ECCV'20 ONLINE
23-28 AUGUST 2020

HDNet: Human Depth Estimation for Multi-Person Camera-Space Localization

Jiahao Lin

Gim Hee Lee

NUS National University of Singapore | School of Computing

CVRP

Multi-Person 3D Pose Estimation

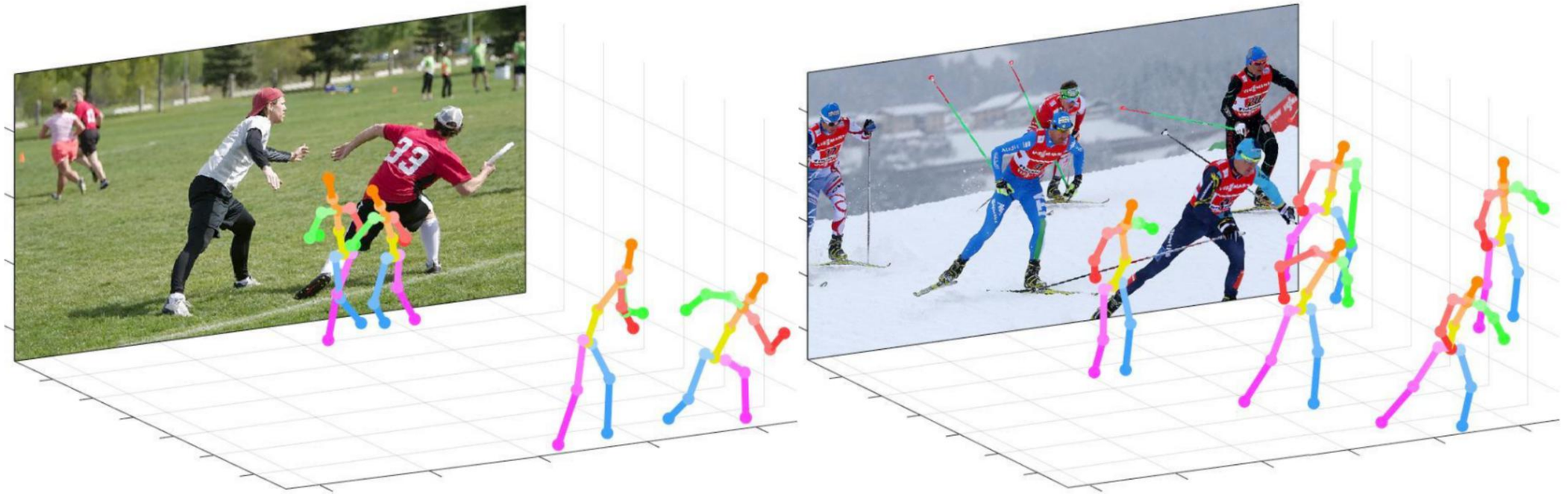


Image source: Moon et al. ICCV 2019

Multi-Person 3D Pose Estimation

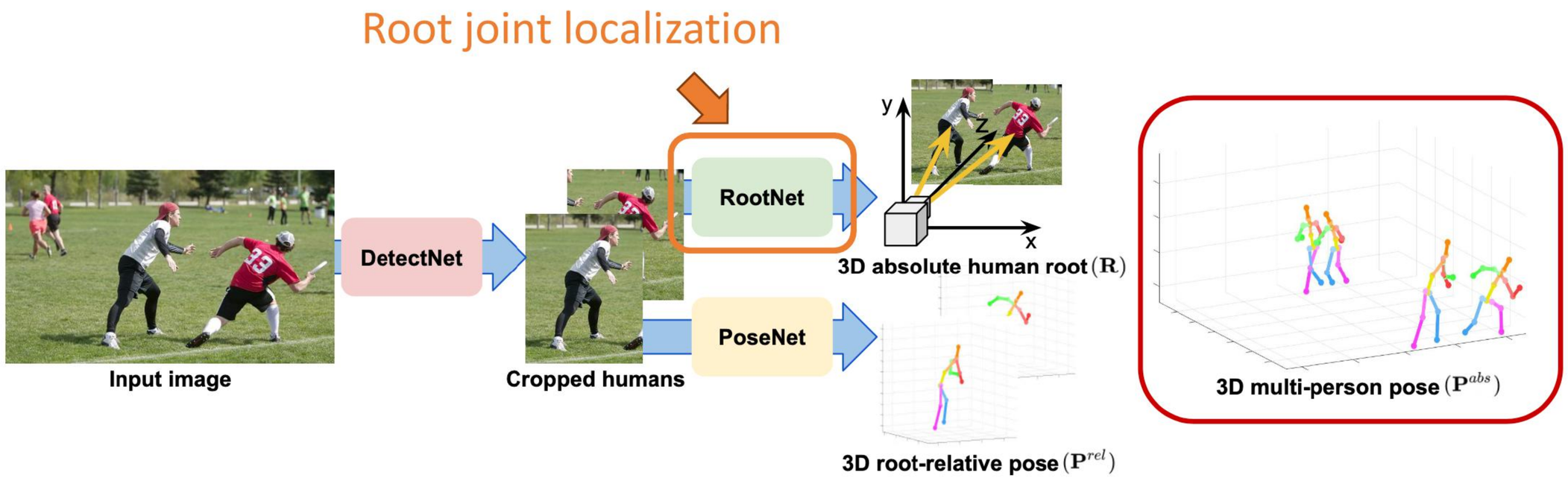
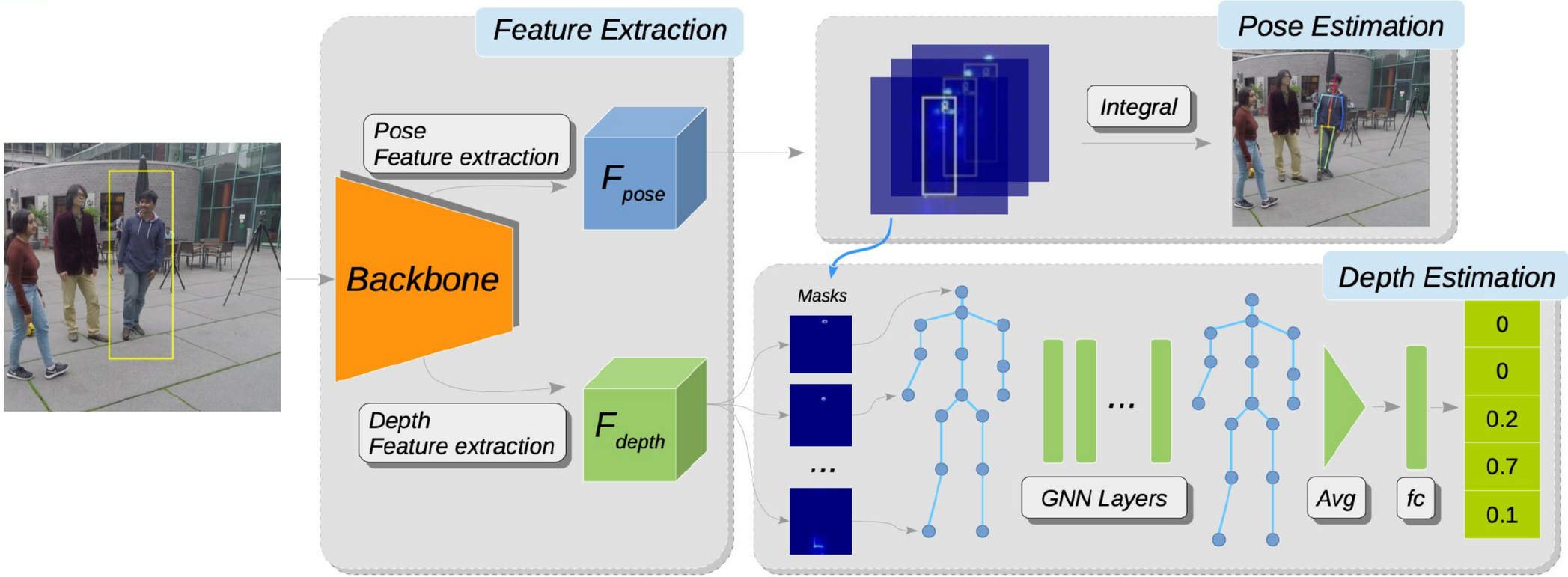
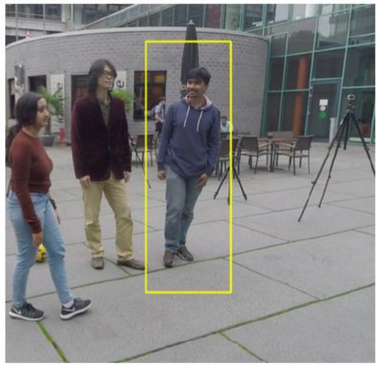


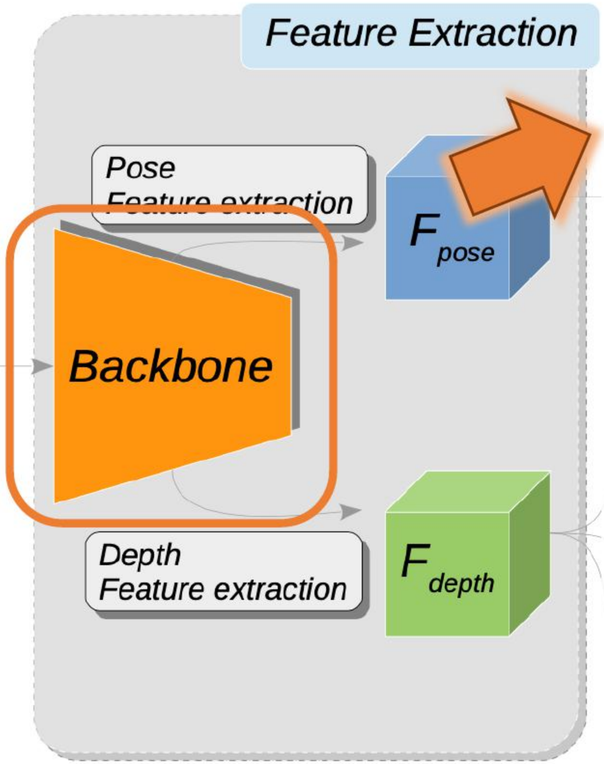
Image source: Moon et al. ICCV 2019

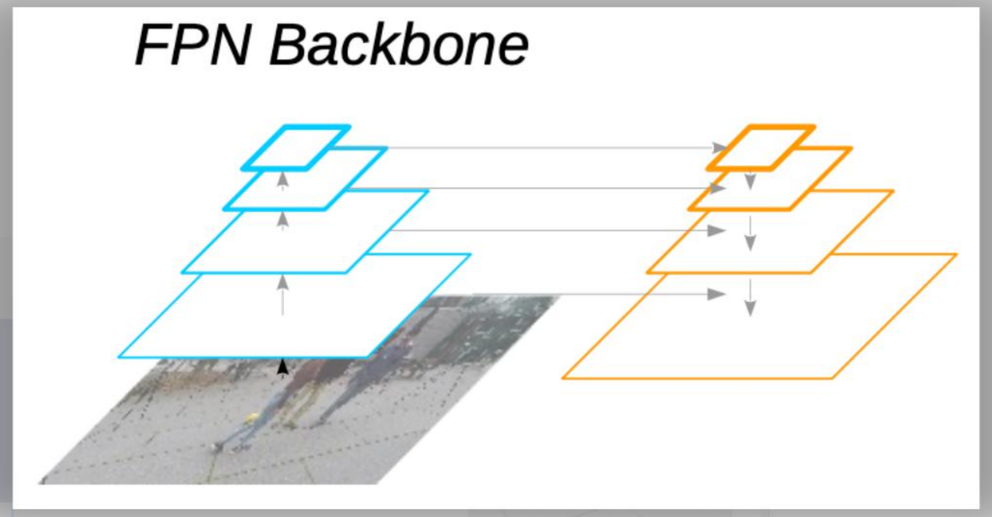
Overview


Feature Extraction
Pose Estimation
Pose
Feature extraction
F_{pose}
Backbone
Depth
Feature extraction
F_{depth}
Integral
Masks
...
GNN Layers
Avg
fc
Depth Estimation
0
0
0.2
0.7
0.1

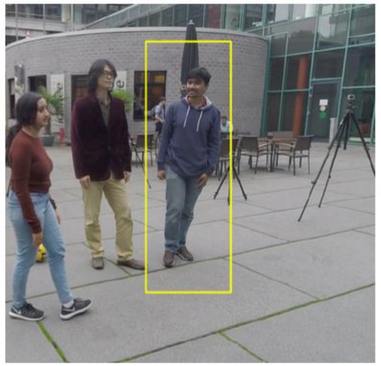
Feature Extraction

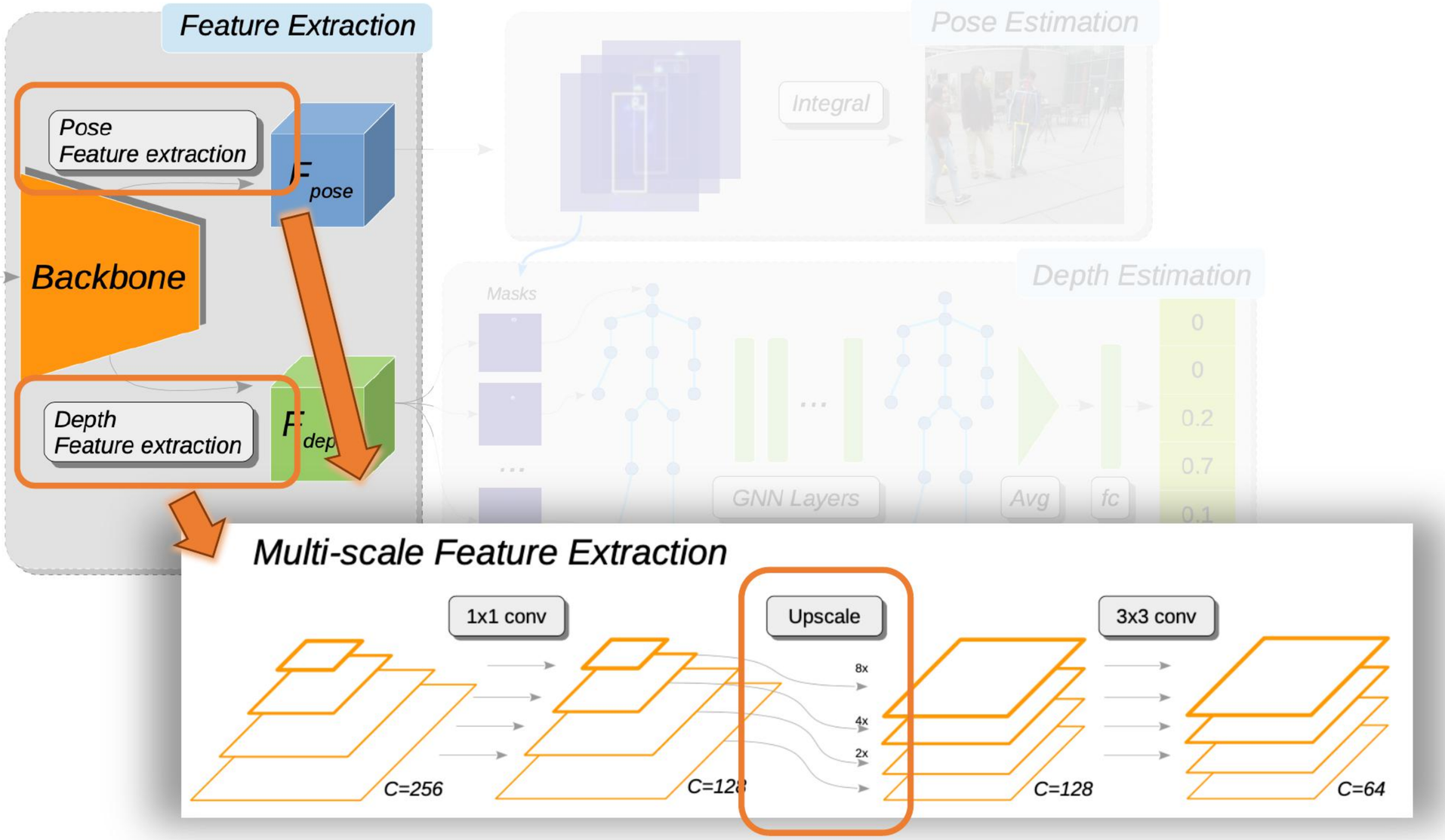






Feature Extraction

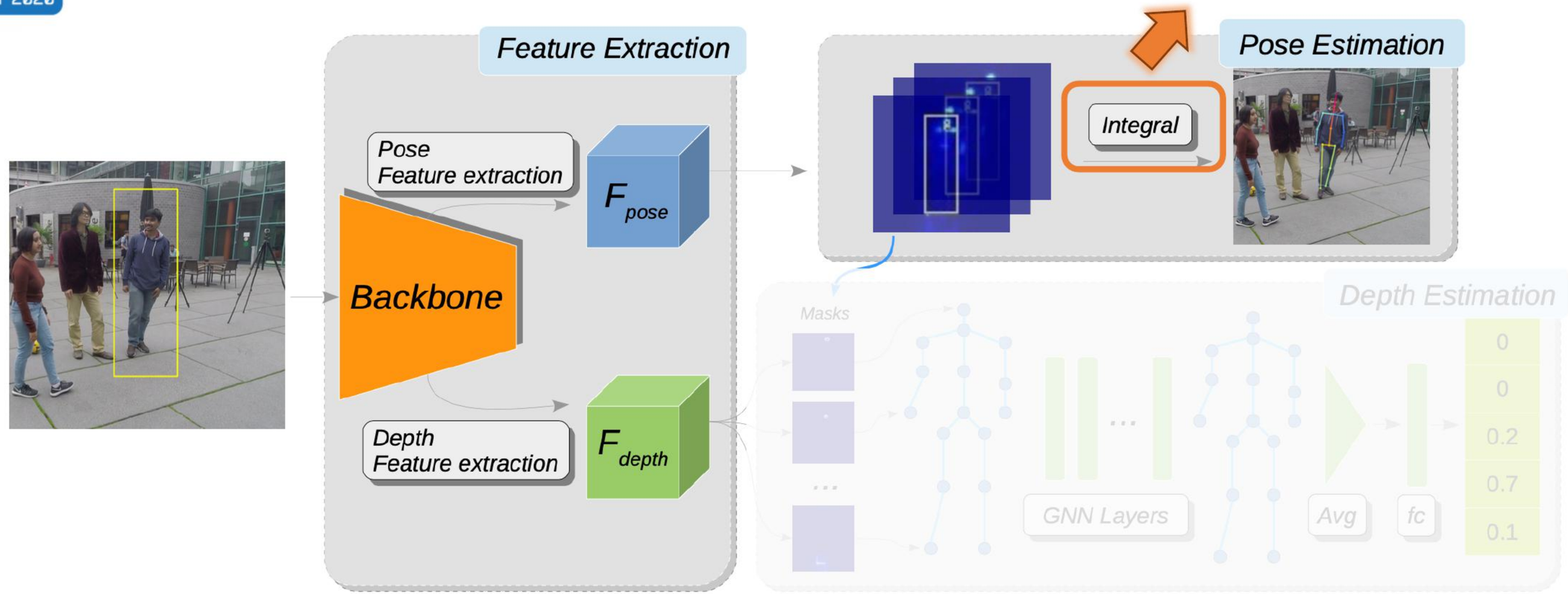




2D Pose Estimation

Soft-argmax

$$(\hat{u}_j, \hat{v}_j) = \sum_{(u,v)=(0,0)}^{(W-1,H-1)} \hat{\mathbf{H}}_{u,v}^{(j)} \cdot (u, v),$$

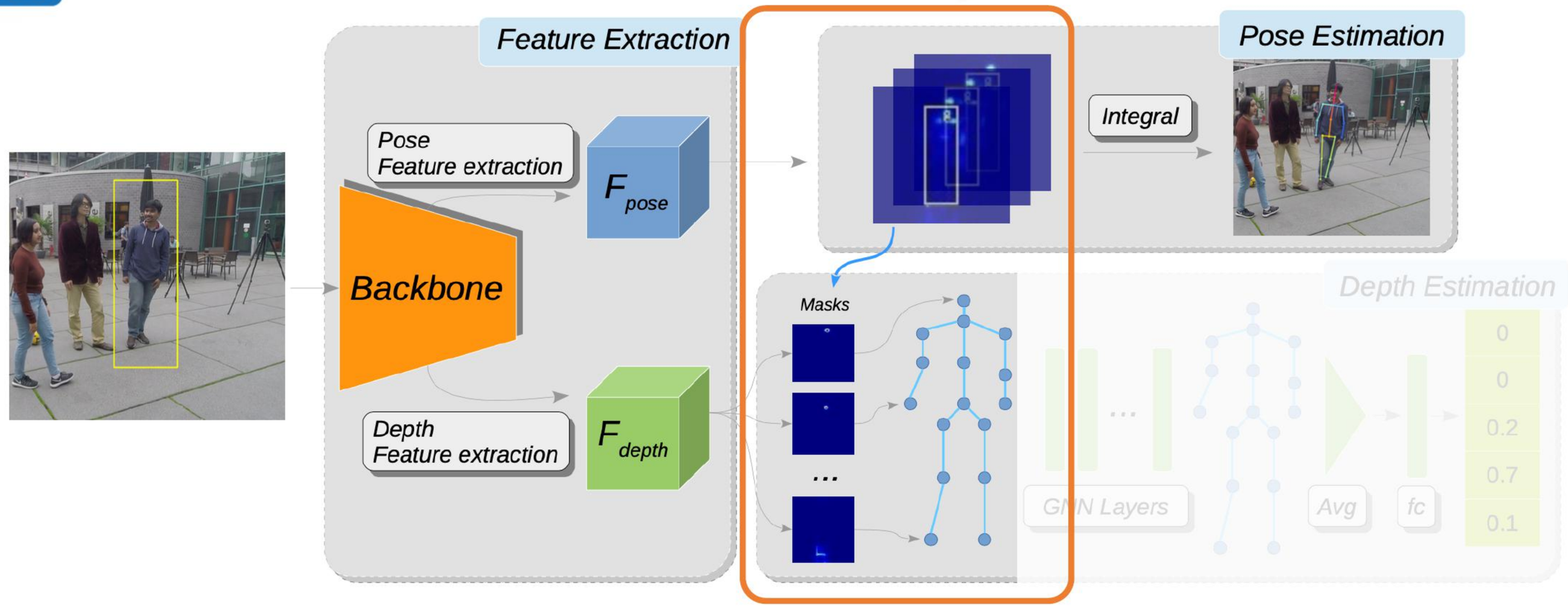


$$\mathcal{L}_{\text{hm}} = \frac{1}{N_J HW} \sum_{j}^{N_J} \sum_{(u,v)=(0,0)}^{(W-1,H-1)} \left\| \mathbf{H}_{u,v}^{(j)\text{GT}} - \hat{\mathbf{H}}_{u,v}^{(j)} \right\|^2$$

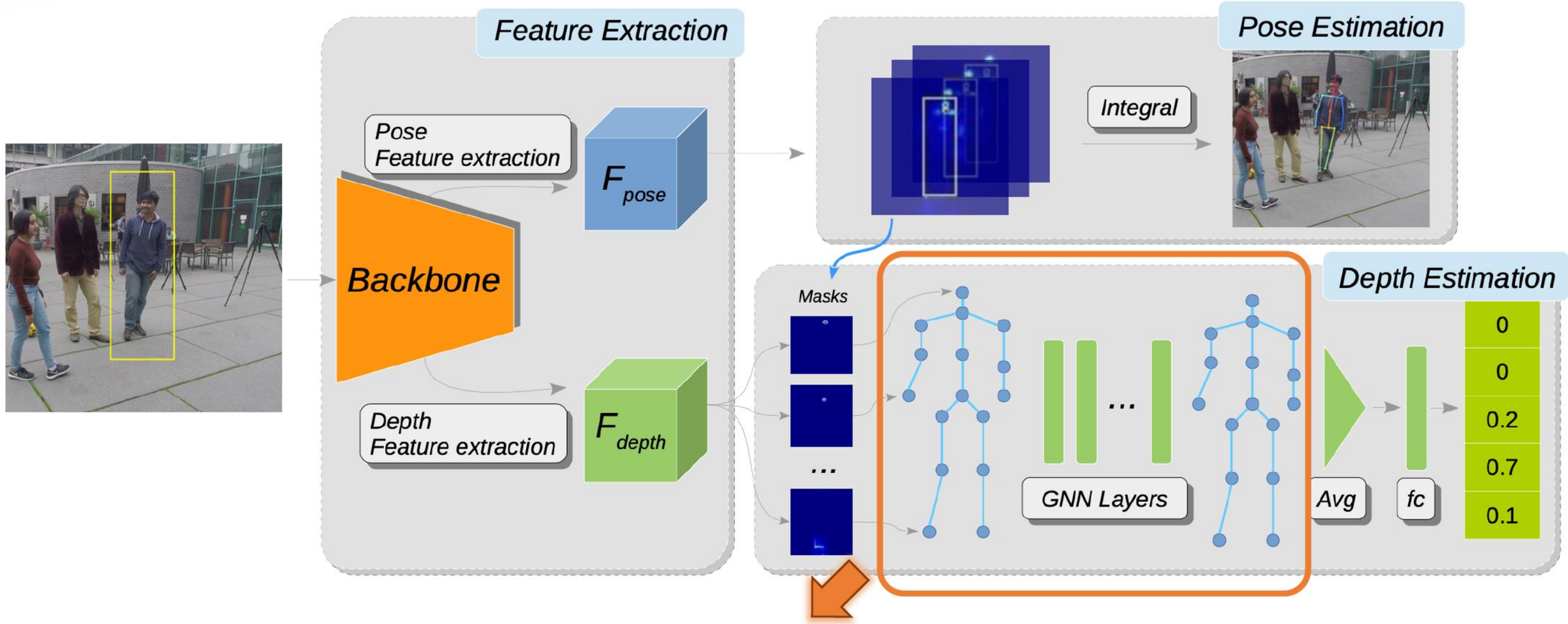
$$\mathcal{L}_{\text{pose}} = \frac{1}{N_J} \sum_{j}^{N_J} \left(\left| u_j^{\text{GT}} - \hat{u}_j \right| + \left| v_j^{\text{GT}} - \hat{v}_j \right| \right)$$

Depth Estimation

$$\mathbf{d}^{(j)} = \sum_{(u,v)=(0,0)}^{(W-1,H-1)} \hat{\mathbf{H}}_{u,v}^{(j)} \cdot \mathbf{F}_{\text{depth}_{u,v}}$$



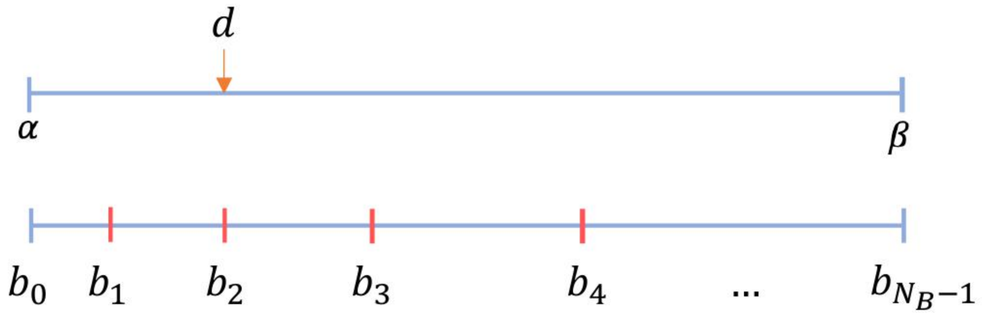
Depth Estimation



$$X_{out}^{(i)} = \sigma\left(\tilde{a}_{ii} f_{self}(X_{in}^{(i)}; \Theta_{self}) + \sum_{j \neq i} \tilde{a}_{ij} f_{inter}(X_{in}^{(j)}; \Theta_{inter})\right)$$

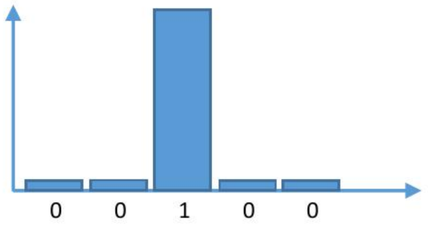
$$A \in \{0,1\}^{N_j \times N_j} \Rightarrow \tilde{A}$$

Depth Estimation

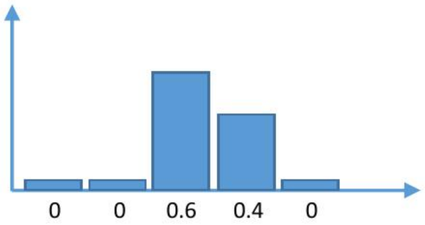


$$b(d) = \frac{\log d - \log \alpha}{\log \beta - \log \alpha} \cdot (N_{\mathbf{B}} - 1)$$

$N_{\mathbf{B}} = 5$ and $b = 2.4$

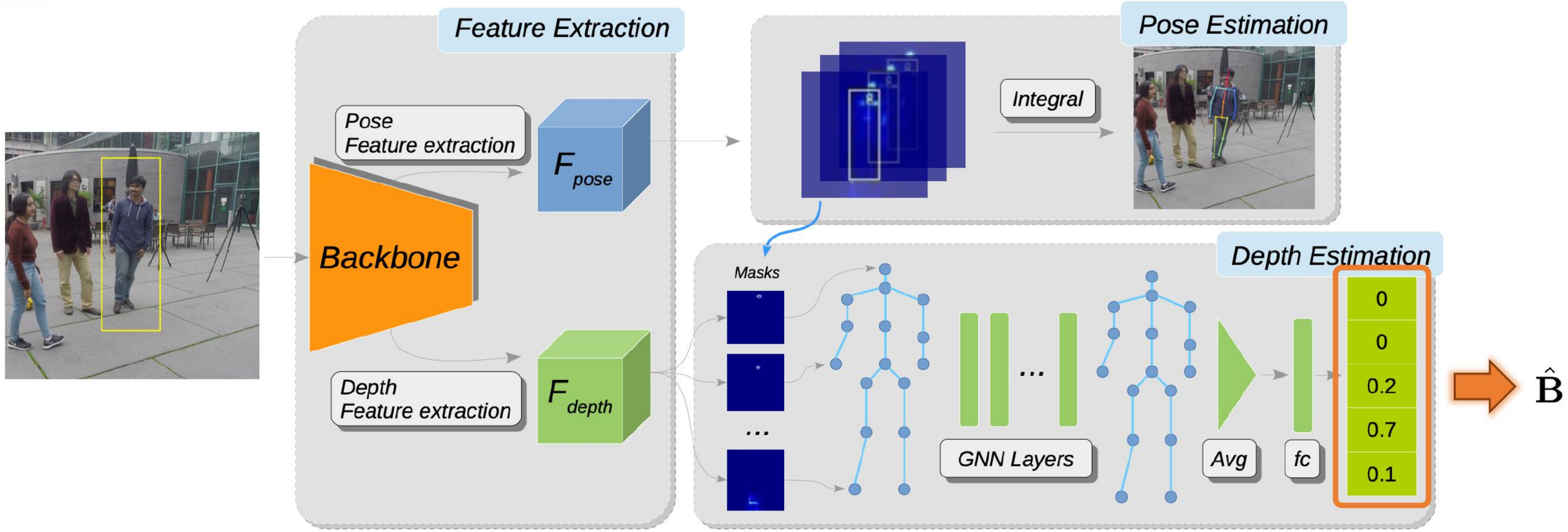


One-hot



Bi-linear

Depth Estimation



$$d = \left[\frac{\hat{b}}{N_{\mathbf{B}} - 1} \cdot (\log\beta - \log\alpha) + \log\alpha\right] \cdot \sqrt{f_x \cdot f_y}, \quad \text{where } \hat{b} = \sum_{i=0}^{N_{\mathbf{B}}-1} \hat{\mathbf{B}}_i \cdot i$$

$$\mathcal{L}_{\text{bins}} = -\sum_{i=0}^{N_{\mathbf{B}}-1} \mathbf{B}_i^{\text{GT}} \cdot \log \hat{\mathbf{B}}_i, \text{ and } \mathcal{L}_{\text{idx}} = \left| b^{\text{GT}} - \hat{b} \right|$$

Soft-argmax

Evaluation

Human3.6M

- Mean Root Position Error (MRPE)

MuCo-3DHP

MuPoTS-3D

- Average Precision (AP) and Recall (AR)
- $3DPCK_{rel}$ and $3DPCK_{abs}$

Image source: http://http://vision.imar.ro/human3.6m

Image source: Mehta et al. 3DV 2018

Results

Table 1. MRPE results comparison with state-of-the-arts on the Human3.6M dataset. MRPE_x, MRPE_y, and MRPE_z are the average errors in x, y, and z axes, respectively.

Method	MRPE	MRPE_x	MRPE_y	MRPE_z
Baseline	267.8	27.5	28.3	261.9
Baseline w/o limb joints	226.2	24.5	24.9	220.2
Baseline with RANSAC	213.1	24.3	24.3	207.1
RootNet [21]	120.0	23.3	23.0	108.1
Ours	**77.6**	**15.6**	**13.6**	**69.9**

Table 2. Root joint localization accuracy comparison in average precision and recall with state-of-the-arts on MuPoTS-3D dataset.

Method	$\text{AP}_{25}^{\text{root}}$	$\text{AP}_{20}^{\text{root}}$	$\text{AP}_{15}^{\text{root}}$	$\text{AP}_{10}^{\text{root}}$	$\text{AR}_{25}^{\text{root}}$	$\text{AR}_{20}^{\text{root}}$	$\text{AR}_{15}^{\text{root}}$	$\text{AR}_{10}^{\text{root}}$
RootNet [21]	31.0	21.5	10.2	2.3	55.2	45.3	31.4	15.2
Ours	**39.4**	**28.0**	**14.6**	**4.1**	**59.8**	**50.0**	**35.9**	**19.1**

Table 4. Joint-wise 3DPCK_{abs} comparison with state-of-the-arts on MuPoTS-3D dataset. Accuracy is measured on matched ground-truths.

Method	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Avg
RootNet [21]	37.6	35.6	34.0	34.1	30.7	30.6	31.3	25.3	31.8
Ours	**38.3**	**37.8**	**36.2**	**37.4**	**34.0**	**34.9**	**36.4**	**29.2**	**35.2**

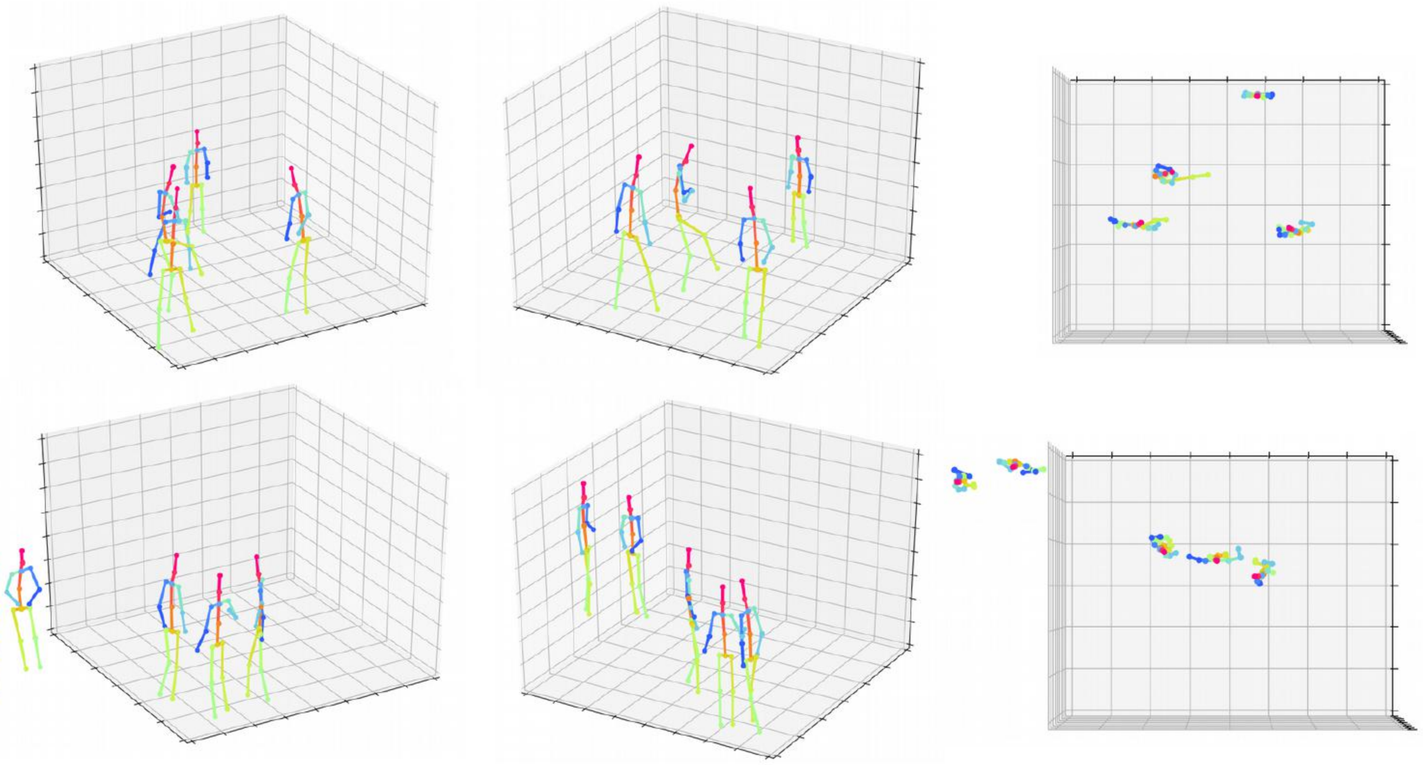
Table 5. Sequence-wise 3DPCK_{rel} comparison with state-of-the-arts on MuPoTS-3D dataset. Accuracy is measured on matched ground-truths.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	-
Rogez *et al.* [26]	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	-
Mehta *et al.* [20]	81.0	65.3	64.6	63.9	75.0	30.3	65.1	61.1	64.1	83.9	-
Rogez *et al.* [27]	88.0	73.3	67.9	74.6	81.8	50.1	60.6	60.8	78.2	89.5	-
RootNet [21]	**94.4**	78.6	79.0	82.1	86.6	72.8	**81.9**	75.8	**90.2**	90.4	-
Ours	**94.4**	**79.6**	**79.2**	**82.4**	**86.7**	**73.0**	81.6	**76.3**	90.1	**90.5**	-
Method	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg
Rogez *et al.* [26]	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
Mehta *et al.* [20]	72.4	69.9	71.0	72.9	71.3	83.6	79.6	73.5	78.9	**90.9**	70.8
Rogez *et al.* [27]	70.8	74.4	72.8	64.5	74.2	84.9	85.2	78.4	75.8	74.4	74.0
RootNet [21]	**79.4**	**79.9**	75.3	81.0	**81.1**	90.7	**89.6**	83.1	81.7	77.3	82.5
Ours	77.9	79.2	**78.3**	**85.5**	**81.1**	**91.0**	88.5	**85.1**	**83.4**	90.5	**83.7**

Table 6. Ablation studies on components of the framework. Depth error MRPE_z (mm) on Human3.6M dataset and $\text{AP}_{25}^{\text{root}}$ (%) on MuPoTS-3D dataset are measured.

Method	$\text{MRPE}_z(\downarrow)$	$\text{AP}_{25}^{\text{root}}(\uparrow)$
RootNet [21]	108.1	31.0
Ours direct regression	94.5	27.3
Ours shared feature branch	72.0	31.9
Ours w/o GNN	72.9	32.7
Ours w/o HM pooling	71.8	26.0
Ours (full)	**69.9**	**39.4**

ECCV'20
ONLINE
23-28 AUGUST 2020



See our project page for more information:
https://github.com/jiahaoLjh/HumanDepth