IVMSP-P18.10

Orchestrating a brighter world

3D POSE ESTIMATION FROM MONOCULAR VIDEO WITH CAMERA-BONE ANGLE REGULARIZATION ON THE IMAGE FEATURE

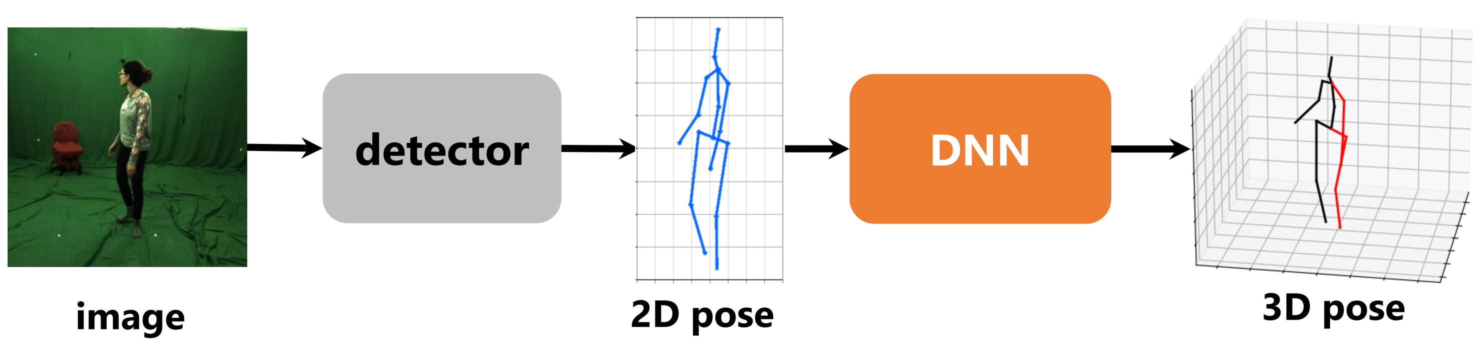
Asuka Ishii, Hiroo Ikeda (NEC)

Appearance information bounds solution space of 2D-to-3D lifting

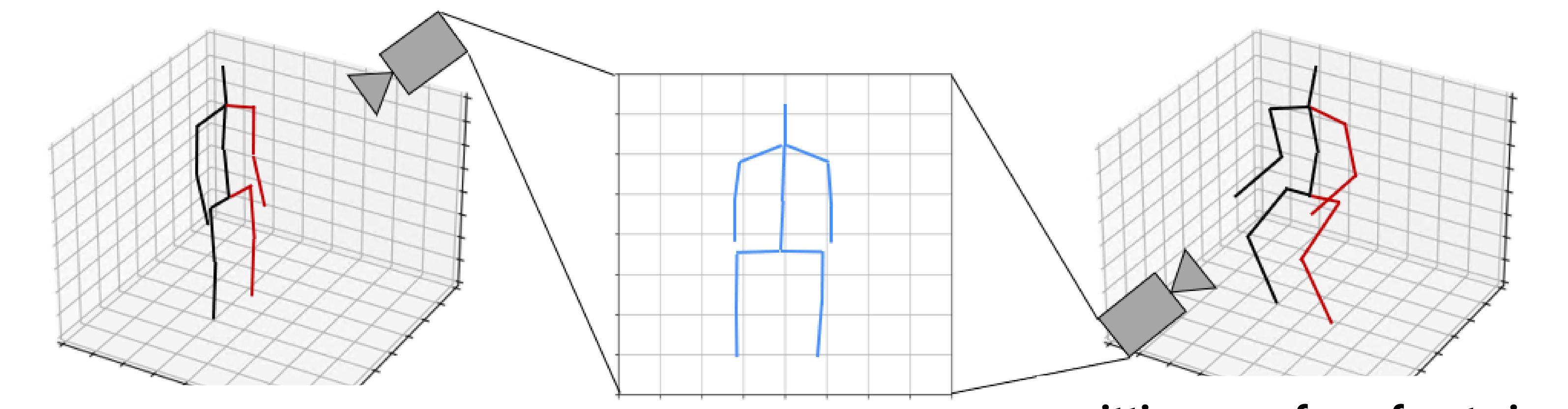
Introduction | 3D human pose estimation from monocular video

2D-to-3D lifting

- estimates 3D pose(s) only from 2D pose(s) detected by another detector
- much lower error than image-based methods [14]
- relative 3D coordinates from root joint (typically center-hip)



ill-posed problem



one 2D pose can be mapped to multiple 3D poses

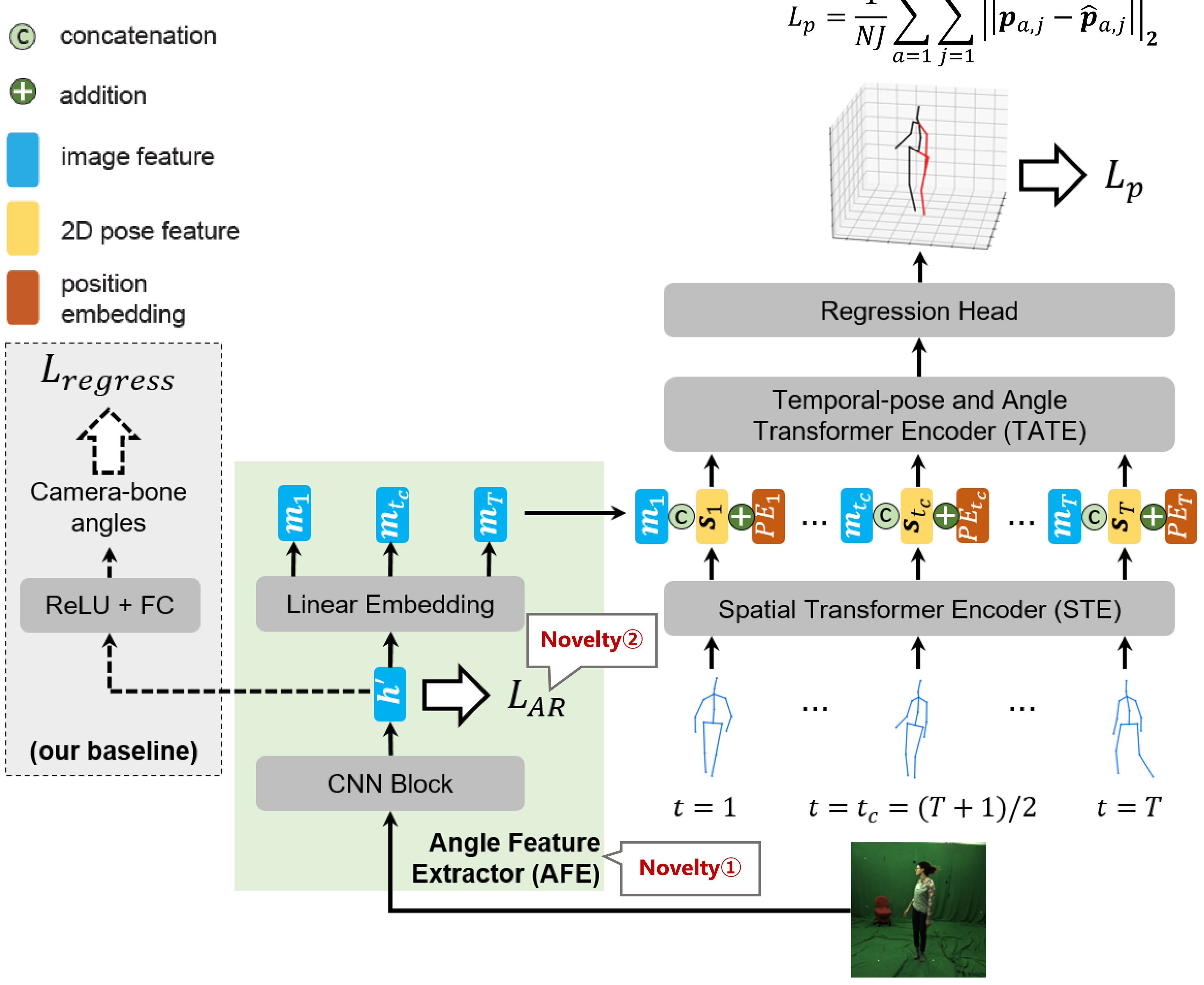
existing methods: consider temporal information only

our method: considers appearance as well

Method | appearance information of subject

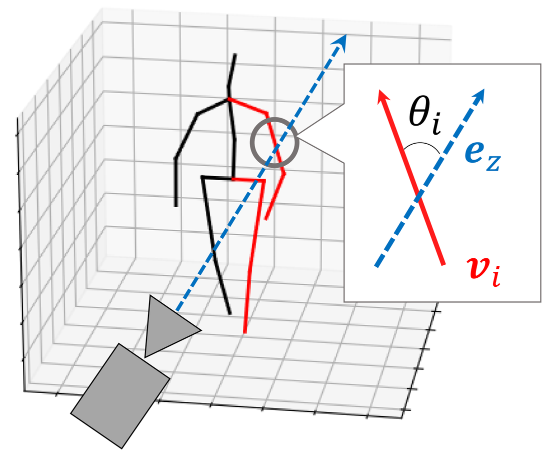
Novelty① CNN (AFE) added on PoseFormer[29], a 2D-to-3D-based network

$$L_p = \frac{1}{NJ}\sum_{a=1}^{N}\sum_{j=1}^{J}\left\|\boldsymbol{p}_{a,j} - \hat{\boldsymbol{p}}_{a,j}\right\|_2$$



Novelty② Regularization on image features using camera-bone angles

- **Camera-bone angles**
  between **camera optical axis** $\boldsymbol{e}_z$ and **bones** $\boldsymbol{v}$ computed from ground-truth



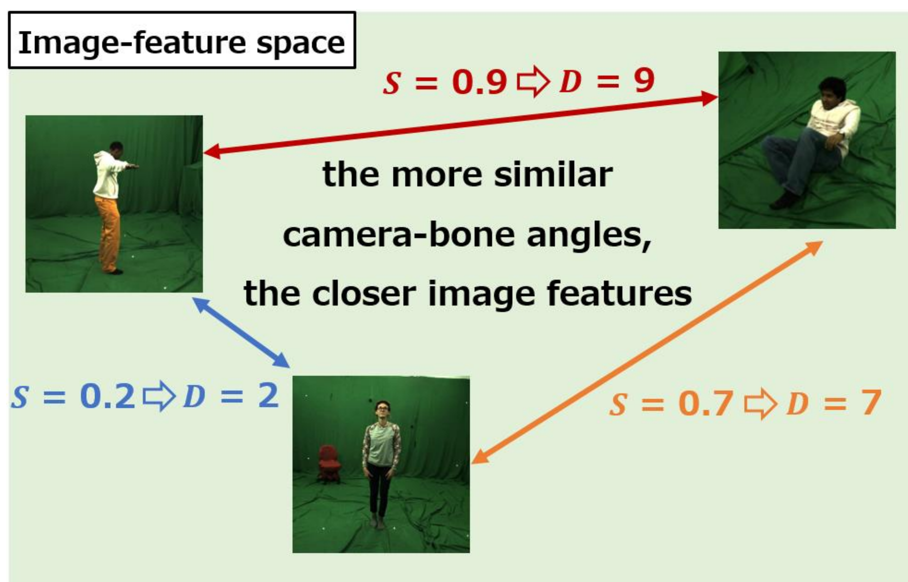
$\boldsymbol{c} = (\theta_1, \theta_2, \ldots, \theta_B)$: camera-bone angle vector

$\theta_i = \arccos\left(\frac{\boldsymbol{v}_i \cdot \boldsymbol{e}_z}{||\boldsymbol{v}_i||\,||\boldsymbol{e}_z||}\right)$: angle between i-th bone and camera optical axis

$\boldsymbol{v}_i = \hat{\boldsymbol{p}}_{i1} - \hat{\boldsymbol{p}}_{i2}$: vector of i-th bone

$\hat{\boldsymbol{p}}_{i1}, \hat{\boldsymbol{p}}_{i2}$: ground-truth 3D coordinates of joints at both ends of i-th bone

- **Regularization loss** L_{AR}
  makes D, **distance between image features** of minibatch samples, **proportional** to S, **unsimilarity of camera-bone angles** of the samples



$$L_{total} = L_p + \lambda L_{AR}$$

$$L_{AR} = \frac{2}{N(N-1)}\sum_{a=1}^{N}\sum_{b=1}^{a} l_{a,b}$$

$$l_{a,b} = \left\|\alpha S(\boldsymbol{c}_a, \boldsymbol{c}_b) - D(\boldsymbol{h}'_a, \boldsymbol{h}'_b)\right\|_2$$

$$S(\boldsymbol{c}_a, \boldsymbol{c}_b) = 1 - \frac{\boldsymbol{c}_a \cdot \boldsymbol{c}_b}{||\boldsymbol{c}_a||\,||\boldsymbol{c}_b||}$$

$$D(\boldsymbol{h}'_a, \boldsymbol{h}'_b) = \left\|\boldsymbol{h}'_a - \boldsymbol{h}'_b\right\|_2$$

Why camera-bone angles?

Given $(||v_i||, K, \theta_i)$, 2D-to-3D lifting solved analytically

(K: camera intrinsic parameters)

However, training model to extract

- $||v_i||$ may cause overtraining
- K requires large extra training data [13]

We focus on camera-bone angles

Experiment

* indicates the experiment was conducted in our environment. Otherwise, values were taken from original papers.

Approach	Method	Human3.6M # frames	Human3.6M MPJPE [mm] ↓	MPI-INF-3DHP # frames	MPI-INF-3DHP MPJPE [mm] ↓
Image-based	Pavlalos+2018 [5]	1	56.2	1	-
	MargiPose [7]	1	55.4	1	85.2
2D-to-3D-based	*PoseFormer [10]	81	49.9	9	50.0
Ours	*PoseFormer + AFE	81	59.8	9	69.6
	*PoseFormer + AFE w/ $L_{regress}$	81	52.4	9	64.8
	*PoseFormer + AFE w/ L_{AR}	81	**44.8**	9	**47.9**

Conclusion & future work

Conclusion

- proposed to bound solution space of 2D-to-3D method, an ill-posed problem, by considering appearance information of subject as well.
- proposed regularization loss using camera-bone angles on image features.
- empirically showed the proposed method improves performance.

Future work

- Replacement based 2D-to-3D network with SOTA
- Evaluation on unseen camera angles