Domain Generalization in Human Pose Estimation via Meta-Learning

You Rim Choi[1], Yejin Hwang[1], Chae Song Park[2], Chan Kyo Kim[3]
[1]Graduate School of Data Science, [2]Program in Artificial Intelligence, [3]Department of Mechanical Aerospace Engineering
{yrchoi, evergreen97, chase121, ckkim10} @ snu.ac.kr

Abstract

Human pose estimation (HPE) aims to understand human posture by localizing body keypoints from images or video sequences. As a basic task in computer vision, HPE is a core component for many practical applications in various fields. However, HPE is still a very challenging task. Acquiring a dataset for application to a new domain is especially big challenge because data collection and annotation are costly, and for many tasks, only a limited amount of data may be available. There are several approaches to solving this problem such as data augmentation, unsupervised or self-supervised learning, and refinement of pose estimation but these methods may cause too much computation or long model-inference latency. Meta-learning, also known as "learning to learn", enables deep learning to achieve higher performance without large datasets and sufficient computational resources by making them learn how to learn. In this team project, we present a novel meta-learning approach that can easily generalize the human pose estimation task to multiple domains with small dataset and low computing cost. Experimental evaluations were performed on whether the proposed approach can quickly learn a new task and can be applied to multiple domains.

1. Introduction

Human pose estimation (HPE) is one of the most important computer vision tasks that includes detecting, associating, and tracking semantic keypoints such as elbows, knees. It aims to understand human posture by localizing body joints from images or video sequences. As a fundamental task in computer vision, HPE is a key component for many practical applications such as human-computer interaction, movies and animation, virtual reality, medical assistance for rehabilitation training and physical therapy, human motion prediction for self-driving, sports motion analysis to automatically track or estimate human movement accuracy. It can also be applied in video surveillance and detecting illegal or inappropriate human behavior.

As well as other vision tasks, HPE also achieves excellent and remarkable progress through the introduction of deep learning. The use of deep convolutional neural networks (DCNNs), advanced computing power, and most importantly, the availability of large amounts of annotated datasets have contributed to improvement in terms of performance [1, 2, 3]. However, HPE is still a very challenging task because the human body appearance changes dynamically due to forms of clothes, occlusions, and background contexts. A good pose estimation has to be robust to these variations. Acquiring a dataset for application to a new domain is also big challenge. Data collection and annotation are time-consuming, difficult, and for many tasks, only a limited amount of data may be available.

HPE enables or acts as a core building block for many vision-based edge AI applications. Thanks to recent researches to enable real-time pose estimation in edge devices [4, 5], it can be applied to more diverse fields. However, in order to flexibly expand the domain of application and to use such techniques in practice, sufficiently large and unbiased datasets are required. This can be especially difficult for extreme motions such as poses in specific sports, which are difficult to infer from typically provided pose examples. In case of video surveillance applications, we need a new dataset for them due to the different angle view, scale and resolution.

There are several approaches to solving the costly data acquisition problem. A typical approach is data augmentation method that is scalable for synthesizing large amount of data [6, 7, 8, 9]. Some researchers address this challenge by proposing unsupervised or self-supervised approach that does not require annotations and be trained from unlabeled data which can be collected relatively easily [10, 11, 12]. Refining the pose estimation result [13] can also be a solution to reduce performance degradation that may occur when HPE is applied to a new domain with insufficient data. However, these methods may cause too much computation or long model-inference latency.

In this team project, we present a meta-learning

approach that can easily generalize the human pose estimation task to multiple domains with small dataset and low computing cost. We tackled the task of 2D multi-person pose estimation based on bottom-up methods. Bottom-up methods directly infer the body keypoints first and then group them to form multiple human poses. Among various meta-learning methods, optimization-based meta-learning represented by Model-agnostic meta-learning (MAML) was applied.

2. Related Works

2D Multi-person Pose Estimation: Prior to our research on HPE, early successful algorithms for human pose estimation introduced inference mechanisms on part-based graphical models [14]. Advanced from this work, a variety of methods have been developed with inference algorithms for detecting body parts in supposed environments such as single person pose estimation or multi-person pose estimation [15, 16, 17]. In recent researches on human pose estimation, 2D multi-person pose estimation algorithms can be divided into classification standards: model-based vs. learning-based. Learning-based human pose estimation uses certain approaches on mapping which learns from given image and joint coordinates with explicit models that infer the relations between body parts in the image with annotated keypoint samples. Although certain approaches in human pose estimation using deep learning may require a great deal of training data and computation time, these methods outperform model-based approaches. With another criteria, in multi-person human pose estimation, it can be classified into top-down methods and bottom-up methods. Top-down methods first detect human instances and estimate the pose of the instances. Examples of top-down method include G-RMI [18], CFN [19], Mask R-CNN [20], and CPN [21]. They all locates joints within bounding boxes previously generated by instance detector such as Fast-RCNN [22], Faster-RCNN [23] or R-FCN [24]. On the contrary, in bottom-up methods, certain model locates all joints of human present in the input images at one time and graphs the estimated results over the given images. Although the computational cost of top-down models increases proportionally with the number of detected human instances, they are scalable detecting diversified poses of human instances. However, bottom-up models have relative advantages that their computational cost are not related to the number of detected human instances, but, severe difficulties on detecting human pose in crowd or occlusion.

MAML: Model-Agnostic Meta-Learning (MAML) [25] is a highly investigated meta-learning algorithm for few-shot learning, achieving competitive performance on several benchmark few-shot learning problems [26, 27]. It is a part of optimization-based meta-learning algorithms, with other members in this group have various approaches on how to train the proper weights of certain classifiers. The MAML algorithm initializes parameters of neural network which can learn new tasks with very few examples in any works containing gradient-based learning rules. k-shot learning refers to k examples from each class in training neural network. This mechanism enables to achieve considerable successes without large datasets and sufficient computational resources. In MAML, there are two loops: inner loop and outer loop. While outer loop updates the model parameters, generally named meta-initialization, into proper setting for fast adaptation to new tasks, inner loop performs a few gradient steps over k samples to optimize the model suitable for separate each task.

3. Method

3.1 Meta-Learning Setup for Human Pose Estimation

We want to train a learning procedure (*i.e*., the meta-learner) that enables the HPE model (*i.e.,* learner) to adapt quickly to various domain images. For the k-shot pose estimation task, each task $i = \{L, D_{train}, D_{test}\}$ aims to estimate a certain human pose from a few (k) examples. It consists of a loss function L, a sampled small training set $D_{train} = \{(X_{train}, Y_{train})\}$ containing k images, and a test set D_{test}. For each task, the meta-learner takes D_{train} as input and produces the HPE model that makes good performance on its corresponding D_{test}.

3.2 Learner: Human Pose Estimation (HPE) Model

We use a box-free bottom-up approach for human pose estimation and instance segmentation of people in [4] as our learner. The model employs a convolutional neural network which learns to detect every keypoints and predict their relative displacements (offsets) followed by grouping them into person instances. First, we produce heatmaps (one channel per keypoint) and short-range offset vectors whose purpose is to improve the keypoint localization accuracy then aggregate them via Hough-voting making Hough score maps. The local maxima in the maps are candidate for person keypoint, but they carry no information about individual instance. Mid-range pairwise offsets are designed to connect pairs of keypoints and

group together the keypoints belonging to each instance. Furthermore, we define the long-range offsets which points from the image position x to the position of k-th keypoint of the corresponding instance to associate each person pixel identified by the instance segmentation with the keypoint detections.

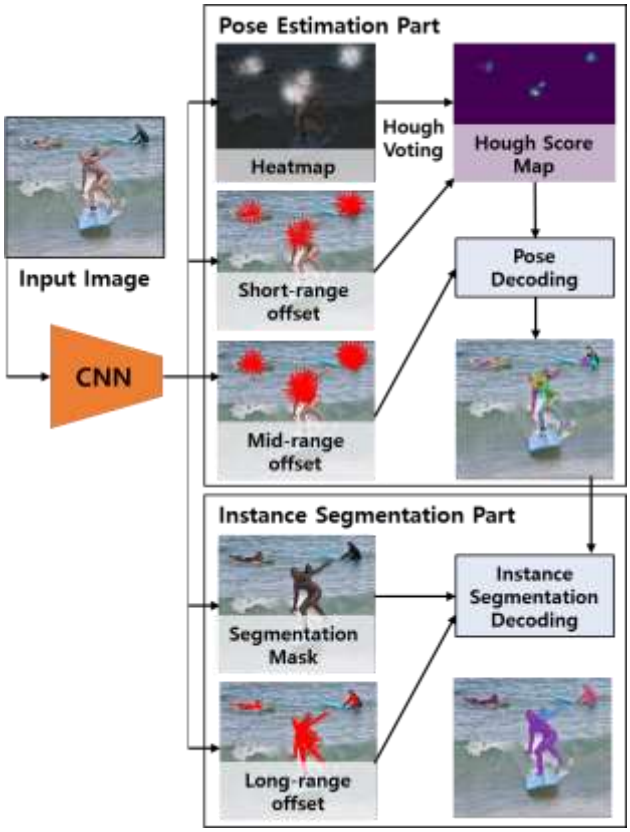

Figure 1. Human pose estimation (HPE) model that predicts: (1) keypoint heatmaps, (2) short-range offsets, (3) mid-range pairwise offsets, (4) person segmentation maps, and (5) long-range offsets. The model estimates human pose using the first three and predicts person instance segmentation masks using the latter two.

3.3 Meta-learner for HPE

For enabling HPE model to adapt quickly even with a few-shot learning, setting an initial model parameter θ is an important key. We learn initial HPE model parameter θ by using model-agnostic meta-learning (MAML). MAML is compatible with any model trained with gradient-based learning rules (*e.g.*, SGD) and aims to learn a model in a way that a few SGD step for a new task can make a rapid adaptation.

We select a set of images for training samples from human pose image in specific action. Each action is considered as a task for the meta-learner H to learn. The learner H is defined as $H(X;\theta)$, where X is input image and θ is the HPE model parameter. In each task i, the initial θ is subject to SGD update for the train set from specific task i. This procedure is called *inner optimization*:

$$\theta_i' = \theta - \alpha\nabla_\theta L(H(X_{train,i};\theta), Y_{train,i}) \quad (1)$$

where L is the loss function and $(X_{train,i}, Y_{train,i})$ is input image – keypoints label pair in the training set from task i. To evaluate whether θ_i' is generalized enough for the task i, we calculate the loss on the test set i with θ_i'. This test loss serves as the training error of the meta-learner. When summed across the tasks sampled from D, we have the meta-objective function:

$$\min G(\theta) = \min \textstyle\sum_{i\sim D} L(H(X_{test,i};\theta_i'), Y_{test,i}) \quad (2)$$

After the inner optimization of tasks, we calculate the meta-objective function $G(\theta)$ to represent overall loss of all tasks. Based on $G(\theta)$, we apply SGD to update initial weight θ. This can be formulated as:

$$\theta = \theta - \beta\nabla_\theta G(\theta) \quad (3)$$

This procedure is called *outer optimization* since this can be done after loop of the inner optimization for all tasks. During each iteration, we sample task mini-batch from specific action image datasets and perform the corresponding inner optimization in Eqn. (1) and outer optimization in Eqn. (3).

3.4. Adaptive Model Generalization

In our tasks, each task contains specific action of humans. Since the model parameter θ_i' of a new task i is updated by few SGD using training set $X_{train,i}$ during the inner optimization, θ_i' is far from the ideal parameter that would be learned from a large dataset.

In fact, we have a large set of annotated images which contains various human actions, and we sample from this original large set to generate few-shot training images. To get a sake of the training on the large dataset, we sample training data $X^*_{train,i}$ which is larger than $X_{train,i}$. $X^*_{train,i}$ contains various actions including task i's action. Thus, $X^*_{train,i}$ can be interpreted as random mini-batch of large dataset which contains task i's action.

Let θ_i' denote the model parameter learned from $X_{train,i}$ and θ_i^* denote the model parameter learned from larger set (*e.g.*, $X^*_{train,i}$). We add model regression networks (MRN) T as model adaptation and generalization strategy. We aim to make the updated θ_i' as close as the desired θ_i^*. MRN model parameter φ is trained to transform θ_i' to θ_i^* in the model parameter space, such that $\theta_i^* \approx T_\varphi(\theta_i')$. We then estimate T_φ based on large set $X^*_{train,i}$ during meta training.

$$\min \textstyle\sum_{i\sim D} \| T_\varphi(\theta_i') - \theta_i^* \|_2^2 \quad (4)$$

Algorithm 1: Meta-training for k-shot HPE

Require: Learner: HPE model $H(X;\theta)$ with input X parameters θ;
MRN adaptation network: $T(X;\varphi)$ with input X parameters φ;
Require: X: dataset over pose estimation tasks from D
Require: α: learning or meta-learning rate hyper-parameters.

Randomly initialize θ and φ
while *not done* **do**
  Sample batch of tasks $i \sim D$
  for *all* i **do**
    Learn θ_i^* from the sampled large dataset $X_{train,i}^*$ contains task i
    Sample k train pairs $X_{train,i}$, $Y_{train,i}$, from task i
    Evaluate $H(X_{train,i};\theta)$ on D_{train}
    Applying adaptation network $T: \theta_i' = T(\theta - \alpha\nabla_\theta L(H(X_{train,i};\theta), Y_{train,i}))$
    Sample $X_{test,i}$, $Y_{test,i}$ from task i for the meta-update
    Evaluate $L_{total,i} = L_i(H) + \lambda L_{T,i}(\theta', \theta^*)$ on $X_{test,i}$
  end for
  Update θ and φ by performing SGD
end while

4. Experiments

4.1 Datasets

Max Plank Institute for Informatics (MPII): MPII Human Pose Dataset [28] is one of the state-of-the-art benchmarks for evaluation of human pose estimation. Images in MPII were extracted from YouTube videos and selected which results in 25K images containing over 40K people with annotated 16 body keypoints. The dataset covers 410 human activities in a variety of human poses and provided with an activity label. We sampled data by activity as task sets for meta-training. For few-shot learning experiments to evaluate domain generalization, additional data in various fields was collected from YouTube video and processed using annotation tool to create new dataset.

Collected Dataset for Few-shot Learning: The main goal of the proposed method is to generalize the human pose estimation task to various domains with small dataset. Thus, dataset preparation from different domains is as important as the model design. We collected datasets from several different domains in which HPE can be effectively utilized. First, we gathered images with low resolutions or images taken from different camera angles such as bird's eye view. Accurate pose estimation for this kind of images can be used to detect illegal behavior in video surveillance. Second dataset collected is the images of unique postures such as yoga or dance choreography. The joints used for postures in these areas are different from common human postures in daily life. Since taking correct posture using appropriate joint or muscle is significant in sports, it will be highly useful if HPE is available. Third dataset includes the images of different luminous intensity such as thermal camera images. Applying HPE to these images can be helpfully used in fire sites or for security detection. Last dataset is collected from animation and movie. Characters in anime or movie look similar with real humans but have more exaggerated shapes in some cases. HPE can be applied to the animation or movie for elaborate motion tracking. **Figure 2** shows the sample image from each dataset.

(a) (b) (c) (d)

Figure 2. Sample image from 4 datasets: (a) dataset of different angles and resolution; (b) dataset of unique postures; (c) dataset of different luminous intensity; (d) dataset of anime and movie

4.2 Domain Adaptation Tests using Baseline Model

We implemented HPE baseline model and trained the model with COCO 2017 dataset [29], another state-of-the-art benchmarks for evaluation of human pose estimation, to test domain adaptation performance. We test whether HPE baseline model outputs the accurate result when images from other domains are used as input. The test is conducted with several images from COCO val2017 and 4 datasets we collected. **Figure 3** shows the results of key points detection, pose estimation and instance segmentation. We found that HPE baseline model lacks the generality across domains, although working accurately for dataset similar with the one used for training (COCO train2017). **Figure 3-(b)** shows that the HPE baseline model produces inaccurate result when there are many

instances in image with low resolution. It does not identify one person sitting on the right side, and estimated poses for the sitting people are imprecise while poses for standing people are reasonable. The estimated result with the image of multi-person yoga posture is noticeable. **Figure 3-(c)** shows that the model only detects 1 person and the pose is partially estimated. We point out that the model needs an improvement for the connected postures between multi-person. It also shows that the HPE baseline model often cannot estimate the pose in lower body part under waist. **Figure 3-(d)** shows that HPE baseline model can perform keypoints detection and segmentation for characters that are non-real human. However, it turns out that if the one hand is overlapped with body or hidden, model can't estimate the posture of the other hand accurately. We will apply meta-learning to improve the generality of the model across domains and overcome the limitations found in these tests.

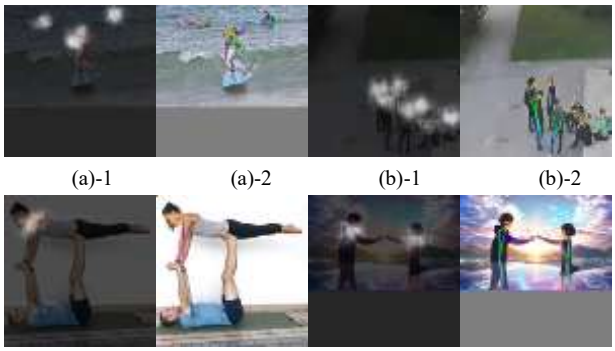

Figure 3. Demo results from PersonLab with sample images: (a)~(d)-1 Right shoulder key point map; (a)~(d)-2 Estimated pose

4.3 Meta-Learning Evaluation: Few-Shot Learning for Domain Adaptation (Examples of expected results)

	Set 1					Set 2					Set 3					Set 3				
Method/Shots	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
Mask-RCNN																				
AlphaPose																				
baseline																				
Ours*																				

Table 1. Few-shot HPE evaluation on novel datasets for few-shot learning. We report the mAP with IoU threshold 0.5(AP50) under 4 datasets with specific domains with a small number of shots. Our model overperforms the other top-down methods and HPE baseline model, showing its effectiveness for few-shot learning

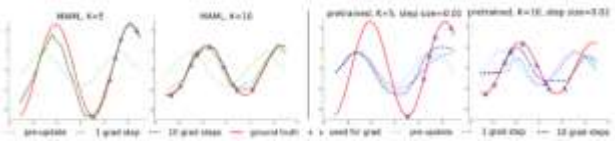

Figure 4. (Sample image from [25]) Few-shot adaptation for the HPE task. Left shows that Our model converges to ground truth as gradient step and *K* increases. Right shows that HPE baseline model cannot converge to ground truth even when gradient step and *K* increases.

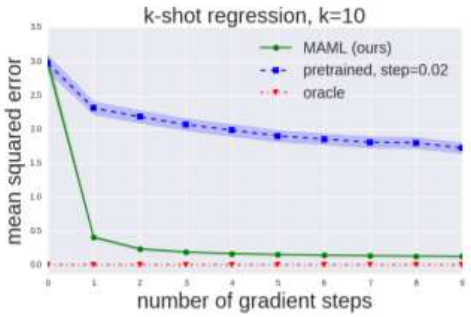

Figure 5. (Sample image from [25]) Change of the MSE by the number of gradient steps. Note that our model continues to improve with additional gradient steps with small dataset during meta-testing, achieving substantially lower loss than the HPE baseline model.

References

[1] Chen Wang, Feng Zhang, Shuzhi Sam Ge. A comprehensive survey on 2D multi-person pose estimation methods. *Engineering Applications of Artificial Intelligence*, 102: 104260, 2021

[2] Yucheng Chen, Yingli Tian, Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding (CVIU)*, 192: 102897, 2020

[3] Miniar Ben Gamra and Moulay A. Akhloufi. A review of deep learning techniques for 2D and 3D human pose estimation. *Image and Vision Computing*, 114: 104282, 2021

[4] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, Kevin Murphy. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. *European Conference on Computer Vision (ECCV)*, 2018

[5] Jinrui Zhang, Deyu Zhang, Xiaohui Xu, Fucheng Jia, Yunxin Liu, Xuanzhe Liu, Ju Ren, Yaoxue Zhang. MobiPose: real-time multi-person pose estimation on mobile devices. *Proc. of the 18th ACM SenSys*, pp. 136-149, 2020

[6] Grégory Rogez and Cordelia Schmid. MoCap-guided data augmentation for 3D pose estimation in the wild. *Conference on Neural Information Processing Systems (NeurIPS)*, 2016

[7] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio Feris, Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2018

[8] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai,

Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, Nong Sang. Adversarial semantic data augmentation for human pose estimation. *European Conference on Computer Vision (ECCV)*, 2020

[9] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, Kwang-Ting Cheng. Cascaded deep monocular 3D human pose estimation with evolutionary training data. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020

[10] Luca Schmidtke, Athanasios Vlontzos, Simon Ellershaw, Anna Lukens, Tomoki Arichi, and Bernhard Kainz1. Unsupervised human pose estimation through transforming shape templates. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021

[11] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R. Venkatesh Babu, Anirban Chakraborty. Self-Supervised 3D Human Pose Estimation via Part Guided Novel Image Synthesis. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020

[12] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, Bodo Rosenhahn. CanonPose: Self-Supervised monocular 3D human pose estimation in the wild. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021

[13] Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee. PoseFix: Model-agnostic general human pose refinement network. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019

[14] Fischler, M.A., Elschlager, R. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 1973

[15] Andriluka, M., Roth, S., Schiele, B. Pictorial structures revisited: People detection and articulated pose estimation. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2009

[16] Eichner, M., Ferrari, V. Better appearance models for pictorial structures. *The British Machine Vision Conference (BMVC)*, 2009

[17] Sapp, B., Jordan, C., B. Taskar. Adaptive pose priors for pictorial structures. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2010

[18] Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C.,Murphy, K. Towards accurate multi-person pose estimation in the wild *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[19] Huang, S., Gong, M., Tao, D. A coarse-fine network for keypoint localization. *International Conference on Computer Vision (ICCV)*, 2017

[20] He, K., Gkioxari, G., Doll´ar, P., Girshick, R. Mask r-cnn. *International Conference on Computer Vision (ICCV)*, 2017

[21] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J. Cascaded pyramid network for multi-person pose estimation. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2018

[22] Ross Girshick. R. Fast R-CNN. *International Conference on Computer Vision (ICCV)*, 2015

[23] Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Conference on Neural Information Processing Systems (NeurIPS)*, 2015

[24] Dai, J., Li, Y., He, K., Sun, J. R-FCN: Object detection via region-based fully convolutional networks. *Conference on Neural Information Processing Systems (NeurIPS)*, 2016

[25] Chelsea Finn, Pieter Abbeel, Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *Proc. of the 34th International Conference on Machine Learning (ICML), PMLR*, 70: 1126-1135, 2017

[26] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. *ICML Deep Learning Workshop,* 2015.

[27] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra. Matching Networks for one shot learning. *Conference on Neural Information Processing Systems (NeurIPS)*, 2016

[28] Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3686–3693, 2014

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision (ECCV)*, 2014