# Active Transfer Learning for Efficient Video-Specific Human Pose Estimation

Hiromu Taketsugu Norimichi Ukita
Toyota Technological Institute
Nagoya, Japan
{sd23426,ukita}@toyota-ti.ac.jp

## Abstract

*Human Pose (HP) estimation is actively researched because of its wide range of applications. However, even estimators pre-trained on large datasets may not perform satisfactorily due to a domain gap between the training and test data. To address this issue, we present our approach combining Active Learning (AL) and Transfer Learning (TL) to adapt HP estimators to individual video domains efficiently. For efficient learning, our approach quantifies (i) the estimation uncertainty based on the temporal changes in the estimated heatmaps and (ii) the unnaturalness in the estimated full-body HPs. These quantified criteria are then effectively combined with the state-of-the-art representativeness criterion to select uncertain and diverse samples for efficient HP estimator learning. Furthermore, we reconsider the existing Active Transfer Learning (ATL) method to introduce novel ideas related to the retraining methods and Stopping Criteria (SC). Experimental results demonstrate that our method enhances learning efficiency and outperforms comparative methods. Our code is publicly available at: https://github.com/ImIntheMiddle/VATL4Pose-WACV2024*

## 1. Introduction

Analyzing Human Poses (HP) in videos has extensive applications in areas such as security [13, 14, 43], sports analysis [1, 3, 23], healthcare [10, 48], computer-aided diagnostics [9, 27], and performance capture [44, 57, 58]. In these applications, a massive amount of HPs in videos are necessary. To collect such many HPs, HP estimation [2, 8, 19, 34, 41, 56] plays a crucial role, as manually annotating every HP in videos is impractical.

Despite huge training datasets, inaccurate HPs may be observed in the test phase due to a domain gap between the training and test datasets [37]. Table 1 shows two examples of such a domain gap. In both two combinations of HP estimators and datasets, a large performance drop is observed. This paper addresses this challenge by enhancing the performance of the HP estimator through test-time adaptation. Specifically, we aim to efficiently adapt a pre-trained HP estimator to each video domain with minimal annotation cost.

Table 1. Performance degradation in the pose estimation accuracy (%) due to a domain gap between the training and test datasets.

| Method | Dataset | Train | Test |
|---|---|---|---|
| FastPose [19] | JRDB-Pose [60] | 95.31 | 39.50 |
| SimpleBaseline [62] | PoseTrack21 [17] | 98.95 | 75.50 |

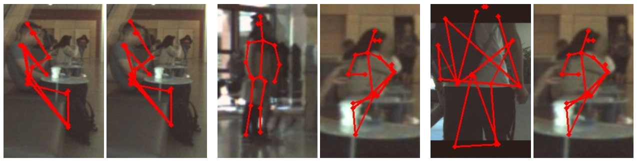

(a) MPE [39] (b) Core-set [53] (c) Ours

Figure 1. Samples with top scores in each selection criterion. Whereas (a) an uncertainty criterion (MPE [39]) selects uncertain but similar samples and (b) a representativeness criterion (Core-Set [53]) selects diverse but uninformative samples, (c) our criteria (THC+WPU+DUW) selects uncertain and diverse samples.

To this end, we propose a novel approach of applying Active Transfer Learning (ATL) [7, 16, 66] on a per-video basis (see Fig. 2). This approach involves the following two schemes, namely (i) selection of a subset of HPs for manual annotation by means of Active Learning (AL) [22, 26, 29, 39, 45, 54] and (ii) retraining of a pre-trained HP estimator with the annotated images by means of Transfer Learning (TL) [11, 18, 40]. Our contributions to ATL are as follows:

1. **ACFT-based Learning (Sec. 3):** To achieve efficient ATL, we refine the existing learning approach "Active, Continual Fine-tuning (ACFT)" [66] with a novel Stopping Criterion (SC) for video-specific ATL.

2. **Novel Criteria (Sec. 4):** To improve the efficiency of ATL for HP estimation, we propose the following criteria for sample selection (see Fig. 1):

   - **Temporal Heatmap Continuity (THC):** Uncer-

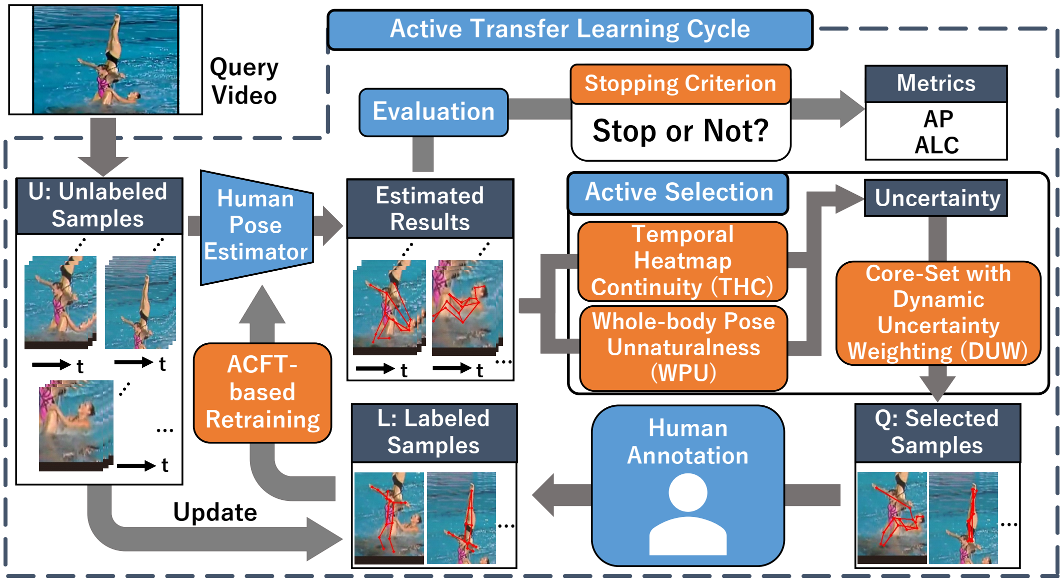


Figure 2. Overview of our proposed method (Sec. 3.1). Given an unlabeled query video, our active selection criteria (Sec. 4) select samples to be annotated based on estimated results. Then, the HP estimator is retrained using labeled samples following ACFT [66] (Sec. 3.2). The learning process terminates when the proposed SC is met (Sec. 3.3). The boxes highlighted in orange represent the novelty of this study.

tainty criterion based on the temporal change in the estimated heatmaps.

- **Whole-body Pose Unnaturalness (WPU):** Uncertainty criterion based on the unnaturalness of the estimated full-body HP.
- **Dynamic Uncertainty Weighting (DUW):** Integration of uncertainty and representativeness criteria via dynamically weighting them in Core-Set sampling [53] using estimated uncertainty.

3. **Comprehensive Evaluation (Sec. 5):** We conducted quantitative and qualitative evaluations to assess our proposed ATL framework. Our ablation study validates that each component of our method (THC, WPU, and DUW) contributes to the performance. The effectiveness of our proposed SC is also demonstrated.

## 2. Related Work

### 2.1. Human Pose (HP) Estimation

Deep learning has improved HP estimation [2, 4, 30, 31, 56, 64]. Simple Baseline [62] and FastPose [19] estimate 2D HPs in still images by estimating the probability distribution of each keypoint as a heatmap. For multi-person 2D HP estimation, two major approaches exist: the top-down approach [19, 24, 41], where all humans in the image are first detected, then each pose is estimated; and the bottom-up approach [8, 34], where all keypoints in the image are first detected, then assembled into individual HPs. Our method can be applied to both single and multi-person HP estimation, regardless of whether it is top-down or bottom-up.

### 2.2. Active Learning (AL)

AL is a learning method that actively selects training data to improve the performance of a learning model, aiming to reduce annotation costs for efficiency. The selection criteria in AL are generally divided into uncertainty criteria [32, 36, 39, 45, 54] and representativeness criteria [29, 53, 65]. Uncertainty criteria evaluate the uncertainty of the estimation results and select samples with high uncertainty for annotation. In contrast, representativeness criteria, including Core-Set [53], consider the distribution of unlabeled samples and select diverse samples.

AL is helpful for tasks with high annotation costs, such as HP estimation [22, 26]. Liu and Ferrari [39] proposed the Multiple Peak Entropy (MPE) as a quantification of uncertainty suitable for HP estimation. Mori *et al.* [45] proposed Temporal Pose Continuity (TPC) for HP estimation in videos, where the temporal change of the estimated pose is considered as uncertainty. We propose THC as a further extension of MPE and TPC (Sec. 4.1).

Shukla *et al.* [54] modeled keypoint-level and pose-level uncertainty in the context of out-of-distribution detection. They quantified the Visual Likelihood for estimated Poses (VL4Pose) based on the skeletal model representation using a Bayesian network. In this method, samples with lower VL4Pose are chosen as the samples with higher uncertainty. In contrast, our proposed WPU defines uncertainty based on the anomaly scores of AutoEncoder (AE) [12, 25, 43], from the perspective of anomaly detection (Sec. 4.2).

### 2.3. Active Transfer Learning (ATL)

The effectiveness of ATL has been demonstrated in tasks such as hyperspectral image classification [16], and handwriting recognition [7]. As an application of medical image analysis, Zhou *et al*. [66] proposed Active, Continual Fine-Tuning (ACFT), an ATL method that sequentially adapts a model pre-trained on a general dataset to a different domain. In this study, we utilize a learning method that further improves upon ACFT. Thereby, we adopt a pre-trained HP estimator from learned domains to unknown individual video domains efficiently. The procedure and our extension of ACFT are described in the next Section 3.

## 3. Video-Specific HP Estimation via ATL

This section introduces our video-specific ATL framework. The overview of our method is described in Sec. 3.1. In Sec. 3.2, we explain our refined retraining method, an improvement upon the existing ATL approach, ACFT [66]. In Sec. 3.3, we propose a novel SC for ATL. Details on the active selection procedure are described in Sec. 4.

### 3.1. Overview

In our ATL framework, the ATL cycle (starts with $c = 0$) is repeated to continuously adapt the HP estimator $M_c$ to a query video. As shown in Fig. 2, a query video capturing human motion can be considered as a set of unlabeled samples $U = \{x_1, x_2, ..., x_N\}$, where $N$ is the total number of samples. For example, if the number of human instances in the 1st, 2nd, and 3rd frames in the 3-frame video are 2, 4, and 1, respectively, $N = 7$. We assume that all $N$ samples in the video are in the same domain.

The $c$-th ATL cycle starts with HP estimation on all $N$ samples within the video ("Human Pose Estimator" in Fig. 2). This estimation serves as the initial phase for the subsequent phases in the ATL cycle. The next phase is the "Active Selection" (also illustrated in Fig. 2), in which each sample's uncertainty $C(x_i)$ where $x_i \in U$ is evaluated based on THC and WPU (described in Sec. 4.1 and Sec. 4.2, respectively). The computed uncertainties are then utilized in Core-Set sampling [53] with DUW, where uncertain and diverse samples are selected (Sec. 4.3).

Following active selection, a human annotator manually annotates the chosen samples $Q$ ("Human Annotation" in Fig. 2). The final phase of each ATL cycle is "ACFT-based retraining" (Sec. 3.2) of the HP estimator. In this phase, both the newly annotated samples $Q$ and already annotated samples $L$ are used for retraining $M_c$ to $M_{c+1}$. The conditions for retraining are adjusted based on the estimated results, which allows us stable continual learning.

At the end of each ATL cycle, the criterion for terminating ATL is evaluated based on the estimated results ("Stopping Criterion" in Fig. 2).

### 3.2. ACFT-based Retraining

In retraining the HP estimator, we adhere to the learning strategy proposed in ACFT [66]. That is, we continually fine-tune a pre-trained model $M_0$ with samples $L$ added via active selection. In the beginning, the number of labeled samples is initialized to zero (i.e., $|L| = 0$), and the HP estimator $M_0$ is pre-trained across a broad domain of a large dataset (i.e., source domain). As repeating the ATL Cycle, the HP estimator $M_c$ is adapted to the domain of each query video (i.e., target domain) continually.

However, in the original ACFT [66], retraining is conducted over a fixed number of epochs $E$, which presents issues in terms of both performance improvement and execution time. That is, when the HP estimator is not adapted to the domain of the query video, it is necessary to increase $E$ to make drastic changes in parameters of $M_c$. Conversely, as the performance of the HP estimator improves, decreasing $E$ can reduce execution time. Following this observation, we determine the number of epochs in our ACFT-based retraining as follows:

$$E_c = \alpha \times (1 - G_c) \in \mathbb{N}, \tag{1}$$

where $G_c \in [0, 1]$ represents the performance of $M_c$ for unlabeled samples $U$, and $\alpha$ is a hyperparameter.

However, in actual operation, it is impossible to correctly evaluate the estimation performance $G_c$ since we do not possess the ground truth for unlabeled samples. Therefore, an alternate metric representing the estimation performance is needed. Hence, we evaluate the performance of the estimated results only with the newly selected samples $Q$ with annotated ground truth. The generalization performance of the HP estimator $M_c$ in the $c$ th cycle is estimated as follows:

$$G_c \approx \frac{1}{|Q|} \sum_{x \in Q} OKS(x), \tag{2}$$

where $OKS(x) \in [0, 1]$ represents the Object Keypoint Similarity (OKS), which is an evaluation metric for HP estimation. The value of OKS increases as the estimated pose becomes more similar to the ground truth.

Additionally, we rethink the retraining of already labeled samples. The original ACFT is applied to a classification task [66], and execution time is reduced by only retraining the misclassified labeled samples. However, since HP estimation is the regression task, it is not possible to determine misestimated data in the same way. We newly define the misestimated labeled samples $R$ as follows:

$$R = \{x | OKS(x) < \theta + m\}, \tag{3}$$

where $\theta$ is a user-defined accuracy threshold and $m$ represents a margin set to ensure the estimation performance stably exceeds the required accuracy $\theta$.

### 3.3. Stopping Criterion

In applications of AL, SC, which determines when to terminate learning, is indispensable. Typically, a common SC involves achieving the desired accuracy on cross-validation data or a hold-out set [61], used in ACFT [66] as well. However, collecting sufficient validation data in AL is impractical because it requires much more extra annotation. SC can be also defined with agreement among multiple estimators [5, 6] so that SC is met if the results of many estimators are equal. However, the computational cost increases in accordance with the number of estimators. While uncertainty-based SCs [35, 68] are also proposed, the reliability of such an SC is questionable since uncertainty merely serves as an estimate of the degree of error.

Zhu and Hovy proposed the Min-error criterion [67], which measures the accuracy of predicted results for newly selected unlabeled samples (i.e., samples in $Q$). If the accuracy surpasses a certain threshold, the learning process is terminated. This criterion does not require additional computation and is a more reliable measure since it uses the actual error. When expressed using OKS, it is represented as follows:

$$SC_{Min} = \begin{cases} 1 & \frac{1}{|Q|}\sum_{x \in Q} OKS(x) > \theta \\ 0 & otherwise. \end{cases} \tag{4}$$

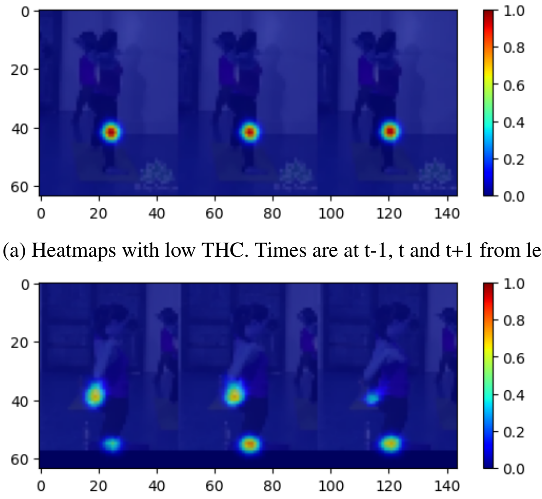
However, as pointed out by Zhu and Hovy [67], there is a risk that $SC_{Min}$ may lead to premature termination when the number of newly added unlabeled samples is small. Therefore, we propose the following $SC_{All}$, a new SC suitable for the practical use in video-specific ATL as follows:

$$SC_{All} = \begin{cases} 1 & \forall x \in \{Q \cup L\}, OKS(x) > \theta \\ 0 & otherwise. \end{cases} \tag{5}$$

We compare the effects of these two criteria in the experiment (Sec. 5).

## 4. Uncertainty Criteria for Efficient Active Transfer Learning

In the active selection phase, we actively select samples for human annotation based on the results of HP estimation for each unlabeled sample. This active selection is depicted in Fig. 2. Our proposed uncertainty criteria, THC (Sec. 4.1) and WPU (Sec. 4.2) calculate the uncertainty from the heatmaps obtained during HP estimation and the estimated whole-body poses, respectively. These uncertainties serve as weights in our DUW Core-Set [53] approach (Sec. 4.3). This approach guides the selection of diverse and uncertain samples effectively.



(a) Heatmaps with low THC. Times are at t-1, t and t+1 from left to right.

(b) Heatmaps with high THC. Times are at t-1, t, and t+1 from left to right.

Figure 3. Qualitative examples of our THC. (a) There is a strong peak at a single point in the heatmap between adjacent frames consistently. (b) In contrast, the estimations are inconsistent and the peaks in the heatmap are dispersed.

### 4.1. THC: Uncertainty Criterion based on Temporal Heatmap Continuity

In this section, we introduce a new uncertainty criterion called THC, an extension of MPE [39] and TPC [45].

Many HP estimation methods output heatmaps for each keypoint and select the maximum probability position in the heatmap as the keypoint position. The rest of the information is discarded although it contains valuable information for assessing the estimation results. For instance, a single peak with low variance might indicate a confident estimation. In contrast, the presence of multiple peaks with high variance could be a signal of an uncertain estimation.

Based on this concept, MPE [39] was proposed as an uncertainty criterion to replace conventional methods such as Least Confidence (LC) [36], which quantify uncertainty only from the maximum values in the heatmap. MPE utilizes not only the maximum values but also local peaks in the heatmap to calculate entropy, making use of the rich information of the heatmap.

Furthermore, when evaluating HP estimation results in videos, we can utilize the property that ‘the correct keypoint position does not change significantly between adjacent frames’. TPC [45] was proposed based on this idea, quantifying the uncertainty of HP estimation in videos by the temporal change of the estimated poses. It sums up the Euclidean distances of estimated keypoint positions between adjacent frames, considering larger values to indicate higher uncertainty.

We introduce a new uncertainty criterion extending upon both MPE and TPC, namely, THC. This quantifies uncer-

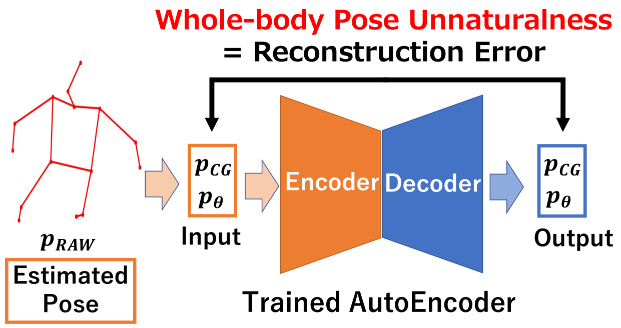


Figure 4. Calculation of Whole-body Pose Unnaturalness (WPU). Given the estimated HP, its hybrid feature [46] is calculated and fed into the trained AutoEncoder (AE). WPU is defined by the reconstruction error between the input and output of AE.

tainty by utilizing spatially rich information in heatmaps. In addition, THC considers the continuity in estimation results between temporally adjacent frames.

Specifically, the uncertainty is quantified by calculating the Sum of the Absolute Difference (SAD) of estimated heatmaps for each keypoint between adjacent frames, as shown below:

$$\begin{aligned} C_{THC}(F_t) = {} & \frac{1}{K}\sum_{k=1}^{K} SAD(H_{t-1}^k, H_t^k) \\ & + \frac{1}{K}\sum_{k=1}^{K} SAD(H_t^k, H_{t+1}^k) \qquad (6) \\ = {} & \frac{1}{K}\sum_{k=1}^{K}\sum_{p_t^k \in H_t^k} (|p_{t-1}^k - p_t^k| \\ & + |p_t^k - p_{t+1}^k|), \qquad (7) \end{aligned}$$

where $K$, $H_t^k$, and $p_t^k$ denote the number of keypoints in each HP, the estimated heatmap for keypoint $k$ in frame $t$, and the probability at each position in $H_t^k$, respectively. Examples of heatmaps with low THC and high THC are shown in Fig. 3.

## 4.2. WPU: Uncertainty Criterion based on Whole-body Pose Unnaturalness

Previous AL for HP estimation [39, 45, 59] computes the uncertainty of an entire body pose by summing the uncertainty of each keypoint. However, this approach may miss unnatural poses where individual keypoints exhibit low uncertainty. Therefore, we introduce WPU, a novel uncertainty criterion that quantifies the unnaturalness of whole-body poses in the context of anomaly detection.

In our approach, we train AE [12, 25, 43], a simple anomaly detection model, on ground-truth HPs from the dataset. The trained AE successfully reconstructs natural HPs but struggles to reconstruct unnatural ones, resulting in larger reconstruction errors. Therefore, this error can measure pose-level uncertainty, as shown in Fig. 4.

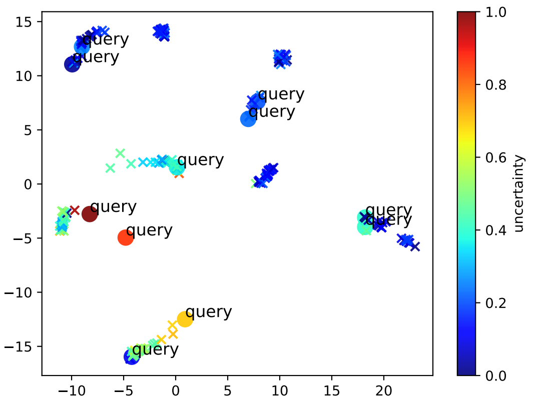


Figure 5. A visualization result of the sample selection of our proposed criterion (THC+WPU+DUW). We have utilized DensMAP [47] to plot feature vectors extracted by the HP estimator. In this plot, circles represent newly selected samples, while cross marks denote unlabeled samples that were not selected for the current ATL cycle. The color of the plot corresponds to the normalized uncertainty.

Although VL4Pose [54] also tackled the issue of pose-level uncertainty using a Bayesian Neural Network architecture, our WPU has a significant advantage in model efficiency. While VL4Pose utilizes ≈ 2.1M parameters, WPU only needs 2.6K parameters. This results in a simple but effective uncertainty measure with fewer computational costs.

To train the AE well, we do not use raw keypoint coordinates from the dataset but rather calculate the Hybrid feature [46] for inputs. This Hybrid feature is a pose representation robust against scaling differences and rotations, consisting of the Center of Gravity (CG) feature $p_{CG}$ [50] and the Angle feature $p_\theta$ [55], which represents eight critical joint angles.

Furthermore, through ATL procedure, the pre-trained AE is retrained on labeled poses in the query video. By learning natural poses in the target domain (i.e., the query video) during ATL, the AE can learn the feature of natural poses that are not included in the source domain (i.e., the training dataset).

## 4.3. DUW: Dynamic Uncertainty Weighting of Core-Set sampling

Many AL studies [39, 45, 54] tend to rely solely on uncertainty for sample selection, but it can fall into selection bias [20, 52]. This is largely because uncertainty criteria tend to select similar data, like redundant samples from continuous video frames in our setting.

Table 2. Quantitative results of our proposed video-specific ATL on PoseTrack21 [17]. Red and blue indicate the best and the second best, respectively. AP@0.6 is the average AP of 170 test videos with a 0.6 OKS threshold. "5%" means the estimation result with 5% labeled samples in the query video. ALC values are also calculated by an average of 170 test videos.

| Criterion | AP@0.6 (%) | | | ALC |
|---|---|---|---|---|
| | 5% | 20% | 40% | (%) |
| Random | 87.76 | 96.09 | 97.39 | 96.91 |
| LC [36] | 77.49 | 94.60 | 96.77 | 95.74 |
| MPE [39] | 78.96 | 95.09 | 97.23 | 96.11 |
| TPC [45] | 83.38 | 95.32 | 97.31 | 96.40 |
| k-means [65] | **93.97** | 96.37 | 98.11 | 97.65 |
| Core-Set [53] | 93.18 | **97.62** | **98.60** | **98.12** |
| **Ours (THC+WPU+DUW)** | **93.35** | **97.90** | **98.77** | **98.21** |

Therefore, in this research, we propose DUW, which balances both uncertainty and diversity by extending the acquisition function of Core-Set sampling [53]. In the original Core-Set sampling, $u$, a new sample to be labeled is sequentially selected according to the following acquisition function:

$$u = \arg\max_{i \in U} \min_{j \in L} \Delta(x_i, x_j), \tag{8}$$

where $\Delta(x_i, x_j)$ is the Euclidean distance between sample's feature vectors $x_i$ and $x_j$.

In contrast, we define the acquisition function in DUW based on each sample's uncertainty score $C(x_i)$ in the following way:

$$u = \arg\max_{i \in U}\{\min_{j \in L}\{(1-G_c)\times\Delta(x_i, x_j)\}+G_c\times\lambda C(x_i)\}, \tag{9}$$

Here, $G_c$ is approximated by Eq. (2) and $\lambda$ is a hyperparameter, respectively.

As depicted in Eq. (9), when $\lambda$ is 0, the sample selection is equal to the original Core-Set sampling [65]. Conversely, when $\lambda$ is large, sample selection is heavily influenced by the uncertainty score. The balance between uncertainty and diversity dynamically changes based on $G_c$ too. When $G_c$ is low, the selection prioritizes coverage over the whole samples, while it emphasizes uncertain estimations when $G_c$ is high. Since $G_c$ increases through ATL, representative samples tend to be selected in the initial phase, and uncertain samples come in selection as ATL progresses. With this formulation, we aim to rapidly cover the data distribution within the query video at the initial cycles of ATL and subsequently promote the identification of remaining hard samples through uncertainty measurement. The example of informative and diverse sample selection using DUW is shown in Fig. 5.

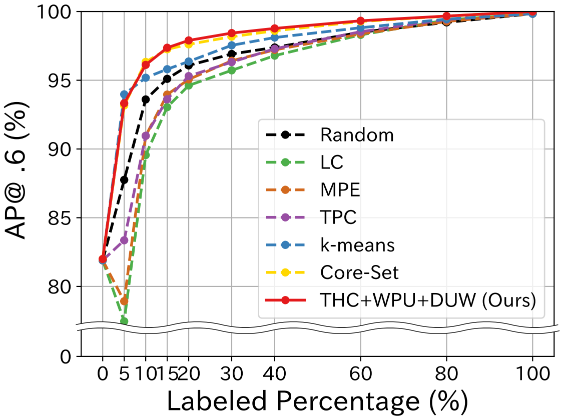


Figure 6. Learning Curve of video-specific ATL on PoseTrack21 [17].

## 5. Experiments

This section is broken down into five parts: (1) We outline evaluation metrics and specify our implementation (Sec. 5.1). (2) We introduce the various selection criteria used, including our proposed method (Sec. 5.2). (3) We contrast our approach with the baseline and several state-of-the-art methods (Sec. 5.3). (4) We conduct an ablation study to verify the effect of each component in our framework (Sec. 5.4). (5) Lastly, we examine the effectiveness of our newly proposed SC (Sec. 5.5).

### 5.1. Evaluation and Implementation Details

**Dataset.** Two large-scale datasets for multi-person 2D HP estimation, PoseTrack21 [17] and JRDB-Pose [60] were employed for our experiments.

PoseTrack21 [17] consists of 593 training videos and 170 validation videos. Since PoseTrack21 does not provide its test data labels, we utilized 579 out of 593 original training videos for training and the remaining 14 videos for validation and conducted evaluations using the original 170 validation videos. For each video evaluation, about 30 annotated frames were utilized. A skeleton-based pose representation in PoseTrack21 consists of 15 keypoints.

In terms of JRDB-Pose [60], as well as PoseTrack21, it does not have available labels for the test split. Thus, we utilized the 27 provided videos, dividing them into 10 for training, 2 for validation, and 15 for testing. For each video, we used the first 150 frames (five times the number for PoseTrack21 [17]) extracted from stitched images. Poses in JRDB-Pose consist of 17 keypoints.

**HP Estimator.** In the video-specific ATL, we followed the top-down approach manner [19, 24, 41]. To simplify the process, we used ground truth bounding boxes and tracking IDs as detection results. Simple Baseline [62] and FastPose [19] were employed to estimate HP in PoseTrack21 [17] and JRDB-Pose [60], respectively. We pre-trained both models on the 579 training videos from PoseTrack21 [17] using the Adam optimizer [33] with a learning rate of $1.0e^{-3}$ and 62/57 epochs terminated by early-stopping, respectively. Training data are augmented by flipping, rotation, and scaling. Subsequently, we fine-tuned this pre-trained FastPose on 10 videos from JRDB-Pose with $lr = 5.0e^{-4}$ and 40 epochs.

**Evaluation Metrics.** The efficacy of our proposed video-specific ATL was evaluated using Average Precision (AP) and Area under the Learning Curve (ALC) [15, 66]. The AP was used to assess the HP estimation results at each ATL cycle. We adopted the MS COCO's calculation [38], using OKS to determine the accuracy of the estimated pose. The ALC was used to evaluate the overall ATL efficiency. The ALC is calculated from a graph, where the vertical axis represents AP (%) and the horizontal axis represents the percentage of labeled samples (%). It should be noted that in the calculation of AP and ALC, poses annotated during ATL are considered correctly estimated. That is to say, AP surely reaches 100% when 100% HPs are labeled.

**Human Annotation.** Consistent with other AL studies [39, 45], we simulated human annotation by automatically providing ground truth annotations for selected samples. The annotated samples are incrementally added for retraining, as illustrated in Fig. 2. The [5%, 5%, 5%, 5%, 10%, 10%, 20%, 20%, 20%] of the total HPs in the query video are sequentially added at each ATL cycle.

**Retraining.** For retraining the HP estimator, we used the AdamW [42] optimizer with learning rate $= 2.5e^{-4}$, weight decay $= 0.7$, and $\gamma = 0.99$, respectively. $\alpha$ in Eq. (1) was set to 250. To prevent overfitting, data augmentation techniques such as flipping, rotation, and scaling were applied to the retrained samples.

## 5.2. Active Selection Criteria

For comparative experiments, our proposed method and several ATL approaches are implemented with the following selection criteria:

- **Random:** Random uniform sampling.
- **LC:** A traditional uncertainty measurement described in [36]. The implementation followed [39].
- **MPE:** An uncertainty criterion in [39].
- **TPC:** An uncertainty criterion in [45].
- **k-means:** A representativeness criterion used in [65].
- **Core-Set:** An original Core-Set sampling in [53]. This implementation followed [28].
- **THC:** Our uncertainty criterion proposed in Sec. 4.1.
- **WPU:** Our uncertainty criterion proposed in Sec. 4.2. For the experiment on PoseTrack21 [17], the AE was trained with the ground truth keypoints from the 579 training videos (300 epochs with learning rate $= 1.0e^{-3}$ by Adam [33]). For JRDB-Pose [60], pre-training was conducted using 10/2 videos in the train/val set. Both the encoder and the decoder of the AE have four layers each, and the dimension of the latent variables is 4. The AE is retrained at each ATL cycle by Adam (20 epochs with $lr = 8.0e^{-4}$).
- **DUW:** The combination of uncertainty criteria and Core-Set sampling [53] proposed in Sec. 4.3. The value of $\lambda$ in Eq. (9) was set to 0.01 and 1000 for PoseTrack21 [17] and JRDB-Pose based on a hyper-parameter search based on the performance of video-specific ATL on validation videos, respectively. We set the weights of THC and WPU at a 1:1 ratio in Sec. 5.3.

Table 3. Ablation study results of video-specific ATL on PoseTrack21 [17]. Red and blue indicate the best and the second best, respectively. AP@0.6 is the average AP of 170 test videos with a 0.6 OKS threshold. "5%" means the estimation result with 5% labeled samples. ALC is also the average of 170 test videos.

| Criterion | AP@0.6 (%) | | | ALC |
|---|---|---|---|---|
| | 5% | 20% | 40% | (%) |
| Core-Set [53] | 93.18 | 97.62 | 98.60 | 98.12 |
| THC | 82.59 | 92.86 | 96.43 | 95.45 |
| WPU | 85.56 | 94.74 | 97.31 | 96.45 |
| THC+WPU | 84.82 | 95.17 | 97.25 | 96.51 |
| THC+DUW | 93.12 | 97.70 | **98.91** | 98.19 |
| WPU+DUW | **93.19** | **97.87** | 98.76 | 98.17 |
| THC+WPU+DUW (fixed) | 93.02 | 97.68 | 98.81 | 98.14 |
| THC+WPU+DUW (increase) | 93.18 | 97.86 | 98.80 | 98.16 |
| THC+WPU+DUW (const) | **93.35** | **97.90** | 98.77 | **98.21** |
| THC+WPU+DUW (decrease) | 93.08 | 97.72 | **98.94** | **98.24** |

For methods that perform sample selection based solely on uncertainty, samples with higher $C(x_i)$ are prioritized to be added to the labeled data. For further details, please refer to our codebase at: https://github.com/ImIntheMiddle/VATL4Pose-WACV2024

## 5.3. Baseline and State-of-the-art Comparison

Figure 6 and Table 2 show quantitative results of the proposed video-specific ATL on PoseTrack21 [17]. While the performances of all uncertainty-based methods [36, 39, 45] are less than the random selection, our method ("THC+WPU+DUW") outperforms other methods throughout the entire ATL process. Video-specific ATL with our methods can efficiently achieve accurate HP estimation (e.g., as shown in Table 2, our method got AP@0.6 $\approx$ 98% with only 20% labeled samples).

Table 4. Quantitative results of our proposed video-specific ATL on JRDB-Pose [60]. Red and blue indicate the best and the second best, respectively. AP is an average of 15 test videos. "5%" means the estimation result with 5% labeled samples in the query video. ALC values are also an average of 15 test videos.

| Criterion | AP@0.6 (%) | | | ALC |
|---|---|---|---|---|
| | 5% | 20% | 40% | (%) |
| Random | 88.16 | 94.19 | 96.46 | 95.42 |
| LC [36] | 65.04 | 89.34 | 94.84 | 92.67 |
| MPE [39] | 81.78 | 95.74 | **98.03** | 95.76 |
| TPC [45] | 74.83 | 92.25 | 95.74 | 93.76 |
| k-means [65] | **88.97** | **95.98** | 97.53 | **96.41** |
| Core-Set [53] | 85.09 | 95.27 | 96.80 | 95.60 |
| **Ours (THC+WPU+DUW)** | **89.76** | **96.48** | **97.59** | **96.52** |

The results for the 15 test videos from JRDB-Pose are presented in Table 4. Here too, our proposed method ("Ours") achieved performance close to 90% with only 5% of the labeling. Furthermore, the ALC performance of the proposed method stably outperforms comparative methods across the entire ATL procedure. For the complete table and evaluation with another metric, please refer to Sec. B in the supplementary materials.

### 5.4. Ablation Studies

Table 3 shows the ablation study results of video-specific ATL on PoseTrack21 [17]. Our proposed methods, THC, WPU, and DUW are all used together, resulting in the highest ALC. THC+DUW and WPU+DUW surpassed the performance of the original Core-Set [53] due to the incorporation of uncertainty in sample selection.

When comparing only uncertainty, THC+WPU achieves the highest ALC including the other methods in Table 2. This suggests the effect of combining THC with WPU.

Next, we compared our method with a case where the balance between uncertainty and representativeness in Eq. (9) is not dynamically adjusted by $G_c$, only using a fixed hyperparameter, $\lambda$. $\lambda$ was 1 based on a hyperparameter search using video-specific ATL on the 14 validation videos. While the results (THC+WPU+DUW (fixed)) surpassed Core-Set [53], it is poorer than other results of THC+WPU+DUW, using the dynamic weighting by $G_c$. This demonstrates the effectiveness of dynamically adjusting the balance between uncertainty and representativeness.

Lastly, we conducted a detailed ablation study on the combination of THC and WPU. In this study, we compared three scenarios: increasing the proportion of THC linearly from 0 to 1 ("increase") based on the labeled percentage, setting the same weight for THC and WPU ("const"), and decreasing the proportion of THC linearly from 1 to 0 ("decrease"). As shown in Tab. 3, it is evident that setting the balance between THC and WPU to "const" enjoys significant performance improvement during the early stages and the midpoint of ATL. Nevertheless, the performance of "decrease" tends to be enhanced in the later stages. These findings further motivate the need to design appropriate strategies for combining THC and WPU.

For further detailed analysis and results, please refer to Sec. A, C and D in the supplementary material.

Table 5. Effectiveness of our SC. $\theta$ is the target value of OKS defined by the user. AP@$\theta$ represents the value of AP calculated at the time learning stopped, with $\theta$ as the threshold. "Stopped" and "Actual" denote the labeled percentage when learning was stopped by the SC and the labeled percentage when all samples actually reached an OKS above $\theta$, respectively.

| SC | $\theta$ | AP@$\theta$ (%) | Stopped (%) | Actual (%) |
|---|---|---|---|---|
| $SC_{Min}$ [67] | 0.5 | **96.86** | 10.04 | 36.90 |
| | 0.6 | **96.19** | 10.85 | 40.26 |
| | 0.7 | **95.75** | 12.55 | 46.55 |
| | 0.8 | **95.58** | 16.94 | 56.54 |
| $SC_{All}$ (Ours) | 0.5 | **99.25** | 29.61 | 36.90 |
| | 0.6 | **99.55** | 33.38 | 40.26 |
| | 0.7 | **99.60** | 39.61 | 46.55 |
| | 0.8 | **99.50** | 49.46 | 56.54 |

### 5.5. Effectiveness of proposed SC

Table 5 compares existing SC, Min-error [67], with our $SC_{All}$ proposed in Sec. 3.3. The labeled samples were increased by 10% increments from 0% and the margin $m$ in Eq. (3) was set to 0.05. As hypothesized in Sec. 3.3, Min-error relies on the average value of a small number of samples and encounters premature stops at $<$ 97% AP. On the other hand, our $SC_{All}$ terminates ATL when AP has almost reached 100% for any threshold, thereby ensuring the HP estimation accuracy that the user demands more precisely.

## 6. Concluding Remarks

In this study, we addressed video-specific HP estimation using ATL for the first time. We revisited the existing ATL method [66] and proposed a retraining method suitable for video-specific ATL along with novel SC. To enhance learning efficiency, we proposed three novel selection criteria: THC, WPU, and DUW.

Our criteria outperformed existing methods by enabling the selection of uncertain and diverse samples. Additionally, we found that proposed $SC_{All}$ can accurately determine the timing to terminate ATL for practical use.

For future work, we suggest integrating video-based HP estimation methods to enhance performance [4, 30, 63, 64]. Additionally, utilizing semi-supervised learning [21, 49, 51, 59] could further reduce annotation cost.

## References

[1] Reza Afrouzian, Hadi Seyedarabi, and Shohreh Kasaei. Correction to: Pose estimation of soccer players using multiple uncalibrated cameras. *Multim. Tools Appl.*, 78(2):2641, 2019. 1

[2] Bruno Artacho and Andreas E. Savakis. Unipose: Unified human pose estimation in single images and videos. In *CVPR*, 2020. 1, 2

[3] Tobias Baumgartner and Stefanie Klatt. Monocular 3d human pose estimation for sports broadcasts using partial sports field registration. In *CVSports (CVPRW)*, 2023. 1

[4] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. Learning temporal pose estimation from sparsely-labeled videos. In *NeurIPS*, 2019. 2, 8

[5] Michael Bloodgood and John Grothendieck. Analysis of stopping active learning based on stabilizing predictions. In *CoNLL*, 2013. 4

[6] Michael Bloodgood and K. Vijay-Shanker. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In *CoNLL*, 2009. 4

[7] Eric Burdett, Stanley Fujimoto, Timothy Brown, Ammon Shurtz, Daniel Segrera, Lawry Sorenson, Mark J. Clement, and Joseph Price. Active transfer learning for handwriting recognition. In *ICFHR*, 2022. 1, 3

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 2

[9] Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel Pierce, Daniel Bogen, Laura Prosser, Michelle Johnson, and Konrad Kording. Computer vision to automatically assess infant neuromotor risk. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 28:2431–2442, 2020. 1

[10] Kenny Chen, Paolo Gabriel, Abdulwahab Alasfour, Chenghao Gong, Werner K. Doyle, Orrin Devinsky, Daniel Friedman, Patricia Dugan, Lucia Melloni, Thomas Thesen, David Gonda, Shifteh Sattar, Sonya Wang, and Vikash Gilja. Patient-specific pose estimation in clinical environments. *IEEE J. Transl. Eng. Health Med.*, pages 1–11, 2018. 1

[11] Shuhong Chen and Matthias Zwicker. Transfer learning for pose estimation of illustrated characters. In *WACV*, 2022. 1

[12] Zhaomin Chen, Chai Kiat Yeo, Bu-Sung Lee, and Chiew Tong Lau. Autoencoder-based network anomaly detection. In *WTS*, 2018. 2, 5

[13] Mickael Cormier, Aris Clepe, Andreas Specker, and Jürgen Beyerer. Where are we with human pose estimation in real-world surveillance? In *RWS (WACVW)*, 2022. 1

[14] Mickael Cormier, Fabian Röpke, Thomas Golda, and Jürgen Beyerer. Interactive labeling for human pose estimation in surveillance videos. In *ILDAV (ICCVW)*, 2021. 1

[15] Matt Culver, Kun Deng, and Stephen Scott. Active learning to maximize area under the ROC curve. In *ICDM*, 2006. 7

[16] Cheng Deng, Yumeng Xue, Xianglong Liu, Chao Li, and Dacheng Tao. Active transfer learning network: A unified deep joint spectral-spatial feature learning model for hyperspectral image classification. *IEEE Trans. Geosci. Remote. Sens.*, 57(3):1741–1754, 2019. 1, 3

[17] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *CVPR*, 2022. 1, 6, 7, 8

[18] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, 2019. 1

[19] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 1, 2, 6, 7

[20] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *ICLR*, 2021. 5

[21] Qi Feng, Kun He, He Wen, Cem Keskin, and Yuting Ye. Rethinking the data annotation process for multi-view 3d pose estimation with active learning and self-training. In *WACV*, 2023. 8

[22] Erik Gärtner, Aleksis Pirinen, and Cristian Sminchisescu. Deep reinforcement learning for active human pose estimation. In *AAAI*, 2020. 1, 2

[23] Brennan Gebotys, Alexander Wong, and David A. Clausi. POOF: efficient goalie pose annotation using optical flow. In *icSPORTS*, 2021. 1

[24] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In *CVPR*, 2018. 2, 6

[25] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, 2019. 2, 5

[26] Jia Gong, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Meta agent teaming active learning for pose estimation. In *CVPR*, 2022. 1, 2

[27] Daniel Groos, Lars Adde, Ragnhild Støen, Heri Ramampiaro, and Espen A. F. Ihlen. Towards human-level performance on automatic pose estimation of infant spontaneous movements. *Comput. Med. Imaging Graphics*, 95:102012, 2022. 1

[28] Kuan-Hao Huang. Deepal: Deep active learning in python. *arXiv preprint arXiv:2111.15258*, 2021. 7

[29] Suyog Dutt Jain and Kristen Grauman. Active image segmentation propagation. In *CVPR*, 2016. 1, 2

[30] Kyung-Min Jin, Gun-Hee Lee, and Seong-Whan Lee. Otpose: Occlusion-aware transformer for pose estimation in sparsely-labeled videos. In *SMC*, 2022. 2, 8

[31] Yuki Kawana, Norimichi Ukita, Jia-Bin Huang, and Ming-Hsuan Yang. Ensemble convolutional neural networks for pose estimation. *Comput. Vis. Image Underst.*, 169:62–74, 2018. 2

[32] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017. 2

[33] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7

[34] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Trans. Intell. Transp. Syst.*, 23(8):13498–13511, 2022. 1, 2

[35] Florian Laws and Hinrich Schütze. Stopping criteria for active learning of named entity recognition. In *COLING*, 2008. 4

[36] David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Mach. Learn.*, pages 148–156, 1994. 2, 4, 6, 7, 8

[37] Yizhuo Li, Miao Hao, Zonglin Di, Nitesh B. Gundavarapu, and Xiaolong Wang. Test-time personalization with a transformer for human pose estimation. In *NeurIPS*, 2021. 1

[38] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 7

[39] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *ICCV*, 2017. 1, 2, 4, 5, 6, 7, 8

[40] Yazhou Liu, Pongsak Lasang, Sugiri Pranata, Shengmei Shen, and Wenchao Zhang. Driver pose estimation using recurrent lightweight network and virtual data augmented transfer learning. *IEEE Trans. Intell. Transp. Syst.*, 20(10):3818–3831, 2019. 1

[41] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *CVPR*, 2021. 1, 2, 6

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 7

[43] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *CVPR*, 2020. 1, 2, 5

[44] Takuya Matsumoto, Kodai Shimosato, Takahiro Maeda, Tatsuya Murakami, Koji Murakoso, Kazuhiko Mino, and Norimichi Ukita. Automatic human pose annotation for loose-fitting clothes. In *MVA*, 2019. 1

[45] Taro Mori, Daisuke Deguchi, Yasutomo Kawanishi, Ichiro Ide, Hiroshi Murase, and Tetsuo Inoshita. Active learning for human pose estimation based on temporal pose continuity. In *IWAIT*, 2022. 1, 2, 4, 5, 6, 7, 8

[46] Bharath Raj N., Anand Subramanian, Kashyap Ravichandran, and N. Venkateswaran. Exploring techniques to improve activity recognition using human pose skeletons. In *HADCV (WACVW)*, 2020. 5

[47] Ashwin Narayan, Bonnie Berger, and Hyunghoon Cho. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nat. Biotechnol.*, 39:765 – 774, 2021. 5

[48] Shunsuke Ochi and Jun Miura. Depth-based in-bed human pose estimation with synthetic dataset generation and deep keypoint estimation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *ACVR (ECCVW)*, 2022. 1

[49] Gaurav Patel, Jan P. Allebach, and Qiang Qiu. Seq-ups: Sequential uncertainty-aware pseudo-label selection for semi-supervised text recognition. In *WACV*, 2023. 8

[50] Marina Pismenskova, Oxana Balabaeva, Viacheslav Voronin, and Valentin Fedosov. Classification of a two-dimensional pose using a human skeleton. *MATEC Web Conf.*, 132:05016, 2017. 5

[51] Aneesh Rangnekar, Christopher Kanan, and Matthew J. Hoffman. Semantic segmentation with active semi-supervised learning. In *WACV*, 2023. 8

[52] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9):1–40, 2022. 5

[53] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 1, 2, 3, 4, 6, 7, 8

[54] Megh Shukla, Roshan Roy, Pankaj Singh, Shuaib Ahmed, and Alexandre Alahi. Vl4pose: Active learning through out-of-distribution detection for pose estimation. In *BMVC*, 2022. 1, 2, 5

[55] Amarjot Singh, Devendra Patil, and S. N. Omkar. Eye in the sky: Real-time drone surveillance system (DSS) for violent individuals identification using scatternet hybrid deep learning network. In *ECV (CVPRW)*, 2018. 5

[56] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 2

[57] Norimichi Ukita, Michiro Hirai, and Masatsugu Kidode. Complex volume and pose tracking with probabilistic dynamical models and visual hull constraints. In *ICCV*, 2009. 1

[58] Norimichi Ukita, Ryosuke Tsuji, and Masatsugu Kidode. Real-time shape analysis of a human body in clothing using time-series part-labeled volumes. In *ECCV*, 2008. 1

[59] Norimichi Ukita and Yusuke Uematsu. Semi- and weakly-supervised human pose estimation. *Comput. Vis. Image Underst.*, 170:67–78, 2018. 5, 8

[60] Edward Vendrow, Duy-Tho Le, Jianfei Cai, and Hamid Rezatofighi. Jrdb-pose: A large-scale dataset for multi-person pose estimation and tracking. In *CVPR*, 2023. 1, 6, 7, 8

[61] Andreas Vlachos. A stopping criterion for active learning. *Comput. Speech Lang.*, 22(3):295–312, 2008. 4

[62] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2, 6

[63] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for $10^{\times}$ efficient 2d and 3d pose estimation. In *ECCV*, 2022. 8

[64] Yuexi Zhang, Yin Wang, Octavia I. Camps, and Mario Sznaier. Key frame proposal network for efficient pose estimation in videos. In *ECCV*, 2020. 2, 8

[65] Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019. 2, 6, 7, 8

[66] Zongwei Zhou, Jae Y. Shin, Suryakanth R. Gurudu, Michael B. Gotway, and Jianming Liang. Active, continual fine tuning of convolutional neural networks for reducing annotation efforts. *Medical Image Anal.*, 71:101997, 2021. 1, 2, 3, 4, 7, 8

[67] Jingbo Zhu, Huizhen Wang, and Eduard H. Hovy. Learning a stopping criterion for active learning for word sense disambiguation and text classification. In *IJCNLP*, 2008. 4, 8

[68] Jingbo Zhu, Huizhen Wang, and Eduard H. Hovy. Multi-criteria-based strategy to stop active learning for data annotation. In *COLING*, 2008. 4