PRINCETON UNIVERSITY

ICCV23 PARIS

Efficient, Self-Supervised Human Pose Estimation with Inductive Prior Tuning

Nobline Yoo, Olga Russakovsky
Princeton University

Published in ICCVW'23

Introduction

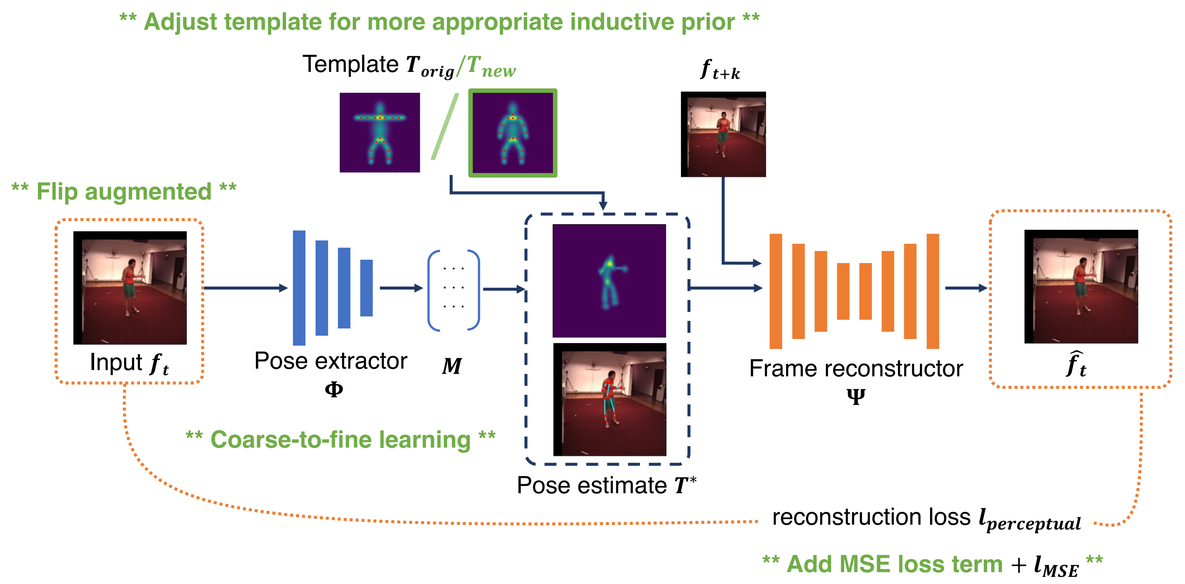
- Many SOTA methods for human pose estimation require labor-intensive labeling.
- Self-supervised models can learn pose from unlabeled data by optimizing for reconstruction, but there is no guarantee that predictions will reflect ground truth.
- Methods to quantify limb length consistency are lacking for unlabeled data settings.

We analyze the relationship between reconstruction optimization and pose estimation accuracy in the absence of ground truth. Our insight is that **inductive prior tuning can better align the reconstruction task with the pose estimation task**.

Building off Schmidtke et al. [1] as the baseline, we develop a model pipeline that **exceeds the baseline performance, using a training set that is less than one-third the size of the original**.

We propose a metric `BPLP-C` to quantify body part length consistency that is suitable for self-supervised settings in the absence of ground truth.

Methods



Baseline [1] model architecture. **Our adjustments in green.** Pose extractor Φ takes input frame f_t, containing the pose to estimate, and outputs transformation matrices $\mathrm{M} = \{M_{1 \le i \le 18}\}$. Template T_{orig} forms the inductive prior. M transforms T_{orig} to generate pose estimate T^*. Frame reconstructor Ψ takes T^* and frame f_{t+k} as input and reconstructs f_t. Baseline reconstruction loss is a perceptual VGG loss.

Our changes:

1) **New reconstruction loss** *to improve reconstruction quality*. $l_{recon} = l_{MSE} + ||VGG(\hat{f}_t) - VGG(f_t)||_1^1$
2) **New template $\mathbf{T_{new}}$** *to align reconstruction with pose estimation*. T_{new} (arms-down) better reflects the natural distribution of poses in the dataset, yielding a more appropriate prior.
3) **Coarse-to-fine learning** *to learn finer pose details*. Expand transformation matrix M; map select matrices at (1) whole-arm granularity, (2) individual-arm-component granularity.
4) **Flip dataset augmentation** *to overcome potential discrepancies in distribution.*
5) **Constrain for consistency.** `BPLP-C` assesses σ(body part length proportions). Formulate M as composition of rotation, localization of individual parts with scaling of whole body.

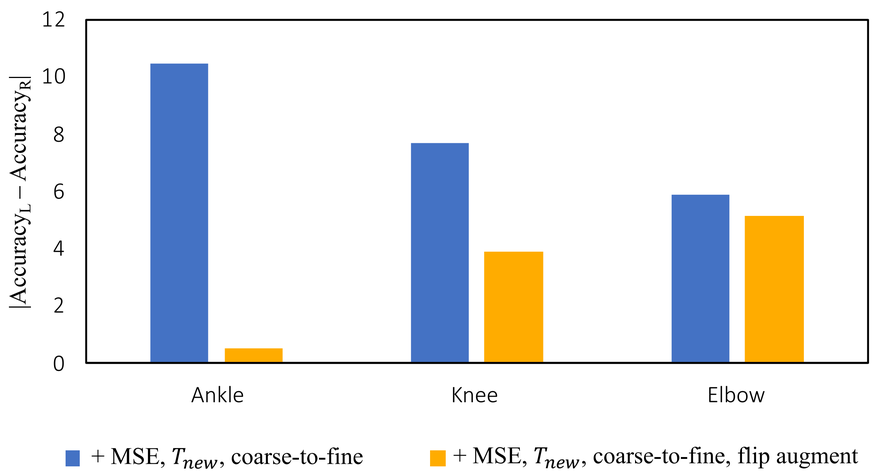
Setup: The baseline uses 600K pairs of frames (f_t, f_{t+k}). For computational efficiency, we keep our dataset to 181,383 pairs pre-augmentation and 181,728 post-augmentation.

Dataset: Human3.6M [2]

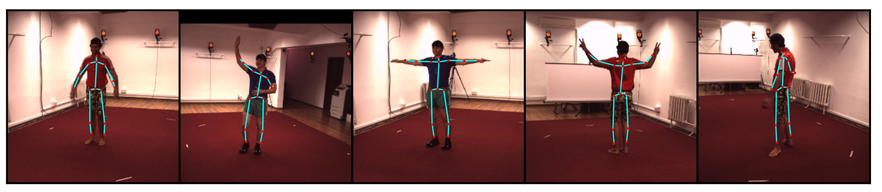
Results

Model	PDJ	L_2 Error
Published checkpoint	40.8	11.0
Baseline	38.5	7.2
+MSE	26.6	9.2
+T_{new}	33.1	7.5
+MSE, T_{new}	37.2	6.7
+MSE, T_{new}, flip augment	34.3	6.9
+MSE, T_{new}, coarse-to-fine	39.0	7.0
+MSE, T_{new}, coarse-to-fine, flip augment	**42.6**	**6.4**
+MSE, T_{new}, coarse-to-fine, flip augment, constrained M	38.7	7.0

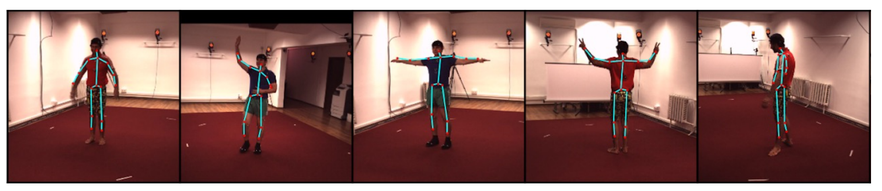
Our +MSE, T_{new}, coarse-to-fine, flip augment model outperforms the baseline.



Discrepancy in keypoint prediction accuracy between left/right side. After flip augmentation, there is noticeably less discrepancy in the ankle, knee, and elbow (the most extreme cases).



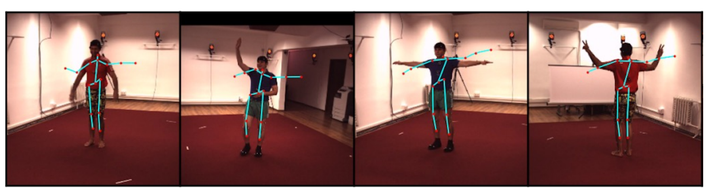
(a) Schmidtke et al.



(b) Ours

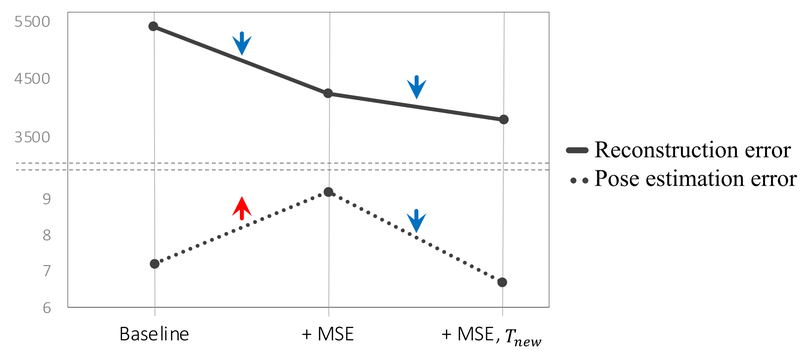
Our predictions more closely follow the outline of the subjects.

Key result #1: Efficient model pipeline that exceeds baseline performance with three times fewer data.



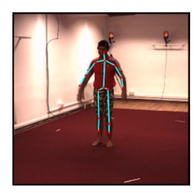
+MSE predictions

The +MSE model consistently predicts an arms-out pose, resembling the arms-out pose in T_{orig}. There is a distribution mismatch between the arms-out template prior (T_{orig}) and test dataset, which has ~40:1 arms-down to arms-out poses.

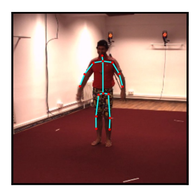


Adding the MSE loss term improves reconstruction but worsens pose estimation accuracy. Redesigning the template to an arms-down pose reduces both reconstruction and pose estimate error, resulting in better alignment.

Key result #2: Providing an appropriate inductive prior is key to aligning reconstruction with pose estimation.



(a) +MSE, T_{orig}, coarse-to-fine, flip augment



(b) +MSE, T_{orig}, coarse-to-fine, flip augment, constrained M

Constraining M yields higher `BPLP-C` and, as pictured, results in left limbs (e.g. left thigh) being relatively more consistent in length with right limbs (e.g. right thigh).

Key result #3: We propose new metric `BPLP-C` and show preliminary evidence of alignment between higher metric and more consistency.

References

[1] Luca Schmidtke, Athanasios Vlontzos, Simon Ellershaw, Anna Lukens, Tomoki Arichi, and Bernhard Kainz. Unsupervised human pose estimation through transforming shape templates. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2484–2494, 2021.
[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

Acknowledgements

This work was done as part of NY's undergraduate senior thesis. We are grateful to Princeton Research Computing for compute resources and to the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (to OR), Johns Hopkins University Applied Physics Laboratory and ICCV WiCV workshop (to NY) for enabling the in-person presentation of this research.