

Supplementary Material for “DeepSportLab: a Unified Framework for Ball Detection, Player Instance Segmentation and Pose Estimation in Team Sports Scenes”

Seyed Abolfazl Ghasemzadeh¹
seyed.ghasemzadeh@uclouvain.be

Gabriel Van Zandycke^{1,2}

Maxime Istasse^{1,2}

Niels Sayez¹

Amirafshar Moshtaghpour¹

Christophe De Vleeschouwer¹

¹ Université catholique de Louvain
ICTEAM/ELEN, Belgium
{firstname.lastname}@uclouvain.be

² SportRadar AG
{f.lastname}@sportradar.com

1 Quality measures

As mentioned in the main document, we tackle a multi-objective task, hence, requires multiple quality measures for expressing the performance of individual sub-tasks. The details of the quality measures used in our work is provided in the following.

Ball Detection Quality (bDQ). As in the BallSeg framework [5], given a threshold τ in the dynamic range of the confidence scores, a predicted ball keypoint is identified as a True Positive (TP) (or False Positive (FP)) detection, if its location lies inside (respectively, outside) the ground truth mask and its predicted confidence is greater than τ : for the predicted ball o^{ball} ,

$$\begin{aligned} o^{\text{ball}} \in \text{TP}^{\text{ball}}(\tau) &\Leftrightarrow o^{\text{ball}} \in \mathcal{I}^{\text{ball}*}, \tilde{S}(\text{ball}, o^{\text{ball}}) \geq \tau \\ o^{\text{ball}} \in \text{FP}^{\text{ball}}(\tau) &\Leftrightarrow o^{\text{ball}} \notin \mathcal{I}^{\text{ball}*}, \tilde{S}(\text{ball}, o^{\text{ball}}) \geq \tau \end{aligned} \quad (1)$$

where $\mathcal{I}^{\text{ball}*}$ denotes the semantic mask of the ball. By repeating this procedure for all images, we obtain the TP and FP sets associated with the full set of images. The TP rate (TPr) and FP rate (FPr) ratios are defined as

$$\text{TPr}(\tau) := \frac{|\text{TP}^{\text{ball}}(\tau)|}{|\{\text{images with annotated ball}\}|}, \quad \text{FPr}(\tau) := \frac{|\text{FP}^{\text{ball}}(\tau)|}{|\{\text{all images}\}|}.$$

The bDQ is then computed as the area under the ROC curve associated with the TPr and FPr.

Player Segmentation Quality (pSQ). Compared with the bDQ, a predicted player mask is identified as a TP mask (otherwise, an FP), if its IoU with one of the ground truth player masks is higher than the threshold of 0.5:

$$(\mathcal{I}_i^{\text{ply}}, \mathcal{I}_j^{\text{ply}*}) \in \text{TP}^{\text{ply}} \Leftrightarrow \exists j, \text{IoU}(\mathcal{I}_i^{\text{ply}}, \mathcal{I}_j^{\text{ply}*}) \geq 0.5. \quad (2)$$

The pSQ is then defined as the averaged IoU over the TP pairs:

$$\text{pSQ} := \frac{1}{|\text{TP}^{\text{ply}}|} \sum_{(u,v) \in \text{TP}^{\text{ply}}} \text{IoU}(u, v). \quad (3)$$

Player Detection Quality (pDQ). Having identified the TP set as in (2), the pDQ is then defined as the F_1 -score:

$$\text{pDQ} := \frac{2|\text{TP}^{\text{ply}}|}{N_{\text{ply}} + N_{\text{ply}}^*}. \quad (4)$$

Note that the pSQ (3) and pDQ (4) criteria are the segmentation and recognition quality components of the Panoptic Quality (PQ) measure introduced in [1], *i.e.*, concretely, the PQ for player segmentation reads $\text{PQ} := \text{pSQ} \cdot \text{pDQ}$.

Pose Estimation Quality (pEQ). For pose estimation task we use the OKS criteria, *i.e.*, for every pair of the predicted pose Υ_i and ground truth pose Υ_j^* , it is defined as

$$\text{OKS}(\Upsilon_i, \Upsilon_j^*) := \text{mean}_k \exp\left(-\frac{\|o_i^k - o_j^{k*}\|_2^2}{2s_j^2 \kappa_k^2}\right), \quad (5)$$

where the mean is taken over the annotated body part keypoints, s denotes the square root of the area of the bounding-box tightly containing all the body parts, and κ_k is the per-keypoint-type scale constant controlling falloff. The predicted skeletons are then sorted according to their confidence scores defined as the average over the body part confidence scores: from the pixel-wise confidence map in (Eq. 2 in the main document)

$$\Upsilon_i^{\text{conf}} = \frac{1}{17} \sum_{k \in \mathcal{K}_{\text{part}}} \tilde{S}(k; o_i^k). \quad (6)$$

Next, the ordered predictions are assigned to the ground truths, with which they have the highest OKS value. Once the matching is complete, the set of TP skeletons $\text{TP}^{\text{skl}}(\tau)$ with respect to the OKS threshold τ is determined. Concretely, for a fixed OKS threshold τ (ranging from 0.5 to 0.95), a pair of predicted and ground truth skeletons is identified as a TP, if their OKS is higher than τ . The Precision (Pr) and Recall (Re) values are then computed as

$$\text{Pr}(\tau) := \frac{|\text{TP}^{\text{skl}}(\tau)|}{N_{\text{ply}}}, \quad \text{Re}(\tau) := \frac{|\text{TP}^{\text{skl}}(\tau)|}{N_{\text{ply}}^*}. \quad (7)$$

Finally, the Average Precision (AP) and Average Recall (AR) values read, respectively,

$$\text{AP} := \text{mean}_{\tau} \text{Pr}(\tau), \quad \text{AR} := \text{mean}_{\tau} \text{Re}(\tau). \quad (8)$$

Remark 1 The quality metrics above are defined per-image; however, in practice, we compute the bdQ , pSQ (3), pDQ (4), and AP and AR (8) over all the images in the validation or test sets.

Remark 2 As required for computing pEQ , the values of κ_k associated with the body parts are set according to the convention of DeepSport dataset, i.e., $\kappa_{\text{head}} = 0.15$, $\kappa_{\text{hip}} = 0.2$, and $\kappa_{\text{foot 1}} = \kappa_{\text{foot 2}} = 0.2$.

2 Ablation studies

2.1 Decoding with Oracle Data

The importance of the accuracy of each output for the decoding process can be obtained by using the oracle data instead of their corresponding network outputs. This study is helpful to find out whether the error propagates from one block to the other. Table 1 compares the metrics when different permutations of oracle data were used on DeepSport dataset. The first message drawn from this study is that our player segmentation is good enough to associate the PIF keypoints because the increase of F_1 Score is only by 0.5% when using all of the oracle data compared to when using none. Thus, for further improve the pose estimation task, the PIF keypoints should be trained better.

player centroid	offset vectors	semantic masks	PQ	pSQ	pDQ	pEQ		
						AP	AR	F_1
–	–	–	72.3	80.3	90.1	87.5	82.1	42.4
✓	–	–	71.4	79.7	89.6	87.2	82.7	42.5
–	✓	–	75.7	82.9	91.3	87	81.8	42.2
–	–	✓	86.7	94.5	91.7	87.2	82.5	42.4
✓	✓	–	77	83	92.8	86.2	82.3	42.1
✓	–	✓	87	94.5	92.1	87.7	83.7	42.8
–	✓	✓	93.5	98.3	95.1	88	82.7	42.6
✓	✓	✓	100	100	100	87.7	83.8	42.9

Table 1: DeepSportLab Decoding with oracle data. Note that $PQ := pSQ \cdot pDQ$.

2.2 Keypoints Error Breakdown

Further improvements can be achieved once we know the source of error. Ronchini and Perona [9] break the estimated body keypoints in 5 different categories based on their calculated KS, i.e. keypoint similarity between the keypoint o of a detection Y and o^* of an annotation Y^* . KS is calculated using (5) without the mean over all keypoints. If KS of o and o^* is higher than 0.85, this prediction is considered Good. Jitter happens when KS drops between 0.5 and 0.85. In case that KS is less than 0.5, o can be either a Miss, Swap, or Inversion. In our case, since we switch the right and left feet in case of wrong detection, the Inversion will never occur. This is because foot 1 and foot 2 labels are interchangeable in DeepSport dataset (See Section 4 in the main document). Next, Swap happens when o is wrongly associated to another skeleton. Miss happens when o is predicted, but not in the right location, and it was not a Swap. Finally, FN KP happens when the keypoint is not detected at all. Fig. 1 shows the examples for each of these categories. Fig. 2 depicts the error breakdown based on the error category and type of keypoint.

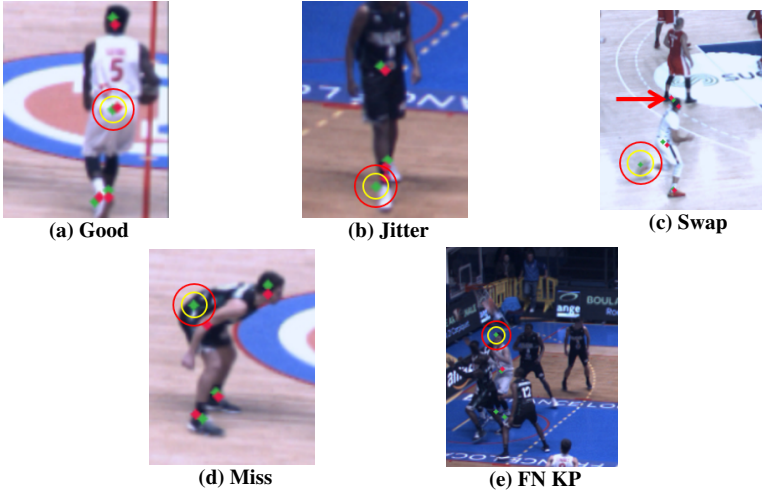


Figure 1: **Error samples.** Green and red dots show the annotated and predicted keypoints, respectively. Yellow and Red circles resemble the borders from which the KS will be less than 0.85 and 0.5, respectively. In (c), the red arrow points toward the wrongly predicted foot where the swap occurs.

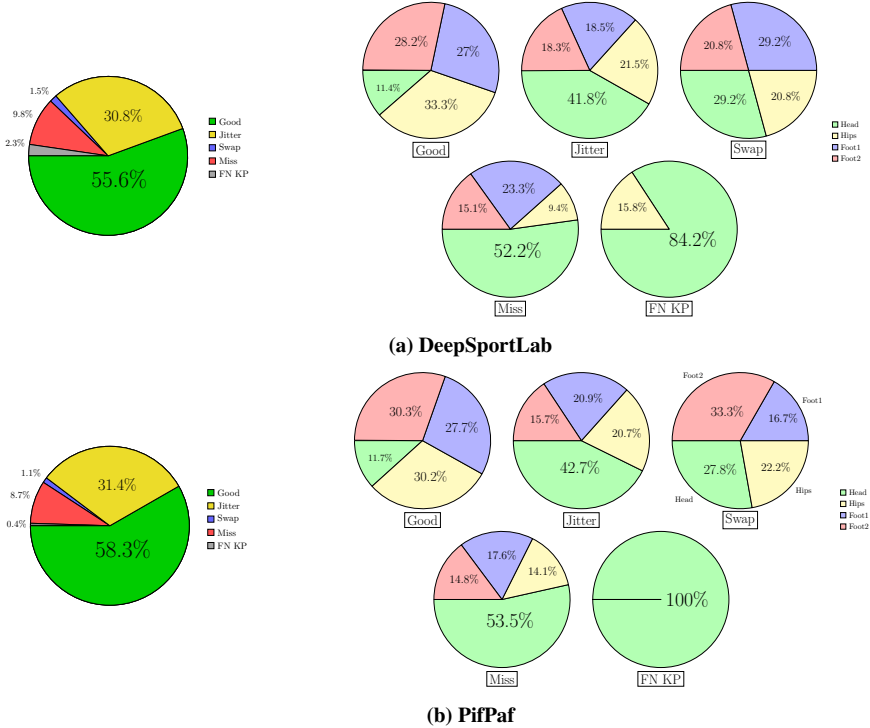


Figure 2: **Error breakdown.** Pie charts on the left show the distribution of keypoints in 5 categories based on their KS. Pie charts on the right show the distribution of each type of error based on the keypoints type.

2.3 Evaluation on COCO dataset

As stated in the main text, our main contribution is to come up with a multi-task framework specific to sports scenes. However, as an ablation study, the model was also evaluated on COCO dataset [9] which is very much diverse in terms of both the scenery and the size of people in images. Table 2 shows metrics evaluated on COCO’s validation set. Note that in this experiment, only the keypoints visible in the image are considered for the evaluation.

DeepSportLab decoder is studied in three different cases: (1) When using the network outputs, (2) When using oracle centroid of humans, and (3) When using the full human mask oracle. In the first case, due to the diversity of people size in COCO images, the segmentation task falls short in terms of pDQ which leads to error in pEQ. When adding the oracle center, the PQ increases significantly, suggesting that the center needs more training. When using the oracle masks, (*i.e.* PQ = 100), pEQ increases by 45.3% and 34.2% in terms of AP and AR, respectively. This shows the importance of the segmentation masks on big and challenging datasets such as COCO. It is worth mentioning that training in this case needs a lot of hyper-parameter tuning and optimization. Our computational resources certainly did not allow to fully explore the parameter space. Obtaining more competitive results on COCO dataset is seen as a future work for this framework.

Method	PQ	pSQ	pDQ	pEQ		
				AP	AR	F_1
DeepSportLab	34.3	75.3	45.5	43.7	44.2	22
DeepSportLab - oracle center	52.1	78.1	66.7	57.1	60.7	29.4
DeepSportLab - oracle segmentation	100	100	100	63.5	67.2	32.6
OpenPifPaf [9]	–	–	–	66.9	70.9	45.4
Pan.-DeepLab [9]	48.4	78.6	61.5	–	–	–

Table 2: **Comparison of different methods evaluated on COCO’s validation set.** Three different cases are considered for DeepSportLab: (1) Decoding with network’s outputs, (2) Decoding with oracle centers, and (3) Decoding with oracle segmentation masks.

References

- [1] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12485, 2020.
- [2] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9404–9413, 2019.
- [3] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *arXiv preprint arXiv:2103.02440*, March 2021.
- [4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [5] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [6] Gabriel Van Zandycke and Christophe De Vleeschouwer. Real-time cnn-based segmentation architecture for ball detection in a single view setup. In *Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, pages 51–58, 2019.