Localization with Sampling-Argmax

Anonymous Author(s)
Affiliation
Address
email

Abstract

Soft-argmax operation is commonly adopted in the detection-based method to localize the target position in a differentiable manner. However, training the neural network with soft-argmax makes the shape of the probability map unconstrained. Consequently, the model lacks pixel-wise supervision through the map during training, leading to performance degradation. In this work, we propose sampling-argmax, a differentiable training method that imposes implicit constraints to the shape of the probability map by minimizing the expectation of the error. To approximate the expectation, we introduce a continuous formulation of the output distribution and develop a differentiable sampling process. The expectation can be approximated by calculating the average error of all samples drawn from the output distribution. We show that sampling-argmax can seamlessly replace the conventional soft-argmax operation on various localization tasks. Comprehensive experiments demonstrate the effectiveness and flexibility of the proposed method. The code is attached in supplementary files and will be published with this paper.

1 Introduction

Localizing the target position from the input is a fundamental task in the field of computer vision. Common approaches to localization can be divided into two categories: regression-based and detection-based. Detection-based methods show superiority over regression-based methods and demonstrate impressive performance on a wide variety of tasks [33, 26, 31, 5, 13, 7, 10, 25, 24]. Probability maps (also referred to as heat maps) are predicted in detection-based methods to indicate the likelihood of the target position. The position with the highest probability is retrieved from the probability map with the *argmax* operation. However, the argmax operation is not differentiable and suffers from quantization error. For accurate localization and end-to-end learning, *soft-argmax* is proposed as an approximation of argmax. It has found a wide range of applications in human pose estimation [26, 16, 17], facial landmark localization [7, 15, 1], stereo matching [33, 11, 2] and object keypoint estimation [24].

Nevertheless, the mechanism of training networks with soft-argmax is rarely studied. The conventional training strategy is to minimize the error between the output coordinate from soft-argmax and the ground truth position. However, this strategy is deficient since it only provides constraints to the expectation of the probability map, not to its shape. As shown in Figure 1, these two maps have the same mean values, but the bottom one is more concentrated. In well-calibrated probability maps, positions that locate closer to the ground truth have higher probabilities. Reliable confidence scores of localization results could be provided, which is essential in unconstrained real-world applications and downstream tasks. Besides, imposing constraints on the probability map can provide supervised pixel-wise gradients and facilitate the learning process.

Prior work [21] attempts to shape the probability map by introducing hand-crafted regularizations. The variance regularization encourages the variance of the probability map to get close to the predefined variance. The Gaussian regularization forces the probability map to resemble a Gaussian

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

distribution. We argue that these variants are overconstrained. The hand-crafted constraints are not always correct in different cases. For example, the underlying shape of the probability map is not necessarily Gaussian, and the underlying variance might change as the input changes. Imposing the model to learn a fixed-variance Gaussian distribution might degrade the model performance.

In this work, we present *sampling-argmax*, a novel training method to obtain well-calibrated probability maps and improve the localization accuracy. To constrain the shape of the map, we replace the objective function of minimizing "the error of the expectation" with minimizing "the expectation of the error". In this way, the network is encouraged to generate higher probabilities around the ground truth position.

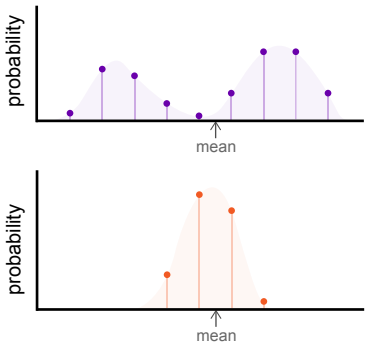

Figure 1: **Top**: an unconstrained probability map. **Bottom**: a well-calibrated probability map. These two maps have different shapes but a same mean value.

A natural way to estimate the expectation is by calculating the probability-weighted sum of the errors at all grid positions. However, we find that the gradient has high variance, and the model is hard to train. To address this issue, we choose to approximate the expectation by sampling. The expectation of the error is calculated as the mean error of all samples. Therefore, the sampling process should be differentiable for end-to-end learning.

In our work, we show that the likelihood of the target position can be modelled in the continuous space with a mixture distribution. Samples can be drawn from the mixture distribution by three steps: i) generate categorical weights from the probability map; ii) draw samples from sub-distributions; iii) obtain a sample by the category-weighted sum. The benefit of using mixture distribution is that differentiable sampling from arbitrary continuous distributions can be resolved by differentiable sampling from categorical distributions, which is less challenging and can be addressed by off-the-shelf discrete sampling methods.

Sampling-argmax is simple and effective. With out-of-the-box settings, it can be integrated into methods that using soft-argmax operation. To study its effectiveness, we conduct experiments on a variety of localization tasks. Quantitative results demonstrate the superiority of sampling-argmax against soft-argmax and its variants. In summary, the contributions of this work are threefold:

1. We propose *sampling-argmax* for improving detection-based localization methods. By minimizing "the expectation of the error", the network generates well-calibrated probability maps and obtains higher localization accuracy.
2. We show the output likelihood can be formulated as a mixture distribution and develop a differentiable sampling pipeline.
3. Comprehensive experiments show that sampling-argmax is effective and can be flexibly generalized to different localization tasks.

2 Preliminary

Given a learned discrete probability map π, the value π_{y_i} indicates the probability of the predicted target appearing at y_i. A direct way to localize the target is taking the position with the maximum likelihood. However, this approach is non-differentiable, and the output is discrete, which impedes end-to-end training and brings quantization errors. Soft-argmax is an elegant approximation to address these issues:

$$\hat{y} = \texttt{soft-argmax}(\pi) = \sum_i \pi_{y_i} y_i. \tag{1}$$

Notice that the soft-argmax operation calculates the probability-weighted sum, which is equivalent to taking the expectation of the probability map π. A conventional way to train the model with the soft-argmax operation is minimizing the distance between the expectation and the ground truth:

$$\mathcal{L} = d(y_t, \mathbb{E}_y[y]) \approx d(y_t, \sum_i \pi_{y_i} y_i), \tag{2}$$

where y_t denotes the ground truth position and $d(\cdot, \cdot)$ denotes the distance function, e.g. ℓ_1 distance. We refer to this objective function as "the error of the expectation".

3 Method

The conventional detection-based method with soft-argmax only supervises the expectation of the probability map. The shape of the distribution remains unconstrained. In well-calibrated probability maps, the positions closer to the ground truth should have higher probabilities. To this end, we proposed a new objective function that optimizes "the expectation of the error" instead of "the error of the expectation". In particular, the objective function is formulated as:

$$\mathcal{L} = \mathbb{E}_y[d(y_t, y)]. \tag{3}$$

The learned distribution tends to allocate high probabilities around the ground truth to minimize the entire loss. In this way, the shape of the probability map is implicitly constrained.

Discrete Distribution. The probability map π predicted by the neural network is discrete. Similar to the soft-argmax operation, the expectation of error can be approximated by calculating the probability-weighted sum of the errors at all grid positions:

$$\mathcal{L} = \mathbb{E}_y[d(y_t, y)] \approx \sum_i \pi_{y_i} d(y_t, y_i). \tag{4}$$

This approximation treats the distribution of the target position as a discrete distribution. The target only appears at the grid positions, i.e. at position y_i with the probability π_{y_i}.

However, because the underlying target lies in a continuous space, modelling the distribution as a discrete distribution is not accurate. The probability map has limited resolution due to the computation complexity. Besides, we find the model is slow to converge by training with Equation 4. For analysis, we derive the gradient from the loss function to the model parameters θ under the discrete approximation:

$$\begin{aligned} \nabla_\theta \mathcal{L} &= \nabla_\theta \mathbb{E}_y[d(y_t, y)] \\ &= \sum_i d(y_t, y_i) \nabla_\theta \pi_{y_i} = \sum_i d(y_t, y_i) \pi_{y_i} \nabla_\theta \log \pi_{y_i} \\ &= \mathbb{E}_y[d(y_t, y) \nabla_\theta \log \pi_y]. \end{aligned} \tag{5}$$

Notice that the form of the gradient is similar to the score function estimator (SF), which is alternatively called the REINFORCE estimator [27]. SF estimator is known to have very high variance and is slow to converge. Therefore, using the discrete approximation for training is not a good solution. This challenge prompts us to explore a better approximation to calculate the expectation of the error.

In the following parts, we present sampling-argmax to estimate the expectation of the error by sampling. We first develop a continuous approximation to the distribution of the target position (Section 3.1). Then we propose a differentiable sampling method (Section 3.2).

3.1 Continuous Mixture Distribution

A differentiable process is necessary to estimate the expectation by sampling. However, since the underlying probability density functions can vary among different input images, it is challenging to draw samples from arbitrary distributions differentiably. In this work, we present a unified method by formulating the target distribution as a mixture distribution.

Let $p(y)$ denotes the underlying density function of the target position, which is defined within the boundary of the input image, i.e. $y \in [0, W]$. As illustrated in Figure 2(a), the interval $[0, W]$ can be divided into n subintervals. The density function can be partitioned into shapes in the subintervals. We could use regular shape (rectangles, triangles, Gaussian functions) in subintervals to form the entire function (as illustrated in Figure 2(b-c)).

Formally, given a finite set of probability density functions $\{f_1(y), f_2(y), \cdots, f_n(y)\}$ and weights $\{w_1, w_2, \cdots, w_n\}$ such that $w_i \geq 0$ and $\sum w_i = 1$, the mixture density function $p(y)$ is formulated as a sum:

$$p(y) = \sum_{i=1}^{n} w_i f_i(y). \tag{6}$$

Here, we can leverage the discrete probability map π to represent the mixture weights, i.e. $w_i = \pi_{y_i}$. In the context of signal processing, the original function can be perfectly reconstructed if the sample

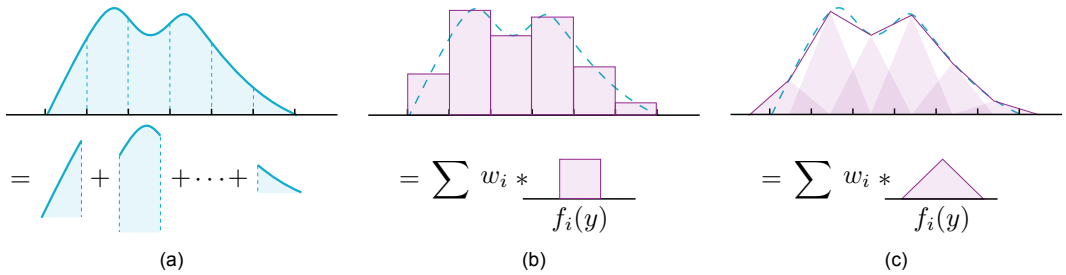

Figure 2: **Representing the continuous distribution as a mixture distribution.** (a) The original probability density function can be viewed as the sum of n sub-functions. Hence, each sub-function can be replaced by standard density functions with proper weights to approximate the original function. (b) Approximate the original function by replacing the sub-functions with uniform distribution. (c) Approximate the original function by replacing the sub-function with the triangular distribution, which is equivalent to the linear interpolation of the discrete weights.

rate (the distance between two adjacent grid points) satisfies the Nyquist-Shannon sampling theorem. However, in our case, the sub-function $f_i(y)$ must be a probability density function, i.e. it has the non-negative values, and its integral over the entire space is equal to 1. Therefore, with these restrictions, the original function $p(y)$ cannot be perfectly reconstructed. For approximation, we study three different types of standard density functions below.

Uniform Basis. For the uniform basis, the sub-function $f_i(y)$ is a uniform distribution centred at the position y_i:

$$f_i(y) = \begin{cases} \frac{1}{c}, & y \in [y_i - \frac{c}{2}, y_i + \frac{c}{2}], \\ 0, & \textit{otherwise}, \end{cases} \tag{7}$$

where c is the distance between two adjacent grid points.

Triangular Basis. For the triangular basis, the sub-function $f_i(y)$ is a triangular distribution:

$$f_i(y) = \begin{cases} \frac{1}{c^2}(y - y_i) + \frac{1}{c}, & y \in [y_i - c, y_i), \\ -\frac{1}{c^2}(y - y_i) + \frac{1}{c}, & y \in [y_i, y_i + c), \\ 0, & \textit{otherwise}. \end{cases} \tag{8}$$

For all y, there exist grid points y_i and y_{i+1} that satisfy $y \in [y_i, y_{i+1}]$. Therefore, we have $p(y) = w_i f_i(y) + w_{i+1} f_{i+1}(y) = \frac{w_{i+1} - w_i}{c^2}(y - y_i) + \frac{w_i}{c}$, which is the linear interpolation of w_i and w_{i+1}. In other words, using triangular bases is equivalent to the linear interpolation of the discrete probability map.

Gaussian Basis. For the Gaussian basis, $f_i(y)$ is the Gaussian function:

$$f_i(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{y - y_i}{\sigma})^2 \Big). \tag{9}$$

where σ denotes the standard deviation. We set $\sigma = c$ by default in the experiments.

3.2 Differentiable Sampling

In this part, we present how to draw a sample from the mixture distribution. We first study the non-differentiable process and then present the differentiable approximation.

Non-differentiable Process. As illustrated in Figure 3(a), the non-differentiable sampling process can be divided into two steps: i) determine which sub-distribution the sample comes from; ii) draw a sample from the selected sub-distribution. In the first step, the sub-distribution can be selected by drawing a random variable from a categorical distribution. The categorical distribution is indicated by the predicted probability map π. The sub-distribution $f_i(y)$ is chosen with the probability π_{y_i}.

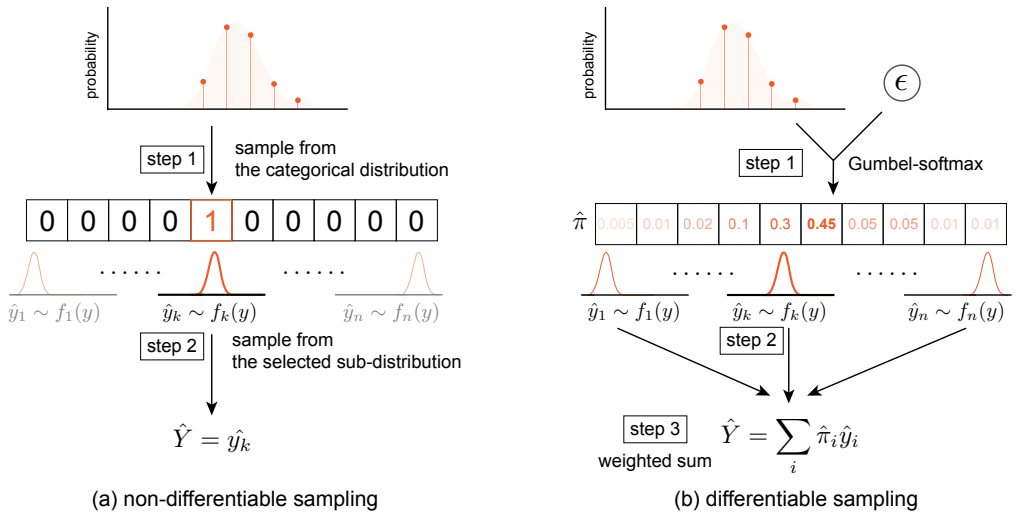

Figure 3: **Illustration of the sampling process.** (a) The non-differentiable process: i) select a sub-distribution by categorical sampling; ii) draw samples from the selected sub-distribution. (b) The differentiable process: i) approximate the categorical sampled weights by Gumbel-softmax; ii) draw samples from all sub-distribution; iii) add all samples together with the sampled weights. Reparameterization allows gradients to flow from the sample to the probability map.

There are a number of methods to draw samples from the categorical distribution. Here, we introduce the Gumbel-Max trick [4, 19]:

$$z = \texttt{one_hot_max}_i[g_i + \log \pi_i], \tag{10}$$

where $g_1, \cdots, g_n$ are i.i.d samples drawn from Gumbel(0, 1), and the sample z is a one-hot vector with the value 1 in the maximum categorical column.

In the second step, sampling from the standard basis function is easy to implement. This step is independent of the predicted probability map π. Therefore, the key to differentiable sampling from the mixture distribution is to make the first step differentiable.

Differentiable Process. The differentiable sampling process consists of three steps. In the first step, we adopt the Gumbel-softmax [9] operation to sample the categorical weight from the probability map. Gumbel-softmax is a continuous and differentiable approximation of the Gumbel-Max trick. We can obtain an $(n-1)$-dimensional simplex $\hat{\pi} \in \Delta$:

$$\hat{\pi}_i = \frac{\exp\left((g_i + \log \pi_i)/\tau\right)}{\sum_{k=1}^{n} \exp\left((g_k + \log \pi_k)/\tau\right)}, \tag{11}$$

where $\hat{\pi} = \{\hat{\pi}_1, \cdots, \hat{\pi}_n\}$ and $\hat{\pi}_i$ denotes the sampled weight of the sub-distribution $f_i(y)$. As the softmax temperature τ approaches 0, the simplex $\hat{\pi}$ becomes one-hot, and its distribution becomes identical to the categorical distribution π.

In the second step, we draw a sample $\hat{y}_i$ from every sub-distribution $f_i(y)$. Note that the sampled weight is not completely one-hot. Therefore, we obtain the final sample $\hat{Y}$ in the third step by adding all samples together with the sampled weight $\hat{\pi}$:

$$\hat{Y} = \sum_{i}^{n} \hat{\pi}_i \hat{y}_i. \tag{12}$$

This process is illustrated in Figure 3(b). With the reparameterization trick, the sample $\hat{Y}$ is computed as a deterministic function of the probability map π and the independent random variables. The randomness of the sampling process is transferred to the variable $g_1, \cdots, g_n$. We denote the sampling process as $\hat{Y} = s(\pi, \epsilon)$, where $\epsilon = \{g_1, \cdots, g_n\}$ follows the multivariate Gumbel(0, 1) distribution. The gradient from the expected error to the model parameters θ is derived as:

$$\nabla_\theta \mathbb{E}_y[d(y_t, y)] = \nabla_\theta \mathbb{E}_\epsilon[d(y_t, s(\pi, \epsilon))] = \mathbb{E}_\epsilon \left[\frac{\partial d}{\partial s} \frac{\partial s}{\partial \pi} \frac{\partial \pi}{\partial \theta}\right]. \tag{13}$$

As we see, the gradient of the continuous sampling process is easy to compute via backpropagation. Therefore, we can relax the objective function by calculating the average error of the samples drawn from the mixture distribution. The objective function is written as:

$$\mathcal{L} = \mathbb{E}_{y\sim p(y)}[d(y_t, y)] \approx \frac{1}{N_s}\sum_{k=1}^{N_s} d(y_t, \hat{Y}_k) = \frac{1}{N_s}\sum_{k=1}^{N_s} d(y_t, s(\pi, \epsilon_k)), \tag{14}$$

where N_s denotes the number of samples. In the testing phase, no randomness is introduced, and sampling-argmax degrades to soft-argmax.

While the sampling process is differentiable, the sample $\hat{Y}$ does not follow the original mixture distribution $p(y)$ for non-zero temperature. For small temperatures, the distribution of $\hat{Y}$ is close to $p(y)$, but the variance of the gradients is large. There is a tradeoff between small temperatures and large temperatures. In our experiments, we start at a high temperature and anneal to a small temperature.

4 Related Work

Localization with Soft-Argmax. Nibali et al. [21] introduced hand-crafted regularization to constrain the shape of the probability map. Unlike them, our objective function does not set pre-defined hyperparameters for the shape of the map, which makes it general and flexible in applying to various applications.

Other works [10, 13] study how to localize target with soft-argmax in different situations. Joung et al. [10] proposed sinusoidal soft-argmax for cylindrical probabilities map. Lee et al. [13] proposed kernel soft-argmax to make the results less susceptible to multi-modal probability map. Our work is compatible with these methods by applying the sinusoidal function to the grid positions or multiplying the Gaussian kernel before obtaining the probability map.

Differentiable Sampling. Differentiable sampling for a discrete random variable has been studied for a long time. Maddison et al. [18] and Jang et al. [9] concurrently proposed the idea of using a softmax of Gumbel as relaxation for differentiable sampling from discrete distributions. Kočiskỳ et al. [12] relaxed the discrete sampling by drawing symbols from a logistic-normal distribution rather than drawing from softmax. In this work, unlike previous methods that study discrete distributions, we focus on continuous distributions. We propose a relaxation of continuous sampling by formulating the target distribution as a mixture distribution.

5 Experiments

We validate the benefits of the proposed sampling-argmax with experiments on a variety of localization tasks, including human pose estimation, retina segmentation, object keypoint estimation and facial landmark localization. Sampling-argmax is compared with the conventional soft-argmax and the variants that using additional auxiliary loss [21]. Training details of all tasks are provided in the supplemental material.

Variance Regularization. Variance regularization is to control the variance of the probability map. It pushes the variance of the probability map close to the target variance σ_t^2:

$$\mathcal{L}_{var} = \|\texttt{Var}(\pi) - \sigma_t^2\|_2^2, \tag{15}$$

where the target variance σ_t^2 is a hyperparameter and the variance of the probability map $\texttt{Var}(\pi)$ is approximated in a discrete manner, i.e. $\texttt{Var}(\pi) = \sum_i \pi_{y_i}(y_i - \sum_k \pi_{y_k} y_k)^2$.

Distribution Regularization. Distribution regularization is to impose strict regularization on the appearance of the heatmap to directly encourage a certain shape. Specifically, [21] forces the probability map to resemble a Gaussian distribution by minimizing the Jensen-Shannon divergence between π and target discrete Gaussian distribution:

$$\mathcal{L}_{JS} = D_{JS}(\pi \| \mathcal{N}(\mathbb{E}(y), \sigma_t^2)). \tag{16}$$

5.1 2D Human Pose Estimation from RGB

We first evaluate the proposed sampling-argmax in 2D human pose estimation. In 2D human pose estimation, the probability map is a typical representation to localize body keypoints. The experiments are conducted on the large-scale in-the-wild 2D human pose benchmark – COCO Keypoint [14]. We adopt the standard model SimplePose [28] for experiments. We follow the standard metric of COCO Keypoint and use mAP over 10 OKS (object keypoint similarity) thresholds for evaluation.

As shown in Table 1, the proposed sampling-argmax significantly outperforms the soft-argmax operation and its variants. The triangular basis brings **5.3** mAP improvement (relative **8.2**%) to the original soft-argmax operation. Besides, we find the auxiliary losses degrade the model performance in COCO Keypoint.

Table 1: Quantitative results on COCO Keypoint.

	Soft	Soft w/ V.R	Soft w/ D.R.	Samp. Uni.	Samp. Tri	Samp. Gau.
mAP	64.5	60.6	55.6	68.2	**69.8**	68.3
mAP@0.5	84.7	81.5	77.8	87.2	**87.9**	87.3
mAP@0.75	70.9	65.7	60.8	75.0	**76.2**	75.2

Number of Samples. In our method, the differentiable sampling process is utilized to approximate the expectation of the error. As the number of samples increases, the approximation will be closer to the underlying expectation. To study how the number of samples affects the final results, we compare the performance of the models that trained with different numbers of samples. In Table 2, we report the results with $N_s = \{1, 5, 10, 30, 50\}$. It shows that a large number of samples might improve the performance but not necessary. Training the model with only one sample can still obtain high performance while saving computation resources.

Table 2: Comparison of different sample numbers.

N_s	1	5	10	30	50
Samp. Uni.	67.8	67.8	67.9	68.2	68.1
Samp. Tri.	69.7	69.7	69.6	**69.8**	**69.8**
Samp. Gau.	68.1	68.1	68.2	68.3	68.3

Correlation with Prediction Correctness. For a well-calibrated probability map, the shape of the map could reflect the uncertainty of the regression output. When encountering challenging cases, the probability map would have a large variance, resulting in a lower peak value. In other words, the peak value establishes the correlation with the prediction correctness. To demonstrate the probability map trained with sampling-argmax is better-calibrated, we calculate the *Pearson correlation coefficient* between the peak value and the prediction correctness. The correctness is represented by the OKS between the predicted pose and the ground-truth pose. Table 3 compares the correlation with prediction correctness among different methods. It shows that sampling-argmax has a much stronger correlation to the correctness than other methods. Compared to the soft-max operation, sampling-argmax with the triangular bases brings **85.4**% relative improvement. It demonstrates that training with sampling-argmax can obtain a more reliable probability map, which is essential to real-world applications and downstream tasks.

Table 3: Correlation testing.

Method	Corr.
Soft	0.233
Soft w/ V.R	0.158
Soft w/ D.R	0.082
Samp. Uni.	0.394
Samp. Tri.	**0.432**
Samp. Gau.	0.423

5.2 3D Human Pose Estimation from RGB

We further evaluate the proposed sampling-argmax on Human3.6M [8], an indoor benchmark for 3D human pose estimation. The 3D probability map is adopted to represent the likelihoods for joints in the discrete 3D space. We adopt the model architecture of prior work [26]. Following previous methods [22, 26, 20], MPJPE and PA-MPJPE [3] are used as the evaluation metrics. Comparisons

with baselines are shown in Table 4. The proposed sampling-argmax provides consistent performance improvements. Different from the experiments on COCO Keypoint, the variance regularization provides performance improvements in Human3.6M.

Table 4: Quantitative results on Human3.6M.

	Soft	Soft w/ V.R	Soft w/ D.R.	Samp. Uni.	Samp. Tri.	Samp. Gau.
MPJPE	50.4	49.7	51.9	49.6	**49.5**	50.9
PA-MPJPE	39.5	39.2	41.4	39.1	39.1	**39.0**

5.3 Retina Segmentation from OCT

Using optical coherence tomography (OCT) to obtain 3D retina images is widely used in the clinic. A major goal of analyzing retinal OCT images is retinal layer segmentation. Previous work [5] proposes a regression method to regress the boundary and obtain the sub-pixel surface positions. One-dimensional probability maps are leveraged to model the position distribution of the surface in each column. In the testing phase, the soft-argmax method is used to infer the final surface positions. The entire surface can be reconstructed by connecting the surface positions in all columns.

The experiments are conducted on the *multiple sclerosis and healthy controls* dataset (MSHC) [6]. Mean absolute distance (MAD) and standard deviation (Std. Dev.) are used as evaluation metrics. Quantitative results are reported in Table 5. It shows that sampling-argmax achieve superior performance to other methods, while the auxiliary losses also provide performance improvements.

Table 5: Quantitative results on MSHC dataset.

	Soft	Soft w/ V.R	Soft w/ D.R.	Samp. Uni	Samp. Tri.	Samp. Gau.
MAD	3.08	0.743	0.746	**0.735**	0.744	0.740
Std. Dev.	0.281	0.114	0.108	0.101	**0.100**	0.104

5.4 Supervised Object Keypoint Estimation from Point Clouds

Detecting aligned 3D object keypoints from point clouds has a wide range of applications on object tracking, shape retrieval and robotics. Probability maps are adopted to localize the semantic keypoints. Different from the RGB input, the probability map indicates the pointwise score of the input point cloud, not the grid position of an image. The distances between the adjacent point-pairs are different. Besides, point clouds are unordered, and each point has a different number of neighbours. Therefore, it is hard to directly apply the uniform bases or linear interpolation, which requires a constant adjacent distance. Fortunately, the Gaussian basis can be adopted. In the experiment, we set the standard deviation σ of the Gaussian bases to 0.01, which is the average adjacent point distance in the input point clouds. PointNet++ [23] is adopted as the backbone network. The experiments are conducted on the large-scale object keypoint dataset – KeypointNet [30]. The percentage of correct keypoints (PCK) [29] is adopted for evaluation. The error distance threshold is set to 0.01.

Table 6 shows the quantitative results on 16 categories. It shows that the proposed sampling-argmax is also effective on the non-grid input data. Table 6 also compare the results of sampling-argmax with different numbers of samples. It is seen that $N_s = 30$ leads to the best average performance.

Table 6: Quantitative results of supervised learning on KeypointNet dataset, reported as PCK.

	Air.	Bat.	Bed	Bot.	Cap	Car	Cha.	Gui.	Hel.	Kni.	Lap.	Mot.	Mug	Ska.	Tab.	Ves.	Avg
Soft	64.9	43.6	44.0	53.9	8.3	40.2	37.2	**45.5**	4.9	43.8	46.6	40.8	23.9	27.7	53.9	32.6	38.2
Soft w/ V.R	64.1	41.6	39.2	53.2	12.5	38.3	37.7	44.5	3.7	39.8	**52.8**	44.0	24.9	25.6	54.4	30.7	37.9
Soft w/ D.R	63.2	42.7	43.9	55.8	16.7	42.2	38.6	43.2	4.9	42.4	48.9	41.9	26.8	28.2	54.0	30.3	39.0
Samp. Gau. ($N_s = 1$)	65.0	43.0	41.2	53.6	6.2	43.4	38.7	42.5	**6.2**	45.4	50.6	43.5	26.3	**37.5**	51.6	**33.3**	39.3
Samp. Gau. ($N_s = 5$)	**65.1**	42.4	43.8	54.7	12.5	43.2	37.1	44.6	1.9	45.4	46.6	44.7	29.7	26.7	**54.6**	31.4	39.0
Samp. Gau. ($N_s = 10$)	64.0	**45.5**	41.7	**58.6**	**20.8**	40.9	37.0	43.4	3.7	45.7	48.3	**46.4**	18.2	34.4	53.5	32.3	39.7
Samp. Gau. ($N_s = 30$)	64.3	45.1	**47.5**	58.4	6.2	**44.6**	**39.2**	45.4	**6.2**	**45.8**	48.7	43.4	**29.9**	30.4	54.1	28.8	**39.9**

5.5 Unsupervised Object Keypoint Estimation from Point Clouds

We then evaluate the proposed method on object keypoint estimation in the context of unsupervised learning. The autoencoder framework is adopted to estimate the keypoint in an unsupervised manner. The encoder first estimates the 3D keypoints, and the decoder reconstructs the object point clouds from the estimated keypoints. We follow the state-of-the-art method [24] that generates 3D keypoints with the soft-argmax operation for differentiable and end-to-end learning. The soft-argmax is replaced with sampling-argmax, where the Gaussian bases with the standard deviation $\sigma = 0.01$ are used.

The experiments are conducted on KeypointNet [30]. Unlike supervised learning, the semantic of each predicted keypoint is unknown in unsupervised methods. Therefore, the PCK metric is not applicable. For evaluation, we adopt the dual alignment score (DAS) following the previous method [24]. Table 7 reports the performance comparison with other methods.

Table 7: Quantitative results of unsupervised learning on KeypointNet dataset, reported as DAS.

	Air.	Bat.	Bed	Bot.	Cap	Car	Cha.	Gui.	Hel.	Kni.	Lap.	Mot.	Mug	Ska.	Tab.	Ves.	Avg
Soft	69.1	56.2	58.0	45.4	59.1	**70.2**	76.8	34.1	55.7	50.0	91.5	53.4	52.2	65.7	72.5	35.8	59.1
Soft w/ V.R	72.0	55.4	57.4	**52.8**	54.7	63.4	70.9	**56.1**	61.6	50.3	82.4	59.8	**71.7**	65.3	**85.1**	38.1	62.3
Soft w/ D.R	47.9	35.5	47.3	46.1	58.3	65.5	60.9	35.3	47.6	**69.3**	64.1	55.0	45.9	44.2	57.6	28.8	50.6
Samp. Gau. ($N_s = 1$)	**73.9**	53.8	**63.5**	43.9	**67.0**	69.3	77.7	46.6	59.1	55.9	87.8	59.0	67.0	66.2	80.3	36.4	62.9
Samp. Gau. ($N_s = 5$)	73.1	54.0	61.9	48.4	64.4	67.0	81.1	50.7	55.2	50.1	87.5	58.2	58.9	65.9	77.9	**41.2**	62.2
Samp. Gau. ($N_s = 10$)	**73.9**	**58.8**	61.7	46.2	60.9	68.6	72.0	53.6	56.5	48.1	**91.6**	**59.8**	68.8	65.8	83.5	34.9	62.8
Samp. Gau. ($N_s = 30$)	71.2	56.7	60.0	51.0	58.4	64.1	**83.8**	47.6	**61.8**	47.8	91.3	55.5	68.5	**70.6**	81.7	37.5	**63.0**

5.6 Facial Landmark Localization from RGB

We further evaluate the proposed sampling-argmax on the facial landmark localization dataset MTFL [32]. Absolute error and relative error (normalized by the two-eye distance) are adopted as evaluation metrics. Quantitative results are reported in Table 8. Consistent with the experiments on other tasks, sampling-argmax provides performance improvement to facial landmark localization.

Table 8: Quantitative results on MTFL dataset.

	Soft	Soft w/ V.R	Soft w/ D.R.	Samp. Uni.	Samp. Tri.	Samp. Gau.
Abs. Err	3.18	3.16	3.15	3.00	**2.98**	2.94
Rel. Err	7.25	7.22	7.20	6.86	**6.82**	6.96

6 Limitation and Future Work

In our method, the underlying density function of the target position is approximated by a mixture of sub-distributions. By comparing the performance of the three proposed bases, we see that a more accurate reconstruction of the underlying function leads to better results. Theoretically, the underlying density function cannot be perfectly reconstructed since the proposed basis distributions are fixed. To address this limitation, learnable sub-distributions could be adopted in future works. For example, *normalizing flow* models can be leveraged to predict sub-distribution at each position according to the corresponding features. In this way, the sub-distributions are no longer fixed, and the mixture distribution has the potential to precisely reconstruct the underlying distribution and further improve the model performance.

7 Conclusion

In this paper, we propose *sampling-argmax*, an operation for improving the detection-based localization. Sampling-argmax implicitly imposes shape constraints to the predicted probability map by optimizing "the expectation of error". With the continuous formulation and differentiable sampling, sampling-argmax can seamlessly replace the conventional soft-argmax operation. We show that sampling-argmax is effective and flexible by conducting comprehensive experiments on various localization tasks.

References

[1] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. Attention-driven cropping for very high resolution facial landmark detection. In *CVPR*, 2020.

[2] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *ICCV*, 2019.

[3] John C Gower. Generalized procrustes analysis. *Psychometrika*, 1975.

[4] Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954.

[5] Yufan He, Aaron Carass, Yihao Liu, Bruno M Jedynak, Sharon D Solomon, Shiv Saidha, Peter A Calabresi, and Jerry L Prince. Fully convolutional boundary regression for retina oct segmentation. In *MICCAI*, 2019.

[6] Yufan He, Aaron Carass, Sharon D Solomon, Shiv Saidha, Peter A Calabresi, and Jerry L Prince. Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls. *Data in brief*, 2019.

[7] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *CVPR*, 2018.

[8] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 2014.

[9] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

[10] Sunghun Joung, Seungryong Kim, Hanjae Kim, Minsu Kim, Ig-Jae Kim, Junghyun Cho, and Kwanghoon Sohn. Cylindrical convolutional networks for joint object detection and viewpoint estimation. In *CVPR*, 2020.

[11] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017.

[12] Tomáš Kočiskỳ, Gábor Melis, Edward Grefenstette, Chris Dyer, Wang Ling, Phil Blunsom, and Karl Moritz Hermann. Semantic parsing with semi-supervised sequential autoencoders. In *EMNLP*, 2016.

[13] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *CVPR*, 2019.

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

[15] Yinglu Liu, Hao Shen, Yue Si, Xiaobo Wang, Xiangyu Zhu, Hailin Shi, Zhibin Hong, Hanqi Guo, Ziyuan Guo, Yanqin Chen, et al. Grand challenge of 106-point facial landmark localization. In *ICMEW*, 2019.

[16] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018.

[17] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 2019.

[18] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017.

[19] Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *NeurIPS*, 2014.

[20] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single rgb image. In *ICCV*, 2019.

[21] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018.

[22] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017.

[23] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017.

[24] Ruoxi Shi, Zhengrong Xue, Yang You, and Cewu Lu. Skeleton merger: an unsupervised aligned keypoint detector. In *CVPR*, 2021.

[25] Riccardo Spezialetti, Federico Stella, Marlon Marcon, Luciano Silva, Samuele Salti, and Luigi Di Stefano. Learning to orient surfaces by self-supervised spherical cnns. In *NeurIPS*, 2020.

[26] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018.

[27] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.

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

[29] Li Yi, Hao Su, Xingwen Guo, and Leonidas J Guibas. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In *CVPR*, 2017.

[30] Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. In *CVPR*, 2020.

[31] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *ICRA*, 2018.

[32] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.

[33] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *ICCV*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We claim that the proposed sampling-argmax can help the model obtain well-calibrated probability maps and improve the localization accuracy. In our experiments, we validate the localization accuracy of sampling-argmax across **six** tasks. We also demonstrate that the probability maps are well-calibrated by conducting correlation testing.
   (b) Did you describe the limitations of your work? [Yes] Please see Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see the supplemental material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]

(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Our code is attached in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Training details are elaborated in the supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Error bars are not reported because it would be too computationally expensive.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The required training resources (number of GPUs) are elaborated in the supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Please see Section 5.
   (b) Did you mention the license of the assets? [Yes] Please see the supplemental material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] All data we used is publicly available. Please see the supplemental material.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Please see the supplemental material.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]