Meta Agent Teaming Active Learning for Pose Estimation

Jia Gong[1] Zhipeng Fan[2] Qiuhong Ke[3] Hossein Rahmani[4] Jun Liu[1*]
[1]Singapore University of Technology and Design, Singapore; [2]New York University, United States
[3]The University of Melbourne, Australia; [4]Lancaster University, United Kingdom
jia_gong@mymail.sutd.edu.sg, zf606@nyu.edu, qiuhong.ke@unimelb.edu.au
h.rahmani@lancaster.ac.uk, jun_liu@sutd.edu.sg

Abstract

The existing pose estimation approaches often require a large number of annotated images to attain good estimation performance, which are laborious to acquire. To reduce the human efforts on pose annotations, we propose a novel Meta Agent Teaming Active Learning (MATAL) framework to actively select and label informative images for effective learning. Our MATAL formulates the image selection procedure as a Markov Decision Process and learns an optimal sampling policy that directly maximizes the performance of the pose estimator based on the reward. Our framework consists of a novel state-action representation as well as a multi-agent team to enable batch sampling in the active learning procedure. The framework could be effectively optimized via Meta-Optimization to accelerate the adaptation to the gradually expanded labeled data during deployment. Finally, we show experimental results on both human hand and body pose estimation benchmark datasets and demonstrate that our method significantly outperforms all baselines continuously under the same amount of annotation budget. Moreover, to obtain similar pose estimation accuracy, our MATAL framework can save around 40% labeling efforts on average compared to state-of-the-art active learning frameworks.

1. Introduction

Human hand (or body) pose estimation, aiming to localize the positions of specific key points in images, is an important task that has a wide range of applications such as augmented reality [11], sign language translation [21], and human-robot interaction [40]. Despite the great success of existing deep learning based pose estimation methods [63, 2, 15, 59, 19, 10, 56], they are notoriously data-hungry. Furthermore, acquiring pose annotation is often very expensive and time-consuming, e.g., annotating a single image in MPII dataset [1] takes around 40 seconds, which limits the development of large-scale datasets. Accordingly, with the limited scale of the dataset, it is essential to develop algorithms to use data more efficiently.

Active learning (AL), which proactively selects the most informative unlabeled images to annotate, is one promising solution to this problem. Recent active learning-based pose estimation frameworks [38, 58, 4, 5, 22] can be categorized into uncertainty-based or distribution-based methods. The uncertainty-based methods [22, 58, 38] query annotations for the samples with the lowest confidence scores. However, as shown in [24], neural networks tend to be over-confident with unfamiliar samples, leading to overestimated model performance and therefore lowering the labeling efficiency. Meanwhile, the distribution-based methods [35, 4] aim to query annotations for representative images from the unlabeled dataset. However, the most representative images w.r.t. the unlabeled set may not always be the most informative ones to the pose estimator, as the estimator may have already learned similar knowledge from earlier samples. As the result, for both types of methods, their image selection strategy does not directly relate to the improvements of the pose estimator, leading to suboptimal performance.

Moreover, these methods suffer in the batch setting, where the active learning algorithm selects multiple images for annotation in one turn. Existing traditional methods [22, 58] rely on selecting the most informative or representative images to construct a batch, disregarding the redundancies in the formed batch. Recently, several works [4, 35] explore the usage of distance-based clustering to identify unique images yet maintain good coverage of the dataset. However, the adopted clustering algorithms tend to be less effective in the high-dimensional space, leading to less effective sample selection processes during AL iterations [34]. Therefore, it is important to construct a batch of samples for annotation in an intelligent way, taking care of both the informativeness of each individual image and the overall diversity of the batch.

* Corresponding Author

To address the aforementioned issues in a single end-to-end learning framework, we propose a novel Meta Agent Teaming Active Learning (MATAL) model for human hand (or body) pose estimation, which leverages an agent team to learn a teaming sampling policy from data. Our main insight is that selecting a batch of informative yet diverse images for annotation can be viewed as a teamwork of a set of agents, where each agent in the team selects one image collaboratively based on the other agents' decisions. Then this active learning procedure can be formulated as a Markov Decision Process (MDP) [45], which could be solved with Reinforcement Learning (RL). The agent team receives a state signal characterizing the distribution of the images in the dataset and cooperatively generates a batch of actions to decide which images should be labeled. To help the agent team to identify informative samples for annotation, we introduce a novel state-action representation by leveraging Kinetic Chain Space (KCS) to encode the topological information of the hand (or body) pose. Finally, as the labeled dataset will expand with the new annotated data, we train our model via meta-learning to facilitate fast adaptation to the iteratively enlarged labeled dataset.

In summary, our main contributions are: **1)** We formulate the pose estimation active learning procedure as a Markov Decision Process (MDP) and develop a Reinforcement Learning (RL) based framework for effective sample selection. **2)** To help the learning of the agents, we propose a state-action representation to characterize the informativeness and representativeness of the samples. **3)** We validate the efficacy of the proposed MATAL framework on both human hand and body pose benchmarks.

2. Related Work

Pose Estimation. Below we briefly review the recent pose estimation methods. More works can be found in [7, 18].

Several approaches [55, 61, 32, 41, 30, 42, 20, 62, 8] have investigated the usage of deep learning to predict hand poses from depth or RGB-D images. These methods employed heatmap [52], pose structure information [44] or hand's shape information [26] to improve the performance. More recent works [63, 29, 17, 64] derived the hand joints' poses from RGB inputs. Similarly, recent human body pose estimation approaches [49, 61, 27, 43, 31, 10] focused more on deriving body joints' poses from RGB images. The state-of-the-art Stacked Hourglass [56] employed an encoder-decoder structure to predict joints' locations as heatmaps, while the HRNet [43] maintained high resolution representations through the process to better localize the joints. Our framework does not assume a specific architecture for the pose estimator and could be used with various existing models to improve their annotation efficiencies.

To reduce the need of labeled data, learning methods with less supervision signal, such as weakly-supervised learning [28, 23, 8], semi-supervised learning [39, 3, 54] and self-supervised learning [9, 51], have attracted much attention recently. These methods utilize the unlabeled data to improve the performance. However, most of the methods still rely on the help of labeled data to distill useful information from the unlabeled images. This means that the quality and informativeness of labeled data are still crucial in their methods. Our active learning approach is parallel to these methods and could be integrated into the labeled data collection process to significantly reduce the annotation cost.

Active Learning for Pose Estimation. Active learning is an important machine learning problem, which has received lots of attentions [35, 58, 6, 22]. In recent years, several works explored applications of active learning for pose estimation. Liu et al. [22] introduced an uncertainty based estimator, utilizing the entropy of the predicted heatmaps to select the informative images. Yoo et al. [58] proposed a loss prediction module, which is learned together with the target model to predict the losses of unlabeled samples. A subset of unlabeled samples with high predicted loss values is selected for annotation. Shukla et al. [38] extended [58] to improve the correlation between the predicted and true loss values. The work in [4] used the Bayesian uncertainty to estimate the confidence of the pose estimator's prediction and combined this with Core-set sampling [35] to perform selection. Caramalau et al. [5] employed Graph Convolutional Networks (GCN) to model the relation between labeled and unlabeled data. They then proposed two GCN-based sampling approaches based on uncertainty and distribution, respectively. Though these methods have achieved increasingly accurate measurements for uncertainty or distribution of the images, their sampling policies are not directly related to the performance of the pose estimator, leading to limited performance improvement. We address this by learning a sampling policy driven by the reward that directly relates to the performance of the pose estimator. To the best of our knowledge, we are the first *Active Learning-based multi-agent* framework to learn a *batch sampling* policy that promotes the learning of the pose estimator.

Reinforcement Learning in Pose Estimation. Reinforcement learning (RL), which is a learning paradigm to solve MDP problems, aims to learn a policy that takes actions to maximize the accumulated reward in MDP [25, 45, 57, 50]. Recently, several works [33, 13] explored different applications of RL in pose estimation tasks. Jianzhun et al. [36] used RL to learn to manipulate the 3D object to match the ground truth mask. Another work [14] considered the multi-camera settings in the human body pose estimation task and leveraged an RL model to select the appropriate viewpoints (or cameras) to improve the performance of the pose estimator. Both of them involved RL into the pose estimation procedure, however, with completely different formulations to ours. Instead of employing RL to directly

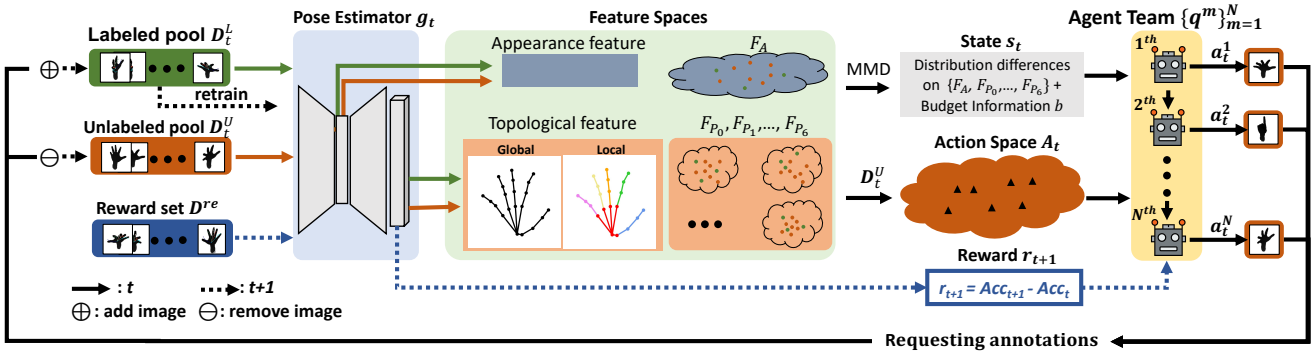


Figure 1. Overview of our MATAL framework for hand pose estimation (MATAL for human body pose estimation shares the similar structure). The solid lines describe the data flow at the t^{th} active learning iteration and dot lines are that of the $(t+1)^{th}$ iteration. Given a labeled sample pool D_t^L and an unlabeled sample pool D_t^U, our active learning framework works as follows: **1)** We first project both the D_t^U and D_t^L to the feature spaces with the pose estimator g_t, then construct the state s_t and action space A_t from the feature spaces. The **state** s_t records the differences between D_t^U and D_t^L in the feature spaces, and the consumption of annotation budget. The **action space** A_t contains the projection of D_t^U in the feature spaces. Each action $a_t \in A_t$ corresponds to a unique image in D_t^U and describes the novelty, representativeness and appearance of the image. **2)** The **agent team** follows the Q-learning [45] framework and evaluates the state-action pair (s_t, a_t) to determine a set of actions $\{a_t^m\}_{m=1}^N$ of raising corresponding images for annotation. **3)** We then update D_{t+1}^U and D_{t+1}^L by moving new annotated images from D_t^U to D_t^L. The pose estimator is retrained on D_{t+1}^L to obtain g_{t+1}. **4)** The **reward** r_{t+1}, which measures the improvement of pose estimator's prediction accuracy on D^{re} as $Acc_{t+1} - Acc_t$, is used to optimize the agent team.

solve pose/camera parameters, we address the task of active learning for annotating selective informative samples under a specific annotation budget, and design a state-action representation with a novel meta agent teaming framework to enable effective batch sampling.

3. Method

Given an unlabeled human hand (or body) dataset with a limited annotation budget, the goal of active learning (AL) is to annotate the most informative images iteratively to maximize the performance of the target pose estimator. We introduce a novel AL framework for human hand (or body) pose estimation, which leverages an agent team to raise a batch of informative images at each active learning iteration as shown in Fig. 1.

In this section, we first show how AL for pose estimation can be formulated as a Markov Decision Process (MDP) (Sec. 3.1). Then we present our cooperative multi-agent framework to perform effective batch selection and introduce a compact representation to facilitate the cooperation between agents (Sec. 3.2). Finally, we introduce the training and deployment pipelines as well as a meta-optimization algorithm, which facilitates the agents' quick adaptation to the enlarged labeled set in AL procedures during deployment (Sec. 3.3).

3.1. Active Pose Estimation as MDP

Existing AL algorithms [38, 58, 4, 5, 22] fall into the paradigm of iteratively selecting a batch of images to label until the annotation budget B runs out. In the t^{th} iteration, given an unlabeled set D_t^U, a labeled set D_t^L and a pose estimator g_t, these AL algorithms take the following steps: (1) Evaluate the informativeness of each image in D_t^U; (2) Select a batch of informative images to query annotation; (3) Move the selected images from D_t^U to D_t^L then retrain the pose estimator g_t on the updated labeled dataset D_{t+1}^L to obtain g_{t+1}.

In this paper, we aim at learning an optimal sampling strategy that directly maximizes the performance of the target pose estimator under a fixed annotation budget, driven by maximizing the designed reward. To ease the understanding, we assume there is a single agent to propose a single image for annotation in this section. In Sec. 3.2, we further discuss the image batch selection by multiple agents. We formulate the AL steps as a MDP: $(s_t, a_t, r_{t+1}, s_{t+1})$ and convert the key AL steps as: (1) Estimate the state s_t which characterizes the distribution difference between the unlabeled set D_t^U and the labeled set D_t^L at the t^{th} iteration. (2) Evaluate each state-action pair (s_t, a_t) to determine an image to be annotated. (3) Update D_t^L, D_t^U to D_{t+1}^L, D_{t+1}^U by moving newly annotated image from D_t^U to D_t^L. Retrain g_t on the updated D_{t+1}^L to obtain g_{t+1} and update the state to s_{t+1} based on D_{t+1}^L and D_{t+1}^U. (4) Compute the reward r_{t+1} based on g_{t+1} and g_t evaluated on a separatedly reserved reward set D^{re} to update the agent.

We adopt the Q-learning algorithm [45] to solve this MDP problem, in which the agent scores each state-action representation pair (s_t, a_t) and takes the action a_t with the highest score (i.e., the Q-value). By deriving reward from the improvement of the pose estimator directly, we can optimize the agent to learn a policy that maximizes the reward as well as the performance of the pose estimator. Below we elaborate on the detailed definition of state s_t, action a_t, and reward r_t.

State. Intuitively, the state s_t should capture the distribution gap between the labeled dataset D_t^L and the unlabeled dataset D_t^U, which helps the agent to pick out the most informative image that could compensate the distribution shift between D_t^L and D_t^U. With an unbiased training set distribution, the pose estimator is more likely to generalize well to unseen cases. Specifically, in pose estimation, we consider two key attributes to characterize the distribution drifts: appearance variation and pose topological variation, which are also key considerations when collecting pose estimation datasets [60].

Based on these intuitions, we propose to collect two kinds of cues including the appearance information and topological information to characterize the distribution difference between D_t^L and D_t^U. Note that the difference is dynamical as it depends on the pose estimator g_t. The design of the state helps the agent to select appropriate samples for the pose estimator g_t during the active learning process.

For the appearance information f_A of the sample x, we collect the average pooled feature from an intermediate layer of the pose estimator g_t, as shown in Fig. 1. This feature depicts the general look of the image sample x.

For the topological information, we encode the topological features such as the bone length and bone rotations via Kinetic Chain Space (KCS) [53, 16]. More precisely, we derive M bone vectors from the estimated pose $\hat{y} = g_t(x)$ and concatenate them to form an $M \times n$ matrix, where n is the dimension of the joint coordinates. Then the KCS is computed as the inner product of the matrix and its transpose. We denote the KCS for all bone vectors of the whole hand (or whole body) as the global topological feature f_{P_0}. Moreover, the performance of pose estimator varies with each joint [56], leading to different pose estimation qualities over various local joints of the hand (or body). To help the pose estimator to achieve good performance on each joint, we additionally track the properties for the local parts of the hand (or body). We decompose the whole hand (or whole body) to six local parts including the palm and five fingers (torso, head, left/right arm, and left/right leg for body). We then compute the KCS for these parts as the local topological features $\{f_{P_1}, f_{P_2}, ..., f_{P_6}\}$ of the image x.

In this way, we extract the appearance feature f_A and the topological features $\{f_{P_0}, f_{P_1}, ..., f_{P_6}\}$ for the image x. Then the appearance features of all data in the labeled and unlabeled datasets form the appearance feature space F_A. Similarly, we can build the topological feature spaces $\{F_{P_0}, F_{P_1}, ..., F_{P_6}\}$. To model the distribution drifts between the labeled dataset D_t^L and the unlabeled dataset D_t^U, we regard the labeled and unlabeled datasets as two domains and measure the domain gap between them. Specifically, we adopt the Maximum Mean Discrepancy (MMD) [47] and compute the gap for each feature space S, where $S \in \{F_A, F_{P_0}, F_{P_1}, ..., F_{P_6}\}$, via MMD as:

$$K_S = \text{MMD}(S^L, S^U) = \sum_{i=1}^{n_L}\sum_{j=1}^{n_L}\frac{k(p_i,p_j)}{n_L^2} + \sum_{i=1}^{n_U}\sum_{j=1}^{n_U}\frac{k(q_i,q_j)}{n_U^2} - \sum_{i=1}^{n_L}\sum_{j=1}^{n_U}\frac{2*k(p_i,q_j)}{n_L n_U}, \tag{1}$$

where S^L and S^U are the distributions of S on D_t^L and D_t^U respectively, and K_S is a scalar representing the distribution difference between S^L and S^U. We denote the samples in S^L and S^U as p and q. n_L and n_U are the numbers of samples in D_t^L and D_t^U, and $k(\cdot)$ corresponds to the radial kernel [47] to measure the distance between two samples.

Moreover, the available budget is another piece of important information for the agent to perform an effective selection. Here, we use the budget consumption ratio b to represent this status. Finally, the state s_t is defined as: $\{K_{F_A}, K_{F_{P_0}}, K_{F_{P_1}}, ..., K_{F_{P_6}}, b\}$, which encodes the distribution drifts between the labeled and unlabeled sets as well as the available budget. It guides the agent to determine which kind of images could benefit the pose estimator most.

Action. The action should ideally captures the potential contribution of a specific unlabeled sample when adding it to the labeled set D_t^L. Intuitively, combining the state and action representations, the agent should have enough information to score each unlabeled sample and select an informative image from the unlabeled set D_t^U to query annotation. To this end, we associate each action a_t in the action space A_t with a unique image x in the unlabeled pool D_t^U.

To assist the selection of the informative sample, we compute three kinds of features from each unlabeled image x: 1) the novelty of the pose in the image x; 2) the representativeness of the image for the unlabeled pool; 3) the general appearance information of the image. Intuitively, these three features characterize the informativeness, the representativeness of the pose as well as the appearance of an unlabeled image x. We detail each representation below.

The novelty of the image helps estimate the potential performance gain brought by adding accurate annotation. However, it is hard to measure without the actual ground truth pose. Therefore, we propose to approximately evaluate it by utilizing the topological features from the labeled set D_t^L. Intuitively, the closeness of the global/local topological information indicates the similarities between the whole/local part of the estimated pose and the ground truth pose. A novel pose will likely have low similarities to any pose from the labeled set D_t^L. Therefore, we compute the maximum cosine similarity between the unlabeled image x and the labeled set D_t^L individually on each topological feature space $\{F_{P_0}, F_{P_1}, ..., F_{P_6}\}$ as $\{s_0, s_1, ..., s_6\}$, and consider it as a proxy for the pose novelty.

We then introduce our parameterization for the representativeness of the sample. The labeled set D_t^L and un-

labeled set D_t^U jointly describe the distribution of the data. Therefore it is also important to sample representative images w.r.t. the unlabeled set D_t^U, which could be characterized by the distribution of the similarity scores. We introduce a histogram-based representation d to record the cosine similarity distribution between x and D_t^U on each topological feature space as $\{d_0, d_1, ..., d_6\}$. Combining with the parameters $\{s_0, s_1, ..., s_6\}$ representing similarity of x to D_t^L, the agent could avoid repeatedly sampling the representative images that our pose estimator has already learned from, leading to improved sampling efficiency.

Finally, we extract the image appearance feature f_A of the unlabeled image x as its appearance property (e.g., clothes texture, skin color, background, etc). The final action representation a_t corresponding to the unlabeled image x is the combination of these features: $a_t = \{s_0, s_1, ..., s_6, d_0, d_1, ..., d_6, f_A\}$, enabling the agent to effectively identify the informativeness of the unlabeled image x and perform selection.

Reward. The reward is a metric that evaluates how much the selected unlabeled image can benefit the target pose model g_t. We reserve a specific subset D^{re} for accurate reward estimation before starting the active learning procedure. Then, we measure the accuracy of the pose estimator on this reward set D^{re}, and the reward r_{t+1} is defined as the difference of the accuracy between g_{t+1} and g_t, as shown in Fig. 1. Note D^{re} is only used for evaluation and not used in any training process of the pose estimator. With the reward r_{t+1}, we can optimize the agent to select the most informative image to maximize the reward, leading to improved pose estimation accuracy during each AL iteration.

3.2. Teaming Sampling Policy Learning

Sampling a single unlabeled image in each active learning iteration to query annotation is inefficient for two major reasons [35]: (1) The performance gain brought by a single sample is often hard to measure; (2) The pose estimator needs to be retrained more frequently due to more iterations involved. To address these issues, the most recent methods [38, 4, 5, 22] often query annotations for a batch of samples at each active learning iteration.

However, it is less trivial to perform batch sampling in our proposed framework. Using a single agent to generate a batch of samples by raising images with high predicted scores (Q-values) disregards the redundancies within the batch, leading to inferior performance as shown in Sec. 4.3.

Therefore, we further introduce a cooperative agent team module, consisting of a set of agents, working collaboratively to select image batch effectively and efficiently. Specifically, the agents in the team sequentially select samples for annotations, and each agent can observe the previous agents' actions to perform the selection cooperatively. For the m^{th} agent in a N-agent team, we denote its policy network as q^m, and the action performed by it as a_t^m. Then, to model the sequential cooperation between agents, we can additionally provide the m^{th} agent with the actions $\{a_t^i\}_{i=1}^{m-1}$ of the previous $m-1$ agents. However, it will require an increasingly deep and wide neural network to process the information of $\{a_t^i\}_{i=1}^{m-1}$ with a large m, leading to undesired high computational complexity. To address this, we use the the expectation of $\{a_t^i\}_{i=1}^{m-1}$, a fixed-length compact representation of the previous agents' actions, as an extra state for the m^{th} agent. Mathematically, the expectation of previous agents' actions h_t^m is computed as:

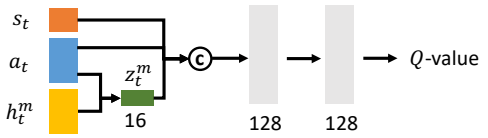


Figure 2. The architecture of the m^{th} agent in the team. Note that each agent shares the similar model architecture but with its own parameters. a_t and h_t^m are first fed into a linear layer with ReLU activation to generate feature $z_t^m \in \mathbb{R}^{16}$. z_t^m, a_t and s_t are then concatenated and passed through three linear layers with ReLU activation in between to output the Q-value.

$$h_t^m = \frac{1}{m-1} \sum_{i=1}^{m-1} a_t^i, \tag{2}$$

and then the action made by the m^{th} agent becomes:

$$a_t^m = \underset{a_t \in A_t}{argmax}\ q^m(s_t, a_t, h_t^m; \theta_m), \tag{3}$$

where a_t^m is the action selected by the m^{th} agent q^m, which is parameterized by θ_m, and $a_t \in A_t$ is the candidate action, which forms into the state-action pair (s_t, a_t) to be evaluated by the agent q^m. The structure of the m^{th} agent is depicted in Fig. 2.

Finally, we build our agent team module as $\{q^m\}_{m=1}^N$, where N is the number of agents in the team. To train the agent team, we follow the Double DQN formulation [50] to optimize our agent team by minimizing the temporal difference (TD) error as:

$$\begin{aligned} TD(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \Big(& \sum_{m=1}^{N} q^m(s_t, a_t^m, h_t^m; \theta_m) - r_{t+1} \\ & - \gamma \sum_{m=1}^{N} q^m(s_{t+1}, a_{t+1}^m, h_{t+1}^m; \hat{\theta}_m) \Big)^2, \end{aligned} \tag{4}$$

where θ_m is the parameters of the m^{th} agent's policy network, $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots \theta_N\}$ is the parameter set of all the agents in the team, $\hat{\boldsymbol{\theta}}$ denotes the parameters of the off-policy network, used to keep the learned Q-value, and periodically updates itself with $\boldsymbol{\theta}$, following the setting of Double DQN [50]. Via such a cooperative mechanism, the agent team performs batch sampling in each iteration effectively.

Algorithm 1: Teaming Sampling Policy Learning

Input: agent team $\{q^m\}_{m=1}^N$, an initial pose estimator g_{init}, an initial set D_{init} with annotation and image batch size N

$D_{init}^L, D_{init}^U, D^{re} \leftarrow \text{RandomPartition}(D_{init})$
while *not done* **do** // Episodes training
  $D_0^L \leftarrow D_{init}^L, D_0^U \leftarrow D_{init}^U, g_0 \leftarrow \text{UPDATE}(g_{init}, D_0^L)$
  for $t = 0$ **to** $T-1$ **do** // AL procedure
    Build the state s_t and action space A_t (Sec. 3.1)
    Use the agent team $\{q^m\}_{m=1}^N$ to select images following Eq. 3: $\{x_m\}_{m=1}^N \Leftarrow \{a_t^m\}_{m=1}^N$
    Annotate data: $\{(x_m, y_m)\}_{m=1}^N \leftarrow \{x_m\}_{m=1}^N$
    Update D_t^U, D_t^L and g_t :
    $D_{t+1}^L \leftarrow D_t^L \cup \{(x_m, y_m)\}_{m=1}^N, D_{t+1}^U \leftarrow D_t^U \setminus \{x_m\}_{m=1}^N, g_{t+1} \leftarrow \text{UPDATE}(g_t, D_{t+1}^L)$
    Compute reward on D^{re}:
    $r_{t+1} = Acc(g_{t+1}) - Acc(g_t)$
  end
  Update $\{q^m\}_{m=1}^N$ following Eq. 4
end

3.3. Model Training with Meta Optimization

With the introduced RL for AL formulation in Sec. 3.1 and the agent teaming framework in Sec. 3.2, we introduce the training and deployment pipelines in this section.

Given an unlabeled dataset D_{full} and an annotation budget B, our MATAL pipeline works as follows. We first randomly sample an initial subset D_{init} to request annotations. With the labeled initial subset D_{init}, we further partition it to simulate the AL procedure and train our agent team $\{q^m\}_{m=1}^N$. Specifically, we partition the labeled initial set D_{init} into the labeled set D_{init}^L, the unlabeled set D_{init}^U, and the reward set D^{re}, and then have our agent team to play the active batch image selection game following Sec. 3.1 and Sec. 3.2. The detailed process is illustrated in Alg. 1. We denote this phase of training the agent team on the initial labeled set as **Training Phase**.

Furthermore, once our agent team is trained on D_{init}, it could be *deployed* to execute the real active learning procedure on the rest of the unlabeled pool $D^U = D_{full} \backslash D_{init}$, until the budget B ran out. We denote this phase as **Deployment Phase**, in which the agent team proposes batch samples $\{x^m\}_{m=1}^N$ for annotation from D^U at each iteration and expands the labeled pool $D^L = D^L \cup \{x^m\}_{m=1}^N$ to update the pose estimator g. We set $D^L = D_{init}$ at the start of this phase and expand it in the Deployment Phase.

With the enlarged labeled set D^L, we can then retrain our agent team on it to improve the performance of the RL agent team, again following Alg. 1. Note that we set D_{init} in Alg. 1 to the most up-to-date D^L each time when we perform retraining in this Deployment Phase.

However, the training of the agent team $\{q^m\}_{m=1}^N$ on the expanded labeled set D^L could be time-consuming due to the growing size of D^L. To reduce the time complexity, we further propose a Meta-Learning based extension of the Alg. 1. Inspired by MAML [12], we consider each retraining process as a task and leverage the Meta-Learning [12] to learn a good initialization for the policy network parameters that could quickly adapt to the new tasks of retraining on the enlarged dataset. We adopt this Meta-Learning based extension in the Training Phase and empirically show that we could reduce the multi-agent team update cost by a half without sacrificing the performance, as shown in Sec. 4.3.

4. Experiment

We conduct extensive experiments on both the human hand and body pose datasets to evaluate the effectiveness of our proposed MATAL framework.

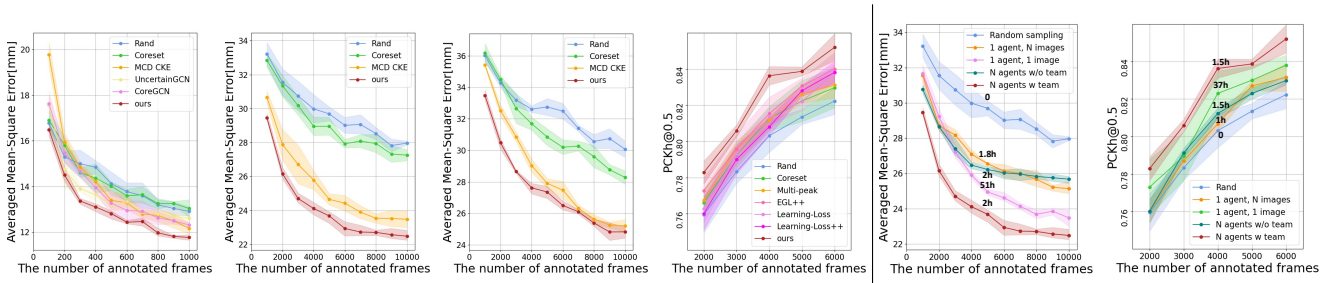
For human hand pose estimation, we follow the experimental settings of [5] and evaluate the performance of MATAL on three widely used datasets, ICVL [46], NYU [48] and BigHand2.2M [60]. ICVL is a depth-based hand image dataset and NYU is a larger RGB-D dataset collected by multiple cameras. Furthermore, to evaluate the efficacy of our method on large-scale datasets, we set up experiments on BigHand2.2M [60], which contains around 2.2 million images collected from ten different subjects. For human body pose estimation, we use MPII [60], which is an RGB dataset widely used in recent works.

4.1. MATAL on Human Hand Pose Estimation

Baseline. We compare the performance of our MATAL on hand pose estimation task with random sampling as well as existing state-of-the-art methods, including Coreset [35], MCD CKE [4], UncertrainGCN [5] and CoreGCN [5], based on their reported results on each dataset.

Implementation Details. Following [5], we use DeepPrior [32] as the backbone of our pose estimator. We extract the feature map from the last convolutional layer of DeepPrior, and perform average pooling by a 5×5 kernel with stride 3, followed by flattening to generate a 128-D appearance feature vector. We use the 21 joints estimated by DeepPrior and compute a 275-D topological feature vector. We use 40 agents to build the agent team for image batch selection on NYU and BigHand2.2M, and use 4 agents for ICVL as it is much smaller than other datasets.

For each dataset, we first randomly sample a small number of images from the training set of the dataset to build the initial set D_{init} and the remaining images form the unlabeled set D^U. The sizes of D_{init} in ICVL, NYU, and BigHand2.2m datasets are 80, 800, and 800, respectively. Then we train our MATAL on D_{init} via Alg. 1, in which the D_{init} is split into three disjoint sets D^{re}, D_{init}^U and D_{init}^L with the ratio of 3:6:1. Later, we deploy the trained MATAL to sample the images from D^U and initialize D^L as D_{init}. In the Deployment Phase, the agent team is frozen to sample informative image batches iteratively while the pose estimator is updated every time a newly annotated batch arrives.



(a) ICVL [46] (b) NYU [48] (c) BigHand [60] (d) MPII [1] (e) NYU [48] (f) MPII [60]

Figure 3. (a)-(d): Active learning results of pose estimation over four datasets. The results of (a) ICVL (b) NYU (c) BigHand are for hand pose estimation and the curves in these figures show the average mean-square error of the joints' poses (lower is better) over different numbers of annotated frames. The result of human body pose estimation on MPII dataset is presented in the sub-figure (d), where the metric is the PCKh@0.5 (higher is better). (e)-(f): Ablation study for agent team on human hand and body benchmarks.

Each time the size of the labeled dataset D^L doubles compared to the previous time the agent team was trained, we go back to the Training Phase to retrain our agent team module via efficient meta optimization with Alg. 1, in which we set D_{init} to the most up-to-date labeled set D^L. With the updated agent team, we resume the AL procedure on D^U. These steps are repeated until the annotation budget B is exhausted.

We set the learning rate of our policy network to 1e-4 and the discount factor γ in Eq. 4 to 0.9. We use the average joint error to measure performance of the pose estimator on the test set of each dataset. To show the robustness of our method, we run our experiments 5 times and report the mean performance and its deviation.

Result on ICVL. Fig. 3 (a) shows the performance of our proposed MATAL on ICVL dataset. Our method constantly outperforms state-of-the-art methods at each active learning iteration by a clear margin. UncertainGCN outperforms other existing methods at the beginning state, but later CoreGCN achieves better performance, which is possibly due to the fact that the fixed criteria based on uncertainty or representativeness could not constantly identify informative samples during the entire AL procedure. Instead, our MATAL selects images that can most benefit the pose estimator with the proposed learning framework, which adapts to the needs of the pose estimator at different stages. As shown in Fig. 3 (a), our MATAL just needs 600 labeled images to reduce the average joint error to less than 12.5 mm, while uncertainGCN [5] and MCD CKE [4] need more than 900 labeled images. At the end of the AL procedure with 1000 labeled images, the average joint error in our model is reduced to 11.89 mm, which is much lower than the minimum value obtained by other methods.

Result on NYU. This dataset was collected by multiple cameras, leading to several images sharing nearly same topological information. Although these images have different appearance features, the redundant topological information significantly decreases the learning efficacy of the pose estimator. As shown in Fig. 3 (b), the performance of Coreset [35] is close to the error of random sampling, as the Coreset mainly relies on the appearance feature information but disregards the topological information. MCD CKE [4] obtains better performance by utilizing the pose estimator's uncertainty. Our method, benefited by learning the sampling policy directly from data, significantly outperforms the MCD CKE baseline. On this dataset, our method only requires 5K images to achieve the nearly same performance (23.5 mm) obtained by other approaches that require around 10K labeled images.

Result on BigHand2.2M. We use the large scale BigHand2.2M [60] dataset to show the scalability of our method. It contains around 2.2 million images of subjects with different hand shapes and contains schemed, random, and egocentric poses. Thus, this dataset is much more diverse and challenging. Figure 3 (c) shows the performance of different AL algorithms. Our method still outperforms other methods. It demonstrates that our MATAL can learn to select informative images even on this diverse dataset.

4.2. MATAL on Human Body Pose Estimation

Baseline. We benchmark our MATAL framework with SOTA active learning frameworks for human body pose estimation, including Coreset [35], LearningLoss [58], LearningLoss++ [38] and EGL++ [37].

Implementation details. Following the previous works [38, 37], we use Stacked Hourglass [32] as the backbone of our pose estimator. We collect the feature map from the bottleneck CNN layer of the last Hourglass block and perform global average on it to build the image appearance feature and use the predicted 16 joints to build the topological features. A team of 40 agents are set up for batch selection, and 800 images are randomly sampled to build the initial dataset D_{init}. Moreover, we follow the previous works [38, 37] and use PCKh@0.5 [31] to measure the performance. Other settings follow the hand pose estimation.

Result on MPII. Figure 3 (d) demonstrates the performance of MATAL on the body pose estimation task. All existing methods achieve better results than random sampling but their PCKh@0.5 scores are close to each other. The EGL++ [37] tends to slightly outperform other exist-

Table 1. Ablation study on the design of the state and the action representation. We ablate the state/action representations by comparing the accuracy of the model with individual component removed for state s_t and action a_t.

Method	MSE (mm) with labeled samples			
	2000	4000	6000	8000
State w/o $K_{F_{P_0}}$	28.44	25.21	23.47	23.01
State w/o $\{K_{F_{P_i}}\}_{i=1}^{6}$	28.30	25.24	23.66	23.23
State w/o K_{F_A}	29.00	25.65	24.49	23.55
State w/o b	28.62	25.22	24.10	23.85
Action w/o $\{s_i\}_{i=0}^{6}$	30.47	26.77	25.36	25.13
Action w/o $\{d_i\}_{i=0}^{6}$	27.60	24.82	24.51	24.12
Action w/o f_A	29.18	25.84	24.44	23.69
MATAL	**26.08**	**24.11**	**22.97**	**22.53**

ing approaches and has a narrow deviation. Our MATAL achieves significantly higher accuracy by learning a sampling policy that directly maximizes the performance of the pose estimator. The proposed MATAL uses around 25% of labels to obtain PCKh@0.5 of 85.1% while using the full annotated data yields PCKh@0.5 of 90.5%. Moreover, the proposed MATAL requires only 4K images to achieve similar performance compared to others that require 6K images, saving the labeling efforts of around 2K images.

4.3. Ablation Study

Effect of the state and action representations We perform an ablation study on NYU dataset to evaluate the contribution of each component in our proposed state and action representations. As the team agent relies on the state to decide the sampling policy, we first investigate the influence of the state information by removing its components individually from the complete model. Similarly, we also discuss the effect of the information in the action vector. As shown in Table 1, the complete MATAL gives the lowest average joint error in all active learning iterations. Removing either global or local topological information in the state will degrade the performance of our method. The largest increase of average joint error occurs when the score of maximum similarity in action representation $\{s_i\}_{i=0}^{6}$ is removed. It further verifies the effectiveness of using the difference in global and local topological features to estimate the novelty of the recovered poses.

Effect of agent team policy learning We further validate the performance of the proposed multi-agent sampling policy on NYU and MPII datasets. We first consider the usage of only one agent to select a single image in each active learning iteration. Then we construct the second baseline as one agent to select a batch of images in one shot. Here, the images with N highest Q-values are sampled. Finally, we present the performance of using N agents to select N images in two different settings: with or without teamwork. Figure. 3 (e) and (f) report the performance of these sampling strategies. As shown in Fig. 3 (e) and (f), selecting multiple images by either a single agent or noncooperative multiple agents gives the worst results. We argue that this is because these methods tend to select similar images whose Q-values are high yet close to each other, leading to several inefficiencies in the batch image selection setting. Introducing cooperation among separate agents helps address this problem, as the proposed expectation of previous actions provides valuable information about the other agents' decisions and the agent could learn to sample with a better coverage of the underlying distribution. Using one agent to select an image at each iteration also provides a competitive performance but still tends to be inferior to our agent team method. The main reason is that the minor improvement of the pose estimator leads to small and noisy rewards, making it difficult for the agent to learn a good sampling policy. Furthermore, the time cost of the method that uses one agent to select one image only is much higher than our agent teaming method.

Table 2. Ablation Study for the meta optimization. We compare MATAL with or without the Meta-Optimization and showcase that Meta-Optimization significantly accelerates the retraining process.

Method	MSE (mm)	Time cost (h)
MATAL w/o meta	23.68	4.5
MATAL w meta	23.74	2

Effect of Meta-Learning We use meta-optimization to update the agent team module more effectively and efficiently. In this experiment, we compare the time cost of collecting 5K informative images by our model with/without meta-learning on the NYU dataset in Table 2. Note that the time cost of sampling is almost the same for both models, but it is the time consumption for the retraining of the agent team that really makes a difference. As shown in Table 2, with our meta-optimization scheme, our model obtains competitive performance while significantly reducing the time consumption by more than a half.

5. Conclusion

In this paper we proposed an RL-based batch selection active learning framework for pose estimation named MATAL. MATAL *directly* learns a *cooperative sampling policy* for *a team of agents* to achieve *effective* image batch selection. Moreover, a *Meta-Optimization* was introduced to significantly accelerate the retraining of our team agent during the Deployment Phase of the active learning procedure. We conducted extensive ablation studies to verify the design of our framework. Furthermore, we compared the performance of our model with existing SOTA works on four widely used datasets and obtained better accuracy on all experiments.

Acknowledgments. The project is supported by AI Singapore under the grant number AISG-100E-2020-065, National Research Foundation Singapore and SUTD Startup Research Grant. This work is also partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019.

[3] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018.

[4] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Active learning for bayesian 3d hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3419–3428, 2021.

[5] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9583–9592, 2021.

[6] Arantxa Casanova, Pedro O Pinheiro, Negar Rostamzadeh, and Christopher J Pal. Reinforced active learning for image segmentation. *arXiv preprint arXiv:2002.06583*, 2020.

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

[8] Yujin Chen, Zhigang Tu, Liuhao Ge, Dejun Zhang, Ruizhi Chen, and Junsong Yuan. So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6961–6970, 2019.

[9] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10451–10460, 2021.

[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*, pages 5386–5395, 2020.

[11] Manuela Chessa, Guido Maiello, Lina K Klein, Vivian C Paulun, and Fabio Solari. Grasping objects in immersive virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 1749–1754. IEEE, 2019.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[13] Guillermo Garcia-Hernando, Edward Johns, and Tae-Kyun Kim. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9561–9568. IEEE, 2020.

[14] Erik Gärtner, Aleksis Pirinen, and Cristian Sminchisescu. Deep reinforcement learning for active human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10835–10844, 2020.

[15] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019.

[16] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8575–8584, 2021.

[17] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[18] Rui Li, Zhenyu Liu, and Jianrong Tan. A survey on 3d hand pose estimation: Cameras, methods, and datasets. *Pattern Recognition*, 93:251–272, 2019.

[19] Shile Li and Dongheui Lee. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11927–11936, 2019.

[20] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16266–16275, 2021.

[21] Xing Liang, Anastassia Angelopoulou, Epaminondas Kapetanios, Bencie Woll, Reda Al Batat, and Tyron Woolfe. A multi-modal machine learning approach and toolkit to automate recognition of early stages of dementia among british sign language users. In *European Conference on Computer Vision*, pages 278–293. Springer, 2020.

[22] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4363–4372, 2017.

[23] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021.

[24] Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. Influence selection for active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9274–9283, 2021.

[25] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2208–2217. PMLR, 06–11 Aug 2017.

[26] Jameel Malik, Ahmed Elhayek, Fabrizio Nunnari, Kiran Varanasi, Kiarash Tamaddon, Alexis Heloir, and Didier

Stricker. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In *2018 International Conference on 3D Vision (3DV)*, pages 110–119. IEEE, 2018.

[27] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[28] Rahul Mitra, Nitesh B Gundavarapu, Abhishek Sharma, and Arjun Jain. Multiview-consistent semi-supervised learning for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6907–6916, 2020.

[29] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020.

[30] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1154–1163, 2017.

[31] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[32] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015.

[33] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics (TOG)*, 37(6):1–14, 2018.

[34] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8505–8514, 2021.

[35] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[36] Jianzhun Shao, Yuhang Jiang, Gu Wang, Zhigang Li, and Xiangyang Ji. Pfrl: Pose-free reinforcement learning for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11454–11463, 2020.

[37] Megh Shukla. Egl++: Extending expected gradient length to active learning for human pose estimation. *arXiv preprint arXiv:2104.09493*, 2021.

[38] Megh Shukla and Shuaib Ahmed. A mathematical analysis of learning loss for active learning in regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3328, 2021.

[39] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 211–228. Springer, 2020.

[40] Srinath Sridhar, Anna Maria Feit, Christian Theobalt, and Antti Oulasvirta. Investigating the dexterity of multi-finger input for mid-air text entry. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3643–3652, 2015.

[41] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2015.

[42] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *European Conference on Computer Vision*, pages 294–310. Springer, 2016.

[43] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[44] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017.

[45] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[46] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3786–3793, 2014.

[47] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29:1930–1938, 2016.

[48] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):1–10, 2014.

[49] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[50] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[51] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Self-supervised 3d hand pose estimation through training by fitting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10853–10862, 2019.

[52] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2018.

[53] Bastian Wandt, Hanno Ackermann, and Bodo Rosenhahn. A kinematic chain space for monocular motion capture. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[54] Rongchang Xie, Chunyu Wang, Wenjun Zeng, and Yizhou Wang. An empirical study of the collapsing problem in semi-supervised 2d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11240–11249, October 2021.

[55] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019.

[56] Tianhan Xu and Wataru Takano. Graph stacked hourglass networks for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16105–16114, 2021.

[57] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 5571–5580. PMLR, 2018.

[58] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019.

[59] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2018.

[60] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4866–4874, 2017.

[61] Ho Yub Jung, Soochahn Lee, Yong Seok Heo, and Il Dong Yun. Random tree walk toward instantaneous 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2474, 2015.

[62] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. *arXiv preprint arXiv:2203.00859*, 2022.

[63] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020.

[64] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017.