Compositional Human Pose Regression

Xiao Sun

Joint work with Yichen Wei

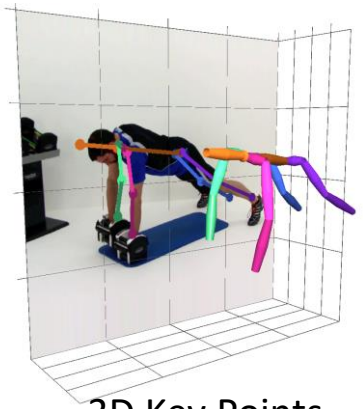
Human Pose Estimation

- Problem: localize key points of a person
- Input: a single RGB image
- Output: 2D or 3D key points

RGB Image (person centered)

Pose Estimator

2D Key Points



3D Key Points

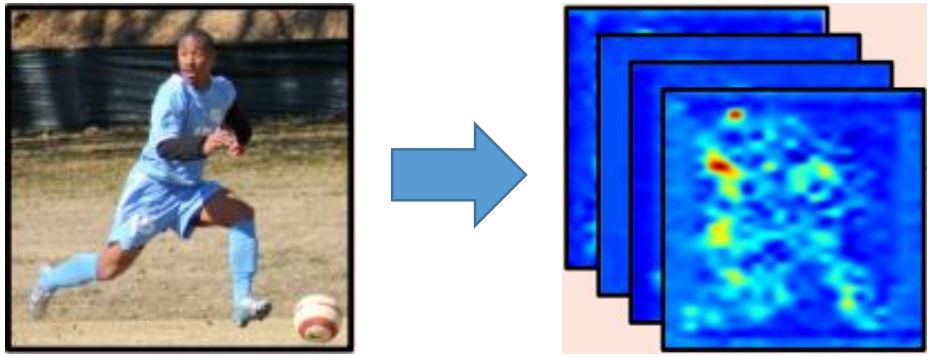
Detection VS. Regression

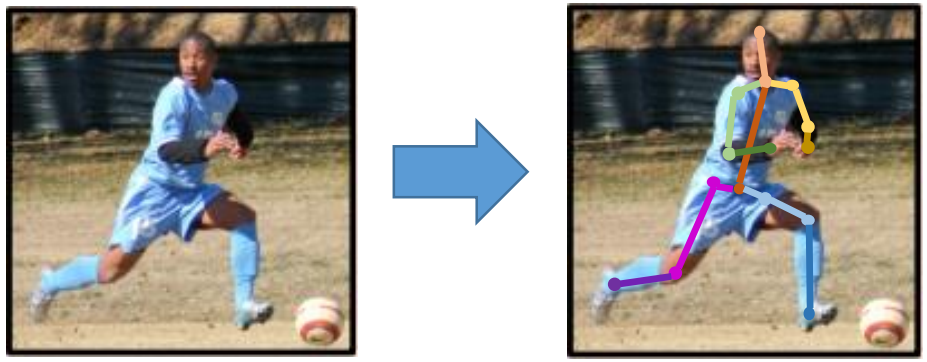
Detection

- Per-pixel classification
- Output: likelihood score maps

Regression

- Location regression
- Output: key points location





Performance

Detection

- Per-pixel classification
- Output: likelihood score maps
- Used in most 2D methods
- State-of-the-art result

Regression

- Location regression
- Output: key points location
- Only used in a few 2D methods
- Unsatisfactory result

2D Pose Benchmark: MPII dataset

- Andriluka et al., 2d human pose estimation: New benchmark and state of the art analysis, CVPR 2014
- YouTube videos, 410 daily activities
- Complex poses and appearances
- 25k images, 40k annotated 2D poses

MPII Leader Board

Metric: percentage of correct keypoints (PCK). The higher, the better.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	PCKh
Pishchulin et al., ICCV'13	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Tompson et al., NIPS'14	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira et al., CVPR'16	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al., CVPR'15	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu&Ramanan., CVPR'16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin et al., CVPR'16*	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz et al., ECCV'16	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary et al., ECCV'16	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi et al., BMVC'16	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Belagiannis&Zisserman, FG'17**	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Insafutdinov et al., ECCV'16	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al., CVPR'16*	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Ning et al., arXiv'2017	97.9	95.4	90.3	85.5	89.3	84.6	78.3	89.3
Bulat&Tzimiropoulos, ECCV'16	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al., ECCV'16	98.2	**96.3**	91.2	87.1	90.1	87.4	83.6	90.9
Chu et al., CVPR'17	**98.5**	**96.3**	**91.9**	**88.1**	**90.6**	**88.0**	**85.0**	**91.5**

Only one regression method

Not competitive to detection

Reason: exploit joint dependency

Detection

- Per-pixel classification
- Output: likelihood score maps
- Used in most 2D methods
- State-of-the-art result
- Scoremap is more expressive

Regression

- Location regression
- Output: key points location
- Only used in a few 2D methods
- Unsatisfactory result
- Dependency not well exploited

Generalization

Detection

- Per-pixel classification
- Output: likelihood score maps
- Used in most 2D methods
- State-of-the-art result
- Scoremap is more expressive
- Hard to generalize to 3D task

Regression

- Location regression
- Output: key points location
- Only used in a few 2D methods
- Unsatisfactory result
- Dependency not well exploited
- General for both 2D and 3D task

Motivation of this work

Detection

- Per-pixel classification
- Output: likelihood score maps
- Used in most 2D methods
- State-of-the-art result
- Scoremaps are more expressive
- Hard to generalize to 3D task

Regression

- Location regression
- Output: key points location
- Only used in a few 2D methods
- Unsatisfactory result
- Dependency not well exploited
- General for both 2D and 3D task

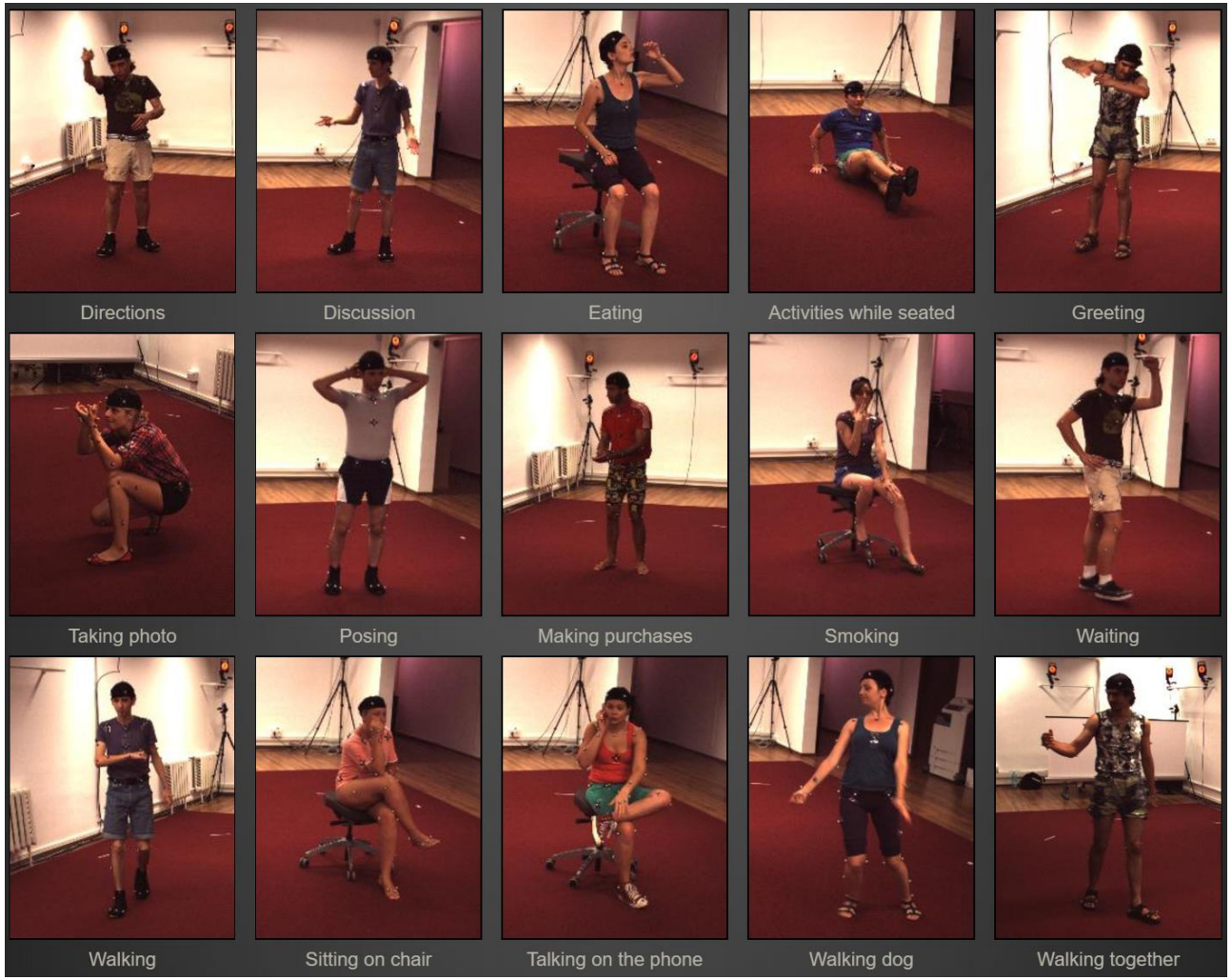
Proposed: structure-aware regression method

- A novel pose representation and novel loss function
  - Better exploit joint dependency
  - Unified framework for 3D and 2D tasks
  - Complementary to network architectures

- State-of-the-art on both 2D and 3D tasks (ICCV2017 submission)

3D Pose Benchmark: Human 3.6M dataset

- Lonescu et al., Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, PAMI 2014
- Ground truth by motion capture
- 7 subjects x 15 actions x 4 cameras
- Millions of RGB frames



Our Performance (3D)

1. Dataset: Human3.6M.
2. Metrics: mean joint position error in *mm*. The lower, the better.
3. Advance the state-of-the-art a large margin, 12.7%.
4. A record of 48.3mm average joint error.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Yasin[52]	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0
Rogez[40]	-	-	-	-	-	-	-	-
Chen[7]	71.6	66.6	74.7	79.1	70.1	67.6	89.3	90.7
Bogo[4]	62.0	60.2	67.8	76.5	92.1	73.0	75.3	100.3
Moreno[30]	67.4	63.8	87.2	73.9	71.5	69.9	65.1	71.7
Zhou[56]	47.9	48.8	52.7	55.0	56.8	49.0	45.5	60.8
Baseline	45.2	46.0	47.8	48.4	54.6	43.8	47.0	60.6
Ours(all)	**42.1**	**44.3**	**45.0**	**45.4**	**51.5**	**43.2**	**41.3**	**59.3**

Method	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg
Yasin[52]	170.8	108.2	142.5	86.9	92.1	165.7	102.0	108.3
Rogez[40]	-	-	-	-	-	-	-	88.1
Chen[7]	195.6	83.5	93.3	71.2	55.7	85.9	62.5	82.7
Bogo[4]	137.3	83.4	77.0	77.3	86.8	79.7	81.7	82.3
Moreno[30]	98.6	81.3	93.3	74.6	76.5	77.7	74.6	76.5
Zhou[56]	81.1	53.7	65.5	51.6	50.4	54.8	55.9	55.3
Baseline	79.0	54.5	56.0	46.7	42.2	51.0	47.9	51.4
Ours (all)	**73.3**	**51.0**	**53.0**	**44.0**	**38.3**	**48.0**	**44.8**	**48.3**

Our Performance (2D)

1. Dataset: MPII.
2. Metrics: percentage of correct keypoints (PCK). The higher, the better.
3. Advance the state-of-the-art regression method 6.3%.
4. Competitive with the state-of-the-art detection methods.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Pishchulin [37]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Tompson[46]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Tompson[45]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu[17]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin[38]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz[25]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary[14]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Raf[39]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Insafutdinov[18]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei[47]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat[5]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell[31]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.0
Chu[11]	**98.5**	**96.3**	**91.9**	**88.1**	**90.6**	**88.0**	**85.0**	**91.5**
Carreira(IEF)[6]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
IEF*	96.3	92.6	83.1	74.6	83.7	74.1	71.4	82.9
Ours (all)	**97.5**	**94.3**	**87.0**	**81.2**	**86.5**	**78.5**	**75.4**	**86.4**

Detection

Regression

Two Key Techniques

- Bone based pose representation
  - Simplify the problem
- Compositional loss function
  - Encodes long range interactions between bones

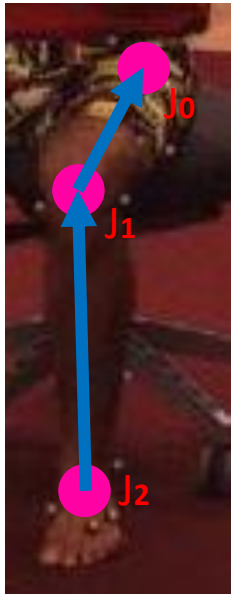
Pose Representation: Joint VS. Bone

Joint

- Relative position to the root joint.
- Joint output:
  $\Delta J_{k,0} = J_k - J_0$
- Joint loss:
  $||\Delta J_{k,0} - \Delta J_{k,0}^{gt}||_2$

Bone

- Relative position to its parent joint.
- Bone output:
  $\Delta J_{k,par(k)} = J_k - J_{par(k)}$
- Bone loss:
  $||\Delta J_{k,par(k)} - \Delta J_{k,par(k)}^{gt}||_2$



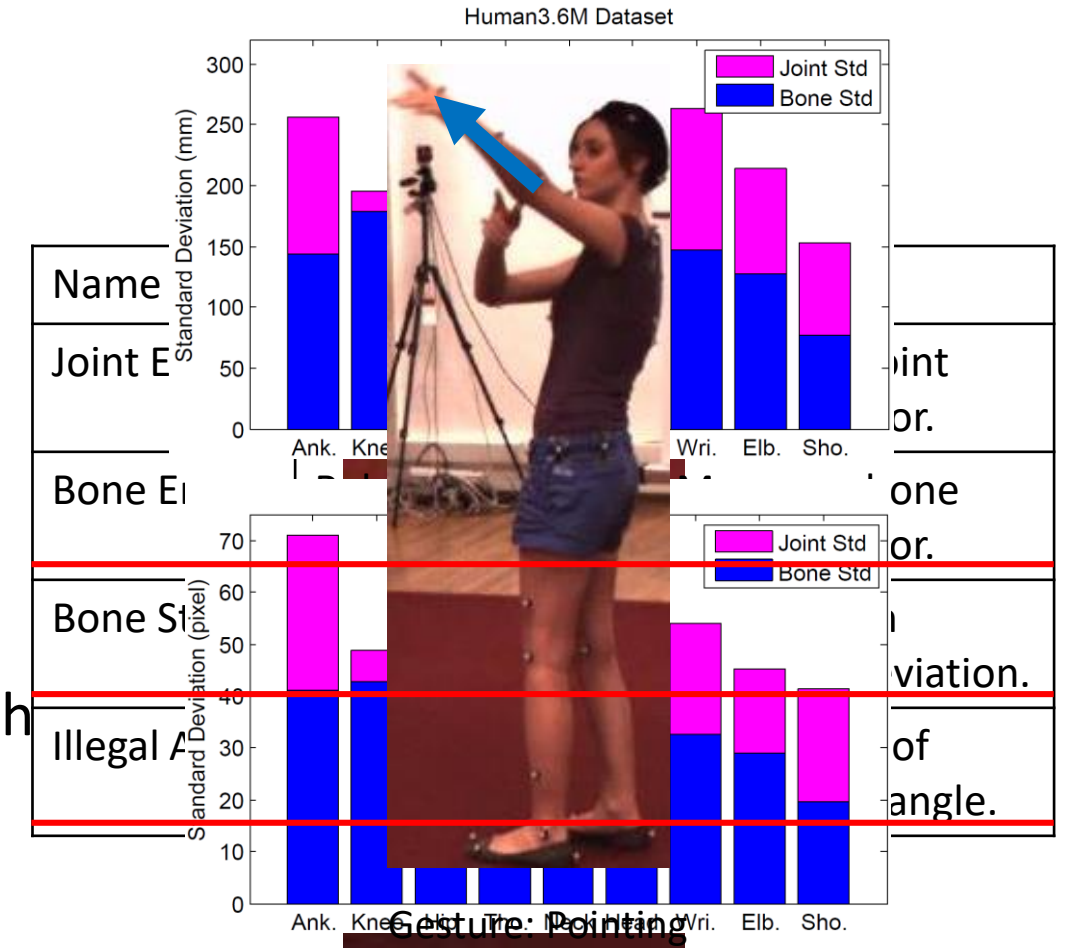
Joint Representation: Drawbacks

- Joints independently estimated
- Internal structure not exploited
- Geometric constraint not satisfied
  - Bone length not constant
  - Joint angle may out of range

Bone Representation: Advantages

- Joints are connected in a tree structure
- Bones are primitive units and local
  - Significantly smaller variance in targets
  - Application convenience: local motion is enough
  - Geometric constraint better satisfied

    (New evaluation metrics)



Standard deviation of bones and joints for the 3D Human3.6M dataset and 2D MPII dataset

Use Bone Loss Only: Drawback

- Joint location of $ankle$ is a summation of thigh and shin:

$$\nabla J_{2,0} = \nabla J_{2,1} + \nabla J_{1,0}$$

bone 2 bone 1

- Joint error of $ankle$: $||\nabla J_{2,0} - \nabla J_{2,0}^{gt}||_2$

$$= ||(\nabla J_{2,1} + \nabla J_{1,0}) - (\nabla J_{2,1}^{gt} + \nabla J_{1,0}^{gt})\,||_2$$

$$= ||\underbrace{(\nabla J_{2,1} - \nabla J_{2,1}^{gt})}_{\text{Error in shin}} + \underbrace{(\nabla J_{1,0} - \nabla J_{1,0}^{gt})}_{\text{Error in thigh}}\,||_2$$

- Errors in bones propagate to joints along the kinematic tree

- Large errors for joints at the far end

Ground truth joint

Ground truth bone

Estimated bone

Motivation

- Besides local bone loss only

- Long-range losses should also be considered and balanced over the intermediate bones.

Add Joint Loss to Bone Outputs

- Bone output:

$$\Delta J_{k,par(k)} = J_k - J_{par(k)}$$

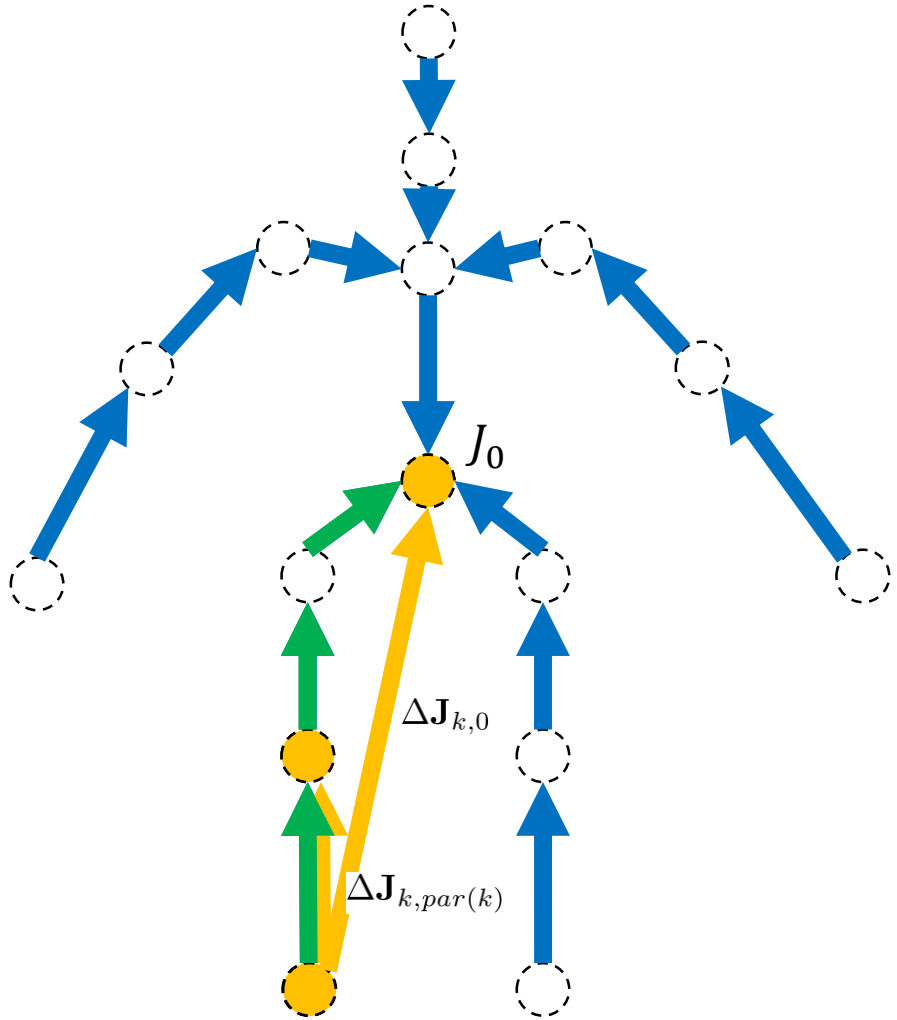
- Bone loss:

$$||\Delta J_{k,par(k)} - \Delta J^{gt}_{k,par(k)}||_2$$

- Add joint loss to bone output:

$$||\Delta J_{k,0} - \Delta J^{gt}_{k,0}||_2$$

- Where, $\Delta J_{k,0}$ is a summation of the bones along the kinematic tree path.



Generalize to Any Joint Pair Loss

- Bone output:

$$\Delta J_{k,par(k)} = J_k - J_{par(k)}$$

- Bone loss:

$$||\Delta J_{k,par(k)} - \Delta J^{gt}_{k,par(k)}||_2$$

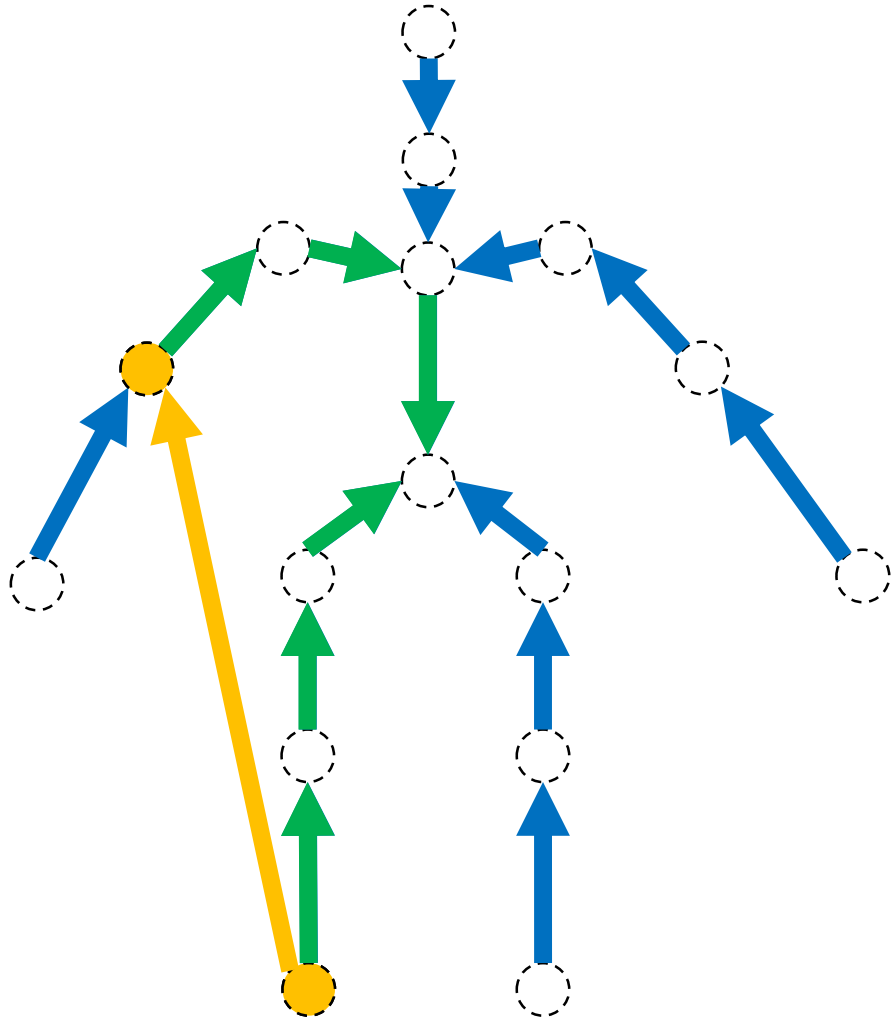
- Add joint loss to bone output:

$$||\Delta J_{k,0} - \Delta J^{gt}_{k,0}||_2$$

- Generalize to any joint pair loss:

$$||\Delta J_{u,v} - \Delta J^{gt}_{u,v}||_2$$

- Where, $\Delta J_{u,v}$ is a summation of the bones along the kinematic tree path.



Compositional Loss Function

Regression output: bones

Ground truth relative position

$$L(\mathcal{B}, \mathcal{P}) = \sum_{(u,v)\in\mathcal{P}} ||\Delta\mathbf{J}_{u,v} - \Delta\mathbf{J}^{gt}_{u,v}||_2^2$$

Joint pair set

Relative position of a joint pair, a summation of the bones along the kinematic tree.

The long-range joint pair losses are considered and balanced over the intermediate bones!

The ground truth is sufficiently exploited!

Comparison Experiments

- Network: 50-layer ResNet
- Dataset:
  - 3D benchmark: Human3.6M
  - 2D benchmark: MPII
- Methods:

Notation	Outputs	Loss
State-of-the-art	-	-
Our Baseline	Joints	Joint position loss
Ours (bone)	Bones	Bone position loss
Ours (all)	Bones	All joint pair position loss

3D Human Pose Results

- A **strong baseline**, already state-of-the-art.
- **Bone representation** is superior to joint.
- **Compositional loss function** is effective.

The lower, the better

Metric	State of the art	Baseline	Ours (bone)	Ours (all)
Joint Error (mm)	78.7 [1]	75.0	75.0 (0.0%)	67.5 (10.0%)
Bone Error (mm)	-	65.5	62.3 (4.9%)	58.4 (10.8%)
Bone Std (mm)	-	26.4	21.9 (17.0%)	21.7 (17.8%)
Illegal Angle (%)	-	3.7%	3.3% (10.8%)	2.5% (32.4%)

- [1] Zhou et al., Deep kinematic pose regression, ECCV 2016.

Apply to 2D Task (Regression Based)

Two stage error feedback

Stage	Metric	State of the art	Baseline	Ours(all)
1	Joint Error (mm)	-	29.7	27.2 (8.4%)
	Bone Error (mm)	-	24.8	22.5 (9.3%)
	PCK (%)	-	76.5%	79.6% (4.1%)
2	Joint Error (mm)	-	25.0	22.8 (8.8%)
	Bone Error (mm)	-	21.2	19.5 (8.0%)
	PCK (%)	81.3% [2]	82.9%	86.4% (4.2%)

Complementary to “multi-stage error feedback”:

a. A two-stage error feedback baseline.
b. Stage1: direct joint regression.
c. Stage2: use joint prediction from stage1.
d. Our method improves both stages.

[2] Carreira et al., Human pose estimation with iterative error feedback, CVPR 2016.

Unified 2D and 3D Pose Regression

- Our method general for 3D and 2D task.
- Easily mixed 3D and 2D data training:
- Decompose the loss into xy part and z part.
  - xy part is always valid for both 3D and 2D samples.
  - z part is only computed for 3D samples and set to 0 for 2D samples.
- Significantly improve 3D pose performance
  - Joint Error 67.5->48.3, 28.4%.
- Plausible and convincing 3D pose on in-the-wild image.

Qualitative Result

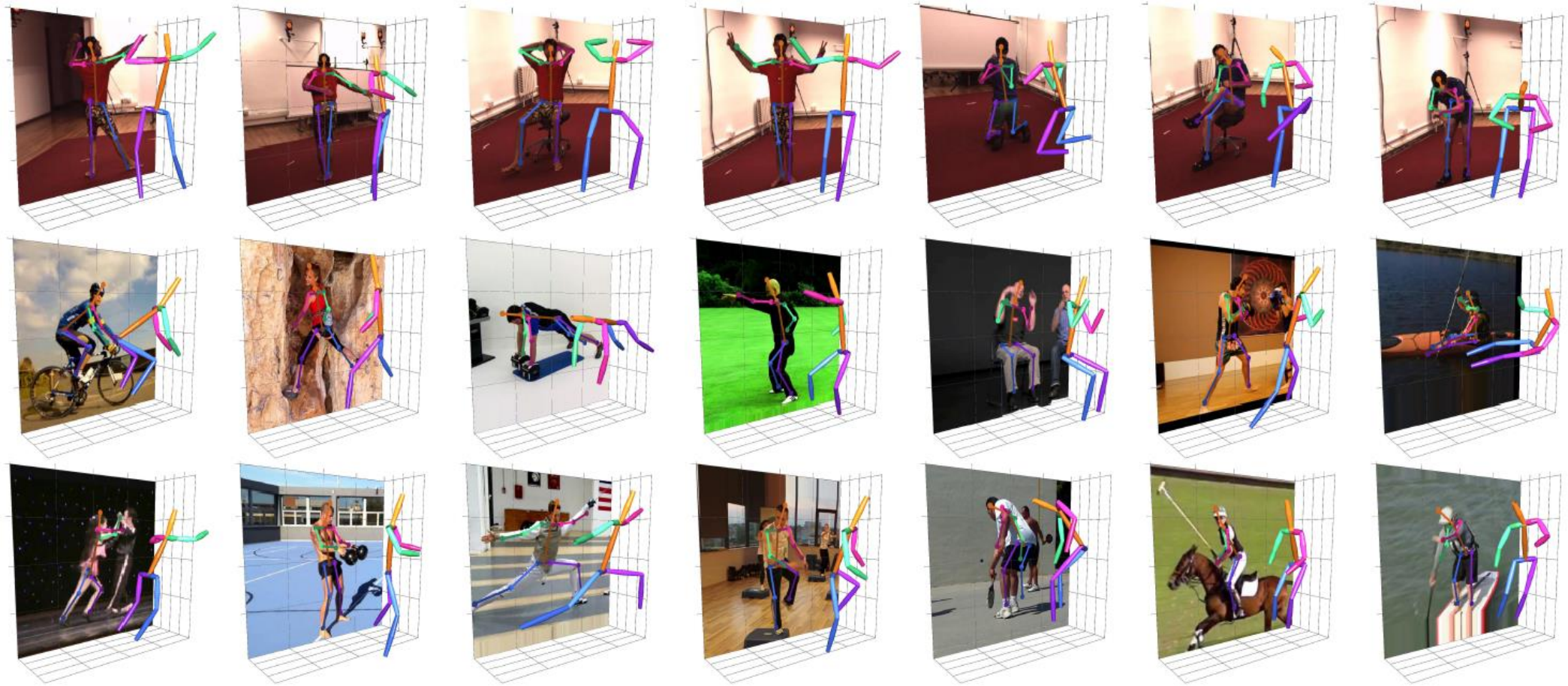


Figure 3. Examples of 3D pose estimation for Human3.6M (top row) and MPII (middle and bottom rows), using *Ours (all)* method in Table 5, trained with both 3D and 2D data. Note that the 3D poses on in the wild MPII images are quite plausible and convincing.

Video Result

Thanks!