SUPPLEMENTARY MATERIAL
Personalizing Human Video Pose Estimation

James Charles
University of Leeds
j.charles@leeds.ac.uk

Tomas Pfister
University of Oxford
tp@robots.ox.ac.uk

Derek Magee
University of Leeds
d.r.magee@leeds.ac.uk

David Hogg
University of Leeds
d.c.hogg@leeds.ac.uk

Andrew Zisserman
University of Oxford
az@robots.ox.ac.uk

1. Graphs for all body joints

Performance graphs from main paper (with corresponding figure numbering) shown for all body joints.

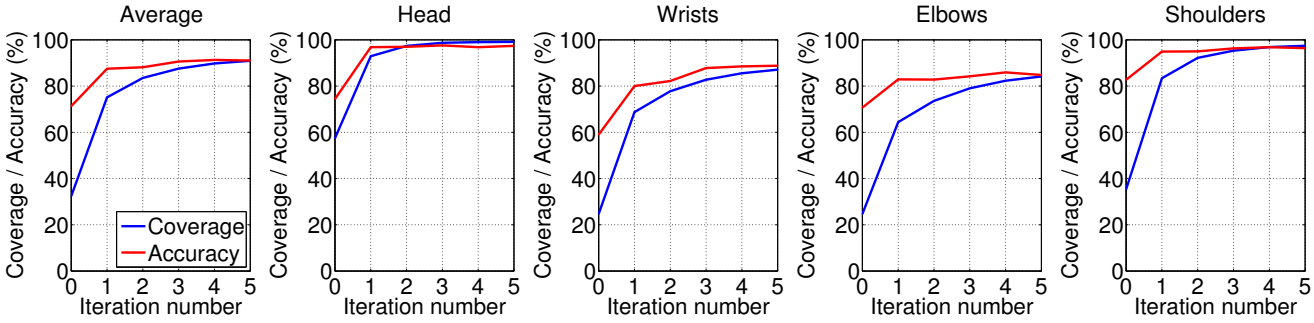

Figure 3. **Annotation accuracy and coverage when iterating on YouTube Pose Subset.** Accuracy of annotation and coverage (% of frames with annotation) across the video increases as the system iterates. Accuracy is measured as the percentage of estimated annotations falling within $d = 20$ pixels from ground truth (approx wrist width is 15 pixels). Results are averaged over videos with ground truth from the YouTube Pose Subset dataset.

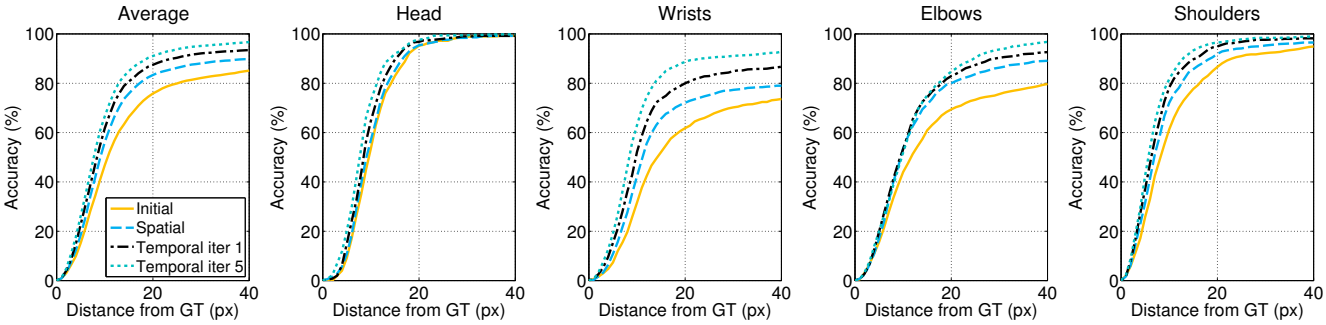

Figure 7. **Component evaluation on YouTube Pose Subset.** The graphs show the improvement from each stage of our algorithm. Notice how each stage leads to a very significant increase in accuracy. Accuracy is shown (averaged over left and right body parts) as the allowed distance from manual ground truth is increased.

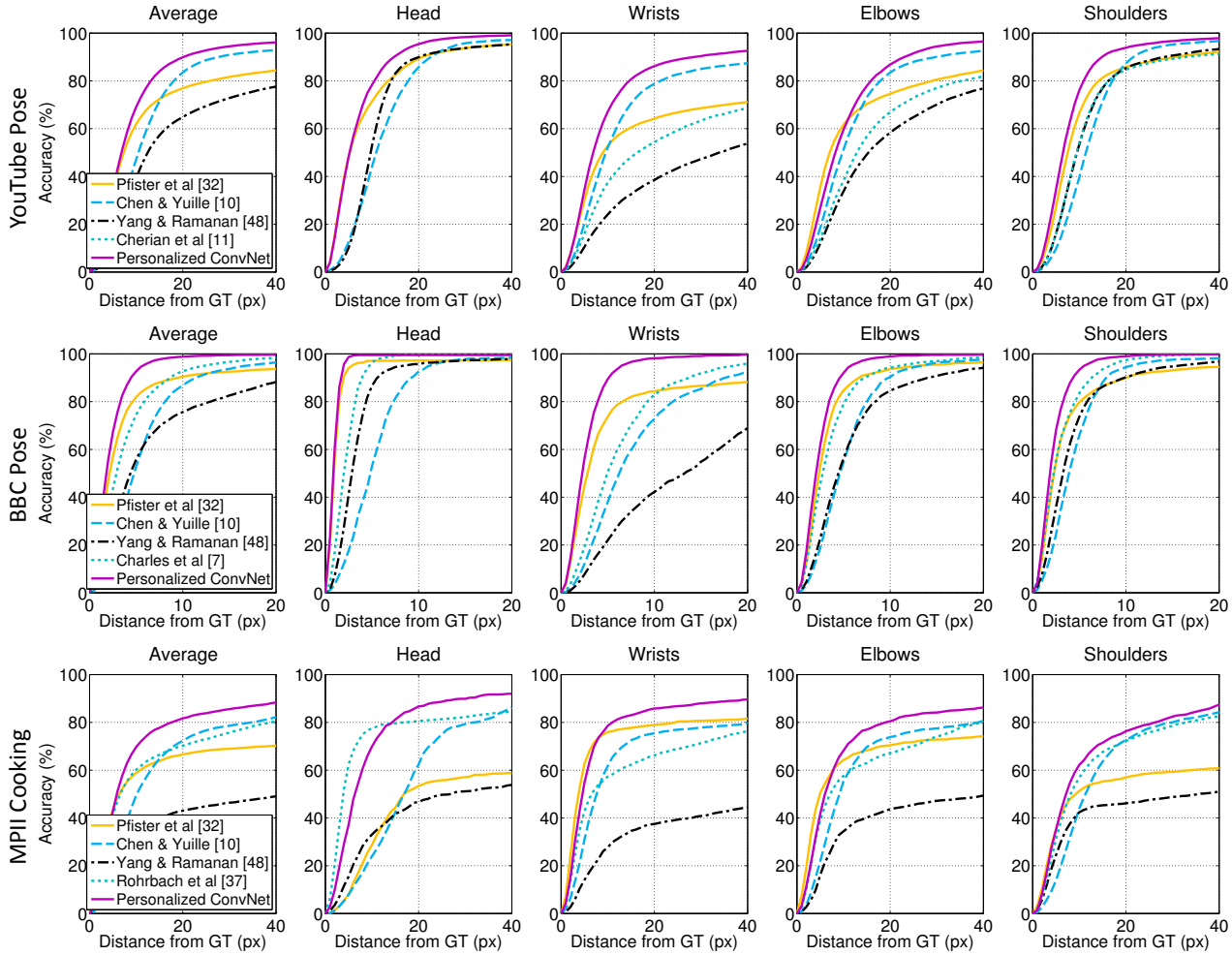

Figure 9. **Comparison to the state of the art.** Accuracy of pose estimation evaluated on three datasets. Accuracy is averaged over left and right body parts and shown as allowed distance from manual ground truth d is increased. Please refer to main paper for method references. Note: for the MPII Cooking performance graphs we compare methods trained/initialized with the FLIC dataset apart from Robrbach et al. [37] which is trained with MPII Cooking training material.

2. Sub-component evaluations

Graphs showing personalization sub-component evaluations.

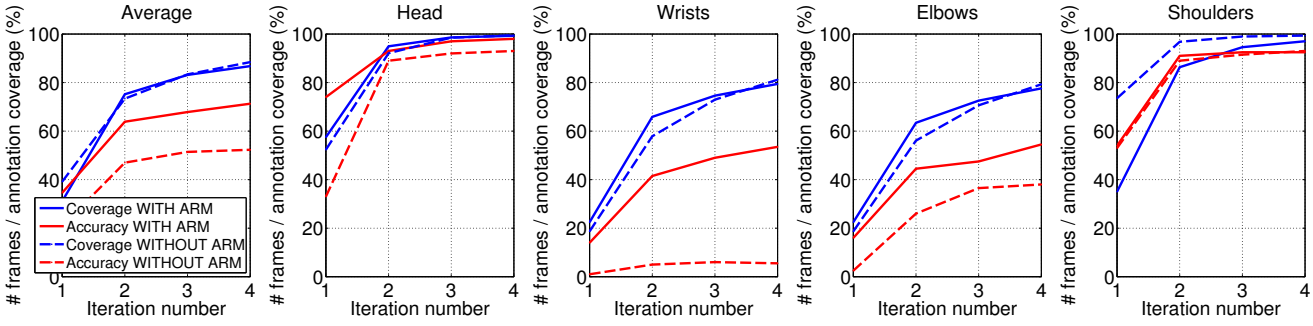

Figure S1. **Coverage and accuracy under different initialization methods on YouTube Pose Subset.** The graphs show the improvement from each stage of our algorithm under two different initialization methods. The first is initialized using the ConvNet [32] and the separate arm detectors (WITH ARM), the second uses only the ConvNet to initialize (WITHOUT ARM). Accuracy curves are produced by training a random forest body part detector (as described in the main paper) from current annotations, and evaluating it on all ground truth frames from YouTube Pose Subset.

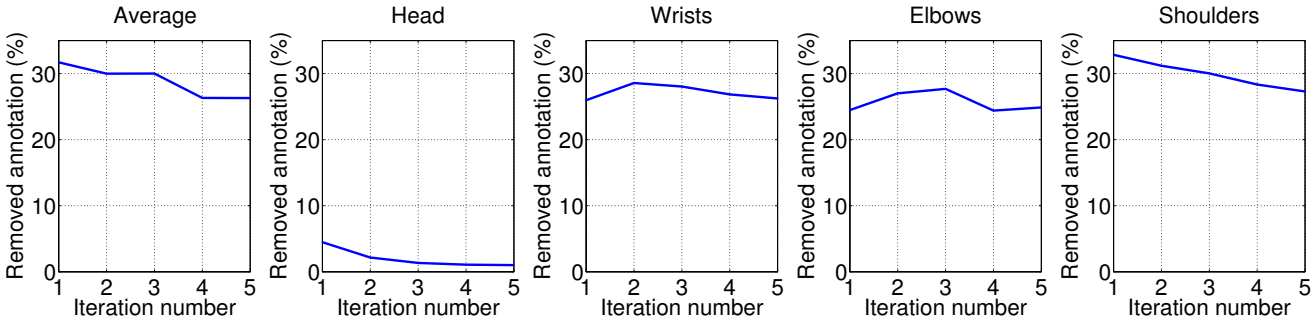

Figure S2. **Puppet evaluator response on YouTube Pose Subset.** For each iteration of our system, the number of per body joint annotations removed by our puppet evaluator are counted. These are expressed as a percentage of the total per body joint annotations prior to applying the puppet evaluator, but after removing some annotations with our annotation agreement measure. The puppet evaluator is shown to remove additional annotations which pass the agreement measure. The graphs demonstrate, on average, a reduction in removed annotation as our system iterates.

3. Boosting a generic ConvNet on FLIC

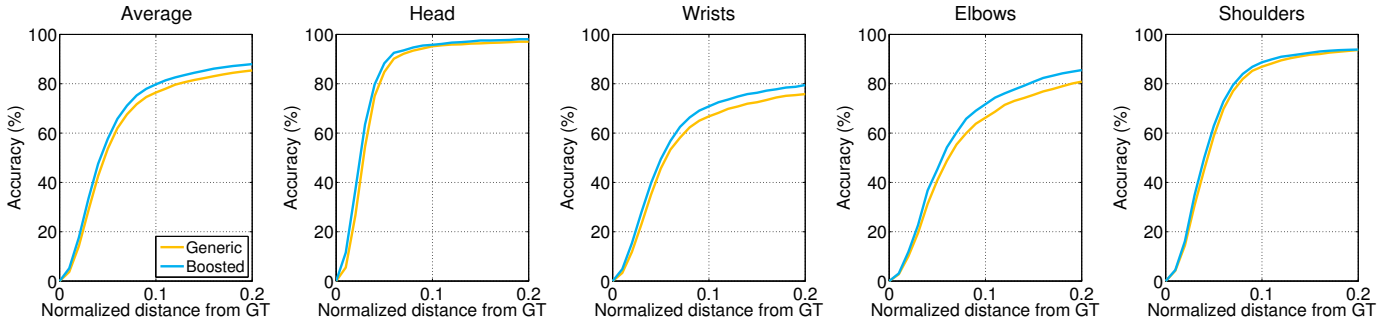

Figure S3. **Improvement when training with automatically annotated videos.** Performance of a generic ConvNet pose estimator [32] trained on the FLIC training set (Generic) is compared against a boosted version produced by fine-tuning with 20 additional automatically annotated YouTube Pose videos (Boosted). Comparison is performed on the FLIC test set. An improvement in performance is observed for all body joints, particularly for the elbows.

4. Experimental details

ConvNet. We use the publicly available ConvNet of Pfister *et al*. [3] both for initialization and for fine-tuning. The available model is pre-trained on the FLIC dataset [5] using the 3987 training video frames. Empirically, for our initialization, we found using body joint estimates with 80% confidence or above produce very good precision.

Arm model training. The second initialization method is for arm pose estimates using the generic pose estimator of Yang and Ramanan [6]. 15 arm pose models are trained on the MPII dataset [2] (by clustering all arm poses into 15 clusters using k-means, and retaining the nearest 150 poses to the cluster centroid for training), making it possible to detect up to 225 different poses. Note, there is no overlap between the MPII dataset [2] used for training and the MPII cooking dataset [4] used for testing. Each model is trained to have high precision (at least 90% detection accuracy) by setting their confidence threshold so as not to fire on pose clusters that they weren't trained on. We use the LSP extended dataset [1] to learn these thresholds.

Parameters. After temporal propagation, annotations are only retained if temporal agreement of overlapping annotation is below 20 pixels *and* overlapping annotation stems from at least three different frames. All videos are scaled to contain a person with width between the shoulders of approximately 100 pixels.

Joint offsets. There exists consistent body joint offsets between manual ground truth annotations on FLIC and those on BBC Pose or MPII Cooking. Therefore, to ensure a fair comparison between all models, pose estimates from those models trained/initialized from FLIC are adjusted by these offsets.

Personalized ConvNet average accuracy on training annotation. Here we report average training error (using all automatically generated annotations as ground truth) of the personalized ConvNet on BBC Pose (98%), YouTube Pose Subset (91%) and MPII Cooking (90%). Interestingly, after training, we found the personalized ConvNet predictions have higher precision than the generated annotation.

5. YouTube Pose dataset pose tracking output

Example video frames and pose tracking output for the YouTube Pose dataset.

Figure S4. **YouTube Pose dataset and pose estimates.** Example frames from videos in the YouTube Pose dataset are shown in each row along with pose estimates (as stick figures) from the personalized ConvNet. Note the variety of poses, clothing, backgrounds and camera angles.

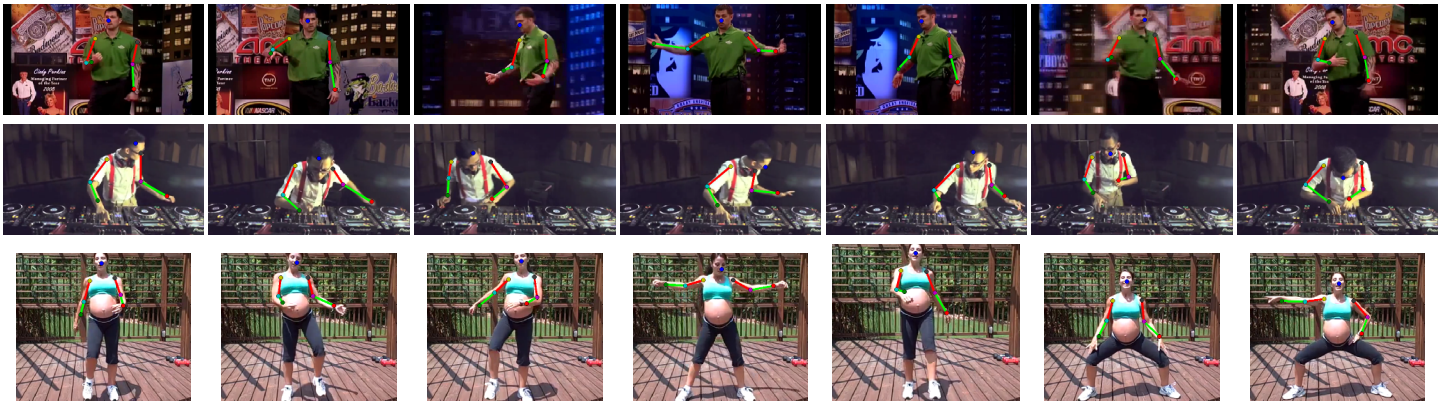

Figure S5. **More YouTube Pose dataset and pose estimates.** Example frames from videos in the YouTube Pose dataset are shown in each row along with pose estimates (as stick figures) from the personalized ConvNet.

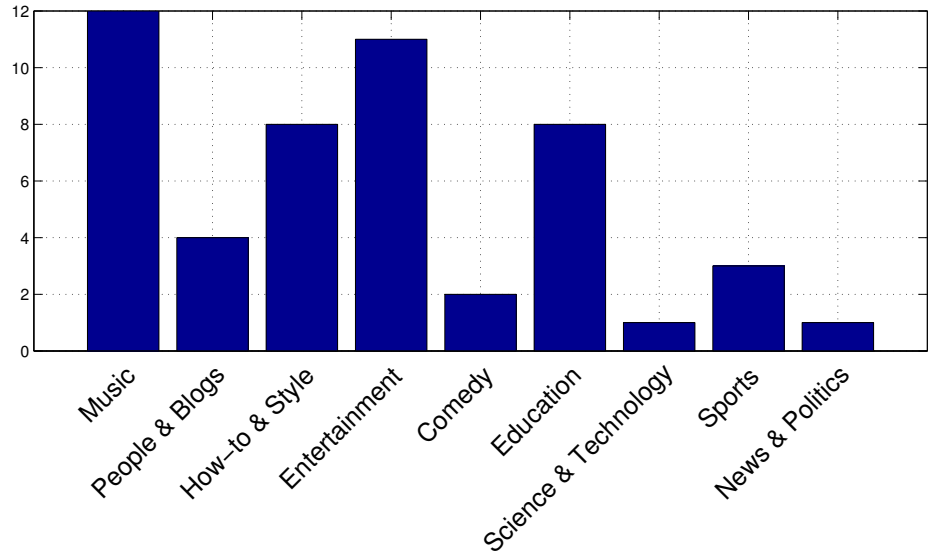

Figure S6. **YouTube Pose category distribution.** Distribution of video categories in the 50 video YouTube Pose dataset.

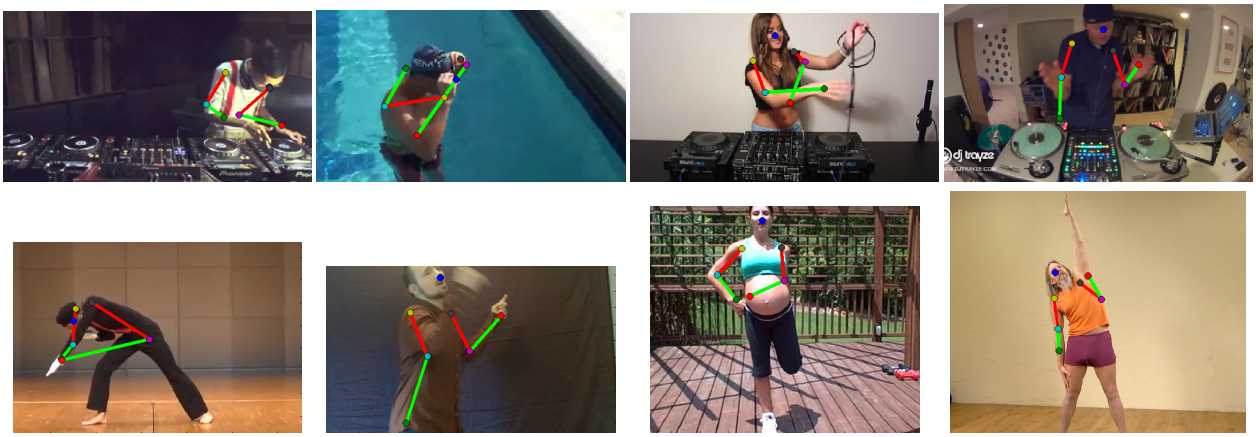

Figure S7. **Failure cases.** Example frames with erroneous pose estimates from personalized ConvNets. There are two main causes of failure: (i) heavy occlusion (including self-occlusion), and (ii) poses which our automated annotation system could not propagate, due to either optical flow error or very few initial annotations.

References

[1] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proc. CVPR*, 2011. 3

[2] A. Mykhaylo, P. Leonid, G. Peter, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proc. CVPR*, 2014. 3

[3] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *Proc. ICCV*, 2015. 3

[4] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *Proc. CVPR*, 2012. 3

[5] B. Sapp and B. Taskar. Multimodal decomposable models for human pose estimation. In *Proc. CVPR*, 2013. 3

[6] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proc. CVPR*, 2011. 3