Supplementary Material for "EAGLE-Eye: Extreme-pose Action Grader using detaiL bird's-Eye view"

Mahdiar Nekoui
University of Alberta
nekoui@ualberta.ca

Fidel Omar Tito Cruz
Universidad Nacional de Ingeniería
ftitoc@uni.pe

Li Cheng
University of Alberta
lcheng5@ualberta.ca

1. Exemplar Human Skeleton

We follow the same pose annotation format as *MPII* dataset[1] to collect the *G-ExPose*. As depicted in Fig.1, this structure considers 16 joints and 15 bones for the human body.

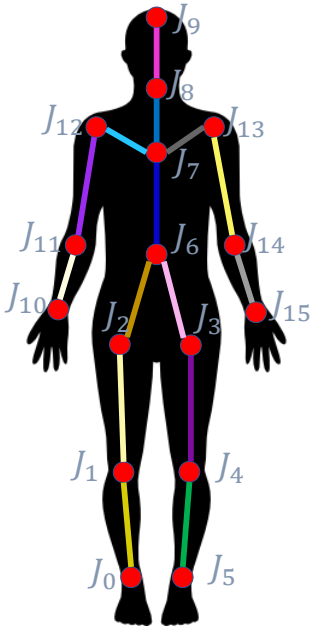


Figure 1: The human skeleton structure in *G-ExPose*

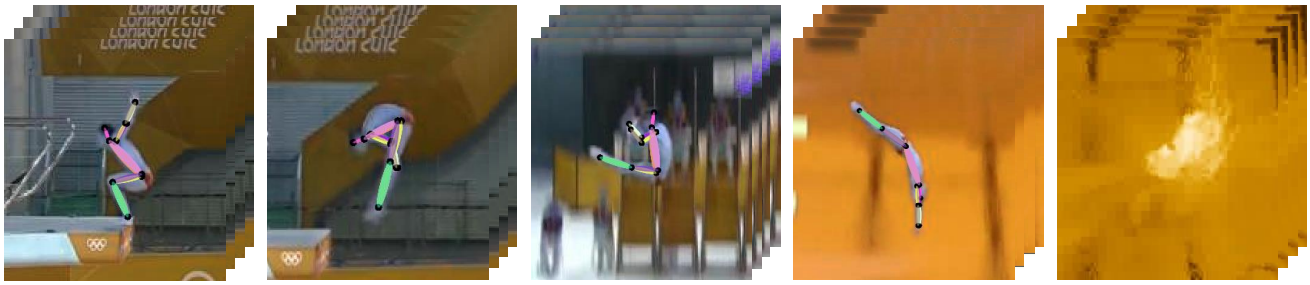
2. Further Implementation Details

Here we provide more details about the experimental settings and architecture of our model. As mentioned in the main manuscript, we make use of the I3D model[3] as the appearance features extractor backbone for the short-term action assessment. Let's consider the input video to the I3D model as a $T \times H \times W \times 3$ matrix. The `mixed-5c` layer of the I3D network outputs a $\frac{T}{8} \times 7 \times 7 \times 1024$ to be fed into the ADA stream. To make the both ADA and JCA streams output features with the same timesteps, the JCA stream is followed by a temporal max pooling with the stride and kernel size of 8. Besides, an average pooling at the end of JCA stream reduces the spatial size of the pose features output to have the same spatial dimension as the ADA stream output. To stabilize the learning process and capture the complex structure of the data, a BatchNorm-ReLU layer is embedded between two successive ADA (JCA) blocks. For assessing long-term actions we utilize the `fc6` layer of C3D network[6] to get the appearance features of the performance. Since the spatial size of the extracted features is 1, the spatial attention block is removed.

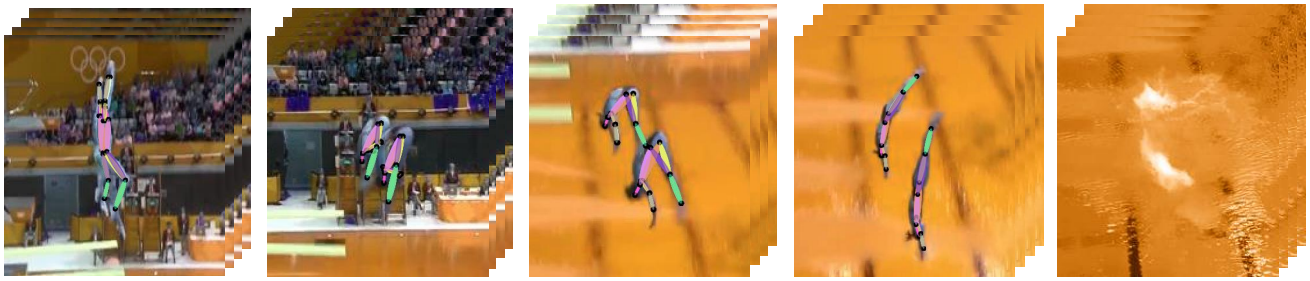
During the training, the backbone networks are frozen. We first train the EAGLE-Eye network on diving samples of AQA-7 dataset. For assessing other short-term sports except skiing, we fine-tune the diving pretrained model on each of the sports separately. The skiing assessor network jump-starts from the model that is fine-tuned on snowboarding samples. In order to assess the long-term figure-skating videos, we first pretrain the EAGLE-Eye on the task of classifying the sports from each other. To this end, we first repeat each short-term sports video (103 frames) to fit 5824 frames of figure-skating samples. We then train the network to classify each sport from the other. Finally, we use the resulted trained weights to fine-tune the EAGLE-Eye for assessing the figure-skating videos.

3. Qualitative Results

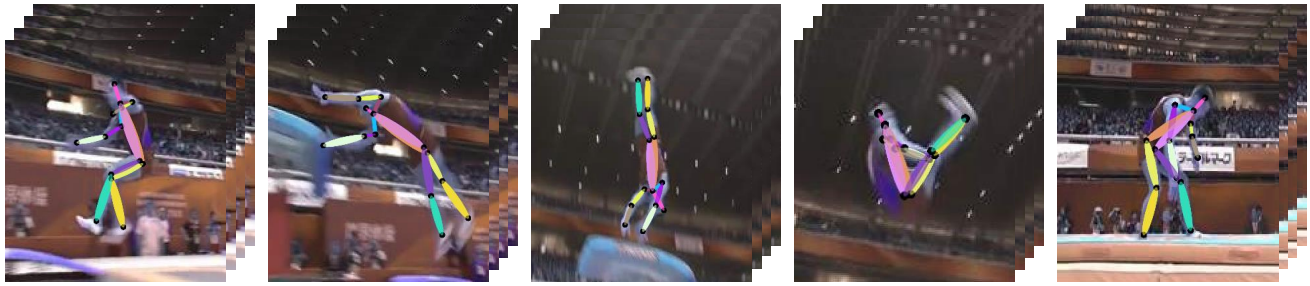
In the end, we present some qualitative results for assessing the quality of both short and long-term activities as well as their estimated pose sequence. It should be noted that in the short-term actions the pose estimator is trained on our *G-ExPose* while for figure-skating it is trained on the COCO+Foot dataset [2]. Fig.2 depicts the qualitative results for five different short-term sports; diving, synchronized diving (3m), gym vault, snowboarding, and skiing. The qualitative results of long-term figure-skating assessment are depicted in Fig.3.



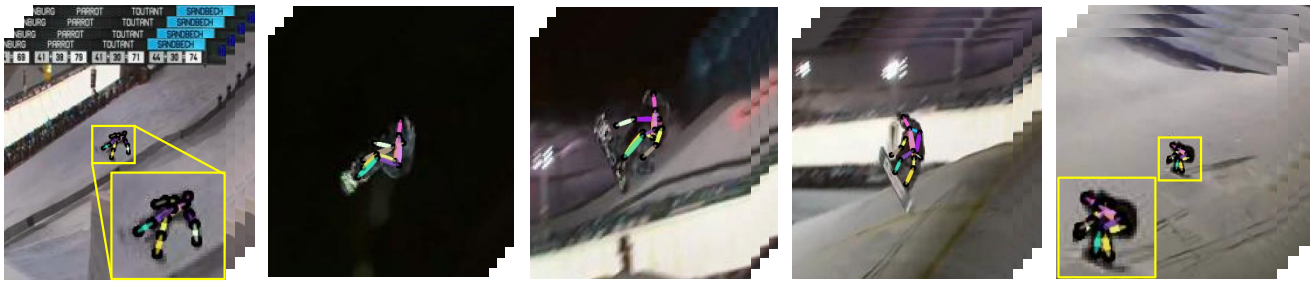
(a) Diving- Ground-truth score: 102.6, Predicted score: 96.91



(b) Sync.3 Diving- Ground-truth score: 69.75, Predicted score: 72.36



(c) Gymnastic Vault- Ground-truth score: 15.7, Predicted score: 15.81



(d) Big air snowboarding- Ground-truth score: 26, Predicted score: 23.19

(e) Big air skiing- Ground-truth score: 47, Predicted score: 44.20

Figure 2: Some qualitative results on short-term actions of AQA-7 dataset[4]

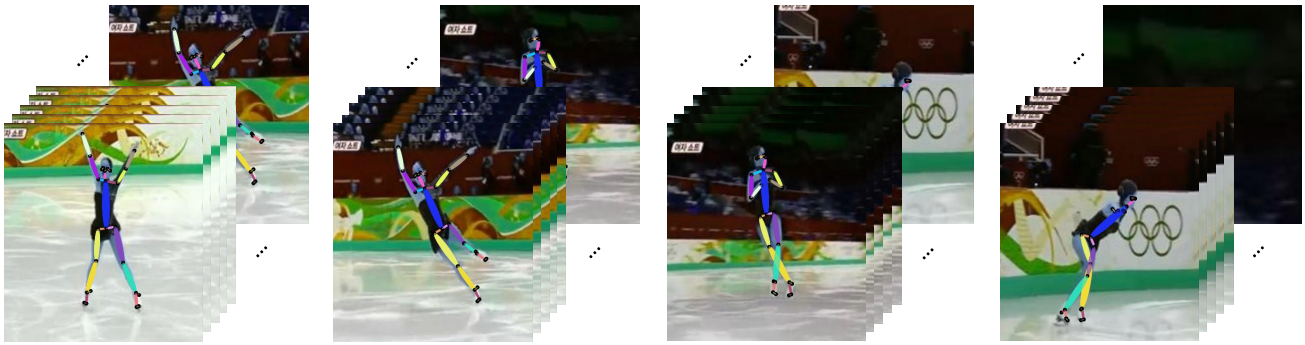


Figure 3: The qualitative results of long-term figure-skating sport (MIT-Skate dataset [5])- Ground-truth score: 49.74, Predicted score: 47.18

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1468–1476. IEEE, 2019.

[5] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 556–571, Cham, 2014. Springer International Publishing.

[6] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.