Self-supervised learning for human pose estimation in sports

Katja Ludwig, Sebastian Scherer, Moritz Einfalt, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Ludwig, Katja, Sebastian Scherer, Moritz Einfalt, and Rainer Lienhart. 2021. "Self-supervised learning for human pose estimation in sports." In *IEEE International Conference on Multimedia and Expo 2021 (ICME), 5-9 July 2021, Shenzhen, China*, 1–6. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICMEW53276.2021.9456000.

SELF-SUPERVISED LEARNING FOR HUMAN POSE ESTIMATION IN SPORTS

Katja Ludwig, Sebastian Scherer, Moritz Einfalt, Rainer Lienhart

University of Augsburg, {katja.ludwig, sebastian1.scherer, moritz.einfalt, rainer.lienhart}@uni-a.de

ABSTRACT

Human pose estimation (HPE) is a commonly used technique to determine derived parameters that are important to improve the performance of athletes in many sports disciplines. This paper proposes two methods to fine-tune a HPE system trained on general poses to a sports discipline specific HPE model using only a few labeled images. We show that 50 labeled 2D poses and additionally unlabeled videos are sufficient to achieve a Percentage of Correct Kexpoints (PCK) of 88.6% at a threshold of 0.1 in the disciplines of triple and long jump, closing the gap between the supervised fine-tuning on the same 50 images and the fully supervised training on 60× more images by 60%. The first proposed method uses pseudo labels as a self-supervised training technique together with a filtering method of the pseudo labels. Furthermore, this paper shows that a mean teacher approach, which is based on consistency between a teacher and a student model, can also improve the results.

***Index Terms*—** computer vision, sports, human pose estimation, self-supervised learning, pseudo labels

1. INTRODUCTION

In many sports disciplines, human pose estimation (HPE) is an important method for performance analysis and improvement of athletes. For example, the poses of soccer players are tracked and evaluated throughout a game [1]. Long and triple jump athletes use HPE for automated detection of landing, jump and stride events [2]. Or, ski jumpers use keypoint detection on their body and their skis to derive flight parameters such as ski angle, lower body angle and upper body angle to perfect their body posture during the flight phase and achieve long flight distances [3]. This paper uses the domain of triple and long jump as an example sports domain but our approach is also easily adaptable to any other discipline. An example for estimated poses during a triple jump is shown in Figure 1.

Annotating a large amount of 2D poses in many images by hand is enormously time consuming, especially when lots of keypoints are necessary. In many sports disciplines, coaches do not have enough time to select images from training or competition videos and annotate them for the purpose of measuring the athletes' performance. Hence, HPE systems based on convolutional neural networks generate a large benefit as they automatically detect the needed keypoints in a fraction of time. However, to train a deep neural network capable of achieving an acceptable performance on images from the desired sports domain, lots of annotations are needed. This paper proposes two techniques that need only a few annotated images and some videos from the sports domain to train a neural network with superior performance based on self-supervised training.

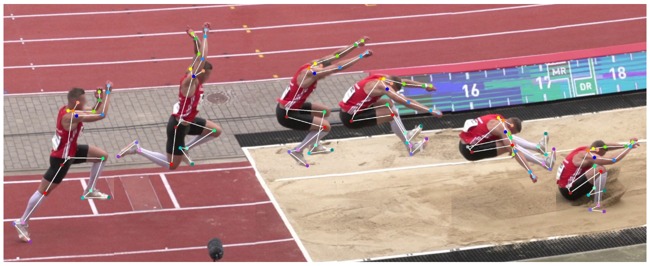

Fig. 1: Human pose estimation for triple jump analysis.

2. RELATED WORK

Human pose estimation is a field of active interest in computer vision. The currently best scoring approaches on common benchmarks like COCO [4] or MPII Human Pose [5] are based on convolutional neural networks [6, 7]. In contrast to multi-stage approaches like Mask R-CNN [8] which find person bounding boxes at first and detect one keypoint of each type in the second step, these recent approaches find all keypoints in a single step. The underlying backbone used in many recent models like [6, 7] is the *High Resolution Net* (HRNet) [9] and its variants for human pose estimation, e.g. [10]. This kind of backbone, which we also use in our experiments, maintains several branches of different resolutions, keeping the highest resolution from the beginning to the end of the network and providing data exchange between the different resolutions.

Computer vision is a popular technique to analyze athletes in various sport disciplines. [11] propose an interactive generative method for estimating and tracking athlete poses from monocular TV sports footage and evaluate it on hurdles and triple jump videos. An approach for automated detection of events like landing and jump in triple and long jump videos is proposed by [2]. Based on the trajectories of basketball players, [12] predict the location of the ball from monocular

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

video footage, no matter if it is visible or not. [3] predict the poses of ski jumpers during their flight phase and use robust estimation techniques to determine the flight angles per camera view. Based on independently detected 2D poses from multiple camera views of e.g., soccer videos, [1] create 3D poses and trackings of the athletes. In order to detect poses of swimmers, [13] propose a convolutional neural network which takes sequences of frames and the swimming style as an input. This additional information and a pose refinement over time improve their detection results.

Self-supervised learning is a highly attractive research field in computer vision. Its goal is to enhance neural networks by exploiting additional unlabeled data for training such that the resulting model performs better than a supervised training using labeled data only. Most common approaches use consistency regularization or pseudo labeling. A survey can be found in [14]. Consistency regularized methods are based on the idea that models should generate consistent predictions under different perturbations such as noise [15] or stronger augmentations [16]. In the domain of 2D HPE, [16] propose cutting out joints to simulate occlusion as a hard augmentation. [15] further uses a model ensemble as a target, as a model ensemble produces better predictions compared to a single model. Other approaches utilize pseudo labels, which means that network predictions are used as annotations [17]. [18] prove that this method is effective for the ImageNet classification task. Furthermore, [19] show that the usage of pseudo labels enhances their joint 2D/3D HPE pipeline for multi-person keypoint detection in operating rooms.

Our contributions: We show that self-supervised learning raises the detection performance of 2D HPE systems fine-tuned for specific sports disciplines with a small labeled training dataset. We propose two different methods, whereby the first leverages pseudo labels in an iterative process to increase network performance. We introduce a pseudo label selection method that selects the most accurate predictions across various augmentations as pseudo labels. Furthermore, we introduce mean teacher training for HPE, a single-step consistency based approach. We show that 50 labeled training images and 122 unlabeled videos are sufficient to generate superior detection results in the domain of long and triple jump.

3. METHODOLOGY

We use *HigherHRNet* [10] as a backbone network, which achieves state-of-the-art performance on HPE benchmarks. As the goal of this paper is to achieve superior performance in keypoint detection of athletes from a given sports discipline with only a few labeled images from that sports domain, we start of with pretrained weights from COCO [4]. As the keypoint definition of long and triple jump athletes is different from the COCO keypoint definition (details can be found in Section 4.1), we load only the backbone weights and not those from the network head.

3.1. Training with Pseudo Labels

3.1.1. Training Procedure

At the beginning, a fully supervised training on a small, in-domain labeled dataset $D_{labeled}$ is executed. Early stopping is used to select the weights of the epoch with the best score on the validation set. The next step is the generation of *pseudo labels*, which means that the labels are not annotations by hand but created from the network itself. Based on the selected weights, pseudo labels are generated for all unlabeled images of the training set, resulting in the pseudo label dataset $D_{PL,1}$ for the first iteration. Details on the generation process are described in the next section. Afterwards, the first self-supervised training iteration is started by training a new network on the generated pseudo label dataset $D_{PL,1}$, starting from pretrained weights from COCO. Hence, the network sees only images different from the labeled training set $D_{labeled}$. Again, the best weights according to the validation score are selected. In the next step, fine-tuning based on the selected weights is executed with $D_{labeled}$. The best weights according to the validation results are determined and used to generate updated pseudo labels $D_{PL,2}$. Then, the next self-supervised training iterations are executed analogous to the first one. Figure 2 visualizes the training procedure. Xie et al. [18] show that this iterative process performs well on ImageNet image classification. This paper uses this approach adapted to human pose estimation.

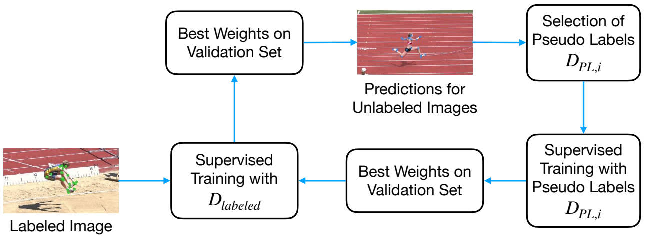
Best Weights on Validation Set
Predictions for Unlabeled Images
Selection of Pseudo Labels $D_{PL,i}$
Labeled Image
Supervised Training with $D_{labeled}$
Best Weights on Validation Set
Supervised Training with Pseudo Labels $D_{PL,i}$

Fig. 2: Training procedure for pseudo label training.

3.1.2. Pseudo Label Selection

In order to improve the network performance with pseudo labels, the selection of the pseudo labels is crucial. The most obvious possibility is to choose the labels based on the network confidence score, keeping all labels with a score larger than a certain threshold. The problem of this method is the non-equal distribution of the pseudo labels across the joints. Hence, the detection performance of joints with a lot of pseudo labels increases, while the scores of joints with fewer labels stagnate. A solution to this imbalance problem is to take the labels with the best $p\%$ confidence scores of each joint, but we found that this method does not output the best labels based on the distance to the ground truth as there is no direct relation between the network confidence score and the distance between the predicted and the ground truth keypoint.

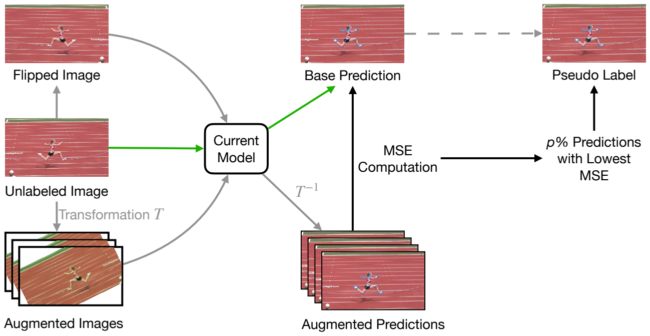

Fig. 3: Pseudo label selection method. Green arrows indicate that the base prediction is the prediction of the unlabeled image. The transformation T is randomly selected from the augmentations described in Section 4.2.

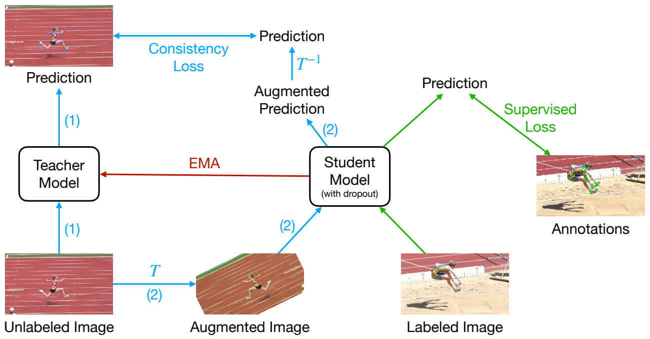

Fig. 4: Mean teacher training. The supervised training part is visualized with green arrows, the consistency regularization part is illustrated in blue. The teacher weights are updated after each training step, symbolized by the red arrow.

As a consequence, we use a different approach to select the best labels, not relying completely on the network confidence scores. As a base prediction, we use the predictions generated with the raw, not augmented input image. We add prediction results of a horizontally flipped image and results from some randomly chosen augmentations (described in Section 4.2) to our prediction set. Predictions with a very low confidence score (below threshold t_{score}) are discarded. The mean squared error (MSE) in pixels between the base prediction and the augmented predictions is calculated. Now, we select the predictions with the lowest $p\%$ MSE for the pseudo label dataset per keypoint, resulting in an equal number of predictions for each keypoint. As pseudo labels, the base predictions are used instead of the mean over all predictions, as single outliers could shift the mean enormously and the predictions on augmented images are less accurate as they are harder for the model. Figure 3 illustrates the pseudo label selection steps.

3.2. Mean Teacher Training

Apart from using pseudo labels, we considered another self-supervised training method based on consistency between a teacher and a student model. Figure 4 visualizes the training steps. At first, the student and the teacher backbone are initialized with pretrained weights from COCO, while all other weights are initialized randomly. Both models are nearly identical with the single difference that we add dropout in all fusion layers of the HRNet backbone (see [9]) in the student model. During a warm-up period, only the supervised training steps are executed based on $D_{labeled}$, visualized in green in Figure 4. After the warm-up period, the consistency loss is taken into account as well. Unlabeled images are fed into the teacher network to generate predictions, this process is marked with *(1)* in Figure 4. The same images are now augmented (see Section 4.2 for details) with a transformation T and given to the student network as an input (marked with *(2)*). The student predictions are transformed back to the original image with the inverse transformation T^{-1}. The consistency loss is now calculated as the MSE of the predicted poses from the student and the teacher model. The loss is masked if the teacher confidence is too low (below threshold t_{score}). Throughout the complete training process, the teacher weights are updated after each training step to be the exponential moving average (EMA) weights of the student model, visualized in red in Figure 4. A step consists of both supervised and consistency loss updates. For inference, the student model is used. Tarvainen et al. [15] show that this method leads to great improvements on common image classification benchmarks. We adapt this approach for the usage in HPE.

4. EVALUATION

4.1. Dataset

The dataset used in this paper contains video recordings of long and triple jump athletes. It consists of 4,522 labeled images from 186 video sequences, whereby 3,154 images from 122 videos are used for training, 200 images from 18 videos for validation and the remaining 1,062 images from 46 videos as the test set. All 3,154 annotated images are only used in the supervised training scenario for comparison. In our self-supervised approach we use a subset of only 50 labeled images for training. The recordings were taken during competitions or training and show various sports sites and athletes. The recordings have a constant frame rate of 200Hz and a length between 670 and 1900 frames. All videos are annotated with 20 keypoints (r./l. eye, r./l. ear, nose, neck, r./l. shoulder, r./l. elbow, r./l.wrist, mid hip, r./l. hip, r./l. knee, r./l. ankle, r./l. big toe, r./l. small toe, r./l. heel), see Figure 1 and Figure 5 for examples.

4.2. Data Augmention and Training Settings

During supervised training, finetuning and pseudo label training, we use common augmentation methods: random horizontal flip, random rotation of up to 30°, random translation of up to 40 pixels (using an input image size of 512×512), random scale in the range $[0.75, 1.5]$ and color jitter. These augmentations are also used during pseudo label generation.

For our experiments, we use 50 randomly selected ground truth labels for our training dataset $D_{labeled}$. During pseudo label creation, we set the score threshold $t_{score} = 0.1$ and we evaluate 10 different augmentations to compute the MSE. We train for 3 iterations and use the best 60%, 70% and 90% pseudo labels in the first, second and third iteration, respectively. As the quality of the pseudo labels improves in each iteration, we increase the number of pseudo labels that we use. In each iteration, we train for $10k$ steps with the pseudo labels and execute the finetuning for 250 steps. For mean teacher training, we use the same score threshold t_{score} of 0.1 to mask the consistency loss. The EMA momentum is set to 0.999 and our warm-up period lasts for 1,500 steps. The dropout rate of the student model is set to 0.2.

4.3. Results

For evaluation, we use the Percentage of Correct Keypoints (PCK) metric. PCK considers a keypoint as correct at a certain threshold t if the distance of the detected keypoint to the ground truth keypoint is less or equal than t times the distance between the left shoulder and right hip. The recall at a certain PCK threshold tells us the percentage of the keypoints that is considered correct at that threshold. During training, we use this metric at a threshold of 0.1 as the performance measure on the validation set, which corresponds to approx. 6cm.

4.3.1. Pseudo Label Results

For the first experiment, we use the subset $D_{labeled}$ with 50 labeled images for training and the remaining 3,104 images that are also annotated as unlabeled images, hence we do not use the labels. This has the benefit that the results are perfectly comparable to the fully supervised results as both networks have seen the exact same images during training. For comparability to the fully supervised training, we use the full validation set of 200 images in these experiments, but we verified that a validation set of 50 images leads to similar results. In practice, this setup requires the coaches, apart from annotating the images for $D_{labeled}$ and the validation set, to select meaningful images from the videos that should be used for training. This image selection process requires a lot less time than annotating all images, but more time than just providing videos without any further work. Table 1 shows the results for three iterations in this training setup in row 3. Figure 5 shows some examples for model predictions after iteration 3.

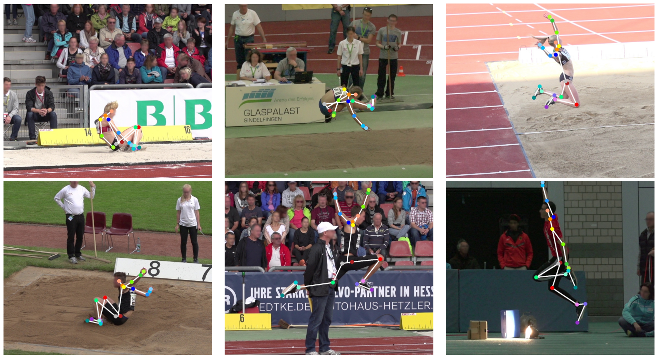

Fig. 5: Example images with predictions from the pseudo label model. A new model is needed as occlusions and keypoints like toetip or heel are not included in COCO. These cases are still harder for the model. Furthermore, the dataset contains some extreme poses.

Table 1: Recall values in % at PCK thresholds of 0.1 and 0.2 on annotated test set images for the pseudo label self-supervised training based on the selected images (row 3) and on every 10th video frame (row 4). The first row shows the results for the fully supervised training and the second row the results for the supervised training on $D_{labeled}$ for comparison. Row 5 displays the mean teacher (MT) results or both variants.

	Images (Labels)	Run	PCK 0.1	PCK 0.2
1	3,154 (3,154)	supervised	91.9	96.5
2	50 (50)	supervised	83.8	90.0
3	3,154 (50)	iteration 1	87.7	93.6
		iteration 2	88.0	93.9
		iteration 3	88.4	94.4
4	17,656 (50)	iteration 1	87.1	93.5
		iteration 2	88.0	94.2
		iteration 3	88.2	94.5
		iteration 4	88.6	94.7
5	3,154 (50)	MT	87.1	93.7
	17,656 (50)	MT	87.1	93.5

The gap between the fully supervised training and the supervised training on 50 images is 8.1% at a PCK threshold of 0.1. After the first iteration, this gap can be narrowed to 4.2%. After the third iteration, the difference shrinks to even 3.5%. This is 40% of the original gap, with using less than 1.6% of the labels. As we use the validation results at threshold 0.1, this threshold is also used as the main evaluation threshold. But regarding PCK threshold 0.2, the gap is be narrowed even further, from 6.5% to 2.1%, which is less than a third.

In the second experiment, we take the same 50 labels as in the first experiment, but no pre-selection of video frames is used. From all 122 videos belonging to the training dataset, every 10th frame is extracted and added to the unlabeled dataset. This results in 17,656 images. We use every 10th frame so that two images are clearly different from each other.

Table 1 shows the results for this experiment in row 4. We use an additional fourth iteration with all pseudo labels here, as further training still improves the results on the validation set, which is not the case in the first experiment. The results after the last iteration of this experiment are similar to the results from the first experiment, but this experiment has slightly better performance at PCK thresholds 0.1 and 0.2. At PCK threshold 0.1, we could close the gap to the fully supervised training from 8.1% to 3.3% and at threshold 0.2 from 6.5 % to 1.8%. The table shows that the improvement from iteration to iteration is slower than in the first experiment, therefore the fourth iteration still gains some improvement.

4.3.2. *Mean Teacher Results*

Identical to the pseudo label evaluation, we conduct two mean teacher experiments. One with the same images as in the fully supervised run and one with every 10th frame from the videos corresponding to the training set. Table 1 shows the results for both experiments in row 5. The results are collected on all annotated test set images using the network weights from the step with the highest validation score (early stopping). The table shows that the mean teacher results are slightly worse than the pseudo label results after the final iterations, but perform a lot better than the supervised training on 50 images. At a PCK threshold of 0.1, the gap to the fully supervised score could be narrowed from 8.1% to 4.8% and at PCK threshold 0.2 from 6.5% to 2.8%, regarding the results from the first experiment.

4.3.3. *Results with more Labels*

To evaluate the benefit of more labeled images, we executed the self-supervised methods also with 100 and 250 labels. See Table 2 for the exact results, using the 3,154 images from the supervised training as the training dataset. For 250 labeled images we changed the percentages of the pseudo labels to 70%, 80% and 95%, otherwise the first iteration does not have an effect as the PCK values are already higher after the first supervised training. The table shows that the gap between the fully supervised result and the supervised training shrinks from 8.1% with 50 labels to 5.9% with 100 labels and 3.5% with 250 labels. With 50 labels, we could close 60% of the gap, with 100 labels, this rate shrinks to 50% and with 250 labels to 45%. Furthermore, after the initial supervised training on the few labels, the difference of the PCK values between 50 and 100 labels is 2.2% and between 100 and 250 labels 2.4%. After the pseudo label training, the differences are a lot smaller, namely 0.6% between 50 and 100 labels and 1.0% between 100 and 250 labels.

For mean teacher training, similar results are observable. Hence, the gain is larger for less labels and the PCK values are closer together after both self-supervised trainings. For all experiments, single-step mean teacher results are in the area of the results from the first or second pseudo label iteration. Hence, the usage of self-supervised training is more effective with less annotated images, but it improves the results in every case. Obviously, the highest absolute score is achieved with the most labels, so there is a trade-off between efficiency of the self-supervised learning and the final absolute score. This should be taken into account when coaches decide how many images they hand-annotate.

Table 2: Recall values in % at PCK threshold 0.1 on annotated test set images for pseudo label and mean teacher training with different numbers of labeled images (first row). The second row contains the results for the fully supervised training and the supervised training on $D_{labeled}$ for comparison.

Labeled Data	3,154	250	100	50
supervised	91.9	88.4	86.0	83.8
iteration 1		89.5	87.7	87.7
iteration 2		89.8	88.3	88.0
iteration 3		90.0	89.0	88.4
mean teacher		89.8	88.0	87.1

5. CONCLUSION AND FUTURE WORK

This paper proposes two techniques for self-supervised learning with a few labeled images in order to train a network for human pose estimation in a new sports domain. One method uses a mean teacher approach like in [15], with a simultaneous training on the labeled and the unlabeled images. The self-supervised training part uses a consistency loss between an EMA teacher model and a student model with dropout layers and stronger augmentation. The other method generates pseudo labels similar to [18] and uses a selected subset of them for the first training step and the labeled images for the finetuning step. This iterative process is continued until convergence.

The evaluation results prove the sufficiency of a training dataset containing 50 labeled images and some video sequences to train a deep neural network for a new sports domain such that it generates acceptable results. The PCK values at a threshold of 0.1 could be raised from 83.8% to 88.4%, which closes the gap between the fully supervised training on 60× more images and the supervised training on 50 images by more than 60%. These methods could open the usage of human pose estimation performance measurements to a wide range of sports disciplines in the future. The expense to collect video material is very low, as it only requires a smartphone or small camera. Furthermore, annotating 50 images by hand is also done quickly.

In this paper, the methods are trained and evaluated for 2D HPE, but they are not limited to this setting. Future work could include the analysis of the methods for more sports disciplines and with other backbone models.

6. ACKNOWLEDGEMENTS

This work was funded by the Federal Institute for Sports Science (BISp) based on a resolution of the German Bundestag. We would like to thank the Olympic Training Center Hessen for collecting and providing the video data.

7. REFERENCES

[1] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton, "Multi-person 3d pose estimation and tracking in sports," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[2] Moritz Einfalt, Charles Dampeyrou, Dan Zecha, and Rainer Lienhart, "Frame-level event detection in athletics videos with pose-based convolutional sequence networks," in *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, 2019, pp. 42–50.

[3] Katja Ludwig, Moritz Einfalt, and Rainer Lienhart, "Robust estimation of flight parameters for ski jumpers," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020, pp. 1–6.

[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*. Springer, 2014, pp. 740–755.

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[6] Junjie Huang, Zengguang Shan, Yuanhao Cai, Feng Guo, Yun Ye, Xinze Chen, Zheng Zhu, Guan Huang, Jiwen Lu, and Dalong Du, "Joint coco and lvis workshop at eccv 2020: Coco keypoint challenge track technical report: Udp+," 2020.

[7] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang, "Adversarial semantic data augmentation for human pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 606–622.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[9] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al., "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[10] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5386–5395.

[11] Mykyta Fastovets, Jean-Yves Guillemaut, and Adrian Hilton, "Athlete pose estimation from monocular tv sports footage," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 1048–1054.

[12] Xinyu Wei, Long Sha, Patrick Lucey, Peter Carr, Sridha Sridharan, and Iain Matthews, "Predicting ball ownership in basketball from a monocular view using only player trajectories," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 63–70.

[13] Moritz Einfalt, Dan Zecha, and Rainer Lienhart, "Activity-conditioned continuous human pose estimation for performance analysis of athletes using the example of swimming," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 446–455.

[14] Jesper E. van Engelen and Holger H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109(2), pp. 373–440, 2020.

[15] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[16] Rongchang Xie, Chunyu Wang, Wenjun Zeng, and Yizhou Wang, "Humble teacher and eager student: Dual network learning for semi-supervised 2d human pose estimation," *arXiv preprint arXiv:2011.12498*, 2020.

[17] Dong-Hyun Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3.

[18] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10687–10698.

[19] Vinkle Srivastav, Afshin Gangi, and Nicolas Padoy, "Self-supervision on unlabelled or data for multi-person 2d/3d human pose estimation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 761–771.