018)*
      - Heatmap을 만드는 방식 : Stacked Hourglass Model을 사용
        - ✓ *Stacked Hourglass Networks for Human Pose Estimation(ECCV 2016)*



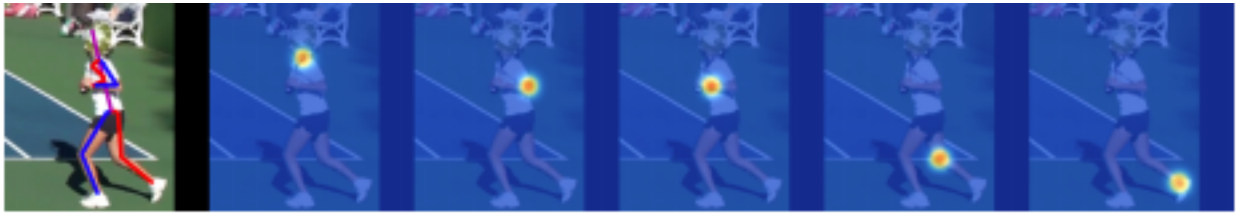
3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Model**
  - PoseNet
    - *Integral Human Pose Regression (ECCV 2018)*
      - Heatmap을 만드는 방식 : Stacked Hourglass Model을 사용
        - ✓ *Stacked Hourglass Networks for Human Pose Estimation(ECCV 2016)*



3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Model**
  - PoseNet
    - *Integral Human Pose Regression (ECCV 2018)*
      - Heatmap을 만드는 방식 : Stacked Hourglass Model을 사용
        - ✓ *Stacked Hourglass Networks for Human Pose Estimation(ECCV 2016)*



3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Model**
  - PoseNet
    - *Integral Human Pose Regression (ECCV 2018)*

Heatmap Representation만 사용할 때의 문제점

✓Non-Diffrentiable

$$\boxed{\mathbf{J}_k = \arg\max_{\mathbf{p}} \mathbf{H}_k(\mathbf{p}).}$$

➢ Maximum Likelihood 방식 : End-to-End Learning 불가능

✓Quantization Error

➢ High Resolution(256 x 256) → Low Resolution(64 x 64)

Joint Regression(CNN + FCN)만 사용할 때의 문제점

✓Low Performance

➢ 성능이 좋지 않음

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Model**
  - PoseNet
    - *Integral Human Pose Regression (ECCV 2018)*
      - 합치는 방식 (Soft-Argmax)

1. 64x64 Heatmap에 Softmax

$$\tilde{\mathbf{H}}_k(\mathbf{p}) = \frac{e^{\mathbf{H}_k(\mathbf{p})}}{\int_{\mathbf{q}\in\Omega} e^{\mathbf{H}_k(\mathbf{q})}}.$$

(64x64)

0.7	0.7	0.5	0.5
0.8	**0.9**	0.7	0.5
0.6	0.7	0.4	0.4
0.3	0.2	0.2	0.1

기존 Heatmap(확률이 가장 큰 지점 선택)

→

0.001	0.025	0.001	0.001
0.025	**0.8**	0.025	0.001
0.001	0.025	0.001	0.001
0.001	0.001	0.001	0.001

(64x64)

Softmax가 적용된 Heatmap (총합: 1)

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Model**

  - PoseNet
    - *Integral Human Pose Regression (ECCV 2018)*
      - 합치는 방식 (Soft-Argmax)

2. Expectation으로 Joint 좌표를 구함

$$\boxed{\mathbf{J}_k = \int_{\mathbf{p}\in\Omega} \mathbf{p} \cdot \tilde{\mathbf{H}}_k(\mathbf{p})}$$

1) Maximum Likelihood 식 → Expectation 식: End-To-End 학습 가능해짐
2) Continuous한 Joint Coord 값 : Quantization Error 해결

- 최종적으로 RootNet output{depth(mm)} + PoseNet Output{3d KeyPoints P_j^{rel}} = $\boldsymbol{P_k^{abs}}$ 추론

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Performance**

DetectNet	RootNet	AP^{box}	AP_{25}^{root}	AUC_{rel}	$3DPCK_{abs}$
R-50	k	43.8	5.2	39.2	9.6
R-50	Ours	43.8	28.5	39.8	31.5
X-101-32	Ours	**45.0**	**31.0**	**39.8**	**31.5**
GT	Ours	100.0	31.4	39.8	31.6
GT	GT	100.0	100.0	39.8	80.2

Table 2: Overall performance comparison for different DetectNet and RootNet settings on the MuPoTS-3D dataset.

- (1)-(2): 조정 상수 γ 유무 → Performance Up
- (2)-(3): Backbone R50 / X-101→ Not Big Gap
- (3)-(4): GT Bbox / Mask RCNN Bbox→ Not Big Gap

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Performance**

Settings	MRPE	MPJPE	Time
Joint learning	138.2	116.7	**0.132**
Disjointed learning (Ours)	**120.0**	**57.3**	0.141

Table 1: MRPE, MPJPE, and seconds per frame comparison between joint and disjointed learning on Human3.6M dataset.

- Disjointed Learning → Performance Much Up
- However, Processing time 비슷
- PoseNet-RootNet 관련성이 없기 때문

3DMPPE- *Camera Distance-aware Top-down Approch for 3D Multi-person Pose Estimation from a Single RBG Image (ICCV 2019)*

- **Performance**
  - 3DPCK (Multi-Persons) : SOTA

MuPoTS-3D				
	Year	Method	3DPCK ↑	
			All people	Matched people
Top down	2019	[189]	70.6	74.0
	2019	[191]	81.8	82.5
	2020	[166]	69.1	72.2
Bottom up	2018	[197]	65.0	69.8
	2019	[198]	70.4	-
	2020	[192]	72.0	-
	2020	[187]	73.5	80.5

  - MPJPE (Single-Person) : Without GT일 때 SOTA

Methods	Dir.	Dis.	Eat	Gre.	Phon.	Pose	Pur.	Sit	SitD.	Smo.	Phot.	Wait	Walk	WalkD.	WalkP.	Avg
With groundtruth information in inference time																
Chen [5]	89.9	97.6	90.0	107.9	107.3	93.6	136.1	133.1	240.1	106.7	139.2	106.2	87.0	114.1	90.6	114.2
Tome [46]	65.0	73.5	76.8	86.4	86.3	68.9	74.8	110.2	173.9	85.0	110.7	85.8	71.4	86.3	73.1	88.4
Moreno [32]	69.5	80.2	78.2	87.0	100.8	76.0	69.7	104.7	113.9	89.7	102.7	98.5	79.2	82.4	77.2	87.3
Zhou [53]	68.7	74.8	67.8	76.4	76.3	84.0	70.2	88.0	113.8	78.0	98.4	90.1	62.6	75.1	73.6	79.9
Jahangiri [17]	74.4	66.7	67.9	75.2	77.3	70.6	64.5	95.6	127.3	79.6	79.1	73.4	67.4	71.8	72.8	77.6
Mehta [28]	57.5	68.6	59.6	67.3	78.1	56.9	69.1	98.0	117.5	69.5	82.4	68.0	55.3	76.5	61.4	72.9
Martinez [26]	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0	94.6	62.3	78.4	59.1	49.5	65.1	52.4	62.9
Fang [7]	50.1	54.3	57.0	57.1	66.6	53.4	55.7	72.8	88.6	60.3	73.3	57.7	47.5	62.7	50.6	60.4
Sun [43]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	63.4	59.1
Sun [44]	**47.5**	**47.7**	**49.5**	**50.2**	**51.4**	**43.8**	**46.4**	**58.9**	**65.7**	**49.4**	**55.8**	**47.8**	**38.9**	**49.0**	**43.8**	**49.6**
Ours (PoseNet)	50.5	55.7	50.1	51.7	53.9	46.8	50.0	61.9	68.0	52.5	55.9	49.9	41.8	56.1	46.9	53.3
Without groundtruth information in inference time																
Rogez [40]	76.2	80.2	75.8	83.3	92.2	79.9	71.7	105.9	127.1	88.0	105.7	83.7	64.9	86.6	84.0	87.7
Mehta [29]	58.2	67.3	61.2	65.7	75.8	62.2	64.6	82.0	93.0	68.8	84.5	65.1	57.6	72.0	63.6	69.9
Rogez [41]*	55.9	60.0	64.5	56.3	67.4	71.8	55.1	**55.3**	84.8	90.7	67.9	57.5	47.8	63.3	54.6	63.5
Ours (Full)	**51.5**	**56.8**	**51.2**	**52.2**	**55.2**	**47.7**	**50.9**	63.3	**69.9**	**54.2**	**57.4**	**50.4**	**42.5**	**57.5**	**47.7**	**54.4**

Table 4: MPJPE comparison with state-of-the-art methods on the Human3.6M dataset using Protocol 2. ∗ used extra synthetic data for training.

감사합니다