Domain Generalization in Human Pose Estimation via Meta-Learning

You Rim Choi[1], Chan Kyo Kim[2], Yejin Hwang[1], Chae Song Park[3]
[1]Graduate School of Data Science, [2]Department of Mechanical and Aerospace Engineering, [3]Program in Artificial Intelligence
{yrchoi, ckkim10, evergreen97, chase121} @ 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 an 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 first 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 key□points such as elbows and 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 body appearances of human change dynamically by change in 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 a 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, 2 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.

To solve this problem, we first applied meta-learning, which learns how to learn quickly, to pose estimation. Our overall contribution is

- By applying a meta-learning algorithm to pose estimation, we show that meta-trained model can adapt well to explicitly different domains.
- We designed dataset applicable for meta-training: various tasks were effectively constructed by appropriate augmentation.
- For the evaluation of the performance in domain generalization, we proposed novel labeled dataset of multi-person 2D pose estimation on the domain of thermal image and unique posture.

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 keypoints 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.

In contrast to top-down approaches, bottom-up approaches detect all the possible keypoints at first, and then assemble these joints into the complete poses for assigned person based on various joint assembly techniques which don't require human bounding box detection

Meta Learning: 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. There are several common approaches in meta learning such as metric based approach and optimization-based approach.

Optimization-based approach is to optimize the model parameter for fast learning. Model-Agnostic Meta-Learning (MAML) [25] is a highly investigated meta-learning algorithm for few-shot learning using optimization-based approach, achieving competitive performance on several benchmark few-shot learning problems [26, 27]. Reptile [28] which is basis of our meta-learning algorithm is a novel first-order optimization-based meta learning algorithm which avoids computational burden of MAML.

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 aims to estimate a 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. In training, we divided the data into 32 size batches after shuffling and considered one batch as one task.

3.2 Learner: Human Pose Estimation (HPE) Model

We use simple and effective model [29] which takes Top-down approach for human pose estimation as our baseline learner. When the pretrained model deploys a

Algorithm 1: Meta-training for *k*-shot HPE

Require: Learner: HPE model $H(X;\theta)$ with input X parameters θ;
Require: X: dataset over pose estimation tasks from D
Require: α: learning rate hyper-parameters for inner optimization
Require: β: hyper-parameters for parameter update

Randomly initialize θ
while *not done* **do**
  Sample batch of tasks $i \sim D$
  for *all* i **do**
    $\theta_{old} = \theta$
    for *inner step* **do**
    $\theta = \theta - \alpha\nabla_\theta L(H(X_{train,i};\theta), Y_{train,i}))$
    end for
  $\theta_{new} = \theta$
  $\theta = \theta + \beta(\theta_{new} - \theta_{old})$
  end for
end while

convolutional neural network which learned to detect person in image (e.g. person detector), the model predictes the pose of each person in detected region (Figure 1). It uses Faster- RCNN for the person detector, and this

detection part is not involved in meta-learning process. Following the common practice in top-down approaches, the location of each keypoints is estimated on the averaged heatmaps of the original and augmented image. And then a quarter offset in the direction from highest response to the second highest response is used to obtain the final location of human body joints. The model uses resnet50 as backbone, and adds few deconvolutional layers over the last feature map to use up-sampling to increase the feature map resolution.

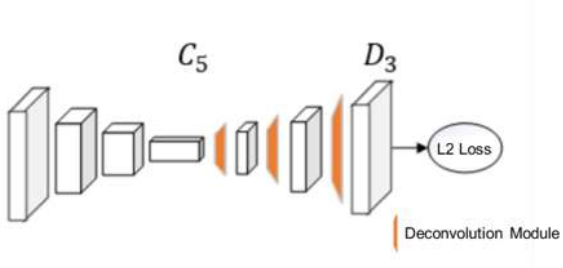

Figure 1. The proposed framework of our baseline model (learner) for human pose estimation

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 meta learning algorithm. Our meta learning algorithm 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 various action. After shuffle, each training samples are 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*. After the inner optimization for the *i th task*, we have a new model learned approximately for the *i* th task, called θ_i'. Our meta learning algorithm updates the parameters of the model using the difference between θ_i' and θ.

Since θ_i' is learned more biasedly to a specific task *i* than θ, $\theta_i' - \theta$ term is added to θ for all *i* o that B learns evenly to all tasks. This can be formulated as:

$$\theta = \theta + \beta(\theta_{new} - \theta_{old}) \quad (2)$$

Since each batch is defined as a task, inner optimization and parameter updates are performed for all batches. When meta-training is over, meta-testing is performed with a few-shot dataset which is in the new domain.

4. Experiments

4.1 Datasets

COCO Dataset: COCO dataset [30] is a comprehensive dataset for object detection, instance segmentation and keypoints detection. COCO 2017 Keypoints dataset is one of the stae-of-the-art benchmarks for evaluation of human pose estimation. It includes more than 200K images containing over 250K person instance labeled with 17 body keypoints. The COCO keypoints evaluation metric defines the object keypoint similarity (OKS) which plays the same role as the IoU in object detection and uses mean average precision (mAP) over OKS [31].

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 two different domains in which HPE can be effectively utilized.

Figure 2. Examples of data augmentation for meta training

First, we gathered images of unique postures such as yoga or dance choreography (UniquePose Dataset). The joints used for postures in these areas are different from

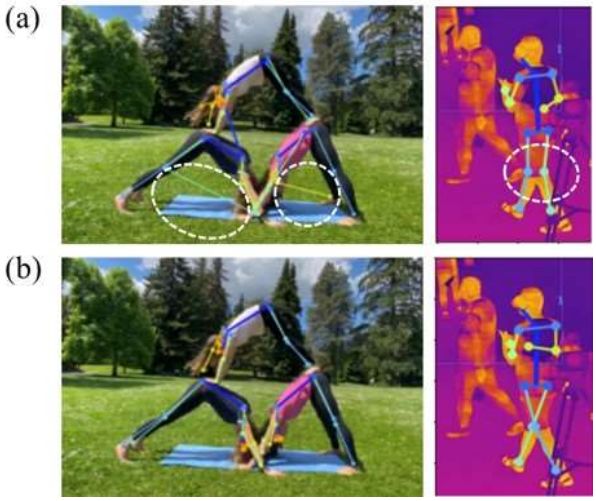

Figure 3. The results of key points detection and pose estimation

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. Second dataset includes the images of different luminous intensity such as thermal camera images (Thermal Dataset). Applying HPE to these images can be helpfully used in fire sites or for security detection.

4.2 Meta Training

We trained the learner in a general gradient descent-based algorithm using COCO 2017 Keypoints dataset as HPE baseline model for comparison with our model. On COCO 2017 Keypoints valid dataset, the HPE baseline model achieves 74.3 of mAP with ground truth bounding boxes.

As our model, we implemented HPE baseline model with meta learning algorithm and trained the model with COCO 2017 dataset to test domain adaptation performance. To be more robust to noise or domain deviation, we supplemented the dataset in three augmentation methods (Figure 2). Considering that the position of the keypoints can be changed to a very free form in a human posture, augmentation was performed with three transforms using affine transformation, elastic transformation, perspective transformation.

Prior to entering the model, the augmented and shuffled datasets are divided into 32 fixed batch sizes, and each batch functions as a task. When the inner optimization for each task is completed, the parameters are updated, and finally, we get a model that becomes a good initial value for any task. Since the parameters of the current task are

Dataset	Model	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
UniquePose *K*=32	HPE Baseline Model	0.51	0.816	0.523	0.616	0.889	0.667
	Our Model	0.546	0.849	0.6	0.636	0.911	0.711
Thermal *K*=16	HPE Baseline Model	0.675	0.896	0.784	0.730	0.919	0.838
	Our Model	0.681	0.948	0.739	0.746	0.973	0.811

Table 1. Comparison of few-shot learning (meta-test) on UniquePose and Thermal dataset of HPE baseline model and Our model. Our model shows slightly higher AP and AR.

updated in a way that adds the parameter difference between the previous task and the current task, the weight, called meta step size, multiplied by the difference between the two parameters is also an important hyperparameter. We set the meta step size to 0.05. Since our dataset, including three augmentations, was 480,000, the number of our tasks, that is, the number of batches, is about 15000. Therefore, our model has been optimized over 15,000 inner optimizations.

4.3 Evaluation: Few-Shot Learning for Domain Adaptation

First, we test whether the HPE baseline model outputs the accurate pose estimation results when images from

explicitly different domains are used as input data.

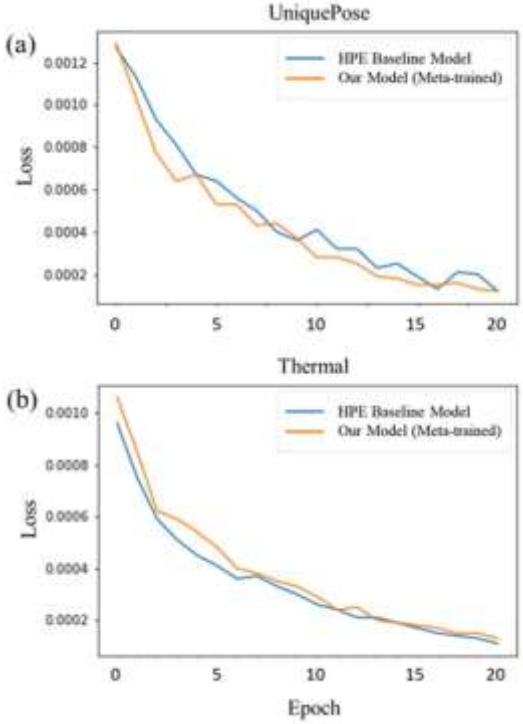

Figure 4. Train loss during meta test

The test is conducted with several images from two datasets we collected. Figure 3 shows the results of key points detection and pose estimation. We found that HPE baseline model lacks the generality across domains, although working accurately for dataset similar with the one used for training.

The goal of our experimental evaluation is to answer the following question: Can meta-trained learner model enable fast learning of new domain task? We evaluate performance by meta-test (fine-tuning) the HPE baseline model and our model on K = {16 for UniquePose dataset, 32 for Thermal dataset} datapoints.

We present the results in Table 1. After training 20 epochs on each dataset, we tested 25 datapoints for evaluation. Our model shows slightly higher AP and average recall (AR). Figure 3 (b) is the result of a well-corrected error by meta-tested our model.

The train loss during meta test is shown in Figure 4 to evaluate how fast the model converges. Although there is not much difference, but in case of (a) UniquePose, the train loss of our model decreased faster. In a relatively short epoch, both models were trained enough to show about 98% of the train accuracy. However, there seems to be an overfitting problem due to small dataset, so further experiments are needed.

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] Nichol, Alex, et al. "On First-Order Meta Learning Algorithms." ArXiv Preprint ArXiv:1803.02999, 2018

[29] Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In Proceedings of the European conference on computer vision (ECCV) (pp. 466- 481).

[30] 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

[31] COCO: COCO Leader Board. http://cocodataset.org