Active Transfer Learning for Efficient Video-Specific Human Pose Estimation Supplementary Material

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

Abstract

This supplementary material covers the additional results and visualizations that were excluded from the main paper due to a lack of space. Specifically, the following contents are included: i) The complete table of our experiment results in the main paper, the impact of hyperparameter changes in our proposed method, and the validation of proposed uncertainty criteria (Sec. A). ii) The experimental results of video-specific ATL on JRDB-Pose (Sec. B), iii) A statement about the limitations of the proposed method (Sec. C), and iv) Additional qualitative examples of sample selection by proposed THC, WPU, and DUW (Sec. D).

A. Additional Results on PoseTrack21

In this section, we provide a complete table showing the experimental results for the video-specific Active Transfer Learning (ATL) on the PoseTrack21 dataset [1], which could not be included in the main paper. All the results are obtained with the same experimental settings as described in the main manuscript. Once again, the selection criteria used in our experiment are as follows:

- **Random:** Random sampling from a uniform distribution.
- **Least Confidence (LC):** A traditional uncertainty measurement described in [4].
- **Multiple Peak Entropy (MPE):** An uncertainty criterion in [5].
- **Temporal Pose Continuity (TPC):** An uncertainty criterion in [6].
- **k-means:** A representativeness criterion used in [13].
- **Core-Set:** An original Core-Set sampling from [10].
- **Temporal Heatmap Continuity (THC):** Our uncertainty criterion based on the temporal change of estimated heatmaps.
- **Whole-body Pose Unnaturalness (WPU):** Our uncertainty criterion based on the unnaturalness of estimated poses.
- **Dynamic Uncertainty Weighting (DUW):** Our criterion combines uncertainty and Core-Set sampling [10].

A.1. Baseline and State-of-the-art Comparisons

Table 1 shows full results of the proposed video-specific ATL on PoseTrack21 [1]. While the performances of all uncertainty-based methods [4–6] are even less than the random selection, our proposed method (THC+WPU+DUW) stably outperforms other methods including k-means [13] and Core-Set [10]. Our proposed method demonstrates high performance consistently. However, during the initial cycles of ATL, representativeness criteria such as k-means clustering [13] and Core-Set sampling [10] show superior performance. This observation aligns with our hypothesis stated in the main paper: during the early stages of ATL, it is important to cover the data distribution of the target domain. In contrast, as ATL progresses, identifying samples with high uncertainty becomes increasingly important. In this context, our DUW effectively enhances performance by identifying these challenging samples. This shows the importance of uncertainty for performance improvement during the later stages of ATL.

A.2. Ablation Studies

Table 2 shows the full results of ablation studies. Our proposed methods, THC, WPU, and DUW, all used together, achieved the highest ALC. THC+DUW and WPU+DUW surpassed the performance of the original Core-Set [10] due to the incorporation of uncertainty in sample selection. In cases where only THC, only WPU, or THC+WPU, the performance is found to be inferior to that of the Core-Set [10]. These lower performances are attributed to a selection bias [3, 9] common in sample selection by uncertainty.

In addition, as mentioned in Sec. 5.4 of the main paper, we investigated the performance of ATL under various con-

Table 1. Quantitative results of our proposed video-specific ATL on PoseTrack21 [1]. 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
	0%	5%	10%	15%	20%	30%	40%	60%	80%	100%	(%)
Random	**81.82**	87.76	93.60	95.10	96.09	96.92	97.39	98.50	99.21	**100.00**	96.91
LC [4]	**81.82**	77.49	89.55	93.03	94.60	95.69	96.77	98.29	99.37	**100.00**	95.74
MPE [5]	**81.82**	78.96	90.96	93.98	95.09	96.44	97.23	98.38	99.28	**100.00**	96.11
TPC [6]	**81.82**	83.38	90.97	93.63	95.32	96.34	97.31	98.54	99.43	**100.00**	96.40
k-means [13]	**81.82**	**93.97**	95.19	95.82	96.37	97.55	98.11	98.82	**99.45**	**100.00**	97.65
Core-Set [10]	**81.82**	93.18	**96.35**	**97.26**	**97.62**	**98.18**	**98.60**	**99.27**	**99.67**	**100.00**	**98.12**
Ours (THC+WPU+DUW)	**81.82**	**93.35**	**96.14**	**97.37**	**97.90**	**98.44**	**98.77**	**99.33**	**99.67**	**100.00**	**98.21**

Table 2. Ablation study results of video-specific ATL on PoseTrack21 [1]. 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 values are also calculated by an average of 170 test videos. (fixed), (increase), (const), and (decrease) denote the video-specific ATL with a fixed balance of uncertainty and representativeness, linearly increasing/decreasing the weight of THC toward WPU, using the same weight for THC and WPU, respectively.

Criterion	AP@0.6 (%)										ALC
	0%	5%	10%	15%	20%	30%	40%	60%	80%	100%	(%)
Core-Set [10]	**81.82**	93.18	**96.35**	97.26	97.62	98.18	98.60	99.27	99.67	**100.00**	98.12
THC	**81.82**	82.59	89.10	91.85	92.86	94.70	96.43	97.74	98.97	**100.00**	95.45
WPU	**81.82**	85.56	91.11	93.39	94.74	96.36	97.31	98.48	99.28	**100.00**	96.45
THC+WPU	**81.82**	84.82	91.72	93.83	95.17	96.38	97.25	98.54	99.35	**100.00**	96.51
THC+DUW	**81.82**	93.12	95.88	97.11	97.70	98.42	**98.91**	**99.35**	**99.74**	**100.00**	98.19
WPU+DUW	**81.82**	**93.19**	95.96	97.26	**97.87**	**98.51**	98.76	99.27	99.65	**100.00**	98.17
THC+WPU+DUW (fixed)	**81.82**	93.02	95.71	97.15	97.68	98.41	98.81	99.28	99.66	**100.00**	98.14
THC+WPU+DUW (increase)	**81.82**	93.18	95.85	97.13	97.86	98.47	98.80	99.27	99.64	**100.00**	98.16
THC+WPU+DUW (decrease)	**81.82**	93.08	96.11	**97.34**	97.72	**98.48**	**98.94**	**99.43**	**99.73**	**100.00**	**98.24**
THC+WPU+DUW (const)	**81.82**	**93.35**	**96.14**	**97.37**	**97.90**	98.44	98.77	99.33	99.67	**100.00**	**98.21**

figurations conceivable for the proposed method's components. As shown in Table 2, without dynamically combining uncertainty and representativeness at each stage of active learning ("fixed"), the performance was not as promising compared to other configurations of our proposed method.

Furthermore, regarding the combination of THC and WPU, generally, the use of THC and WPU with the same weight ("const") led to significant performance improvements in the initial phase of active learning. On the other hand, by gradually decreasing the weight of THC and increasing the weight of WPU ("decrease"), performance efficiently improved from the mid-phase of ATL, resulting in the highest value for ALC. This suggests that not only high performance can be achieved by simply using THC and WPU with the same weight ("const"), but also designing an appropriate method for combining THC and WPU can lead to a more effective video-specific ATL.

A.3. Impact of Hyperparameter Changes in DUW

In this section, we investigate the influence of the hyperparameter λ in the following objective function used in DUW:

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

where $\Delta(x_i, x_j)$ is the Euclidean distance between the sample's feature vectors x_i and x_j, G_c is the approximated generalization performance of Human Pose (HP) estimator, $C(x_i)$ is each sample's uncertainty score, L represents la-

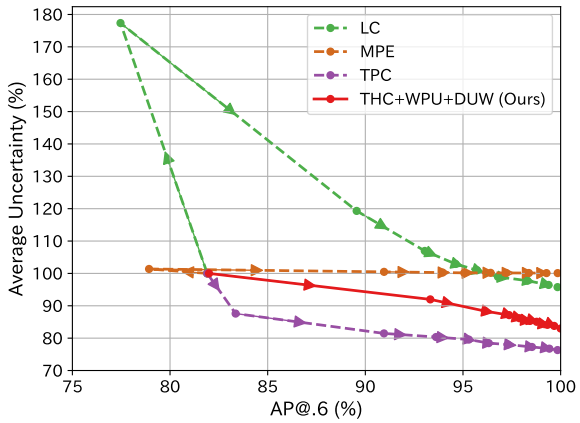


Figure 1. The change in the average uncertainty accompanying the AP@0.6 transition in video-specific ATL on PoseTrack21 [1], which are shown in Tables 1 and 2. The uncertainty at the beginning of ATL is used as a baseline (100%).

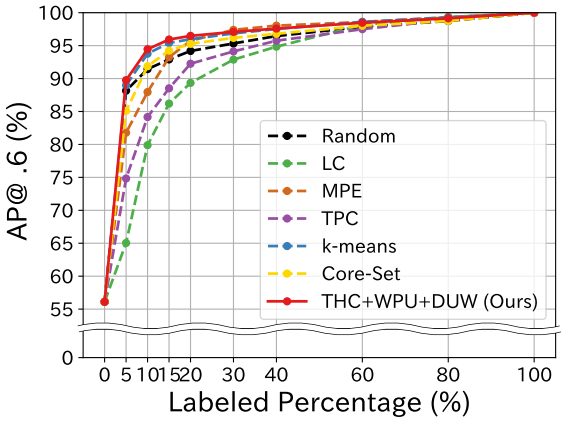


Figure 2. Learning Curve of video-specific ATL on JRDB-Pose [11].

beled samples set, and U represents an unlabeled set, respectively. As explained in the main manuscript, a larger value of λ leads to a sample selection that emphasizes uncertainty more. Conversely, a smaller value of λ results in a sample selection more similar to the original Core-Set sampling [10]. When λ equals zero, the sample selection is the same as the original Core-Set sampling.

Table 3 shows the results of video-specific ATL when the order of λ in Eq.(9) is changed. The active selection criterion used for this experiment was THC+WPU+DUW, and the detailed experimental setup remains the same as Sec. A.1 and A.2. According to the results presented in Table 3, the highest ALC is achieved when λ equals 0.01, followed closely by λ values of 0.1 and 0.001. On the other hand, performance is degraded when λ equals 0, corresponding to the original Core-Set [10] that tends to select samples with low informativeness due to not considering uncertainty. Similarly, performance decreases with λ values of 1 and 10, which overly prioritize uncertainty and consequently lose diversity in sample selection.

A.4. Validation of Uncertainty Criteria

We validate our proposed uncertainty criteria, THC and WPU. Uncertainty is expected to be low when HP estimation is correct and high when the result is incorrect. Hence, as the accuracy of estimation improves, the uncertainty should decrease. Fig. 1 shows the uncertainty change accompanying the transition of HP estimation performance in video-specific ATL. In MPE [5], the uncertainty remains almost constant regardless of changes in performance. On the other hand, THC and WPU show a desirable change in uncertainty that decreases with the improvement of AP.

B. Additional Results on JRDB-Pose

In this section, we report a complete table of experimental results on the JRDB-Pose dataset [11] in the main paper, additional experimental results with another metric, and the validation of uncertainty criteria on JRDB-Pose. The basic experimental conditions are the same as those described in Sec. 5.1 of the main paper.

B.1. Baseline and State-of-the-art Comparison

The results of the video-specific ATL for the 15 test videos from JRDB-Pose are presented in Tables 4 and 5. Fig. 2 is the plotted learning curve. Here too, while the initial AP@0.6 is 56.11%, our proposed method ("Ours") achieved performance close to 90% with only 5% of the labeling. Furthermore, our ALC performance across the entire ATL outperformed all comparison methods.

Moreover, when evaluated using OSPA, our method displayed performance comparable to the best-performing method ("k-means [13]"). This suggests that our method can efficiently adapt the HP estimator even for challenging datasets like JRDB-Pose [11].

B.2. Validation of Uncertainty Criteria

Figure 3 illustrates the change in uncertainty with the improvement of the FastPose's [2] performance in the video-specific ATL using the JRDB-Pose dataset [11]. As with the results in PoseTrack21 [1] (Sec. A.4), as the performance of FastPose improves, the value of uncertainty decreases. This suggests that our proposed uncertainty criteria (THC+WPU) can reflect the error of the prediction results.

Table 3. Impact of variation in hyperparameter λ of DUW. 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 values are also calculated by an average of 170 test videos. When $\lambda = 0$, the sample selection is equivalent to the original Core-Set sampling [10].

Criterion	AP@0.6 (%)										ALC (%)
	0%	5%	10%	15%	20%	30%	40%	60%	80%	100%	
$\lambda = 0$ (Core-Set [10])	**81.82**	93.18	**96.35**	97.26	97.62	98.18	98.60	99.27	**99.67**	**100.00**	98.12
$\lambda = 0.001$	**81.82**	93.30	**96.33**	**97.34**	97.78	**98.50**	**98.82**	**99.28**	99.62	**100.00**	**98.20**
$\lambda = 0.01$	**81.82**	**93.35**	96.14	**97.37**	**97.90**	**98.44**	98.77	**99.33**	**99.67**	**100.00**	**98.21**
$\lambda = 0.1$	**81.82**	**93.37**	**96.35**	97.29	**97.82**	98.36	**98.80**	99.26	**99.70**	**100.00**	**98.20**
$\lambda = 1$	**81.82**	93.31	95.77	97.05	97.55	98.10	98.43	99.12	99.60	**100.00**	98.01
$\lambda = 10$	**81.82**	93.34	95.64	96.64	97.12	97.99	98.36	99.07	99.58	**100.00**	97.91

Table 4. Quantitative results of our proposed video-specific ATL on JRDB-Pose [11]. Red and blue indicate the best and the second best, respectively. OSPA 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	OSPA ↓			ALC ↓
	5%	20%	40%	$\times 10^{-}4$
Random	**0.129**	0.074	**0.047**	5.22
LC [4]	0.287	0.122	0.068	7.80
MPE [5]	0.196	0.071	**0.040**	5.43
TPC [6]	0.239	0.110	0.066	7.59
k-means [13]	**0.132**	**0.067**	**0.040**	**4.72**
Core-Set [10]	0.157	0.077	0.052	5.70
Ours (THC+WPU+DUW)	0.134	**0.068**	**0.047**	**5.05**

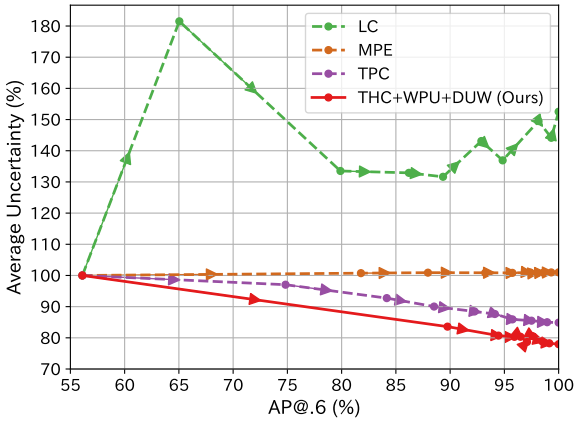


Figure 3. The change in the average uncertainty accompanying the AP@0.6 transition in video-specific ATL on JRDB-Pose [11], which are shown in Table 5. The uncertainty at the beginning of ATL is used as a baseline (100%).

C. Limitations

Despite the positive results observed in our study, it is important to acknowledge some limitations.

First, the tuning of learning conditions can be challenging. In particular, the optimal hyperparameters of the proposed methods could be dataset-specific and still require some tuning to achieve acceptable results.

Second, our Temporal Heatmap Continuity (THC) method has an inherent limitation. The THC might be high wrongly, especially when the object's movements are drastic. This could lead to a skewed selection towards instances with more intense movements.

Lastly, determining an upper bound for learning efficiency in ATL is still challenging. Mainly due to the too many cases of possible combinations of sample selections [12], it is difficult to define an optimal sampling strategy during the ATL process even if ground truth labels are available.

These limitations offer potential areas to further improve the proposed method.

D. Visualization of Proposed Active Selection Criteria

In this section, we present qualitative results of sample selection by THC (Fig. 4), WPU (Fig. 5), and DUW (Fig. 6) on PoseTrack21 [1]. The experimental settings for each criterion are the same as those in the main paper.

Figs. 4 and 5 demonstrate that our proposed THC and WPU accurately capture incorrect pose estimation results. Furthermore, Fig. 6 clearly shows that the balance between uncertainty and representativeness changes dynamically with the parameter λ. This also suggests that when $\lambda = 0.01$, we can achieve a selection of uncertain and diverse samples.

Table 5. Quantitative results of our proposed video-specific ATL on JRDB-Pose [11]. Red and blue indicate the best and the second best, respectively. AP@0.6 is the average AP of 15 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 15 test videos.

Criterion	AP@0.6 (%)										ALC
	0%	5%	10%	15%	20%	30%	40%	60%	80%	100%	(%)
Random	**56.11**	88.16	91.44	92.91	94.19	95.33	96.46	97.86	98.74	**100.00**	95.42
LC [4]	**56.11**	65.04	79.89	86.20	89.34	92.88	94.84	98.14	**99.32**	**100.00**	92.67
MPE [5]	**56.11**	81.78	87.95	93.24	95.74	**97.39**	**98.03**	**98.59**	**99.32**	**100.00**	95.76
TPC [6]	**56.11**	74.83	84.16	88.52	92.25	94.12	95.74	97.50	98.94	**100.00**	93.76
k-means [13]	**56.11**	**88.97**	**93.78**	**95.41**	**95.98**	96.86	97.53	**98.61**	**99.28**	**100.00**	**96.41**
Core-Set [10]	**56.11**	85.09	91.87	94.24	95.27	96.18	96.80	98.01	98.75	**100.00**	95.60
Ours (THC+WPU+DUW)	**56.11**	**89.76**	**94.45**	**95.93**	**96.48**	**97.08**	**97.59**	98.47	99.14	**100.00**	**96.52**

References

[1] 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, 2, 3, 4

[2] 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. 3

[3] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *ICLR*, 2021. 1

[4] David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Mach. Learn.*, pages 148–156, 1994. 1, 2, 4, 5

[5] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *ICCV*, 2017. 1, 2, 3, 4, 5

[6] 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

[7] Bharath Raj N., Anand Subramanian, Kashyap Ravichandran, and N. Venkateswaran. Exploring techniques to improve activity recognition using human pose skeletons. In *HADCV (WACVW)*, 2020. 7

[8] Ashwin Narayan, Bonnie Berger, and Hyunghoon Cho. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nat. Biotechnol.*, 39:765 – 774, 2021. 7

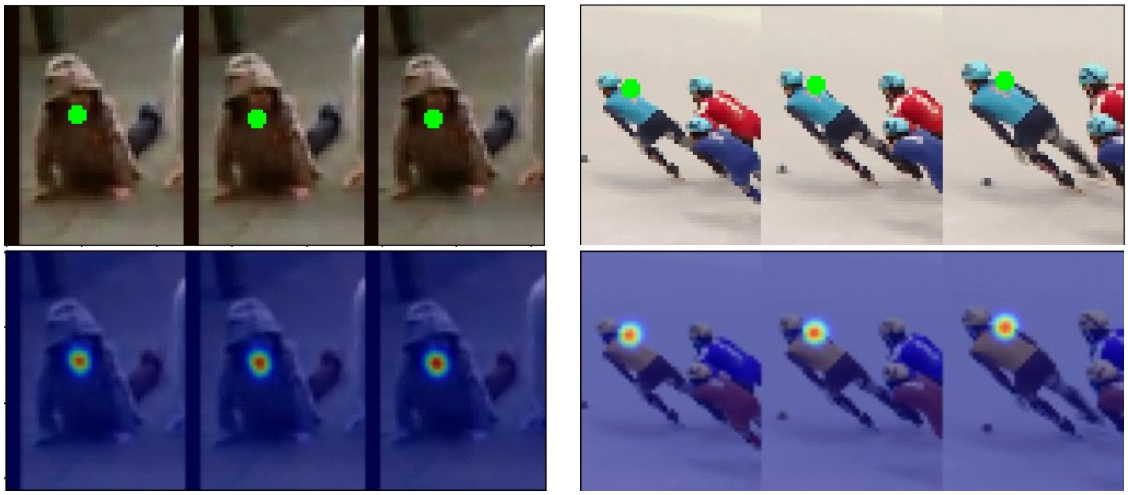
[9] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9):1–40, 2022. 1

[10] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 1, 2, 3, 4, 5, 7

[11] 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. 3, 4, 5

[12] Yuexi Zhang, Yin Wang, Octavia I. Camps, and Mario Sznaier. Key frame proposal network for efficient pose estimation in videos. In *ECCV*, 2020. 4

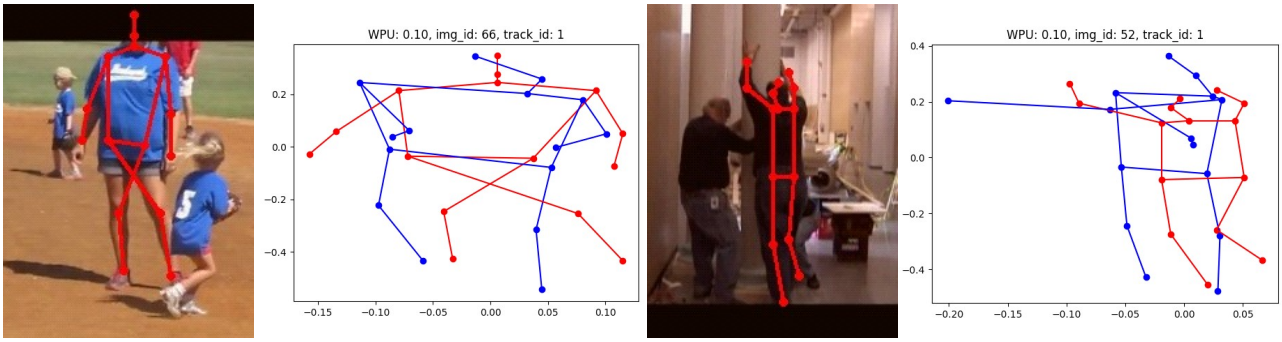
[13] Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019. 1, 2, 3, 4, 5



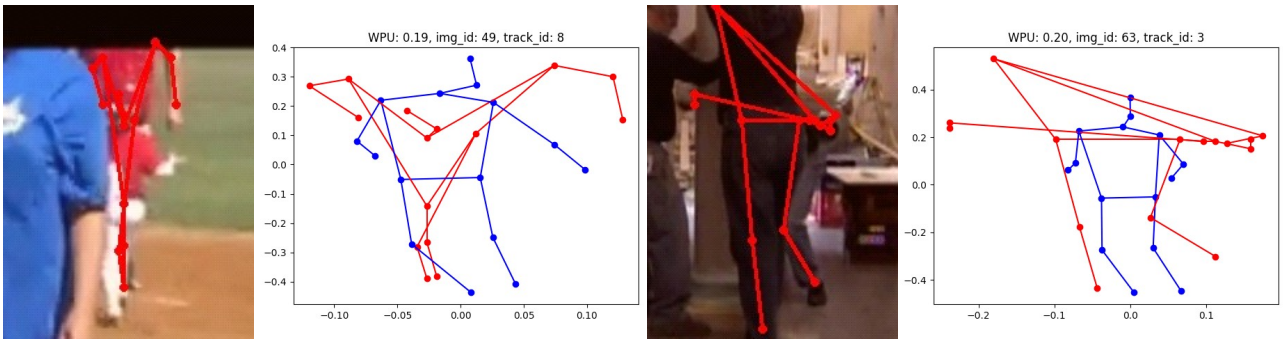
(a) Heatmaps with low THC. Times are at t-1, t and t+1 from left to right in each scene.

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

Figure 4. Additional qualitative examples of our THC. The top row of the figure shows the original images, with the estimated keypoint positions marked by green circles. The bottom row presents the heatmaps estimated for each of the three adjacent frames, where a color closer to blue indicates a lower probability of keypoint presence, while a color closer to red suggests a higher probability. (a) There is a strong peak at a single point in the heatmap between adjacent frames consistently. As a result, estimated keypoint positions are accurate. (b) In contrast, the estimations are inconsistent and the peaks in the heatmap are dispersed. It results in an erroneous estimation.



(a) Estimated pose with low WPU.



(b) Estimated pose with high WPU.

Figure 5. Examples of samples selected by WPU. In the figure, the red lines represent the estimated pose and its Hybrid feature [7], and the blue lines represent the Hybrid feature output by the AE trained on natural poses. In the case of (a), where the WPU is low, the red Hybrid feature, which is the input to the AE, and the blue Hybrid feature, which is the output, are close to each other, and the estimated pose is also close to the correct one. On the other hand, in (b), due to an incorrect pose estimation input, the Hybrid features are far apart from each other, resulting in a high WPU value.

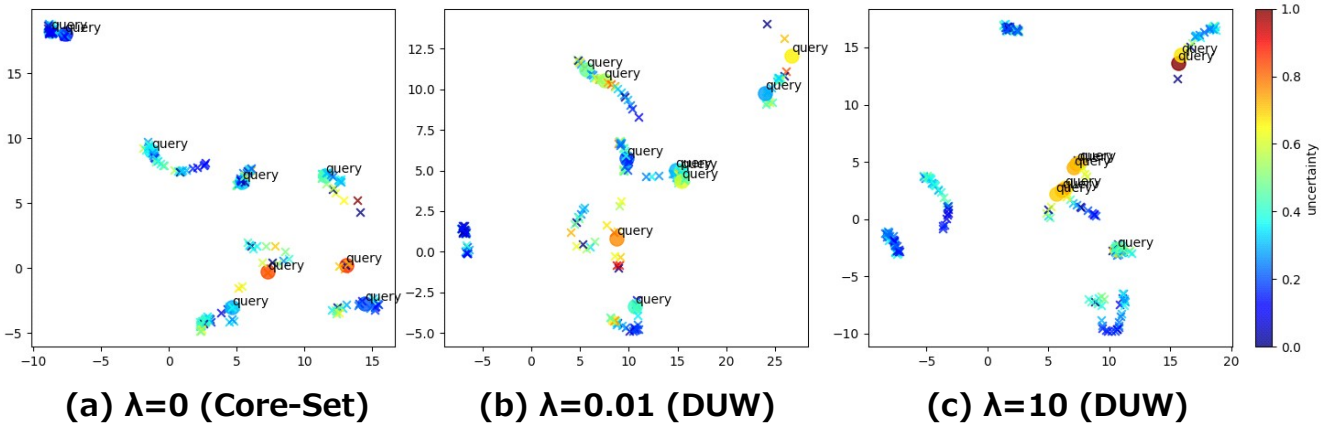


Figure 6. A visualization result of the sample selection of our DUW criterion. We have utilized DensMAP [8] to plot feature vectors extracted by the HP estimator. In this plot, circles represent newly selected samples, while cross marks denote unlabeled samples that were not selected for the current ATL cycle. The color of the plot corresponds to the normalized uncertainty. (a) represents the results when $\lambda = 0$ (i.e., equivalent to the original Core-Set [10]), which tends to select diverse but uninformative samples with low uncertainty. (b) represents the selection for $\lambda = 0.01$, which yielded the best results in Sec. A.3. It can be seen that uncertain and diverse samples are selected. (c) represents the case when $\lambda = 10$. Although the uncertainty of selected samples is high, data points located within a limited range in the feature space are selected in a biased manner.