A Practical Hybrid Active Learning Approach for Human Pose Estimation

Kaplan Sinan, Juvonen Joni, Lensu Lasse

DOI: 10.1007/978-3-030-73973-7_32

A Practical Hybrid Active Learning Approach for Human Pose Estimation *

Sinan Kaplan[1], Joni Juvonen[2], and Lasse Lensu[1]

[1] Computer Vision and Pattern Recognition Laboratory, Department of Computational and Process Engineering, School of Engineering Science, LUT University, Lappeenranta, FINLAND
[2] University of Turku, Department of Future Technologies, Turku, FINLAND

Abstract. Active learning (AL) has not received much attention in deep learning (DL) for human pose estimation. In this paper, a practical hybrid active learning strategy is proposed for training a human pose estimation model, and it is tested in an industrial online environment. The conducted experiments show that the active learning strategy to select diverse samples to be annotated outperforms the baseline method with random sampling. As a result, the strategy enables a significant improvement in the performance of pose estimation.

Keywords: active learning · human pose estimation· human in the loop artificial intelligence (AI) .

1 Introduction

DL has caused a paradigm shift and its success has been regularly seen in supervised learning tasks where a large amount of labeled data is available [18, 9, 35]. For applications where the amount of data is limited, methodological remedies exist in the form of data augmentation, transfer learning and few-shot learning. Even with these approaches, however, a relatively large amount of data is needed for rapid adaptation of deep models in complex domains where the initial data is limited. In applications where the acquisition of data and labelling them can be an expensive and laborious task, one may consider an iterative approach to sample and label the data. AL is a family of such techniques to appropriately select data samples to be annotated next [27]. These methods enable collaboration of a human and AI to annotate a subset of data without resorting to fully annotating the data or purely random selection of samples.

The ultimate goal in AL is to reduce the annotation and training effort/cost while making models as accurate as possible with less amount of data. To achieve this, the data to be annotated is sampled by an acquisition function and is brought to an oracle (human annotator) to review and perform the annotation.

There are two settings where AL strategies can be applied and tested: 1) online (live) and 2) offline (simulated) environments [16]. In offline environments,

* Supported by PintaWorks Oy.

there is already a pool of labeled data available and the goal is to evaluate the performance of a AL strategy on this labeled data pool. Offline environments are more common than the online ones in the AL research area.

An online environment is a setting where machine learning (ML) research makes an impact on real-world problems. It connects research and applications in real life to assess whether an improvement of a particular ML model makes a difference outside the common benchmark data sets [33].

In this paper, an AL strategy is presented for an online environment involving human pose estimation [8]. Compared to object detection and image classification tasks, the annotation cost is much higher as labeling of a number of keypoints on a human body image is required. For instance, COCO-style keypoint annotation requires 18 keypoints to be labeled [7].

To use active learning in this context, a hybrid approach combining model-based uncertainty sampling and diversity sampling [2] is proposed in this study. The method takes advantage of transfer learning and approximate nearest neighbors as parts of the solution. The aims are as follows: 1) To improve the accuracy of the pose estimation model with diversely picked data samples, 2) to avoid adding another level of complexity, such as training another model for sampling, to the AL pipeline, 3) to reduce the sampling time by using approximate nearest neighbors instead of exhaustive search methods, and 4) to provide an analysis of practical challenges in the online environment.

The paper is organized as follows: Section 2 reviews the studies of active learning and human pose detection area. Section 3 covers the proposed method in detail. Section 4 focuses on experiments with the results. The findings are further evaluated in Section 5 and Section 6 presents the conclusion[3].

2 Related Work

Human pose estimation focuses on providing reliable estimates of human poses in various applications, including person tracking and analysis of sports activities [5, 8]. In the literature, it is studied under two categories: 1) top-down, and 2) bottom-up approaches. The top-down approaches first detect person candidates and then perform pose extraction. For instance, the regional multi-person pose estimation framework [11] and the cascaded pyramid network [6] fall into this category. On the other hand, the bottom-up approaches first extract features from a given image and construct bipartite graphs to produce a human pose estimate. A widely used framework in this category is OpenPose [4] that is based on part affinity fields described in the work by Cao et al. [5]. Other important studies apply long short term memorys (LSTMs) [1] and different variations of convolutional neural networks (CNNs) [34, 14] to improve the reliability of human detection.

Active learning is widely studied in the literature and the methods vary depending on the application context [27, 2]. For instance, [23] focuses on appli-

[3] Initial results of this study have been presented in the 2nd ICML 2020 Workshop on Human in the Loop Learning as a work-in-progress paper.

cations of active learning techniques in DL models in particular. The techniques applied for active learning are utilized under two categories: 1) pool-based strategies and 2) query synthesizing strategies [27, 2].

The pool-based strategies utilize information, ensemble and uncertainty based methods to select samples from an unlabeled data pool [27, 2, 28]. Bayesian models are also studied among the pool-based methods [12]. As they embrace Bayesian inference principles, they provide a natural basis for the uncertainty estimation compared to more traditional uncertainty approximation methods.

The strategies relying on query synthesis take advantage of generative adversarial networks (GANs) [36] and variational autoencoders (VAEs) [29, 21]. The methods in this category are based on learning a latent representation for both labeled and unlabeled data for a discriminator module to classify whether a given data sample is from the labeled or unlabeled data pool. They are difficult to generalize for an online setting [16], since the data distribution changes by the time in online environments. Besides, this would add another level of complexity and cost by requiring an additional model to be trained on both labeled and unlabeled data pool, which this study aims to avoid.

Human pose estimation and active learning. In the field of computer vision, active learning methods are mainly studied and tested on image classification and object detection tasks [24, 26]. There is a limited number of studies that cover human pose estimation tasks. Liu et al. [19] studied active learning for human pose detection. They applied uncertainty measurement and the principle of influence [10] for sampling, and convolutional pose machine (CPM) [34] was used as the pose model. One limitation of the CPM-based methods is that they provide heatmaps with the same resolution as the image size. This causes heatmaps to be diffuse making the application of non-maximum suppression (NMC) more difficult. To solve this issue, in this paper, the model provides low resolution heatmaps for each keypoint to extract peak points with NMC. In addition to that, in contrast to greedy representative sampling strategy used by Liu et al., approximate nearest neighbors approach [22] is applied to reduce sampling time (more details are given in sections 3 and 4).

3 Methods

The proposed AL approach for human pose estimation is shown in Figure 1. For the development, OpenPose-plus [32] is used as the pose model, which is based on the popular OpenPose framework [4]. It is an optimal framework for our study, since it provides real time pose estimation with simplicity and flexibility of switching between different backbones (MobileNet, Resnet18, VGG and VGGTiny) stated in the repository. By considering the given inference time, model size and accuracy in OpenPose-plus repo, the model with VGGTiny backbone is trained from scratch with initially annotated 2000 data samples.

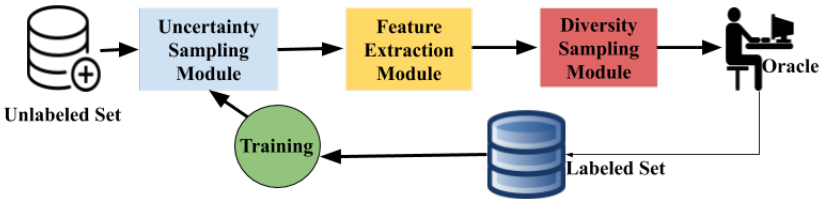


Fig. 1: An overview of the active learning procedure.

3.1 Active Learning Framework

The OpenPose-plus model provides confidence maps (heatmaps) and part affinity fields (PAFs). Heatmaps represent activations of the last layer of the model as confidence scores, whereas PAFs represent connection vectors between the keypoints. In total, there are 19 heatmaps (18 specific body points and one background). Based on the heatmap activations, AL is designed module-by-module as follows[4]: (1) **Uncertainty Sampling Module** is responsible for filtering data samples based on heatmap activations. Since the heatmaps correspond to confidence scores, a lower activation value implies higher uncertainty. (2) **Feature Extraction Module** takes filtered data from the uncertainty module and computes embedding features using the pretrained Resnet50 model on ImageNet data. Given the speed, accuracy level and amount of operations required for a single forward pass in the Resnet50 model, it is chosen to extract embedding features [3]. (3) **Diversity Sampling Module** takes the filtered data with features from the preceding module and constructs an approximate nearest neighbor tree [22] to apply diversity sampling. To shorten the sampling time on high-dimensional embedding features, an approximate nearest neighbor search is used. (4) **Oracle** is a human annotator who reviews and labels filtered data from the diversity sampling module and updates the training set. (5) **Training Module** resumes the training task with the updated training set.

As the model is deployed in an online environment, the whole procedure is an iterative process and it repeats itself depending on the model performance, the available unlabeled data size and available resources, such as the oracle and/or hardware resources, to complete the task.

3.2 Baseline Method and Evaluation Metric

To compare the effectiveness of our model for human pose estimation, *random sampling* is used as the baseline. For the evaluation of experiments, *person count accuracy vs. size of training set* is chosen to report the results for comparing the AL approach and the baseline method. Person count accuracy is a percentage of correctly detected people out of the total number of people.

[4] For a detailed flow of each module refer to Algorithm on https://github.com/kaplansinan/S-SPRR2020ALpose.

4 Experiments

The experiments run in an online environment are presented in this section. First, the AL method is initially tested on the COCO validation set to qualitatively assess whether the strategy selects diverse samples after the uncertainty sampling procedure. Afterwards, experiments are run in the online environment and report results for each training session. The experiments are conducted in collaboration with PintaWorks Oy. The company provides solutions for person tracking and activity recognition for the healthcare industry.

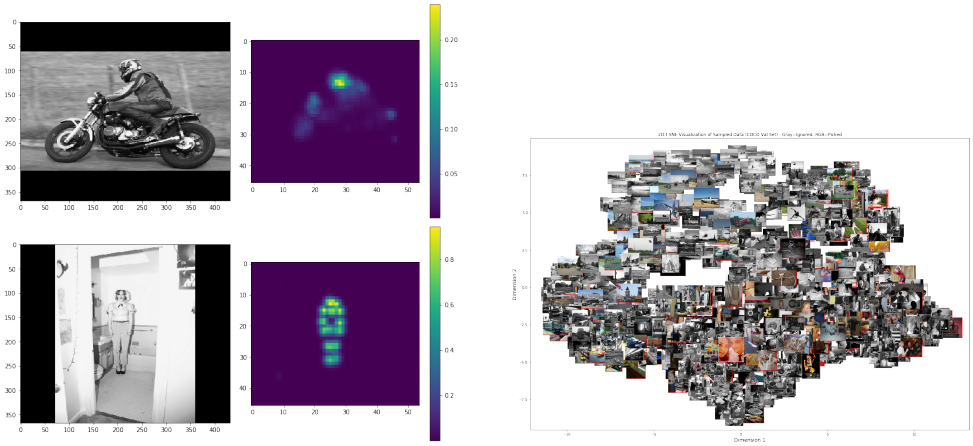
Data and Model. The data are provided by the company and there are environment-dependent variations in the data, such as location, lighting conditions and camera angle. The data consists of grayscale images with size of 368x368 (WxH). A limited amount of data (2000 samples) was available to train the model, thus, data augmentation was used to increase the original data set. The augmentation techniques applied are as follows: rotation with limit of [-30, 30] degrees, translation with limit of [-0.62, 0.62] in width or height, scaling factor with range of [0.6, 1.4], random brightness factor with limit of [0.7, 1.3], and random contrast factor with limit of [0.7, 1.3].

The OpenPose-plus model [32] is used with the VGGTiny backbone. Two different libraries were experimented with for the approximate nearest neighbor search: Annoy [31] and FLANN [20]. Based on initial tests with both of them, as a major difference, Annoy was found faster than FLANN and it was selected for the experiments.

Validation of AL strategy. Before testing the proposed method, a qualitative assessment was performed on a benchmark set to see whether it is capable of selecting diverse samples. To do so, the trained model on the first batch of the data provided by the company is used to apply uncertainty and diversity sampling on the COCO validation set. In Figure 2a, randomly picked samples from the COCO validation set with low and high activations from the model outputs are shown. The heatmap threshold th_{HM} was set to 0.3 and the threshold for the number of keypoints th_{KP} was set to 6 after initial tests in the development environment. After the uncertainty module, the embedding features of the size of 1×2048 are extracted for each selected sample by the feature extraction module. Next, 20% of the samples (2K images in total) are identified by the diversity sampling module to be labeled. The chosen diverse samples from the COCO validation set are shown in Figure 2b. The figure reveals that the chosen samples are scattered across the image set. Hence, it is viable that the proposed strategy can identify diverse samples successfully.

Training Details. Tensorflow stack is used for training the models and monitoring each training session. Experiments are conducted on NVIDIA GeForce GTX 1060 with 6GB and CUDA Development Toolkit (CUDA 10.2). The data is divided into training and validation sets at each session and the aforementioned augmentation techniques are applied during training. Early stopping criteria is used on validation set to halt the training when the model performance stops improving.

After initial tests with the training pipeline, the hyperparameters for each



(a) (b)

Fig. 2: (a) Examples of low and high heatmap activations. The first row shows an image with corresponding low activation values and the second row demonstrates an image with high activations. The first image will be filtered by uncertainty module for annotation.; (b)The results from diversity sampling module on the COCO validation set. The images shown in color are the ones chosen for the annotation based on approximate nearest neighbors search.

training session are set as follows: batch size of 4, step decay learning rate scheduler with $learningrate = 0.0001$, early stopping patience 35, and Adam optimizer [17] with $learningrate = 0.0001$, $beta_1 = 0.9$, $beta_2 = 0.999$.

Four training sessions were run for each method. It is important to note that the baseline method with random sampling was evaluated twice to provide information on the performance variation. This was necessary to see variation within the baseline method. In total, 12 training sessions were executed. At each training session, 1000 data samples are selected from the available unlabeled data pool and added to the training set after annotated by the oracle. The next training session continues from the previous one. Each training iteration takes 6-8 hours on our training setup above.

Testing Details and Results. The proposed AL method and the baseline method were evaluated on the test set, separate from the training set. The evaluation was done based on the metric given in Section 3.2. The comparison of the methods is shown in Figure 3 (A). The proposed method outperforms the baseline method at each iteration with a clear margin based on the person count accuracy. This is because of that the AL method can pick data samples to be annotated diversely, which leads the pose model to generalize well on the data set. In other words, the samples that the model underperforms at previous iterations are successfully selected to be annotated for the next training iteration. On the other hand, due to its randomness, the baseline method fails to diversify data selection part.

To test further the robustness of the proposed method under environmental variations, also test time augmentations (TTAs) was utilized on the test set. The variations in the data occur due to person size, lightning conditions, and camera noise. Thus, the following TTAs were applied: blur with kernel size of 5 (BLUR), brightness factor of 1.3 (BRIGHT) and 0.7 (DARK), scale factor of 1.4 (ZOOM IN) and 0.6 (ZOOM OUT). The evaluations on each of these augmented sets are presented in Figure 3 (B-F). One can see the performance variation of the baseline method at every step as a result of it's randomness. For instance, at one iteration it samples the bright images more than dark ones and at another iteration it selects the dark ones more than bright images.

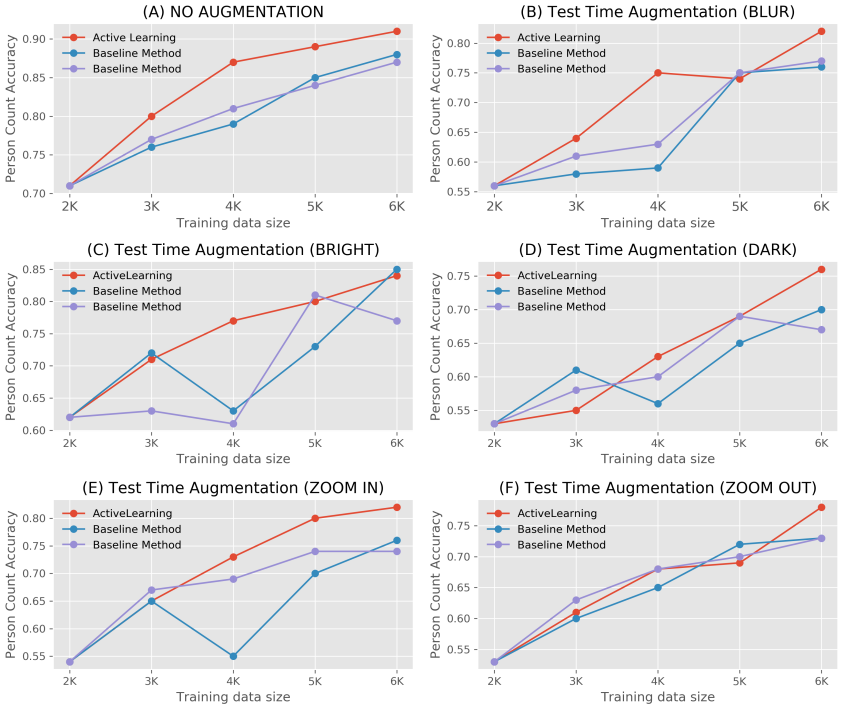


Fig. 3: Comparison of the AL method and the baseline method: (A) No Augmentations; (B-F) Test time augmentations.

5 Discussion

Data labeling is a costly and time consuming process in the development and training of human pose estimation models. For instance, COCO-style keypoint

annotations for a single person takes 15 - 20 seconds per one image. To shorten this process in practical applications and get an accurate model with less data, AL methods can be taken into use.

In this study, an AL strategy was proposed to sample data for human pose estimation. The proposition is a hybrid procedure that makes use of both uncertainty and diversity sampling. To evaluate the method, it was compared with a baseline method based on random sampling. The experiments showed that the proposed method is able to select diverse data samples successfully and boost the performance of the human pose estimation better than the baseline method. The robustness of the method is further tested and proved in a production environment of the company.

The main issue of the proposed AL method is adversarial data samples. For instance, adversarial samples, in which an object assembles a human shape, are discarded. Since they have high activation values, the uncertainty sampling module filters out these samples. To the best knowledge of the authors, this issue has not been not studied as part of AL in deep models and it may bring an advantage.

As this is an empirical study, it is equally important to state hidden gems and issues encountered during training process. In terms of the training procedure, the following methodologies were found useful: 1) learning rate finder [30] is an important technique to use for getting starting value for the learning rate of the optimizer, 2) weight initialization is an important factor to consider when training DL methods from scratch. In this study, better success was achieved with Glorot uniform [13] than Kaiming uniform [15], 3) Tensorflow Profiler is quite helpful to identify data loading/pipeline bottlenecks during the training, 4) One should not try an augmentation that cannot be observed in an environment, where the model is in use. It reduces the generalization capability of the model, and 5) as noted by Ruggero et al. [25], small person size and occlusions are two main issues of human pose models that were encountered in the testing and production pipeline.

Further development of the AL method will be focused on substituting the diversity sampling module with a hierarchical clustering method. Also, it can be worth using both local and global image features to perform diversity sampling. One can combine global features, such as image hashes, with the local features extracted by the feature extraction module. To improve uncertainty sampling in catching hard negatives, it can help to use average of multiple TTAs as a validation step (this is because of that true person detections likely have higher activations than adversarial ones).

6 Conclusion

In this study, an active learning strategy for human pose estimation that is implemented in an online development environment of the company was proposed. The method combines both uncertainty and diversity sampling. The uncertainty sampling is applied on the heatmap activations of the pose model, while the di-

versity sampling is further carried with the embedding features extracted from the pre-trained feature extraction model on ImageNet. To reduce sampling time on the high dimensional embedding features, an approximate nearest neighbor method is used. The experiments revealed that the proposed AL strategy improves the performance of the pose model by introducing a smart data sampling framework.

References

1. Artacho, B., Savakis, A.: Unipose: Unified human pose estimation in single images and videos. arXiv preprint arXiv:2001.08095 (2020)
2. Budd, S., Robinson, E.C., Kainz, B.: A survey on active learning and human-in-the-loop deep learning for medical image analysis. arXiv preprint arXiv:1910.02923 (2019)
3. Canziani, A., Paszke, A., Culurciello, E.: An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678 (2016)
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7291–7299 (2017)
6. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018)
7. COCO: Coco keypoints (Jun 2020), http://cocodataset.org/keypoints-2019
8. Dang, Q., Yin, J., Wang, B., Zheng, W.: Deep learning based 2d human pose estimation: A survey. Tsinghua Science and Technology **24**(6), 663–676 (2019)
9. Dargan, S., Kumar, M., Ayyagari, M.R., Kumar, G.: A survey of deep learning and its applications: A new paradigm to machine learning. Archives of Computational Methods in Engineering pp. 1–22 (2019)
10. Dutt Jain, S., Grauman, K.: Active image segmentation propagation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2864–2873 (2016)
11. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2334–2343 (2017)
12. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1183–1192. JMLR. org (2017)
13. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256 (2010)
14. Golda, T., Kalb, T., Schumann, A., Beyerer, J.: Human pose estimation for real-world crowded scenarios. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–8. IEEE (2019)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)

16. Kagy, J.F., Kayadelen, T., Ma, J., Rostamizadeh, A., Strnadova, J.: The practical challenges of active learning: Lessons learned from live experimentation. arXiv preprint arXiv:1907.00038 (2019)
17. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
19. Liu, B., Ferrari, V.: Active learning for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4363–4372 (2017)
20. Mariusmuja: Flann - fast library for approximate nearest neighbors (May 2020), https://github.com/mariusmuja/flann
21. Mottaghi, A., Yeung, S.: Adversarial representation active learning. arXiv preprint arXiv:1912.09720 (2019)
22. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. VISAPP (1) **2**(331-340), 2 (2009)
23. Munro, R.: Human-in-the-Loop Machine Learning. Manning Publications (2020), https://books.google.fi/books?id=LCh0zQEACAAJ
24. Roy, S., Unmesh, A., Namboodiri, V.P.: Deep active learning for object detection. In: BMVC. p. 91 (2018)
25. Ruggero Ronchi, M., Perona, P.: Benchmarking and error diagnosis in multi-instance pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 369–378 (2017)
26. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
27. Settles, B.: Active learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2009)
28. Shao, J., Wang, Q., Liu, F.: Learning to sample: an active learning framework. arXiv preprint arXiv:1909.03585 (2019)
29. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5972–5981 (2019)
30. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 464–472. IEEE (2017)
31. Spotify: Annoy (approximate nearest neighbors oh yeah) (May 2020), https://github.com/spotify/annoy
32. Tensorlayer: Hyperpose(python training library, c++ inference library) (Jun 2020), https://github.com/tensorlayer/hyperpose
33. Wagstaff, K.: Machine learning that matters. arXiv preprint arXiv:1206.4656 (2012)
34. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4724–4732 (2016)
35. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems **30**(11), 3212–3232 (2019)
36. Zhu, J.J., Bento, J.: Generative adversarial active learning. arXiv preprint arXiv:1702.07956 (2017)