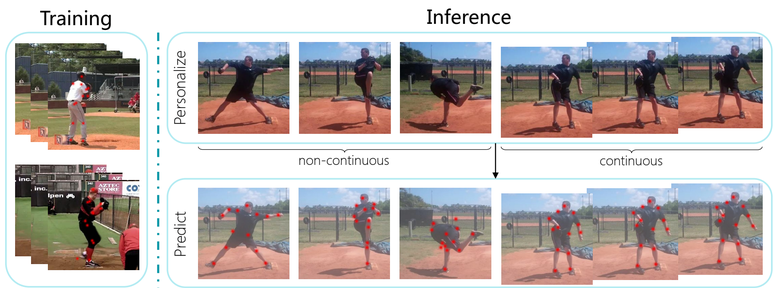
# Test-Time Personalization with a Transformer for Human Pose Estimation

Miao Hao[1]*, Yizhuo Li[2]*, Zonglin Di[1]*, Nitesh Gundavarapu[1], Xiaolong Wang[1]

[1]UC San Diego, [2]Shanghai Jiao Tong University

## Introduction




We propose to personalize a human pose estimator given a set of test images of a person without using any manual annotations. To help the model generalize to different unknown environments and unseen persons. Instead of using a fixed model for every test case, we adapt our pose estimator during test time to exploit person-specific information.

## Method

Our model is firstly trained with diverse data on both a supervised and a self-supervised keypoint estimation task, using a Transformer modeling their relation. During inference, the model conducts Test-Time Personalization which only requires the self-supervised keypoint estimation task.

**I. Joint Training for Pose Estimation with a Transformer**

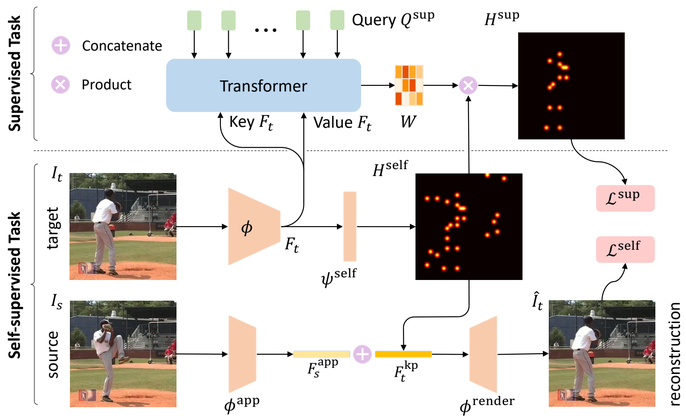
*a. Self-supervised Keypoint Estimation:* For the self-supervised task, we use an image reconstruction task to perform disentanglement of human structure and appearance, which leads to self-supervised keypoints as intermediate results.

*b. Supervised Keypoint Estimation with a Transformer:* For the supervised part, we use a transformer to compute an affinity matrix $W$, which models the relation between the self-supervised and supervised keypoints. The supervised heatmap is acquired by $H^{\text{sup}} = H^{\text{self}} W^T$.

**II. Test-Time Personalization**

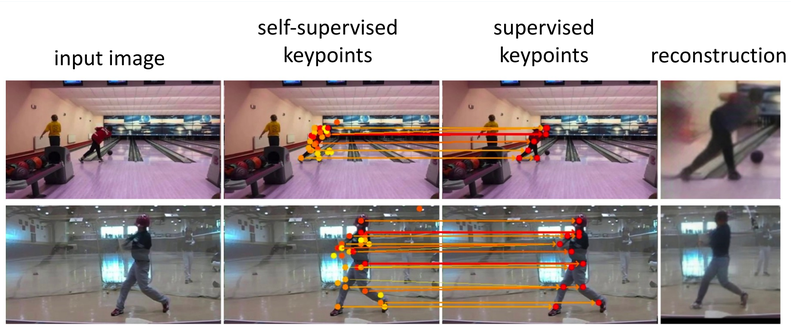
During inference, we fine-tune the model relying solely on the self-supervised task. The supervised output will be improved thanks to the fine-tuned self-supervised keypoints. We also propose two scenarios: (i) *online* scenario, which takes input as a sequence and update the model at each incoming image. (ii) *offline* scenario, where we have access to the whole person domain and has no requirement in real-time inference.

## Method




Architecture of our proposed model. We use a reconstruction task to generate self-supervised keypoints. We further predict the supervised keypoints from them utilizing an affinity matrix $W$ computed by the Transformer.

## Visualization

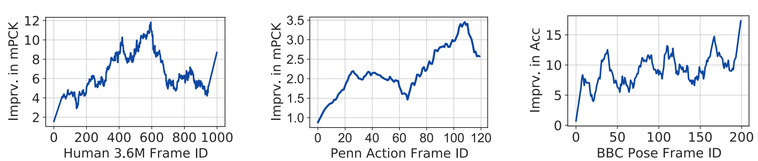



Visualization of the proposed method on Penn Action validation set. We use the lines to indicate the weight in the affinity matrix, which indicates the correspondence between the two sets of keypoints. Warmer colors indicate larger weight. Weight less than 0.1 are not shown.

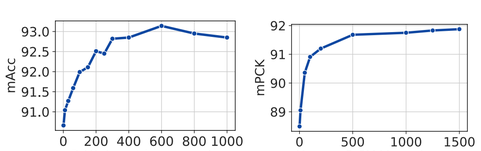
## Experiments

| Method | TTP Scenario | Human 3.6M | Penn Action | BBC Pose |
|---|---|---|---|---|
| Baseline | w/o TTP | 85.42 | 85.23 | 88.69 |
| Feat. shared (*rotation*) | w/o TTP | 87.37 (+1.95) | 84.90 (–0.33) | 89.07 (+0.38) |
| | Online | 88.01 (+2.59) | 85.86 (+0.63) | 89.34 (+0.65) |
| | Offline | 88.26 (+2.84) | 85.93 (+0.70) | 88.90 (+0.21) |
| Feat. shared (*keypoint*) | w/o TTP | 87.41 (+1.99) | 85.78 (+0.55) | 89.65 (+0.96) |
| | Online | 89.43 (+4.01) | 87.27 (+2.04) | 91.48 (+2.79) |
| | Offline | 89.05 (+3.63) | 88.12 (+2.89) | 91.65 (+2.96) |
| Transformer (*keypoint*) | w/o TTP | 87.90 (+2.48) | 86.16 (+0.93) | 90.19 (+1.50) |
| | Online | 91.70 (+6.28) | 87.75 (+2.52) | **92.51 (+3.82)** |
| | Offline | **92.05 (+6.63)** | **88.98 (+3.75)** | 92.21 (+3.52) |

① Results of pose estimation on three different datasets. *Feat. Shared (rotation)* is a counterpart using the naïve self-supervised task of rotation prediction. *Feat. Shared (keypoint)* is a counterpart using self-supervised keypoints, but does not use the Transformer design. *Transformer (keypoint)* is our proposed method. The proposed method shows significant improvement over baseline thanks to TTP and also surpasses other alternative methods.




② Improvement vs. Frame ID in *online* scenario. We plot the gap between the Test-Time Personalization and the baseline model for each frame step. We adopt the averaged metric across all test videos. In most cases, we observe TTP improves more over time.




③ TTP with different video length. Our method benefits from more unlabeled test samples.

| Iters | Penn Action | BBC Pose |
|---|---|---|
| 1 | 87.75 | 92.51 |
| 2 | **88.01** | **92.64** |
| 3 | 76.27 | 92.59 |
| 4 | 76.13 | 92.53 |

④ Our method does not need excessive update iters during TTP.