Supplementary Material for Widely Applicable Strong Baseline for Sports Ball Detection and Tracking

Shuhei Tarashima[1,2]
tarashima@acm.org
Muhammad Abdul Haq[2]
muhabdulhaq@gmail.com
Yushan Wang[2]
yushanwang218@gmail.com
Norio Tagawa[2]
tagawa@tmu.ac.jp

[1] Innovation Center
NTT Communications Corporation
Tokyo, Japan

[2] Faculty of System Design
Tokyo Metropolitan University
Tokyo, Japan

A Details of Existing SBDT Methods

As is mentioned in §4.2, we re-implemented 6 state-of-the-art (SOTA) sports ball detection and tracking (SBDT) algorithms in our codebase, 4 of which have been proposed in the recent literature [4, 5, 7, 10, 12] and the remaining 2 of which are their variants. We basically followed the default implementation settings proposed by authors, meanwhile we found that their performance can be boosted by simple modifications. In the following we describe the details of SOTA SBDT methods including modifications made by us.

DeepBall [4, 5]. This is a small convolutional neural network (CNN) that is originally proposed to detect a soccer ball. Unfortunately, its official implementation has not been publicly available. DeepBall takes a single frame to produce the heatmap representing ball position via aggregating multi-scale intermediate feature maps. At inference time, a ball position is determined by simply detecting a peak from the heatmap. Model training is performed by minimizing the pixel cross-entropy (CE) loss between model predictions and ground truth (GT) binary maps. The GT binary map is produced by setting a true ball position and its nearest neighbours as foreground. Adam optimizer [3] is used to train the model, and hard negative mining [8] is employed to mitigate the effect of foreground-background class imbalance. Notice that we directly followed the above settings for our re-implementation.

DeepBall-Large. Through the re-implementation of DeepBall, we found that the original model is too small (< 0.1M parameters) to be applied to other ball-game datasets (*cf.* Table 2 in our main body). To increase the model capacity, we made the following two modifications to the original DeepBall model: (1) The depths of block $\{1, 2, 3\}$ are increased from $\{8, 16, 32\}$ to $\{48, 96, 192\}$, (2) a kernel size of the stem is set to 3. Here we call the resulting variant of DeepBall as DeepBall-Large. Its model training is the same with the original.

BallSeg [12]. This is a variant of ICNet [14] originally proposed to detect a basketball. Its official implementation has not been publicly available. BallSeg takes two consecutive

frames by concatenating a frame of interest with its difference to another frame. The model is trained using the Stochastic Gradient Descent (SGD) applied on the pixel-wise CE loss. Since the specific ICNet architecture used to build BallSeg is not described in the original paper, we chose to adapt the smallest model provided in the official ICNet repository[1]. Also, we found that model training is failed when the proposed loss and optimizer are used. Instead, we employed the focal loss [6] and Adam optimizer [8] to successfully train BallSeg, then evaluated the performance of resulting models in our experiments (*cf.* §4 in our manuscript).

TrackNetV2 [10]. This is a UNet-based [9] SBDT model originally proposed to detect a shuttlecock from badminton videos. The authors proposed multiple-in multiple-out (MIMO) design to efficiently capture ball dynamics: Three consecutive frames are concatenated along the channel dimension, then the resulting tensor is fed into the model that generates corresponding three heatmaps. The model is trained using the Adadelta [13] optimizer applied on the the focal loss [6]. Though its official implementation has been public[2], unfortunately it is strongly tied up with the badminton dataset thus is difficult to adapt to other sports datasets. Therefore, we re-implemented TrackNetV2 following the above settings while being applicable it to various sports datasets.

ResTrackNetV2. We found that there is a public SBDT repository[3] that extends TrackNet [2] by introducing residual connections [1]. Based on this idea, we also added a residual connection to each encoder/decoder block in TrackNetV2 [10] to promote the model training. Also, we decreased the channel dimension of encoder/decoder blocks, which results in almost one-tenth model parameters compared to the original TrackNetV2. Here we call this variant as ResTrackNetV2. We trained this model with the same manner with TrackNetV2.

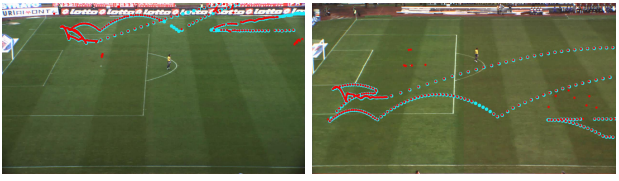
MonoTrack [7] is another variant of TrackNetV2 [10], which removes some convolution layers while adding skip connections. One notable difference from TrackNetV2 is that they adopt the combo loss [11] in model training. Since its official implementation has not been publicly available, we also re-implemented this method following settings described in [7].

B Qualitative Results and Error Analysis

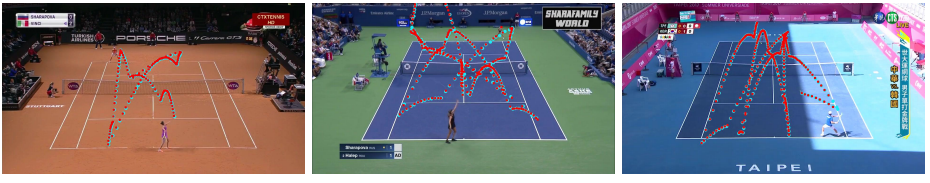
Figure 1 shows typical SBDT results of our proposed method, WASB (*cf.* §3 in our manuscript). These results demonstrates that WASB correctly track balls from video clips of different sports categories. Interestingly, we can see that sports balls can be tracked from video clips with very different viewpoints (*e.g.*, (d) Volleyball), and also from video clips including fast camera motion (*e.g.*, (e) Basketball).

Figure 2 shows some error modes of our proposed method. For example, the result (a) (*i.e.*, Soccer) represents a false positive, while the result (e) (*i.e.*, Basketball) shows a false negative. We can see that in (a) the model detection is not precisely aligned due to the noisy background (*e.g.*, player shoes), while in (e) a ball cannot be detected because it is blurry and ambiguous. The results (b), (c) and (d) (*i.e.*, Tennis, Badminton, Volleyball) also represent false positives. Interestingly, however, in these examples model detections (red circles) seem to capture true ball positions (light blue) more correctly than manually annotated ground truths. There results indicate a potential of WASB surpassing human ball localization performance.

[1] https://github.com/hszhao/ICNet
[2] https://nol.cs.nctu.edu.tw:234/open-source/TrackNetv2
[3] https://github.com/Chang-Chia-Chi/TrackNet-Badminton-Tracking-tensorflow2



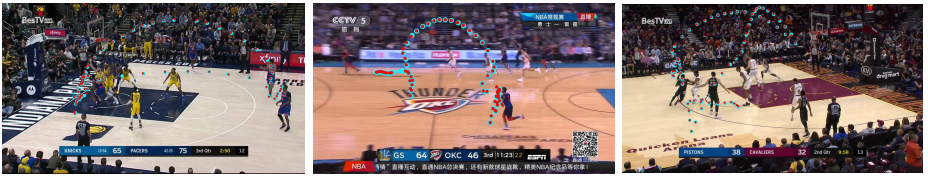
(a) Soccer



(b) Tennis

(c) Badminton

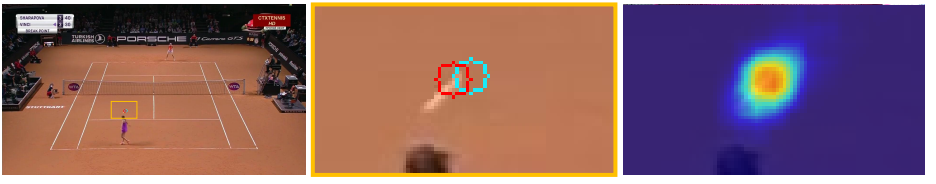
(d) Volleyball



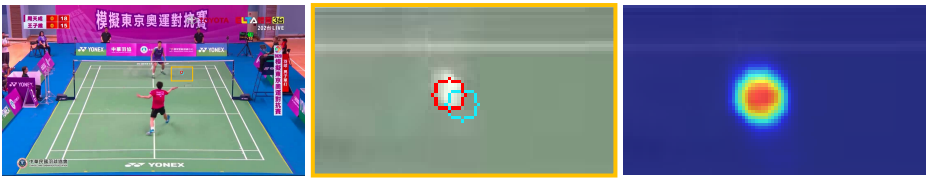
(e) Basketball

Figure 1: Exemplar qualitative results of our proposed method on each sports category in our dataset collection. A red circle represents a detection result while a light blue circle represents a ground truth ball position. The ball trajectory is overlaid on the first frame in each video clip. Best viewed in color.

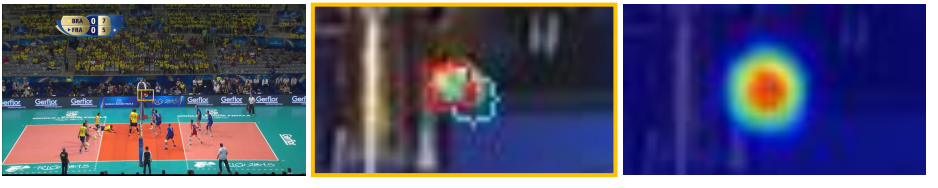
(a) Soccer



(b) Tennis



(c) Badminton



(d) Volleyball

(e) Basketball

Figure 2: Exemplar error modes of our proposed method. A red circle represents a detection result while a light blue circle represents a ground truth ball position. Results in the second column is the zoom of yellow rectangle areas in the first column, and the third column shows the corresponding heatmaps produced by our model. Best viewed in color.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Yu-Chuan Huang, I-No Liao, Ching-Hsuan Chen, Tsì-Uí İk, and Wen-Chih Peng. TrackNet: A Deep Learning Network for Tracking High-speed and Tiny Objects in Sports Applications. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2019.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] Jacek Komorowski, Grzegorz Kurzejamski, and Grzegorz Sarwas. BallTrack: Football Ball Tracking for Real-time CCTV Systems. In *2019 16th International Conference on Machine Vision Applications (MVA)*, 2019.

[5] Jacek Komorowski, Grzegorz Kurzejamski, and Grzegorz Sarwas. FootAndBall: Integrated Player and Ball Detector. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, 2020.

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[7] Paul Liu and Jui-Hsien Wang. MonoTrack: Shuttle Trajectory Reconstruction From Monocular Badminton Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022.

[8] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*, 2016.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[10] Nien-En Sun, Yu-Ching Lin, Shao-Ping Chuang, Tzu-Han Hsu, Dung-Ru Yu, Ho-Yi Chung, and Tsì-Uí İk. TrackNetV2: Efficient Shuttlecock Tracking Network. In *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, 2020.

[11] Saeid Asgari Taghanaki, Yefeng Zheng, S. Kevin Zhou, Bogdan Georgescu, Puneet Sharma, Daguang Xu, Dorin Comaniciu, and Ghassan Hamarneh. Combo Loss: Handling Input and Output Imbalance in Multi-organ Segmentation. *Computerized Medical Imaging and Graphics*, 2019.

[12] Gabriel Van Zandycke and Christophe De Vleeschouwer. Real-Time CNN-Based Segmentation Architecture for Ball Detection in a Single View Setup. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, 2019.

[13] Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method, 2012.

[14] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *Computer Vision – ECCV 2018*, 2018.