VL4Pose: Active Learning Through Out-Of-Distribution Detection For Pose Estimation

Megh Shukla ✉ [1]
megh.shukla@epfl.ch
Roshan Roy *[3]
roshan.roy@lmco.com
Pankaj Singh *[2]
pankaj.singh@mercedes-benz.com
Shuaib Ahmed [2]
shuaib.ahmed@mercedes-benz.com
Alexandre Alahi [1]
alexandre.alahi@epfl.ch

[1] Visual Intelligence for Transportation Lab
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

[2] Mercedes-Benz Research and Development India
Bengaluru, India

[3] Lockheed Martin Corporation
New Jersey, USA

Abstract

Advances in computing have enabled widespread access to pose estimation, creating new sources of data streams. Unlike mock set-ups for data collection, tapping into these data streams through *on-device active learning* allows us to directly sample from the real world to improve the spread of the training distribution. However, on-device computing power is limited, implying that any candidate active learning algorithm should have a low compute footprint while also being reliable. Although multiple algorithms cater to pose estimation, they either use extensive compute to power state-of-the-art results or are not competitive in low-resource settings. We address this limitation with VL4Pose *(Visual Likelihood For Pose Estimation)*, a first principles approach for active learning through out-of-distribution detection. We begin with a simple premise: pose estimators often predict incoherent 'poses' for out-of-distribution samples. Hence, can we identify a distribution of poses the model has been trained on, to identify incoherent poses the model is unsure of? Our solution involves modelling the pose through a simple parametric Bayesian network trained via maximum likelihood estimation. Therefore, poses incurring a low likelihood within our framework are out-of-distribution samples making them suitable candidates for annotation. We also observe two useful side-outcomes: VL4Pose in-principle yields better uncertainty estimates by unifying joint and pose level ambiguity, as well as the unintentional but welcome ability of VL4Pose to perform pose refinement in limited scenarios. We perform qualitative and quantitative experiments on three datasets: MPII, LSP and ICVL, spanning human and hand pose estimation. Finally, we note that VL4Pose is simple, computationally inexpensive and competitive, making it suitable for challenging tasks such as on-device active learning.

*Pankaj and Roshan contributed equally. Work originated when Megh and Roshan were at Mercedes-Benz.

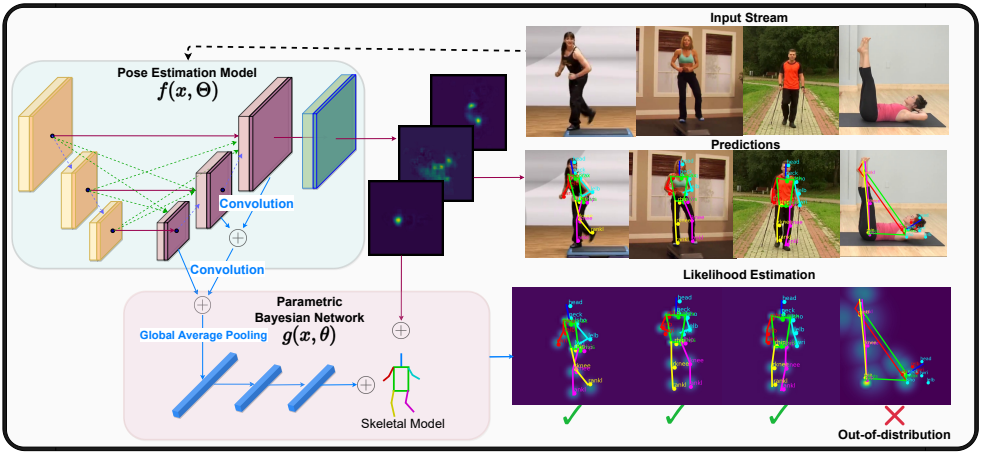


Figure 1: *On-Device Active Learning*: Keeping on-device resource constraints in mind, we use a small (~25% size of the pose model) auxiliary network consisting of convolutional + fully connected layers to parameterize our skeletal model. The skeletal model is a simple Bayesian Network, trained via maximum likelihood to learn poses from the training distribution. Poses that incur a low likelihood are out-of-distribution and are added to the training dataset upon relabelling. VL4Pose delivers real time active learning at ~30FPS.

1 Introduction

Data-centric methods rely on developing models which are robust under a wide range of scenarios. However, training datasets are usually made in staged setups and may not capture the complexity associated with real world use cases. So how can we adapt our models to address different real world scenarios? One approach harnesses data streaming in from multiple end users, which represents the real world distribution. However, the resulting volume of data would overwhelm both, end user bandwidth and our data processing pipelines. Instead, can we turn to active learning [45], to sample only those images which the model considers informative?

Active learning is a cyclical process of *train* $\Leftrightarrow$ *sample and label new images* to improve the spread of the training distribution. Formally, the goal is to identify a subset of unlabelled data, which if labelled, imparts maximum information to the model thereby improving its performance. One could argue that the very essence of active learning is to perform out-of-distribution sampling. Indeed, out-of-distribution images will impart maximum information to a model trained on the training distribution. Active learning facilitates lower annotation costs as well as faster training and prototyping due to reduced data volumes. While active learning has been well studied in literature, keypoint estimation, and in particular human pose estimation presents a unique challenge. Specifically, popular human pose estimation architectures are fully convolutional and directly regress 2D heatmaps [28, 58]. This prevents [8, 47] the use of many algorithms which rely on ensembles, dropouts or logits. As a result, a new wave of active learning methods [8, 14, 34, 47, 57, 59] address keypoint estimation.

However, recent literature overlooks a key question: *Where is active learning taking place?* The process can either be centralized in a computing cluster with no resource constraints, or decentralized on the end-user's device in a resource constrained environment.

The latter, typically referred to as on-device active learning [15, 41], allows us to directly tap into the real world data distribution, while also opening new possibilities in customized machine learning for the end user. This is in contrast to centralized active learning which has indirect or no exposure to real time samples from the end user. *While existing work focuses on pushing state-of-the-art in active learning for keypoint estimation, they impose high compute requirements which are infeasible for on-device learning.*

With *VL4Pose*, we propose an algorithm for on-device active learning. We investigate a first principles approach to active learning for pose estimation; can we leverage simple pose constraints to identify out-of-distribution samples? Specifically, we model the skeletal structure through a Bayesian Network which captures simple conditional relationships between joints. These relationships are parameterized by a small auxiliary neural network which uses visual cues from images to maximize the likelihood of poses in our training data. Consequently, we expect that out-of-distribution images will have lower likelihood values, making them suitable candidates for annotation. We also show that our maximum likelihood formulation derives Multi-Peak Entropy [34] and seamlessly unifies joint and pose level ambiguity, making it a better representative of uncertainty is comparison to [8, 17, 29]. Surprisingly, modelling simple skeletal constraints also facilitates pose refinement in limited scenarios, a first for an active learning algorithm. We validate our claims on two different tasks (human / hand pose) using two different architectures (direct keypoint regression / heatmaps) across three different datasets - MPII [2], LSP [22, 23] and ICVL [50]. Our experiments show that VL4Pose has lower compute costs, interpretable and competitive with state-of-the-art, making it suitable for on-device real-time active learning.

2 Related Work

Active Learning. Settles' survey [45] is a comprehensive work on classical active learning covering algorithms based on diversity, ensemble and uncertainty sampling. The Query by Committee method of sampling (or ensembles) [6, 26, 36] uses a family of hypothesis for active learning selection. Diversity based approaches include Core-Set [43] which sequentially selects non-semantically similar points. Uncertainty estimation [10, 12, 13, 16, 25, 31] provides quantitative measures to model ambiguity in the prediction. Bayesian Neural Networks have been traditionally used to estimate uncertainty; however recent approaches [27, 53] explore computing uncertainty using a single forward pass through the network. Empirical approach to estimate ambiguity of the model include Learning Loss [48, 57] which similar to our approach uses an auxiliary neural network to predict the 'loss' for an unlabelled image. Approaches that measure model change include expected gradient length [45], which uses the model's gradient as a directly proportionate measure of informativeness. Application domains [6, 19, 46, 47] of expected gradient length include image and text analysis.

Pose Estimation. Keypoint estimation (HPE) has been widely studied [5, 30, 38, 49, 52] in literature, with popular architectures regressing two dimensional heatmaps denoting the location of the joint. Heatmaps are preferred over direct keypoint regression since they retain positional accuracy which may otherwise have been lost with fully connected layers. While our work does not address multi-person pose estimation, our goal shares similarity with affinity fields [7, 27, 28] used to associate parts with individual persons. Jain *et al.* [20] and Tompson *et al.* [51] have previously leveraged Markov Random Fields to validate the configuration of poses predicted by the model. However, the priors defined by the

approach do not scale well with rotations as well as changes in scale of the person. Moreover, belief propagation is expensive and takes a significant amount of time to converge. *Unlike these methods, our goal is not to improve human pose estimation but instead detect out-of-distribution samples for active learning*. This allows us to use simpler models which converge faster and benefit from a low compute footprint.

Active Learning for Pose Estimation. Challenges posed by popular human pose architectures (detailed in [8, 47]) have lead to the development of new algorithms. Liu and Ferrari [34] proposed an intuition driven framework to model heatmap ambiguity. With VL4Pose, we provide a mathematical framework which not only incorporates heatmap ambiguity but also models spatial relationships between joints. Learning loss [48, 57] explored an idea parallel to out-of-distribution detection and identified which images were the most difficult for the model to learn. However, the method has limited ability in identifying out-of-distribution samples. The uncertainty modelled by EGL++ [47] has high compute requirements. Recent state-of-the-art methods such as MATAL [14] and UncertainGCN [9] use powerful techniques such as reinforcement learning and graph convolutional networks, however they are computationally very expensive. Uncertainty in 3D human pose [1, 44, 56] utilizes depth information or stereo images, both of which are not available for general 2D pose estimation. We reserve our discussion on uncertainty in 2D for later.

3 Methodology

VL4Pose proposes active learning through out-of-distribution (OOD) for pose estimation. However, how do we define OOD in the context of pose estimation? For instance, [18, 33, 42] study OOD for classification, and limited literature addresses the same for pose estimation. Hence, VL4Pose frames OOD detection as a maximum likelihood problem; samples with a low likelihood in our framework have limited representation in the training set. Our premise is that pose estimators f_Θ may not generalize well beyond the training distribution. While the pose model accurately predicts the pose for images from the training set, more often than not the model makes errors on images from an unseen distribution. Formally, let $\hat{Y} = f(x,\Theta)$ where $\hat{Y} = \hat{y_1} \dots \hat{y_N}$ represents the predicted joints and x the input image. Can we identify images x for annotation where $\hat{Y}$ is invalid?

3.1 Visual Likelihood Estimation

Our solution lies in flipping the question: we learn to identify when $\hat{Y}$ is valid since the space of valid poses is much smaller than the set of invalid poses. However, what makes a pose valid? Indeed, a random collection of keypoints rarely makes a recognizable pose, the key to a valid pose is *structure defined by the keypoints*. We represent this skeletal structure with a simple parametric Bayesian network (Fig. 2, human and hand models), with singular parent-child chains. Such an approach allows us to exploit the Markov blanket principle of independence and simplify our analysis as well as model relations between various joints in an interpretable manner. Let $q(Y|x,\theta) = q(y_1, y_2 \dots y_N|x,\theta)$ represents the distribution of joints y_i and θ represents the parameters of the Bayesian network. Using the chain rule of probability and the Markov blanket, we can now decompose this distribution into a linear chain capturing pairwise dependencies:

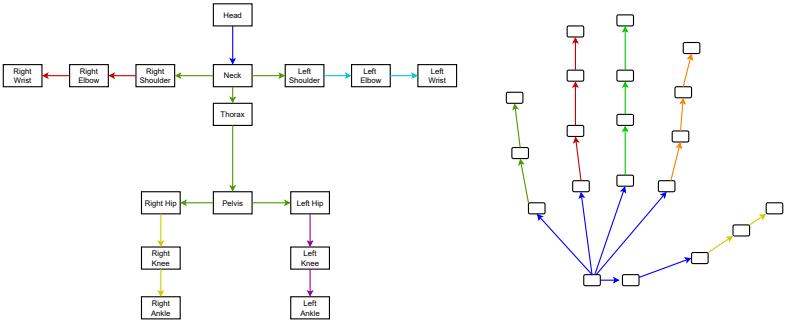


Figure 2: *Skeletal Model*: While the pose estimator models joint localization $p_{pose}(Y = y)$, the parametric Bayesian network models the much simpler $q_{BN}(Y_1 = y_i | Y_2 = y_j)$ where the child is conditioned only on the parent joint. VL4Pose incorporates both p_{pose} and q_{BN} to model the distribution of poses for an image.

$$q_{BN}(y_1, y_2 \dots y_N | x, \theta) = q(y_1 | y_2 \dots y_N, x, \theta) q(y_2 | y_3 \dots y_N, x, \theta) \dots q(y_N | x, \theta)$$
$$q_{BN}(y_1, y_2 \dots y_N | x, \theta) = \left[\prod_{i=1}^{N-1} q(y_i | y_{i+1}, x, \theta) \right] q(y_N | x, \theta) \tag{1}$$

Here y_N represents the joint corresponding to the root node. We specifically note that this formulation is simple; for instance intuition suggests that the subset of joints forming the torso are not independent of each other, a fact that is oversimplified by our formulation. For instance, complex graphical models have been successfully used [1, 32, 40, 55] to improve pose estimation. *However, we recall that our goal is to solve the simpler task of out-of-distribution detection*, and the proposed framework succeeds in achieving this objective.

Traditional application of maximum likelihood involves finding a set of parameters that maximizes the likelihood of our observations X, Y, where both X and Y are deterministic. However, the observation X, Y need not be deterministic; for example in human pose estimation a joint can have multiple plausible locations based on the local maxima in the heatmap. Let $p_{pose}(Y) = p_{pose}(y_1, y_2 \dots y_N) = \prod_{i=1}^{N} p(y_i)$ represent the pose estimator's (human/hand) distribution over the joints $y_1 \dots y_N$. Note that our assumption of independence is in line with the training objective of popular pose estimators [8, 33, 49]. The expected log-likelihood *w.r.t* the set of keypoints is:

$$\mathbb{E}_Y \left[\log \, q_{BN}(y_1, y_2 \dots y_N | x, \theta) \right] \tag{2}$$

Substituting Eq. 1 in Eq. 2 and expanding, we get (full derivation in supplementary):

$$\sum_Y \left[p_{pose}(y_N) \log \, q_{BN}(y_N | x, \theta) + \sum_i^{N-1} p_{pose}(y_i) \log \, q_{BN}(y_i | y_{i+1}, X, \theta) \right] \tag{3}$$

Since Y represents the set of keypoint random variables, expectation over Y is the weighted likelihood over all possible pose configurations. Therefore, computing the expected likelihood allows us to incorporate a distribution over Y into the framework. *Intuitively, p_{pose} allows us to model keypoint ambiguity whereas q_{BN} models the the ambiguity associated*

with the entire pose. The auxiliary network uses visual cues from the image to fit the parameters θ such that the likelihood over the training distribution is maximized. Therefore, poses incurring a low-likelihood correspond to out-of-distribution samples. Annotating such samples increases the spread of the training distribution resulting in better performing models. Since gradients of θ are detached from the pose estimator Θ, we train the two networks simultaneously. During the training phase, we can directly observe both the parent y_{i+1} as well as child joints y_i to compute the likelihood. In the absence of ground truth (while performing active learning), we rely on the predictions of the pose estimator $\hat{Y} = f(x,\Theta)$ to estimate the likelihood.

The computation of $\sum_Y$ and $q_{BN}(y_i|y_{i+1},X,\theta)$ depends upon the architecture (direct keypoint regression or heatmap) which we discuss now in greater detail.

3.1.1 Direct Keypoint Regression

Modelling $q_{BN}(y_i|y_{i+1},X,\theta)$. We refer to the DeepPrior [8] architecture, which uses fully connected layers to directly regress 3D keypoints ($y_i \in \mathbb{R}^{\text{joints}\times 3}$) for hand pose estimation. To impose conditional dependency, we predict the *offset* to obtain the child joint given the parent joint: $y_i = y_{i+1} + \hat{o}_i$. The offset is learnt by the parametric network $\hat{o}_i = g_1(x,\theta)$. Our belief is that the y_i can be completely recovered given the parent y_{i+1} and visual cues from image X. Further, we also learn the covariance matrix $\Sigma_i = g_2(x,\theta)$ [11, 17, 25, 35] that determines the spread around the offset for the child joint. Therefore:

$$q_{BN}(y_i|y_{i+1},x,\theta) = \mathcal{N}\Big(y_i - [y_{i+1} + \hat{o}_i], \Sigma_i\Big) \tag{4}$$

Modelling $\sum_Y$. For hand pose estimation $Y = f(x,\Theta)$ is a point estimate which implies there exists exactly one pose configuration; $p(y_i)$ = {1 at ground truth location, 0 otherwise} $\forall i$. This eliminates the need to sum over all possible pose configurations in Eq. 3, simplifying the computation.

3.1.2 Heatmap Regression

Modelling $q_{BN}(y_i|y_{i+1},X,\theta)$. A significant portion of errors in human pose are due to incorrect association of left-right joints and keypoint swaps within the pose [21]. If we treat the offset as vectors, these sources of errors would cause some components of the offsets to average out during training leading to poor results. Hence we use an euclidean distance based measure which intuitively tries to find the optimal bone length $d_i = g(x,\theta)$ between the parent and child joint (y_{i+1} and y_i respectively). Specifically:

$$q(y_i|y_{i+1},x,\theta) = \mathcal{N}(\texttt{dist}(y_i,y_{i+1}) - \hat{d}_i\,,\,\sigma_i) \tag{5}$$

Learning the bone length is easier and is less strict in comparison to offsets, with our observations confirming that convergence is better for a distance based modelling approach for human pose estimation.

Modelling $\sum_Y$. During the training phase, we have access to the ground truth values which are point estimates for $Y = f(x,\Theta)$. Therefore, the summation over all poses is reduced to one pose configuration to train the auxiliary network θ via maximum likelihood. However, during the active learning phase (no ground truth) we need to rely on the pose estimator's predictions which consists of heatmaps $h \in \mathbb{R}^{\text{joints}\times 64\times 64}$ that represents a spatial probability

distribution $p(y_i)$ for a joint $y_i \in \mathbb{R}^2$. Yet, computing the expectation over the entire heatmap is infeasible. Instead, we limit the domain of $p(y_i)$ to only the local maxima of the heatmap h_i and thereby showing that Multi-Peak Entropy [34] can be viewed as a consequence of maximizing the likelihood of poses. We then apply softmax normalization to provide a probabilistic interpretation of various local maxima being the true location of joint i. Therefore, we replace $p(y_i)$ with $\hat{p}(y_i) = \texttt{softmax}(\texttt{local\_maxima}(h_i))$ and the domain of $\hat{p}(y_i)$ now restricted to the location of the local_maxima of h_i.

3.2 Discussion

Uncertainty For Pose Estimation. Although multiple recent works [8, 17, 29, 34, 35] explore uncertainty for pose estimation, they suffer from a few limitations. Inspired by Kendall and Gal [25], some approaches [8, 17, 35] model heteroscedastic aleatoric uncertainty by learning a covariance matrix of the order $O(n^2)$ where n represents the number of keypoints. This leads to two drawbacks: First, a much larger network is required to learn a fit for the covariance matrix. Second, larger networks tend to learn spurious correlations. Additionally, we believe that the definition of aleatoric uncertainty is misinterpreted in pose estimation (detailed in supplementary material) and hence we do not specify the uncertainty learnt by VL4Pose. Caramalau *et al.* [8] assumes independence between joints which is incorrect. Intuition suggests that observing a variable leads to a decrease in uncertainty for a correlated variable. While both Liu and Ferrari [34] and Kundu *et al.* [29] share our approach of modelling joint level ambiguity, [34] does not model pose uncertainty, whereas [29] uses a domain adaptation specific approach of modelling pose uncertainty. In contrast, VL4Pose provides a mathematical framework that incorporates both joint and pose level uncertainty for different network architectures. The method employs a linear chain of probabilities, therefore reducing the order of the variance / covariance matrix to $O(n)$. Further, VL4Pose directly captures the correlation between neighboring joints, limiting the possibility of learning spurious correlations.

Auxiliary Network. The Bayesian network is parameterized through an auxiliary neural network as shown in Fig. 1. Initially the network captures features from the pose estimator at various scales by using an appropriate convolutional kernel to downsize and add the larger feature map to the next smaller feature map. This is progressively done till we reach the smallest feature map, and subsequently perform global average pooling to obtain a one-dimensional feature. This is followed by simple fully connected layers, with the final layer predicting the parameters for each link in the skeleton. The network is trained to minimize the negative log-likelihood (Eq. 3). The two stage hourglass has $\approx 8.4M$ parameters, with the auxiliary network adding a further $\approx 2.1M$ parameters.

Hardware and Time Complexity. We use two setups: Mobile computing (AMD Ryzen 5000, NVIDIA RTX 3060 Mobile) and server grade (Intel Xeon, NVIDIA V100). VL4Pose takes $\approx$ 30ms to process each image on both the setups (also implying that our hardware is not fully utilized). There are four real-time algorithms: VL4Pose, Learning Loss, Aleatoric uncertainty and Multi-Peak entropy. The former three have similar processing times ($\approx$ 30FPS) as they have near identical parametric networks, whereas Multi-Peak entropy is marginally faster since it benefits from a highly vectorized implementation. The inference time does not depend on the number of samples and have a complexity of $O(1)$ for a fixed model. In contrast, other algorithms are not designed for real-time use: graph based methods such as CoreSet, MCD-CKE, EGL++, GCN have a time complexity of at least $O(mn)$ since

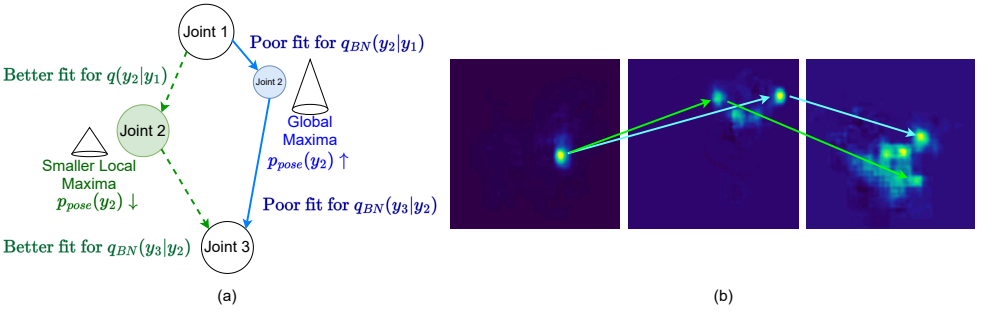


Figure 3: *Pose Refinement*: Smaller but well placed local maxima are more likely to define a valid pose in comparison to poorly positioned global maxima.

they model the interaction between all m unlabelled and n labelled samples. MATAL uses reinforcement learning which is compute intensive; the most efficient method takes 2 hours for active learning [14].

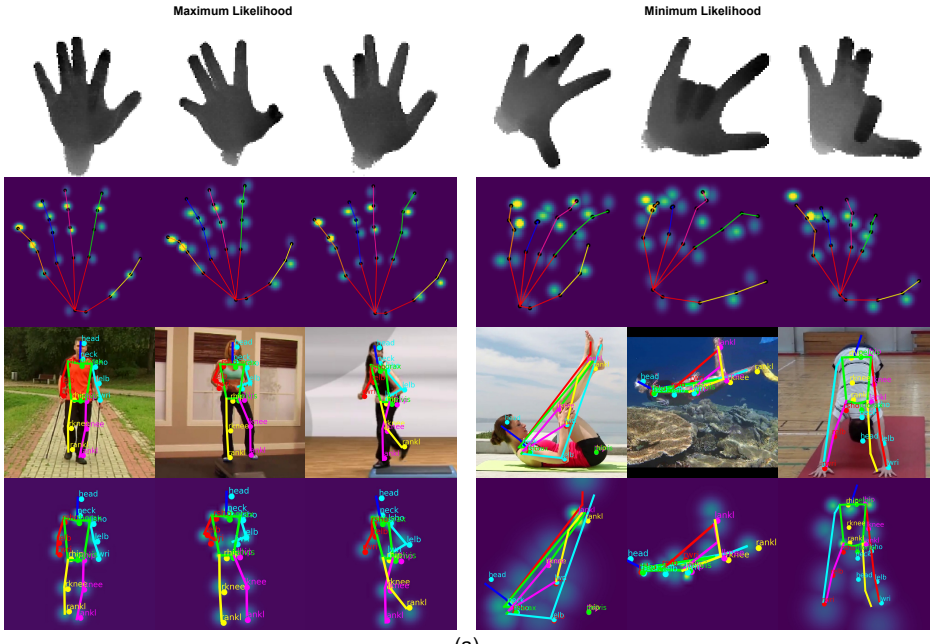
Pose Refinement: Human Pose Estimation. Surprisingly, a minor tweak to VL4Pose can also allow us to perform pose refinement [24, 37, 39, 54] in certain scenarios, which to the best of our knowledge is the first active learning algorithm to do so. We categorically state that we do not intend to compete with state-of-the-art in pose refinement, but our intention is to highlight the versatility associated with VL4Pose in comparison to other active learning algorithms. Fig. 3 provides some intuition into the interplay between skeletal structure q_{BN} and heatmap ambiguity p_{pose}. Conventional human pose estimation approaches rely on inferring keypoints as the global maxima of their respective heatmaps. *However, certain poses may have a higher likelihood by incorporating local maxima that better explain the skeletal structure.* Mathematically, instead of computing the expected likelihood over all poses (Eq. 3), we find the pose configuration that results in the highest likelihood:

$$Y^* = \arg\max \left[p_{pose}(y_N) \log\, q_{BN}(y_N|x,\theta) + \sum_{i}^{N-1} p_{pose}(y_i) \log\, q_{BN}(y_i|y_{i+1},X,\theta) \right] \quad (6)$$

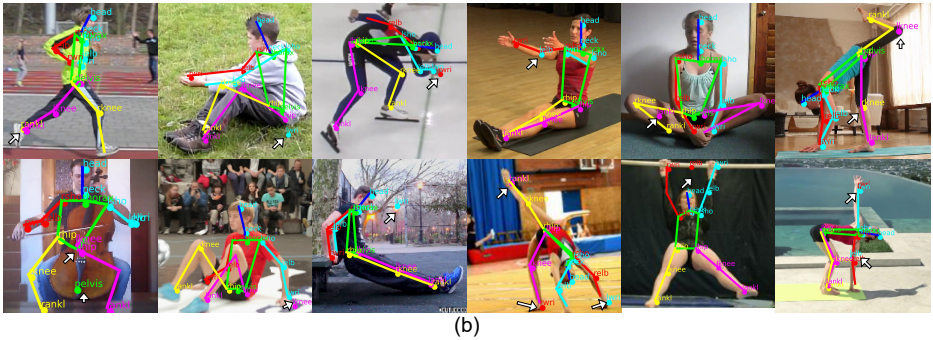
4 Experiments

Goal: VL4Pose is an algorithm for *on-device active learning*, which is not only competitive but also works in real-time due to lower compute costs. Our experiments cover uncertainty and likelihood estimation, which forms the core of VL4Pose. We provide qualitative visualizations as well as quantitative comparisons through active learning to highlight how the uncertainties learnt by the model play a role in out-of-distribution detection. Our code is available at: https://github.com/meghshukla/ActiveLearningForHumanPose

Qualitative analysis. Fig. 4(a) is generated by visualizing the distributions corresponding to *offsets* for hand pose and *distances* for human pose (visualization method described in the supplementary material). We observe that images with a low log-likelihood are a case where the joint predictions do not match the conditional distribution and are out-of-distribution. We observe that the accuracy of model predictions is highly correlated with the log-likelihood value; poor scores are correlated to poor predictions. Another important observation is that highly correlated parent-child joints have a higher degree of certainty given



(a)



(b)

Figure 4: [*Please zoom in*] (a) *Visualizing* $q_{BN}(y_i|y_{i+1},x,\theta)$: The skeleton represents the pose estimator's predictions $\hat{Y} = f(x,\Theta)$ and filled circles are the ground truth Y. We highlight the correlation between pose uncertainty and likelihood, and likelihood with actual model performance (b) *Pose refinement*: The skeleton represents the optimal pose configuration Y^* that maximizes the likelihood, and filled circles are the the pose estimator's predictions $\hat{Y} = f(x,\Theta)$. We highlight VL4Pose's potential to identify the correct pose Y^* even when $\hat{Y}$ has minor errors (marked in arrows). *Additional images in supplementary.*

the parent. Previous methods discarded correlation, and hence double counted the uncertainty for highly correlated joints. Fig. 4(b) highlights VL4Pose's ability to refine poses when the model gets a few joints wrong. VL4Pose is the first active learning algorithm to perform pose refinement since pose is explicitly being modelled by our method.

Quantitative analysis. We use the same active learning experiment setup as in [8, 47, 48, 57]. We conduct our experiments on human pose using two single person datasets: MPII

MPII: (PCKh, Percentage Correct Keypoints - head) Mean±Std. Dev. (5 runs)

#images →	2000		3000		4000		5000		6000	
Methods	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Random	78.05	0.55	81.59	0.33	84.2	0.27	85.94	0.23	87.43	0.21
Core-set [34]	75.66	1.22	80.52	1.15	84.1	0.54	86.08	0.35	87.68	0.31
Learning Loss [36, 57]	77.09	0.96	82.47	0.46	85.15	0.36	86.83	0.28	87.95	0.24
EGL++ [34]	78.54	0.71	82.03	0.61	84.39	0.22	86.11	0.32	87.78	0.48
Aleatoric [34]	76.07	1.12	82.03	0.52	84.62	0.62	86.76	0.62	88.0	0.25
Multi-peak [34]	82.49	0.72	84.62	0.46	85.88	0.31	87.47	0.51	88.54	0.71
VL4Pose	82.3	0.93	84.71	0.72	86.16	0.66	87.71	0.33	88.96	0.38

LSP and LSPET: Mean±Sigma (5 runs)

#images →	2000		3000		4000		5000		6000	
Methods	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Random	74.24	0.68	76.91	0.91	79.11	0.64	80.56	0.35	81.47	0.65
Core-set [34]	74.26	0.61	76.89	0.96	79.06	0.39	80.14	0.47	80.94	0.50
Learning Loss [36, 57]	73.99	0.28	76.71	0.63	78.53	0.37	79.91	0.37	80.77	0.24
EGL++ [34]	74.51	1.02	77.32	0.69	79.26	0.69	80.68	0.45	81.76	0.24
Aleatoric [34]	74.24	0.60	76.94	0.47	79.15	0.62	80.11	0.40	80.91	0.49
Multi-peak [34]	77.24	0.61	79.56	0.46	81.29	0.31	82.81	0.5	83.11	0.71
VL4Pose	77.36	0.68	79.71	0.50	81.48	0.46	82.75	0.49	83.69	0.47

ICVL: (MSE, Mean Square Error) Mean±Std. Dev. (5 runs)

#images →	200		400		600		800		1000	
Methods	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Random	16.83	0.84	15.31	0.45	14.12	0.53	13.68	0.44	13.21	0.35
Core-set [34]	16.86	0.74	14.73	0.46	14.02	0.81	13.69	0.49	13.43	0.44
MCD-CKE [8]	19.88	0.38	15.06	0.43	13.61	0.46	12.85	0.71	12.54	0.59
CoreGCN [9]	17.84	0.44	14.68	0.53	13.27	0.56	12.91	0.81	12.69	0.41
VL4Pose	16.87	0.47	14.89	0.52	13.64	0.51	13.02	0.77	12.84	0.54

Table 1: *Active Learning Simulation*: Human Pose: MPII, LSP-LSPET and Hand Pose: ICVL. Both PCK/PCKh and Mean Square Error (MSE) indicate accuracy of predictions, with higher values being better for PCK/PCKh and lower the better for MSE.

[2] and LSP/LSPET [22, 23]. For hand pose, we use the ICVL dataset [50]. Following [34] we subsample MPII to contain images with persons having all joints. Since VL4Pose does not model diversity, we follow [8, 57] by performing an initial round of random sampling every active learning cycle followed by VL4Pose (for ICVL dataset only). We differ from [47] in two aspects: we train the network to predict occluded joints and take larger crops around persons to add variability. We use different model architectures (Stacked Hourglass [Human Pose][38] - following [47, 57] and DeepPrior [Hand Pose] - defined and followed by [8, 9]). Tab. 1 shows that VL4Pose performs favourably amongst all algorithms with low compute (capable of running at 30FPS): Learning Loss [57], Aleatoric Uncertainty [8], Multi-peak entropy [34]. Multi-Peak is architecture specific and does not extend to hand pose. Learning Loss and Aleatoric lack in performance since they do not explicitly model the pose. In comparison to state-of-the-art (MATAL [14] and GCN [9]), VL4Pose reports competitive but slightly inferior numbers. However, both methods are compute intensive and cannot be used for real-time on-device active learning.

5 Conclusion

VL4Pose is about making the small things click; how far can incorporating simple domain knowledge take us? We answer this with a framework that models simple skeletal constraints to identify out-of-distribution samples. The method seamlessly unifies joint and pose level uncertainty, allowing for better uncertainty estimates. We qualitatively and quantitatively assess VL4Pose, noting that the method is interpretable, real-time, architecture agnostic and competitive with the state-of-the-art, making it suitable for on-device active learning. While pose refinement was an unintended consequence, we lay foundation for future work to incorporate complex skeletal models to push the barriers for out-of-distribution, active learning and pose refinement simultaneously.

References

[1] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1446–1455, 2015.

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

[3] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018.

[4] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6861–6871, 2019.

[5] Adrian Bulat, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. Toward fast and accurate human pose estimation via soft-gated skip connections. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 101–108. IEEE Computer Society.

[6] W. Cai, Y. Zhang, and J. Zhou. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60, 2013. doi: 10.1109/ICDM.2013.104.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[8] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Active learning for bayesian 3d hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3419–3428, 2020.

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

[10] Stefan Depeweg, Jose Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, Stockholm, Sweden, 2018.

[11] Garoe Dorta, Sara Vicente, Lourdes Agapito, Neill DF Campbell, and Ivor Simpson. Structured uncertainty prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5477–5485, 2018.

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[13] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[14] Jia Gong, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Meta agent teaming active learning for pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11089, 2022.

[15] Gautham Krishna Gudur, Prahalathan Sundaramoorthy, and Venkatesh Umaashankar. Activeharnet: Towards on-device deep bayesian active learning for human activity recognition. In *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications*, pages 7–12, 2019.

[16] Théo Guénais, Dimitris Vamvourellis, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. Bacoun: Bayesian classifers with out-of-distribution uncertainty. *arXiv preprint arXiv:2007.06096*, 2020.

[17] Nitesh B Gundavarapu, Divyansh Srivastava, Rahul Mitra, Abhishek Sharma, and Arjun Jain. Structured aleatoric uncertainty in human pose estimation. In *CVPR Workshops*, volume 2, 2019.

[18] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[19] Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. Active learning for speech recognition: the power of gradients, 2016.

[20] Arjun Jain, Jonathan Tompson, Mykhaylo Andriluka, Graham W Taylor, and Christoph Bregler. Learning human pose estimation features with convolutional networks. In *International Conference on Learning Representations*, pages 1–14. Cornell University, 2014.

[21] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2021.

[22] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[23] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[24] Aouaidjia Kamel, Bin Sheng, Ping Li, Jinman Kim, and David Dagan Feng. Hybrid refinement-correction heatmaps for human pose estimation. *IEEE Transactions on Multimedia*, 23:1330–1342, 2020.

[25] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[26] Christine Körner and Stefan Wrobel. Multi-class ensemble-based active learning. In *European conference on machine learning*, pages 687–694. Springer, 2006.

[27] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986, 2019.

[28] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–14, 2021. doi: 10.1109/TITS.2021.3124981.

[29] Jogendra Nath Kundu, Siddharth Seth, Pradyumna YM, Varun Jampani, Anirban Chakraborty, and R Venkatesh Babu. Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20448–20459, 2022.

[30] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, et al. Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods*, 19(4):496–504, 2022.

[31] Quoc V Le, Alex J Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 489–496, 2005.

[32] Andreas M Lehrmann, Peter V Gehler, and Sebastian Nowozin. A non-parametric bayesian network prior of human pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1281–1288, 2013.

[33] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

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

[35] Changsheng Lu and Piotr Koniusz. Few-shot keypoint detection with uncertainty learning for unseen species. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19416–19426, 2022.

[36] Prem Melville and Raymond J Mooney. Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 74, 2004.

[37] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7773–7781, 2019.

[38] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46484-8.

[39] Xuecheng Nie, Jiashi Feng, Junliang Xing, Shengtao Xiao, and Shuicheng Yan. Hierarchical contextual refinement networks for human pose estimation. *IEEE Transactions on Image Processing*, 28(2):924–936, 2018.

[40] Sangho Park and Jake K Aggarwal. A hierarchical bayesian network for event recognition of human actions and interactions. *Multimedia systems*, 10(2):164–179, 2004.

[41] Jia Qian, Sarada Prasad Gochhayat, and Lars Kai Hansen. Distributed active learning strategies on edge computing. In *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pages 221–226. IEEE, 2019.

[42] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

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

[44] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16094–16104, 2021.

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

[46] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, 2008.

[47] Megh Shukla. Bayesian uncertainty and expected gradient length - regression: Two sides of the same coin? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2367–2376, January 2022.

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

[49] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[50] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[51] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014.

[52] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014. doi: 10.1109/CVPR.2014.214.

[53] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020.

[54] Jian Wang, Xiang Long, Yuan Gao, Errui Ding, and Shilei Wen. Graph-pcnn: Two stage human pose estimation with graph pose refinement. In *European Conference on Computer Vision*, pages 492–508. Springer, 2020.

[55] Yuan-Kai Wang and Kuang-You Cheng. A two-stage bayesian network method for 3d human pose estimation from monocular image sequences. *EURASIP Journal on Advances in Signal Processing*, 2010:1–16, 2010.

[56] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11199–11208, 2021.

[57] D. Yoo and I. S. Kweon. Learning loss for active learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102, 2019. doi: 10.1109/CVPR.2019.00018.

[58] Duncan Zauss, Sven Kreiss, and Alexandre Alahi. Keypoint communities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11057–11066, 2021.

[59] Wuqiang Zhang, Zijie Guo, Rong Zhi, and Baofeng Wang. Deep active learning for human pose estimation via consistency weighted core-set approach. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 909–913. IEEE, 2021.