Rethinking the Data Annotation Process for Multi-view 3D Pose Estimation with Active Learning and Self-Training

Qi Feng, Kun He, He Wen, Cem Keskin, and Yuting Ye

Meta Reality Labs

{fung,kunhe,hewen,cemkeskin,yuting.ye}@meta.com

Abstract

Pose estimation of the human body and hands is a fundamental problem in computer vision, and learning-based solutions require a large amount of annotated data. In this work, we improve the efficiency of the data annotation process for 3D pose estimation problems with Active Learning (AL) in a multi-view setting. AL selects examples with the highest value to annotate under limited annotation budgets (time and cost), but choosing the selection strategy is often nontrivial. We present a framework to efficiently extend existing single-view AL strategies. We then propose two novel AL strategies that make full use of multi-view geometry. Moreover, we demonstrate additional performance gains by incorporating pseudo-labels computed during the AL process, which is a form of self-training. Our system significantly outperforms simulated annotation baselines in 3D body and hand pose estimation on two large-scale benchmarks: CMU Panoptic Studio and InterHand2.6M. Notably, on CMU Panoptic Studio, we are able to reduce the turn-around time by 60% and annotation cost by 80% when compared to the conventional annotation process.

1. Introduction

Pose estimation is a fundamental problem in computer vision. Accurate pose estimations of the human body/hands allow automated systems to perform markerless motion capture [9, 37], recognize actions [6, 43], understand social interactions [18] and sign languages [16], and so on.

While supervised learning methods using deep neural networks have achieved considerable success for pose estimation [27, 38, 39, 40], the annotation of pose data is time-consuming and costly. For example, the creators of MPII [1], a popular body pose estimation benchmark, reported that on average it takes an annotator *one minute* to annotate all body keypoints on an image. Human labels can also have inconsistent quality, especially for the difficult occluded cases. On the other hand, multi-view camera systems [18, 45, 47] are increasingly being used to generate pose labels automatically, which is a major motivation for our work. However, training the underlying labeling models still requires significant upfront annotation.

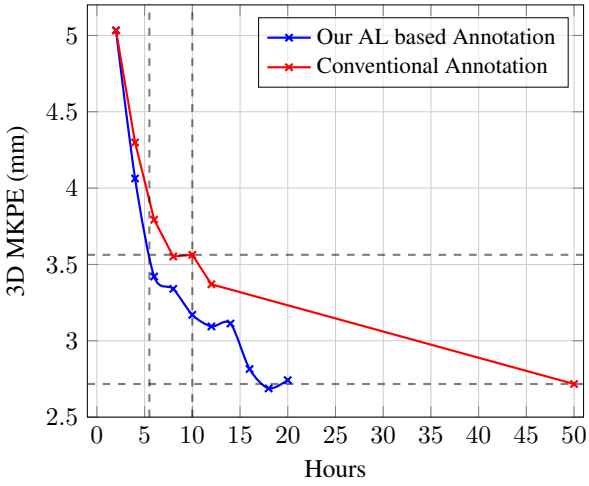


Figure 1: Model test accuracy vs. annotation turn-around time. We use the estimate of 1 minute per frame for annotation and 1 hour for training the model. Conventional Annotation does not require training during annotation. Our AL based annotation system saves the overall turn-around time by 45% for PoseResNet with a test performance of 3.5 mm, and more than 60% with a test performance of 2.7mm.

In this paper, we propose an annotation process based on Active Learning (AL) [7, 33] to make the data annotation process for learning deep pose estimation models faster and more cost-effective. Our AL based approach focus the annotation budget (time and cost) on the most valuable samples.

We study AL formulations in the context of pose estimation; in particular, we consider *3D body and hand pose estimation from multi-view RGB images*. By exploiting the use

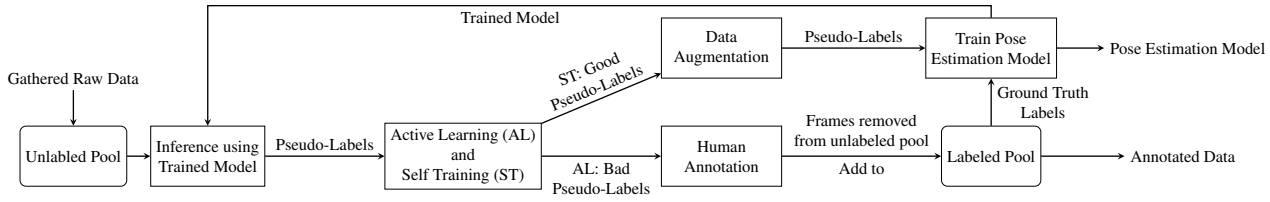


Figure 2: An overview of the proposed active learning (AL) system for multi-view 3D pose estimation. While prior works have only considered AL for single-view pose estimation, our system is the first to work in the multi-view setting (Sec. 3.3), and we propose two effective strategies that make full use of multi-view geometry. Additionally, by incorporating pseudo-labels in the proposed self-training process (Sec. 3.4), we show further improvement in annotation efficiency without extra annotation or computational cost.

of multi-view geometry, we propose two novel AL strategies that are geometrically inspired, and easy to compute. To our knowledge, other existing AL systems for pose estimation [5, 21, 44] do not consider multi-view input, and the proposed single-view strategies therein do not generalize well to the multi-view setup.

In addition to the main AL formulation, we also explore further improvements to the annotation efficiency with *self-training*, which has been a successful strategy for image classification [28, 41, 48]. To this end, during each AL iteration we augment the human-annotated labels with pseudo-labels computed from the model's prediction, inspired by the multi-view bootstrapping method [36]. Our experiments show that, with careful selection, pseudo-labels can further boost pose estimation performance without additional annotation or computational cost.

We conduct annotation simulations, experiments, and ablation studies on two large-scale benchmarks, CMU Panoptic [18] for body pose estimation, and InterHand2.6M [25] for hand pose estimation. Our proposed multi-view AL strategies, together with the self-training strategies, consistently outperform baseline strategies by significant margins. Notably, as shown in Fig. 1, on CMU Panoptic our complete system reduces the annotation turn-around time by 60% and the annotation labor cost by 80% when compared to existing data annotation processes. In summary, our contributions in this paper are threefold:

- We propose a data annotation process based on Active Learning for 3D pose estimation from multi-view RGB images, and propose AL strategies that utilize multi-view geometry to reduce the annotation time and cost.
- We explore self-training for pose estimation in the proposed AL framework, and show that further gains can be realized by including pseudo-labels.
- We show that the proposed AL and self-training strategies significantly improve annotation efficiency over baselines, and establish the state-of-the-art in AL for multi-view pose estimation.

2. Related Work

3D Pose Estimation: Pose estimation is one of the fundamental tasks in computer vision. To model human bodies that can undergo articulation and deformation, early approaches mostly take inspiration from the classical Pictorial Structures [2, 10]. Following the success of deep neural networks, and facilitated by benchmarks such as Human 3.6M [15] and MPII [1], deep CNNs have been widely applied to body and hand pose estimation. Representative methods include Convolutional Pose Machines [39], Stacked Hourglass Networks [27], PoseResNet [40], HRNet [38], *etc*. These methods typically work by predicting the locations of body/hand keypoints, formulated as a heatmap regression problem. Single-view 3D pose estimation methods [19, 20, 23, 47] on the other hand, directly lift 2D image evidence into 3D keypoints or mesh representations, but need more high-quality training data in order to resolve the inherent 2D-3D ambiguity.

With the increasing availability of multi-camera setups, *multi-view* pose estimation has gathered increased interest [12, 17, 29]. A key motivation is that these systems can be used to automatically or semi-automatically generate "ground truth" for single-view 3D pose estimation, and significantly reduce labeling cost. In fact, such a procedure has been adopted in benchmarks like CMU Panoptic [18] and HUMBI [45] for body pose estimation, as well as FreiHand [47] and InterHand2.6M [25] for hands. However, training multi-view models still requires large amounts of annotated 3D pose data, which strongly motivates cost-saving strategies such as active learning.

Active Learning: Active Learning (AL) [7, 33] considers a dynamic environment where an ML system selects unlabeled examples to acquire labels for, and iteratively re-trains itself using newly labeled data. This is critical in many real-world scenarios with constrained annotation budgets. A large AL literature exists for classification, including uncertainty-based sampling [30], diversity maximization [42], Bayesian methods [32], *etc*. Despite years of progress, in practice, the best AL strategies are often problem-dependent, and heuristics such as random sampling remain strong base-

lines [24, 34].

In computer vision, AL has also been widely studied for problems such as semantic segmentation [22, 35] and object detection [31]. Siddiqui *et al*. [35] demonstrate that incorporating multi-view geometry can improve the effectiveness of AL for semantic segmentation. Yet, the pixel classification formulation in semantic segmentation makes it easier to adapt AL approaches designed for classification, while for the keypoint localization task, multi-view adaptations are less straightforward.

For pose estimation, Yoo *et al*. [44] applied task-agnostic loss prediction as an AL strategy but with marginal gains over random sampling. Liu and Ferrari [21] propose the Multi-Peak Entropy metric to guide the sampling of single-view images for annotation. As we demonstrate later however, extending this metric to multi-view is a non-trivial task.

More recently, Caramalau *et al*. [5] extend the CoreSet [32] AL algorithm to hand pose estimation with a Bayesian formulation. While we also propose an extension to CoreSet in this paper, our AL strategy relies on geometric intuitions and does not require expensive Bayesian inference. Additionally, [5] estimates 3D pose from a single depth camera, while we take RGB images from multiple calibrated cameras as input.

Self-Training and Pseudo-Labeling: Besides active learning, self-training [3, 28, 41, 48] is another prominent approach to increasing annotation efficiency. Building on the principle of knowledge distillation [13], these methods perform iterative pseudo-labeling and re-training with unlabeled data. For image classification, this paradigm has been shown to improve model generalization and robustness without increasing the amount of human-annotated labels.

For the keypoint localization task in pose estimation, similar ideas have been explored in the form of semi-supervised learning and pseudo-labeling [4, 14, 26]. In this paper, inspired by the seminal work of multi-view bootstrapping [36], we also develop a pseudo-labeling method. When applied in conjunction with our AL framework, it leads to even greater gains in efficiency.

3. Methods

The overview of our proposed Active Learning with Self-training system is shown in Fig. 2. The whole iterative system consists of two main branches: the active learning branch which selects the unlabeled frames for human annotation and the self-training branch for pseudo-labeling on the unlabeled frames. In this section, we first formally define the multi-view pose estimation problem we are addressing(Sec. 3.1). Next, we extend prior works on AL for single-view pose estimation (Sec. 3.2). Then, two effective strategies that make full use of multi-view geometry in the multi-view setting are proposed (Sec. 3.3). Additionally, by incorporating pseudo-labels in the proposed self-training process (Sec. 3.4), we show further improvement in annotation efficiency without extra annotation or computational cost.

3.1. Pose Estimation Problem Formulation

We assume a multi-view capture setup with N synchronized and calibrated cameras, and we use the term *frame* F to denote the collection of images from all cameras (*views*) V at a particular time instance t, *i.e.* $F(t) = \{V_1(t), V_2(t), \cdots, V_N(t)\}$. In the following, we drop t from the notation unless necessary. The entire dataset, which is a set of frames (possibly infinite), is denoted as $\mathcal{D} = \{F(1), F(2), \cdots\}$.

The task of 3D pose estimation is to estimate the 3D locations of a set of keypoints on the human body/hand from an input frame. In this work, we focus on a well-established approach, where the 3D keypoints are obtained by triangulating 2D predictions on each camera view, using robust triangulation techniques [11], *e.g*. RANSAC. In particular, the 2D keypoint prediction problem is formulated as heatmap regression, where the ground truth heatmaps are commonly constructed by placing a 2D isotropic Gaussian at the ground truth location. We use K to denote the number of keypoints.

Note that unlike entropy based AL methods [21], our AL and self-training system does not limit the pose estimation model to predict a heatmap for 2D keypoints. Instead, any pose estimation model that performs 2D keypoint localization and then triangulation would be sufficient.

3.2. Extending Single-View AL for Pose Estimation

Active Learning starts with an initial labeled set $\mathcal{L}_0$, and trains an initial pose estimator. Afterwards, in each iteration $i \geq 1$, an AL strategy samples a set of frames from the remaining unlabeled set $\mathcal{U}_i$ following an AL metric $\mathcal{M}$, queries human annotators, and obtains labels for them. This enlarges the labeled set $\mathcal{L}_i$ into $\mathcal{L}_{i+1}$, with which the pose estimation model is re-trained. Note that $\forall i, \mathcal{L}_i \cup \mathcal{U}_i = \mathcal{D}$, and that $\mathcal{L}_1 \subset \mathcal{L}_2 \cdots \subset \mathcal{D}$.

An intuitive approach to AL is to sample examples that receive the most *uncertain* predictions, and the definition of uncertainty is usually problem-dependent. The BSB and MPE strategies introduced by Liu *et al*. [21] fall into this category.

To our knowledge, no prior work has applied AL to *multi-view* pose estimation, and the closest work is Liu and Ferrari [21], who focused on the single-view case. We thus extend BSB and MPE to multi-view, and use them as baselines. We extend these single-view strategies by aggregating the per-view uncertainty metrics, without taking geometry into consideration. In particular, we focus on the average[1]: if the per-view predictions have higher uncertainty on average, then, heuristically, the frame will have higher uncertainty.

[1]We also experimented with other aggregation functions such as variance, and found them to perform worse.

We define the metric for the aforementioned entropy-based metrics as

$$\mathcal{M}_{\text{BSB}}(F) = \frac{1}{N} \sum_{V \in F} \mathcal{M}_{\text{BSB}}(V), \tag{1}$$

$$\mathcal{M}_{\text{MPE}}(F) = \frac{1}{N} \sum_{V \in F} \mathcal{M}_{\text{MPE}}(V), \tag{2}$$

where $\mathcal{M}_{\text{BSB}}(V)$ and $\mathcal{M}_{\text{MPE}}(V)$ are the per-view metrics introduced by Liu *et al*. [21]. A visualization of these metrics are shown in the supplementary material.

3.3. Multi-View AL for Pose Estimation

We now discuss AL strategies under the multi-view setting. However, beyond simple aggregation, the multi-view setting provides extra information to define geometrically-inspired AL strategies. Recall that the 3D prediction for any keypoint k, denoted as P^k, is obtained through robust triangulation; we will build on this fact to define novel AL strategies. Below, we propose two AL strategies: *CoreSet-Poses* which is based on pose diversity, and *Multi-View Consistency* which is based on 3D prediction uncertainty.

CoreSet-Poses: CoreSet [32] is a state-of-the-art AL strategy based on selecting diverse representative examples from the unlabeled set, formulated as solving a combinatorial set-cover problem. Critical to the effectiveness of CoreSet is modeling the distance between unlabeled examples; in the case of image classification, Sener *et al*. [32] uses the Euclidean distance between pretrained convolutional features. Caramalau *et al*. [5] introduced an CoreSet based AL strategy that only applies to Bayesian pose estimation models. Unlike this prior work, our proposed CoreSet-Poses strategy can be used on any pose estimation models.

Our first strategy, CoreSet-Poses, builds on CoreSet by supplying it with a distance metric tailored for pose estimation. Specifically, given a pair of frames (F, F'), we define their distance Δ to be the average Euclidean distance between 3D keypoint predictions $(P_F, P_{F'})$ with the current model. While more sophisticated distance metrics could be defined with respect to the underlying sets of 2D heatmap predictions, the 3D predictions have already been filtered through robust triangulation, and have much lower dimensions so distance computation can be efficient. In practice, we align P by shifting the root keypoint to the origin, *e.g*. if the root keypoint is 0, the aligned pose would be: $\hat{P} = P - P^0$.

Given the distance metric, CoreSet-Poses solves a set-cover problem in order to maximize coverage in the pose space. While this problem is NP-hard, prior works [5, 32] show that it can be approximately solved by a greedy k-center algorithm. Specifically, for each candidate unlabeled frame $F \in \mathcal{U}$, we define the CoreSet-Poses AL metric as

$$\mathcal{M}_{\text{CS}}(F) = \min_{F' \in \mathcal{L}} \Delta(\hat{P}_F, \hat{P}_{F'}), \tag{3}$$

Algorithm 1: AL for multi-view pose estimation

Input: Labeled set $\mathcal{L}$, unlabeled set $\mathcal{U}$, AL metric $\mathcal{M}$, annotation budget B;
Sampled Data $S \leftarrow \{\}$;
for $F \in \mathcal{U}$ **do**
  $\mathcal{H}_F = \{H_V | \forall V \in F\} \leftarrow$ Model Inference;
  $P_F, \varepsilon_F \leftarrow$ triangulate$(\mathcal{H}_F)$;
repeat
  $F_{\text{greedy}} \leftarrow \arg\max_{F \in \mathcal{U}} \mathcal{M}(F)$; ▷ $\mathcal{M}_{\text{CS}}$, $\mathcal{M}_{\text{MC}}$, *etc.*
  $S \leftarrow S \cup \{F_{\text{greedy}}\}$;
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{F_{\text{greedy}}\}$;
  $\mathcal{U} \leftarrow \mathcal{U} \setminus \{F_{\text{greedy}}\}$;
until $|S| = B$;
return S

Algorithm 2: AL + self-training w/ pseudo-labels

Input: Unlabeled set $\mathcal{U}$, previous pseudo-label set $\mathcal{P}$, target amount M;
Output: New pseudo-label set $\mathcal{P}'$;
$\mathcal{P}' \leftarrow \{\}$, $\mathcal{U}' \leftarrow \mathcal{U}$; ▷ Make a copy of $\mathcal{U}$.
repeat
  $F_{\text{min}} \leftarrow \arg\min_{F \in \mathcal{U} \setminus (\mathcal{P} \cup \mathcal{P}')} \varepsilon_F$; ▷ No re-labeling.
  $\mathcal{U} \leftarrow \mathcal{U} \setminus \{F_{\text{min}}\}$;
  if $c_{F_{min}} = N$ **then** ▷ All views are inliers.
    $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{F_{\text{min}}\}$;
until $|\mathcal{P}'| = M$ OR $|\mathcal{U}| = 0$;
$\mathcal{U} = \mathcal{U}' \setminus \mathcal{P}'$;
return $\mathcal{P}'$

which measures how "close" F is to the current labeled set. Then, the greedy algorithm samples frames with the largest $\mathcal{M}_{\text{CS}}$ values. Despite the improved efficiency, CoreSet-Poses would still take $O(|\mathcal{U}|^2)$ time to compute the pairwise distances, making it potentially impractical for large datasets.

Multi-view Consistency: We now present an uncertainty measure that is intrinsic to the 3D pose predictions. Our reasoning is that given a frame with multiple views, it is less likely for the frame-level prediction to be wrong if the per-view 2D predictions agree with each other. This agreement is in the geometric sense, *e.g*. for two views, we say two keypoint predictions exactly agree if their epipolar distance is 0. The corresponding AL strategy is then to sample frames with the largest disagreement. Additionally, we would like to compute this in $O(|\mathcal{U}|)$ time, to make it practical for large datasets. We call this the Multi-View Consistency strategy.

Specifically, we take the triangulation error, or the average Euclidean distance between the 2D keypoint predictions and the reprojected 3D triangulation, as the AL metric. Note that, since this is exactly the minimization objective for trian-

gulation, a high error directly indicates strong disagreements between 2D predictions. Formally, let the predicted 2D location of the k-th keypoint in view V be l_V^k, and its reprojected location from the triangulated P^k be $\hat{l}_V^k$. The triangulation error metric can be written as:

$$\mathcal{M}_{\text{MC}}(F) = \frac{1}{N}\frac{1}{K}\sum_{V\in F}\sum_{k=1}^{K}\|l_V^k - \hat{l}_V^k\|^2. \quad (4)$$

For simplicity, we use ε_F to denote $\mathcal{M}_{\text{MC}}(F)$.

Alg. 1 presents a unified view of AL for multi-view pose estimation, where different sampling strategies are realized by choosing the corresponding metric $\mathcal{M}$.

3.4. Improvement via Self-Training

AL is shown to benefit from the addition of techniques like data augmentation and semi-supervised learning [24]. In this work, we want to explore a novel direction to improve it further. We leverage the fact that our measure of geometric inconsistency can also help us identify reliable frames with good *pseudo-labels*, which can be directly injected into the training set. In fact, this is a form of self-training, which has shown great success recently for image classification tasks [28, 41, 48]. These methods use soft pseudo-labels assigned to unlabeled frames directly and show that the richness of predictions (compared to a one-hot encoding) is crucial. In the pose estimation task, the heatmaps can play a similar role, as was demonstrated by Zhang *et al*. [46] in their work that distills heatmaps from an 8-stack hourglass model to a 4-stack one. However, this approach is not suitable to make the best use of multi-view predictions, which is the direction we explore. To take full advantage of multi-view predictions, we project the 3D keypoints formed by triangulation back to each camera view, and assign psuedo-heatmaps to a set of frames with the most inliers and with the smallest triangulation error (Equation 4). These predictions are the most likely to be closest to the actual ground truth, thus they are excellent candidates to be used in self-training.

We call this the pseudo-label set $\mathcal{P}$, and we augment the training set to be $\mathcal{P}\cup\mathcal{L}$ in each AL iteration. Similar to multi-view bootstrapping [36], our motivation is that by adding $\mathcal{P}$ to the training set, the model is exposed to more varied data and can learn to generalize better. However, the proposed self-training algorithm is able to avoid "model drifting" in iterative training, using an entirely *automated* strategy, as shown in Fig. 7. This is in contrast to multi-view bootstrapping [36], which requires human verification in the loop.

Contrary to AL, self-training requires the pseudo-labels to be confident and accurate, and careful selection is key. Simon *et al*. [36] uses heuristics specific to hand anatomy to filter candidate frames, and conducts additional human verification. Instead, our approach is fully automated. Specifically, for a pseudo-labeled frame to be considered for selection, we require that all views for all keypoints to be inliers during triangulation. Then, we take candidate frames with the smallest triangulation error ε_F, that are *not already selected* in the previous AL iteration, to form the pseudo-label set $\mathcal{P}$. We found the latter heuristic to be critical in preventing drifting of the pseudo-labels. Our self-training algorithm is summarized in Alg. 2.

4. Experiments

4.1. Datasets and Evaluation

To simulate the data annotation process, we use two large-scale multi-view benchmarks in our experiments: CMU Panoptic [18] for the body pose estimation problem, and InterHand2.6M [25] for the hand pose estimation problem.

The CMU Panoptic dataset has 9 sequences each having 31 camera views, and over 160,000 frames in total. We split them into 7 sequences for training, 1 sequence for validation and 1 sequence for test. We use 8 eye-level cameras for training and validation and 30 cameras[2] for testing, including the 8 eye-level cameras used during training and validation. Sequences are temporally sub-sampled at 1 frame per second, and we end up with 5,008 training frames (40,064 images), 891 validation frames (7,128 images) and 771 test frames (23,130 images). We use the 5fps version of InterHand2.6M, and sub-sample the dataset into 10 Captures for training, 1 Capture for validation and another 1 Capture for testing. For each capture, we use 16 cameras that are distantly-located during training and validation. Moreover, we use 32 cameras during testing. We end up with 12,123 training frames (193,968 images), 1,900 validation frames (30,400 images) and 1,762 test frames (56,384 images).

For each experiment, we conduct 3 randomized trials, and report the average and variance for the 3D Mean Key Point Error (MKPE) in millimeter (mm). As our backbone models predict 2D heatmaps for each view, to obtain the 3D prediction P^k we perform RANSAC triangulation with the 2D keypoint predictions l_1^k (argmax of the heatmap).

4.2. Annotation Simulation Details

We use two backbone models in our experiments: PoseResNet-50 [40] and HRNet [38]. For body pose estimation, both backbones are pretrained on the MPII [1] dataset. As MPII and CMU Panoptic define different sets of keypoints, we initialize the weights of all layers except the output layer of the PoseResNet-50. For HRNet, we use the pretrained weights of the first 4 layers and randomly initialize the remaining layers. For hand pose estimation, as no pretrained models for our setting are available, we randomly initialize all parameters from a normal distribution.

The annotation amount in each AL iteration is set to 100 frames for CMU Panoptic, and 1,000 frames for InterHand2.6M. Regardless of the AL strategy, frames in the

[2] Video from 1 test camera is missing on CMU website.

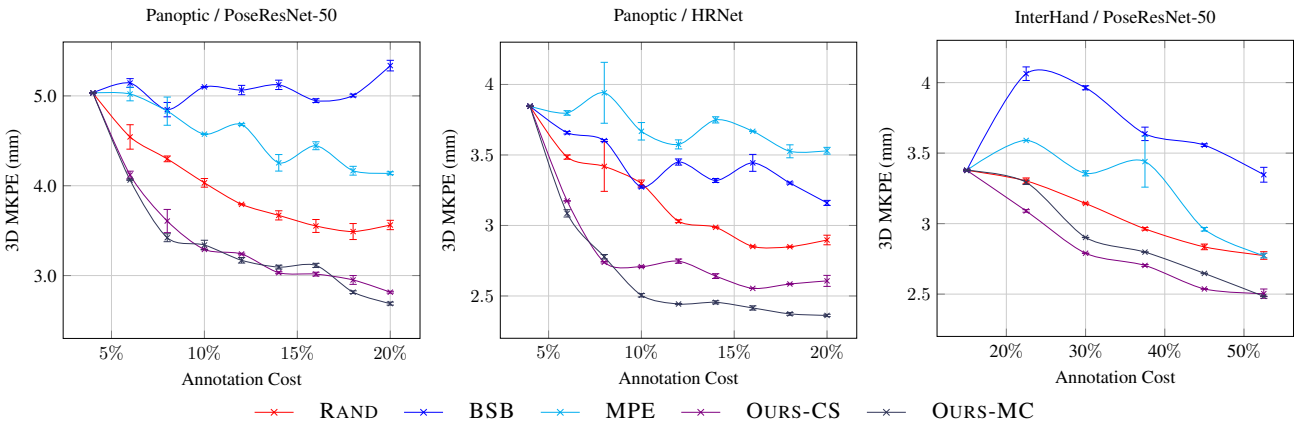


Figure 3: Comparison of AL strategies on CMU Panoptic and InterHand2.6M. X-axis: percent of dataset labeled. BSB and MPE [21], developed for single-view pose estimation, do not perform better than RAND when extended to multi-view. Our proposed strategies (OURS-CS and OURS-MC) significantly outperform random sampling. Best viewed in color.

initial labeled set $\mathcal{L}_0$ (200 frames for CMU Panoptic, 2000 frames for InterHand2.6M) are always randomly sampled, to provide a reasonable starting point. Furthermore, for the sake of reproducibility, all strategies start with the same set of randomly sampled frames. For self-training, the pseudo-label amount is set to 10%-20% of the annotation amount.

In each AL iteration, we train the model from scratch on the current labeled dataset, as well as the pseudo-labeled dataset if available. Both backbones are trained with a batch size of 32 images per GPU for a total of 5000 optimization steps. We use Adam optimizer with a learning rate starting at 0.001, and decayed by 1/10 at the midpoint. All experiments are carried out with this training procedure, and all reported results are evaluated on the same held-out set.

Following Mittal *et al*. [24], we also experiment with data augmentation. We use RandAugment [8] to augment the training images for CMU Panoptic. On the other hand, RandAugment does not result in better performances on InterHand2.6M, which contains more diverse poses.

4.3. Results

We experiment with both PoseResNet-50 and HRNet on CMU Panoptic, while for the much larger InterHand2.6M we report results from PoseResNet-50. Below, we refer to random sampling as RAND, Multi-Peak Entropy strategy [21] as MPE, Best vs. Second Best strategy [21] as BSB, our proposed CoreSet-Poses strategy as OURS-CS, and Multi-View Consistency strategy as OURS-MC.

4.3.1 Active Learning

Results with PoseResNet-50 and HRNet on CMU Panoptic and PoseResNet-50 on InterHand2.6M are reported in Fig. 3. We do not use data augmentation in this experiment in order to highlight the differences in sampling strategies.

As we mentioned earlier, the RAND strategy can be a very strong baseline for difficult tasks like pose estimation. Although MPE has been reported to outperform RAND in single-view pose estimation [21], we observe that extending MPE or BSB to multi-view by aggregating per-frame uncertainty measures fails to beat RAND. Furthermore, simple forms of aggregation also fail to account for the geometric structure in the problem: it is possible that all 2D predictions are highly confident, while being geometrically inconsistent. In such cases, the frame would fail the triangulation, yet still score low enough with MPE and BSB to evade selection.

Next, our proposed strategies OURS-MC and OURS-CS outperform RAND consistently by a large margin in all scenarios. OURS-MC is on par with OURS-CS with the PoseResNet-50 backbone, but outperforms the OURS-CS with the HRNet backbone, despite only taking a fraction of the computational cost on the unlabeled set ($O(|\mathcal{U}|)$ vs. $O(|\mathcal{U}|^2)$).

The improvement of OURS-MC compared to RAND on InterHand2.6M is smaller than that on CMU Panoptic, as the variations of poses in InterHand2.6M is much higher than CMU Panoptic. Additionally, we sample frames from InterHand2.6M more sparsely than CMU Panoptic. As mentioned above, InterHand2.6M is much larger and contains more diverse poses than CMU Panoptic, *i.e*. diverse samples can be achieved by random sampling when the unlabeled set is diverse. Therefore, we conduct our ablation studies mainly on CMU Panoptic.

4.3.2 AL + Self-Training

For this experiment, we focus on building a complete system: we use pseudo-labels to augment the training set in AL iterations, and we add data augmentation (except for InterHand2.6M as previously mentioned). For clarity, we

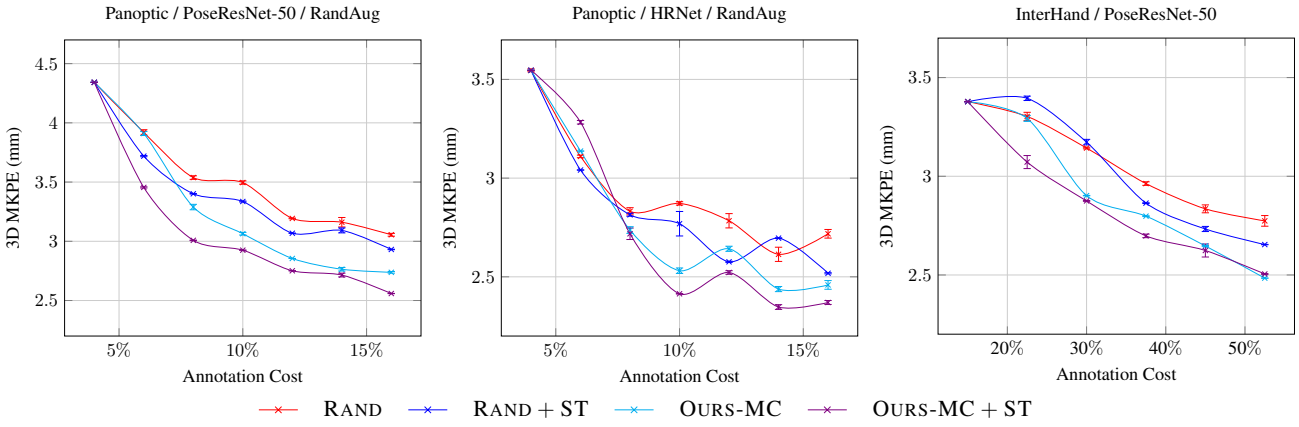


Figure 4: AL + self-training (ST) on CMU Panoptic and InterHand2.6M. X-axis: percent of dataset labeled. When combined with AL, our automated self-training strategy enables additional label efficiency gains at no extra computational cost, especially during the early stages of training. Best viewed in color.

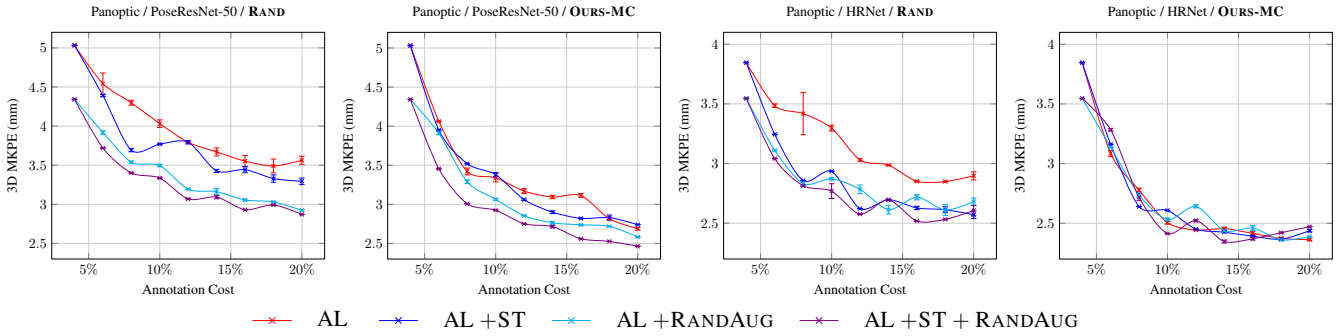


Figure 5: Comparison between AL, AL + self-training (ST), AL +RANDAUG, and AL +ST + RANDAUG on CMU Panoptic. X-axis: percent of dataset labeled. Our self-training strategy provides a large improvement on PoseResNet with the RAND AL strategy. Although RandAug further improves the generalization of the PoseResNet and HRNet for both RAND and OURS-MC AL strategies, our self-training strategy still shows a minor improvement from the AL only baseline.

pick the overall best method from the previous experiment, OURS-MC, and compare it against RAND. Results are shown in Fig 4.

Similar to the findings in multi-view bootstrapping [36], the additional self-training process provides consistent improvements to active learning. In our problem setting, we also observe the benefits to be more pronounced at the early stages: for example, on CMU Panoptic with 10% data annotated, pseudo-labels reduce the gap between 10% and 20% annotated data amount by 20% with the PoseResNet-50 backbone, and by around 50% for HRNet.

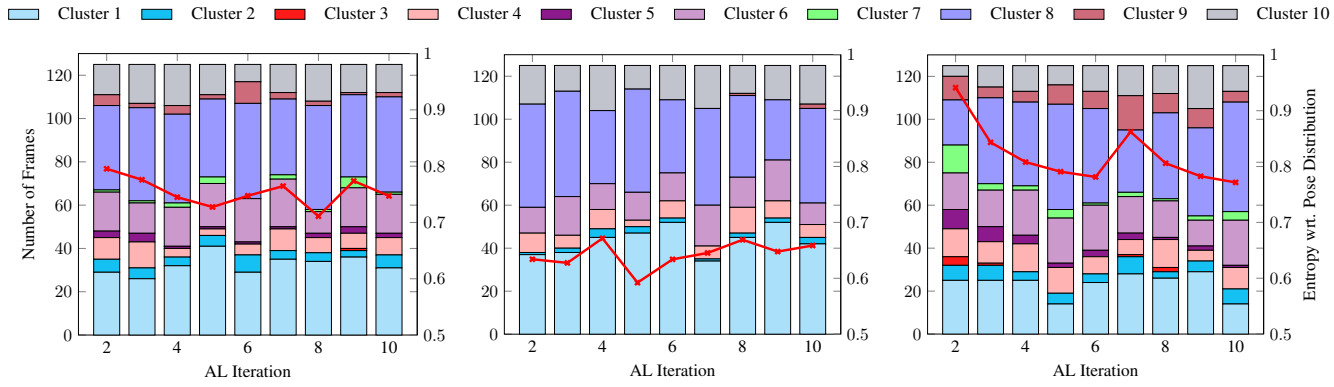
We find that pseudo-labels would negatively drift if the pseudo-labeled frames are sampled from $\mathcal{U}$ instead of $\mathcal{U} \setminus \mathcal{P}$ in each iteration. Essentially, the same frames would keep re-entering $\mathcal{P}$ and their labels become worse over each AL iteration. The number of frames to include, M, is also a crucial parameter. We present more ablative studies regarding these design choices in the supplementary material.

In summary, the above results show that our proposed AL strategies outperform the baselines steadily by a large margin, for both body and hand pose estimation. Additionally, with a carefully tuned self-training process, we can further improve annotation efficiency, with no extra cost.

4.3.3 Data Augmentation, Self-Training, and AL

The comparison between AL, AL+ST, AL+RANDAUG, and AL+ST+RANDAUG for different backbones and AL strategies are shown in Fig. 5. Data augmentation would improve the efficiency of our AL based annotation process, especially at an earlier stage. Larger performance gains can be observed on RAND and PoseResNet-50 based AL systems.

Self-training shows possible additional gains for all variations of experiments with different AL strategies and data augmentation. However, performance improvements from self-training saturates with higher performance models, *i.e.* at late stages of the AL process. Nonetheless, self-training can provide additional gains "for free" in our AL based an-



(a) RAND (b) MPE (c) OURS-MC

Figure 6: On CMU Panoptic, for three AL strategies, we visualize the pose distribution of sampled frames. Colors represent different clusters, and the red curve tracks the entropy of poses (wrt. cluster IDs) over AL iterations. OURS-MC produces diverse samples (higher entropy) and focuses more on under-represented clusters, leading to consistently superior performance.

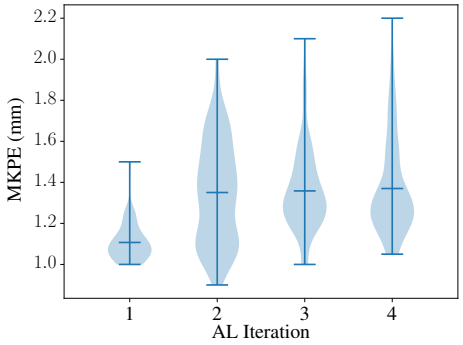


Figure 7: Self-training (Alg. 2): the deviation between sampled pseudo-labels and corresponding ground truth, measured in MKPE, for CMU Panoptic. Our selection strategy ensures that the pseudo-labels are accurate, and prevents "drifting" over time.

notation process, since it does not incur additional computational or annotation costs.

Finally, the choice of AL strategy outweigh data augmentation and self-training in terms of the label-efficiency. Our proposed OURS-MC and OURS-CS would outperform other compared ALs under all different setups.

4.4. Ablation Studies

Diversity of Samples: We take one trial of our experiments with PoseResNet-50 on CMU Panoptic where the annotation amount is 50 for each iteration, and study the distribution of sampled poses. Intuitively, sampling more diverse poses (while still following the data distribution) should help generalization. The ground truth 3D poses are shifted to have keypoint 2 (waist) at origin, and clustered into 10 clusters using K-means. We visualize the distribution of frames sampled by each AL strategy based on this clustering in Fig. 6, along with the entropy computed from the discrete distributions. The long-tail nature of the pose distribution can be seen from Fig. 6(a): samples from RAND are unevenly distributed, and dominated by clusters 1 and 8 in particular, which are common standing poses. Compared to RAND, the MPE strategy actually samples common pose clusters more heavily, and frames from minority clusters (5, 7, 9) are almost never sampled. In contrast, the proposed OURS-MC, being based on an uncertainty measure, attains much better pose diversity (higher entropy), especially in the early iterations. This is because OURS-MC looks for *geometric* disagreements in the predictions, which are largely decoupled from the prediction targets and their distribution.

Accuracy of self-training pseudo-labels: The main challenge with pseudo-labels is to ensure their accuracy and avoid drifting. In Fig. 7, we visualize the distribution of MKPE between pseudo-labeled frames and their actual ground truth, over several AL iterations. Our selection strategy maintains high accuracy (< 1.5 mm MKPE on average), resulting in consistent improvements over the course of AL.

5. Conclusion

In this paper, we propose an active learning framework for the data annotation process of multi-view pose estimation. We first extend existing entropy-based single-view AL strategies to multi-view, and then propose two AL strategies utilizing 3D keypoint triangulation. The proposed CoreSet-Poses and Multi-View Consistency strategies consistently outperform all AL and conventional annotation baselines, for both body and hand pose estimation problems. In addition, we introduce a self-training procedure using pseudo-labels, and further improve the annotation efficiency with minimal cost. Our complete system achieves state-of-the-art data annotation efficiency on CMU Panoptic and InterHand2.6M, while using a fraction of the annotation cost and turn-around time.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. 1, 2, 5

[2] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1014–1021. IEEE, 2009. 2

[3] Javad Zolfaghari Bengar, Joost van de Weijer, Bartlomiej Twardowski, and Bogdan Raducanu. Reducing label effort: Self-supervised meets active learning. In *IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 1631–1639, 2021. 3

[4] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3

[5] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Active learning for bayesian 3d hand pose estimation. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision(WACV)*, pages 3419–3428, 2021. 2, 3, 4

[6] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3218–3226, 2015. 1

[7] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994. 1, 2

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 702–703, 2020. 6

[9] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[10] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005. 2

[11] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[12] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[13] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems (NeurIPS) Workshops*, 2015. 3

[14] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz. Improving landmark localization with semi-supervised learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1546–1555. IEEE Computer Society, 2018. 3

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013. 2

[16] Jason Isaacs and Simon Foo. Hand pose estimation for american sign language recognition. In *Proceedings of the Thirty-Sixth Southeastern Symposium on System Theory.*, pages 132–136. IEEE, 2004. 1

[17] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 7718–7727, 2019. 2

[18] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, and Iain Matthews. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):190–204, 2017. 1, 2, 5

[19] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[20] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[21] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 4363–4372, 2017. 2, 3, 4, 6

[22] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals - cost-effective region-based active learning for semantic segmentation. *Proc. British Machine Vision Conference (BMVC)*, 2018. 3

[23] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 2

[24] Sudhanshu Mittal, Maxim Tatarchenko, Özgün Çiçek, and Thomas Brox. Parting with illusions about deep active learning, 2019. 3, 5, 6

[25] 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 *Proc. European Conference on Computer Vision (ECCV)*, pages 548–564. Springer, 2020. 2, 5

[26] Olga Moskvyak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Semi-supervised keypoint localization. In *International Conference on Learning Representations*, 2021. 3

[27] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016. 1, 2

[28] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proc. IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR), pages 11557–11568, 2021. 2, 3, 5

[29] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[30] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pages 413–424. Springer, 2006. 2

[31] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *Proc. British Machine Vision Conference (BMVC)*, page 91, 2018. 3

[32] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations Recognition (ICLR)*, 2018. 2, 3, 4

[33] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009. 1, 2

[34] Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. Deep active learning: Unified and principled method for query and training. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1308–1318. PMLR, 2020. 3

[35] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[36] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1145–1153, 2017. 2, 3, 5, 7

[37] Denis Tome, Matteo Toso, Lourdes Agapito, and Chris Russell. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In *International Conference on 3D Vision (3DV)*, pages 474–483. IEEE, 2018. 1

[38] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, and Xinggang Wang. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 5

[39] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 1, 2

[40] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proc. European Conference on Computer Vision (ECCV)*, pages 466–481, 2018. 1, 2, 5

[41] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10698, 2020. 2, 3, 5

[42] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015. 2

[43] Angela Yao, Juergen Gall, Gabriele Fanelli, and Luc Van Gool. Does human action recognition benefit from pose estimation? In *Proc. British Machine Vision Conference (BMVC)*, 2011. 1

[44] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102, 2019. 2, 3

[45] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2990–3000, 2020. 1, 2

[46] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3517–3526, 2019. 5

[47] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 1, 2

[48] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. 2, 3, 5