Towards Unbiased Label Distribution Learning for Facial Pose Estimation Using Anisotropic Spherical Gaussian

Zhiwen Cao[1(✉)], Dongfang Liu[2], Qifan Wang[3], and Yingjie Chen[1]

[1] Purdue University, West Lafayette, USA
{cao270,victorchen}@purdue.edu
[2] Rochester Institute of Technology, Rochester, USA
dongfang.liu@rit.edu
[3] Meta AI, Menlo Park, USA
wqfcr@fb.com

Abstract. Facial pose estimation refers to the task of predicting face orientation from a single RGB image. It is an important research topic with a wide range of applications in computer vision. Label distribution learning (LDL) based methods have been recently proposed for facial pose estimation, which achieve promising results. However, there are two major issues in existing LDL methods. First, the expectations of label distributions are biased, leading to a *biased pose estimation*. Second, *fixed* distribution parameters are applied for all learning samples, severely limiting the model capability. In this paper, we propose an Anisotropic Spherical Gaussian (ASG)-based LDL approach for facial pose estimation. In particular, our approach adopts the spherical Gaussian distribution on a unit sphere which constantly generates *unbiased expectation*. Meanwhile, we introduce a new loss function that allows the network to learn the distribution parameter for each learning sample *flexibly*. Extensive experimental results show that our method sets new state-of-the-art records on AFLW2000 and BIWI datasets.

Keywords: Facial pose estimation · Anisotropic spherical Gaussian · Label distribution learning

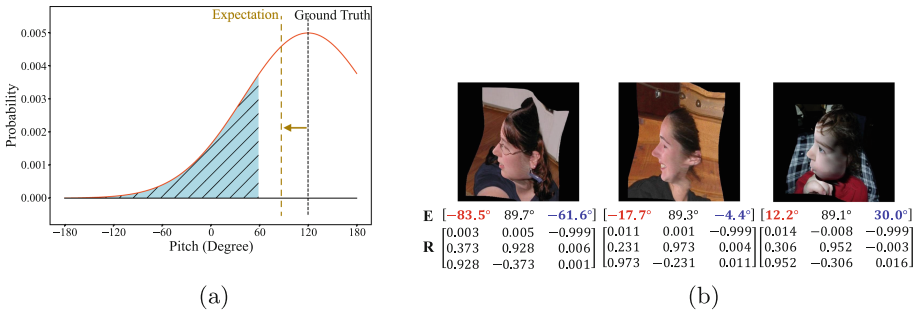
1 Introduction

The task of facial pose estimation is to estimate the orientation of the face from a single RGB image. It plays an important role in many real-world applications, including driver's monitoring system [16,33], human-computer interaction [4,29] and face alignment [3,44]. With the recent advance of deep learning in computer vision [5,6,19,25,26,43], learning-based facial pose estimation has become

Z. Cao and D. Liu—Equal contributions.
Q. Wang—The analysis and all work described in this paper was performed by the authors at Purdue and RIT. Qifan Wang served as an advisor to the project.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13672, pp. 737–753, 2022.
https://doi.org/10.1007/978-3-031-19775-8_43

a dominant approach, achieving promising results [1,2,35,38]. However, as a general problem in deep learning, data shortage also limits the concurrent methods for facial pose estimation to achieve superior performance. How to effectively estimate the facial pose with limited data remains a challenge, which is the focus of this work.



(a) (b)

Fig. 1. (a) Example of Biased Expectation for 1D Gaussian label distribution. The distribution from original pitch = 120° in the range $(-180°, +180°]$ gives a biased expectation of the label. The condition becomes worse as the original angle gets closer to 180° or −180°. **(b) E** and **R** denote ***Euler angles*** and ***rotation matrix***, respectively. Euler angles have inconsistent representations for profile faces. The **Red** and **Blue** values show evident discrepancies when denoting similar profile poses. Applying LDL on Euler angles inevitably introduces heavily noisy supervision. In contrast to Euler angles, the corresponding elements in rotation matrices are close to each other. All samples are from the 300W-LP dataset [45]. (Color figure online)

Recently, label distribution learning (LDL) has been introduced to address the issue of insufficient training data. These LDL methods aim at reconstructing new labels of the distribution around the original ones for training, which promote the learning of facial images not only from their own labels but also the adjacent ones. LDL has shown its effectiveness in tasks such as facial age estimation [15], facial attractiveness estimation [9] and crowd counting [49]. However, the exploration of LDL application to facial pose estimation is insufficient.

To date, LDL in the task of facial pose estimation are mainly applied on Euler angles which are known as pitch, yaw and roll. A seminal work from [14] proposed to use a 2D Gaussian Distribution to describe the probability distribution between pitch and yaw in the range of $(-90°, +90°)$. Liu *et al.* [28] followed the track and converted each Euler angle label to a 1D Gaussian distribution. They also expand the task to the one of wild range. Therefore, each face image corresponds to three 1D Gaussian distributions (*i.e.*, ρ_{pitch}, ρ_{yaw} and ρ_{roll}). For instance, an original label of the pitch angle (120°) can be used to generate a Gaussian distribution in $(-180°, 180°]$ (see Fig. 1a). Through predicting the probability of each integer degree in set $\mathcal{S} = \{-179°, -178°, \cdots, 179°, 180°\}$ and compute the cross entropy loss, the task can be considered as the combination of both regression and classification.

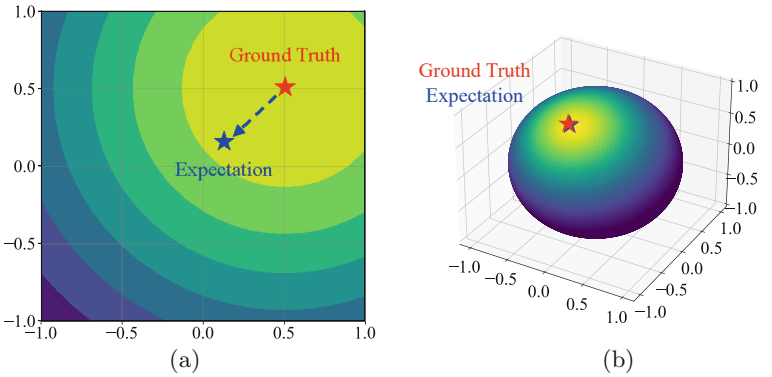


Fig. 2. (a) 2D distribution [2] generated by the first two elements of a column vector of rotation matrices (we omit the third element for visualization). The limit of range $[-1, 1]$ in two directions results in biased expectation. **(b) Spherical Gaussian Distribution** (ours) guarantees ***unbiased expectation***.

Albeit being simple and effective, applying LDL on Euler angles has several obstacles: ① Euler angle is not a continuous rotation representation and LDL deteriorates the issue by contributing the learning of adjacent angles. The discontinuity is embodied in the Euler angle labels of profile faces (see Fig. 1b). Since similar profile images have very different Euler angle labels, converting the angles to distributions cannot help the learning of adjacent labels; ② Gaussian distribution labeling on Euler angles leads to biased expectations. Since the angle is limited to a certain range $(-180°, 180°]$, probabilities assigned in the shadow area make the expectation of labels incorrectly shift to left (see Fig. 1a); and ③ concurrent LDL methods utilize the variance of Gaussian distribution as a hyper-parameter which is fixed during training. This is computationally inefficient because they need to perform an exhaustive search to cherry pick the best parameter setting. Besides, using the same distribution for all the poses is not aligned with the real situation. Since faces at different poses have diverse contributions to adjacent faces, the network should learn the distribution parameters adaptively.

The first issue was studied in [2], which identifies the discontinuity issue of Euler angle and proposed a vector-based representation to train the network. In other words, they let the neural network learn the rotation matrices from facial images. Rotation matrices can form a continuous special orthogonal group $SO(3)$ and can circumvent the problem of discontinuity. However, they still failed to recognize the issue of biased expectation. Since every element of the rotation matrix stays in the range of $[-1, 1]$, they convert each element to a Gaussian distribution in range $[-1, 1]$ and let the network learn the distribution in an element-wise manner. Consequently, the issue of biased expectation is inevitably similar to Euler angles. To our best knowledge, the second and third issues remain largely under-explored in existing literature.

In light of the foregoing discussions, we are motivated to present our Anisotropic Spherical Gaussian (ASG)-based label distribution learning method for facial pose estimation. Specifically, we treat each column vector of the rotation matrix as an entity and map them to a spherical Gaussian distribution respectively. Due to the symmetric distribution of ASG, our approach guarantees for an unbiased expectation of label distributions. The difference between ASG and the method in [2] is demonstrated in Fig. 2. Armed with the spherical Gaussian distribution, we further design a new loss function for the network to learn the distribution parameters adaptively during the training stage. This enables every facial image to adjust contribution to adjacent poses based on its pose. Ablation studies show that it can transcend the cherry-picked parameter by at least 4.0% when trained on 3000W-LP and tested on AFLW2000 dataset.

Our method enjoys a few attractive qualities: ❶ it ensures the network learns the distribution with ***unbiased expectation***. Since most existing methods have biased expectations (unless the original ground truth is exactly in the middle), we observe significant performance gain from our method; ❷ the capacity of learnable ASG distribution parameters allows the network to adjust the parameter for each pose, enabling a ***fine-grained*** prediction; ❸ all the performance achievement comes from optimization on ***representation of rotation*** without increasing the size of neural networks. Our approach achieves state-of-the-art performance with a very light-weighted backbone network, *i.e.*, ResNet18 [19]. Specifically, we decrease the Mean Absolute Error (MAE) by 0.27° (6.9% ↓) compared to [1] and 0.19° (5.0% ↓) compared to [38] when tested on AFLW2000 dataset [51]; and ❹, our method is the ***first attempt*** that adopts directional statistics in the task of pose estimation. We believe it can help invoke more thoughts for further exploration in the community. Our contributions are summarized below:

- We propose a novel ASG-LDL method which encodes each column vector of the rotation matrix as an anisotropic spherical gaussian on a unit sphere. Our method addresses the issue of biased expectation that is under-explored in previous works.
- We propose a novel training paradigm that allows the network to learn the distribution parameters adaptively. The flexibility allows the network to learn individual distribution parameters for each pose.
- We conduct extensive experiments on two benchmarks. Experimental results show the effectiveness of our method. With a light-weight ResNet-18 as the backbone, our method achieves state-of-the-art results and outperforms many strong baselines with a heavier backbone (*i.e.*, ResNet-50).

2 Related Work

This section summarizes the recent progresses in the related fields regarding facial pose estimation, label distribution learning and spherical Gaussian distribution.

Facial Pose Estimation. Recently, landmark-free learning based methods have become popular. By training an end-to-end deep neural network, it can estimate the face poses using global information and can be more robust to the environment variations. [35] puts forward a CNN with a multi-loss function that performs binned classification to regress three Euler angles. [46] proposes a fine-grained structure by learning global spatial feature importance that improves the results. [20] formulates face pose estimation using quaternion-annotated labels to avoid the ambiguity problem in Euler angle representation. [1] proposes a Faster RCNN based network to regress 6DoF pose of faces by performing pose estimation and face alignment simultaneously. [38] puts forward a multi-modal network that can perform three tasks of head pose estimation, landmark-based face alignment and localization of face simultaneously. By the combination of three tasks they achieve state-of-the-art results. All of the above methods perform the training process through direct regression. Differently, we approach the problem as a label distribution learning task.

Label Distribution Learning. Label distribution learning [30] is a learning paradigm that is first proposed for facial age estimation [15,27]. [15] finds that faces at close ages look similar. Therefore, they map each face image to a label distribution which covers a certain number of ages. Through this way one face image can contribute to not only the learning of its chronological age, but also the learning of its adjacent ages. LDL also shows it effectiveness in similar tasks such as facial attractiveness estimation [9], crowd counting [49] and movie rating prediction [13] *etc.* [8] applies a similar approach on ordinal regression such as image ranking and monocular depth estimation. [2] shows that the evaluation metric, mean absolute error of Euler angles (MAE), cannot reflect the actual performance especially for profile faces. Instead, they propose to use mean absolute error of vectors (MAEV) as a new metric. However, all the methods give biased expectation from the distribution, which severely limits the performance of neural network.

Spherical Gaussian Distribution. Spherical Gaussian (SG) distribution, also known as von Mises-Fisher distribution [11], is commonly used to simulate the properties of illumination and reflection in computer graphics. [18] uses SG to estimate multiple light sources and reflectance properties. [39] approximates the normal distribution function (NDF) by a mixture of SGs. [7] uses SG for the approximation of Bidirectional Transmittance Distribution Function (BTDF) for real-time estimation of environment lighting. However, SG only describes an isotropic distribution. [42] further proposes the ASG distribution which can describe an anisotropic distribution for rendering applications. Inspired by the above work, we successfully extend ASG to the field of head pose estimation.

6D Object Pose Estimation. 6D object pose estimation includes estimation of 3D location and 3D orientation. The latter task resembles our head pose estimation. The approaches for 6D object pose estimation can be generally classified into two categories. The first type such as [34,37,47] first capture instance information and keypoints from images to determine locations of objects, then build

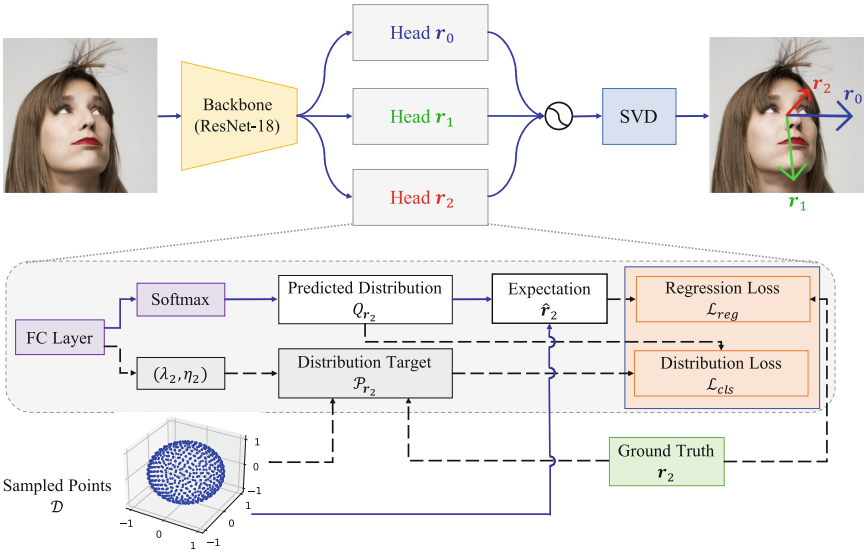


Fig. 3. The overall framework. The dashed lines are only used in the training stage. For simplicity, we only visualize one head $\boldsymbol{r}_2$ but each head has the same working pipeline.

the correspondence between the 2D and 3D keypoints. After that, they obtain 6D pose estimation by solving the PnP problem [24]. The other category of methods such as [12,31,41] use neural networks to estimate orientation of objects directly. To our knowledge, the use of spherical Gaussian is a new attempt in the task of pose estimation.

3 Proposed Method

3.1 Overview

Our overall framework is illustrated in Fig. 3. The network learns the pose information from a cropped human facial image. To demonstrate the advantage of our approach, we choose light-weighted ResNet-18 as our backbone network. We append three heads to the ResNet-18 backbone as each head corresponds to one pose vector. They work collectively to perform the facial pose estimation. During training, the backbone first extracts the features from the input image and then feeds them to each of the heads, which is supervised by the classification and regression loss respectively. During inference, the three heads work collaboratively to predict the rotation matrix through singular value decomposition (SVD). We elaborate our method in the following sections.

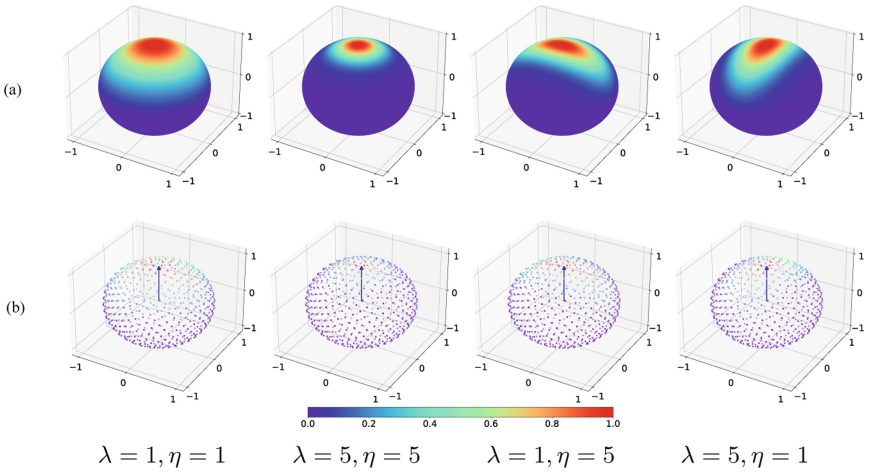


Fig. 4. Visualization of ASG distributions of different λ and η when $\boldsymbol{r} = [0, 0, 1]^T$. a) ASG distribution on a unit sphere; b) visualization of sampled points with probabilities when $M = 600$.

3.2 Motivation

When we use a rotation matrix $\boldsymbol{R}_{3\times3} = [\boldsymbol{r}_0, \boldsymbol{r}_1, \boldsymbol{r}_2]$ to describe facial poses, the three column vectors $\boldsymbol{r}_0$, $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ are equivalent to three pose vectors in Fig. 3, *i.e.* left (blue), down (green) and front (red) vectors respectively [2]. Therefore, for a ground truth pose vector $\boldsymbol{r}_i, i = \{0, 1, 2\}$, any direction $\boldsymbol{v}$ surrounding it can also be regarded as an alternative legitimate label. The smaller the angle difference is, the more likely that the vector $\boldsymbol{v}$ is a valid label. Therefore, all the probabilities of a direction $\boldsymbol{v}$, that can be considered as a legitimate label, constitute a probability distribution on a unit sphere. Intuitively, the probability distribution can be represented using an isotropic spherical Gaussian (SG) model, since the probability is only related to the angle between $\boldsymbol{v}$ and $\boldsymbol{r}_i$. However, human faces change at different rates when rotating along different axes. For example, rolling a face with 45° does not change the observed area of the face, while nodding or raising the face for 45° makes large portion of facial area self-occluded. Based on this observation, we propose to use ASG distribution which is able to capture the anisotropic features along different axes.

3.3 Label Distribution Construction

All three pose vectors constitute an orthogonal coordinate system. For each ground truth pose vector $\boldsymbol{r}_i$, we can calculate the portion G^i that a direction $\boldsymbol{v}$ accounts for a full class description of the sample:

$$
\begin{aligned}
G^i(\boldsymbol{v}; \boldsymbol{R}, [\lambda, \eta]) &= c \cdot \mathrm{S}(\boldsymbol{v}; \boldsymbol{r}_i) \cdot e^{-\lambda(\boldsymbol{v}\cdot\boldsymbol{r}_j)^2 - \eta(\boldsymbol{v}\cdot\boldsymbol{r}_k)^2} \\
\text{where } i &= \{0, 1, 2\} \\
j &= (i+1) \bmod 3 \\
k &= (i+2) \bmod 3.
\end{aligned} \tag{1}
$$

Here, $\boldsymbol{R} = [\boldsymbol{r_0}, \boldsymbol{r}_1, \boldsymbol{r}_2]$. λ and η are the parameters that control the decreasing speed of possibility along $\boldsymbol{r}_j$ and $\boldsymbol{r}_k$. Figure 4a illustrates the spherical Gaussian distribution of different λ and η. c is the normalization term that ensures the sum of probability distribution to be 1. $S(\boldsymbol{v}; \boldsymbol{r}_i) = \max(\boldsymbol{v} \cdot \boldsymbol{r}_i, 0)$ is the smooth term. Since the exponential part $B(\boldsymbol{v}) = e^{-\lambda(\boldsymbol{v}\cdot\boldsymbol{r}_j)^2 - \mu(\boldsymbol{v}\cdot\boldsymbol{r}_k)^2}$, also known as Bingham distribution [23], is antipodally symmetric and has two peaks at $\boldsymbol{v} = \pm\boldsymbol{r}_i$. We keep only the peak of $\boldsymbol{v} = \boldsymbol{r}_i$ with the smooth term $S(\boldsymbol{v}; \boldsymbol{r}_i)$.

To convert a vector to a distribution, we first adopt spherical Fibonacci lattice algorithm [17] to sample M near-equidistant points from an unit sphere, denoted by $\mathcal{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_M\}$ where $\boldsymbol{d}_i \in \mathbb{R}^3$ (see Fig. 3). Note that we only perform the sampling once, thus all pose vectors share a same set of sampled points. During the training stage, for any ground truth vector label $\boldsymbol{r}_i$, the network first predicts parameters λ and η and then use them to calculate the probabilities for all the sampled points $\mathcal{P}_{\boldsymbol{r}_i} = \{p_1^i, p_2^i, \cdots, p_M^i\}$. The probability of point k can be obtained by the normalization:

$$
p_k^i = \frac{\exp\{\mathrm{G}^i(\boldsymbol{v}_k; \boldsymbol{R}, [\lambda, \eta]\}}{\sum_{j=1}^{M} \exp\{\mathrm{G}^i(\boldsymbol{v}_j; \boldsymbol{R}, [\lambda, \eta]\}}. \tag{2}
$$

The process of label distribution generation is applied on all three column vectors $\boldsymbol{r}_0, \boldsymbol{r}_1$ and $\boldsymbol{r}_2$. Therefore, we can obtain three sets of probability distribution $\mathcal{P}_{\boldsymbol{r}_0}, \mathcal{P}_{\boldsymbol{r}_1}$ and $\mathcal{P}_{\boldsymbol{r}_2}$ with the same size of M. The probability distribution on sampled points are visualized in Fig. 4b.

3.4 Working Pipeline

Training. In the training stage, the backbone-encoded features are first fed into three heads separately (See Fig. 3). Each head has one fully connected (FC) layer, which outputs a vector with size of $M+2$. The first M elements denote the ASG probabilities of sampled points for the corresponding pose vector, which is normalized by a softmax layer to generate the probability distribution $\mathcal{Q}_{\boldsymbol{r}_i} = \{q_1^i, q_2^i, \cdots, q_M^i\}$. Therefore the expectation of the distribution is given by:

$$
\hat{\boldsymbol{r}}_i = \mathbb{E}_{\mathcal{Q}_{\boldsymbol{r}_i}}[\mathcal{D}] = \sum_{k=1}^{M} q_k^i \boldsymbol{d}_k. \tag{3}
$$

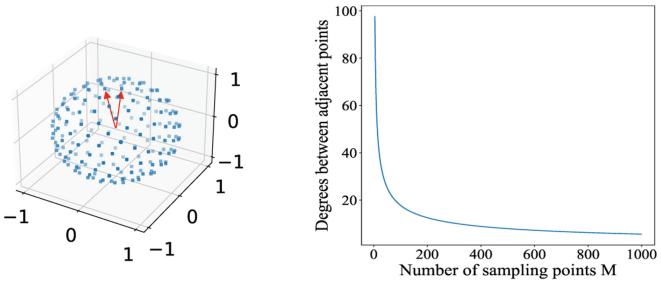
The last two elements of the output vector from the FC layer correspond to the parameters (λ_i, η_i). In conjunction with the sampled point set $\mathcal{D}$ and the ground truth vector $\boldsymbol{r}_i$, the network is able to generate the distribution target $\mathcal{P}_{\boldsymbol{r}_i}$ using Eq. 1 and 2.

Loss Function. To supervise our method, our training loss consists of two terms: classification loss $\mathcal{L}_{cls}$ and regression loss $\mathcal{L}_{reg}$. The overall loss $\mathcal{L}$ is given by:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha L_{reg}. \tag{4}$$

More concretely, we adopt mean square error (MSE) loss function for regression $\mathcal{L}_{reg} = \text{MSE}\ (\boldsymbol{r}_i, \hat{\boldsymbol{r}}_i)$ and Kullback-Liebler (KL) divergence for classification $\mathcal{L}_{cls} = \text{D}_{\text{KL}}(\mathcal{P}_{\boldsymbol{r}_i} || \mathcal{Q}_{\boldsymbol{r}_i})$. The value of the trade-off parameter α is in the range of $[0, 1]$. We find its optimal value in our experiments.

Inference. In the inference stage, we first concatenate the three pose vectors $\hat{\boldsymbol{r}}_0, \hat{\boldsymbol{r}}_1, \hat{\boldsymbol{r}}_2$ generated by the three heads from the learned network to obtain matrix $\hat{\boldsymbol{R}} = [\hat{\boldsymbol{r}}_0, \hat{\boldsymbol{r}}_1, \hat{\boldsymbol{r}}_2]$. We then obtain its closest rotation matrix through singular value decomposition (SVD). Given a matrix $\hat{\boldsymbol{R}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, its closest rotation matrix is obtained by $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{V}^T$.



(a) (b)

Fig. 5. (a) Visualization of the angle between two adjacent points. **(b)** Relationship between the number of sampling points M and the angle between two adjacent points.

4 Experiments

4.1 Datasets and Metrics

We conduct an extensive set of experiments to evaluate our approach on three benchmarks: 300W-LP [51], AFLW2000 [52] and BIWI [10]. **300W-LP** is a synthesized dataset which contains 122,450 images with large varieties in facial poses and identities. Image samples in 300W-LP are synthesized from 300W dataset [36] which includes around 4,000 images. **AFLW2000** contains the first 2,000

images of the popular AFLW [32] dataset with diverse facial poses in the wild. The dataset is commonly used as the test set to evaluate model performances. **BIWI** is collected in an indoor environment with an RGB-D camera. It provides accurate ground truth labels. This dataset is also widely used for depth-based facial pose estimation. Since bounding boxes of human heads are not provided in BIWI, we use MTCNN [48] to detect and crop the face areas.

To ensure a fair comparison with different methods, we follow the same experiment scenarios applied in [20,35,46] and discard the test samples with Euler angles beyond the range of $[-99°, 99°]$. **Scenario 1:** We train our network on 300W-LP and test on both AFLW2000 and BIWI datasets. **Scenario 2:** We perform the 3-fold cross validation on the BIWI dataset. We randomly split the BIWI dataset into 3 groups. Each group contains 8 videos and the videos of the same person appear only in one group. We use mean absolute error of Euler angles (MAE) as our metric.

Table 1. MAE and MAEV results of **different representations** of rotation under scenario 1 and 2. All use the ResNet-18 as backbone. We highlight the **best** results.

Train	Test	Representation	Euler Angle Errors				Vector Errors			
			Pitch	Yaw	Roll	MAE	Left	down	front	MAEV
300W-LP	AFLW 2000	Euler angles	6.36	4.64	4.84	5.28	6.71	5.97	7.62	6.76
		Lie algebra	5.62	3.92	4.04	4.52	5.84	5.13	6.52	5.83
		Quaternion	5.77	4.01	4.20	4.66	5.63	5.62	6.57	5.94
		Rotation matrix	**5.46**	**3.71**	**3.77**	**4.31**	**5.52**	**4.97**	**5.92**	**5.47**
	BIWI (all)	Euler angles	6.43	4.22	4.08	4.91	6.08	5.72	6.13	5.98
		Lie algebra	5.87	**3.39**	3.73	4.33	5.82	5.66	5.42	5.63
		Quaternion	6.11	3.54	**3.61**	4.42	5.79	5.88	5.61	5.76
		Rotation matrix	**5.43**	3.52	3.63	**4.19**	**5.74**	**5.10**	**5.12**	**5.32**
BIWI (70%)	BIWI (30%)	Euler angles	4.07	3.76	3.73	3.85	4.52	4.89	4.57	4.66
		Lie algebra	3.46	3.21	3.11	3.26	4.31	4.22	4.18	4.24
		Quaternion	3.52	3.35	3.24	3.37	4.51	4.32	4.20	4.34
		Rotation matrix	**3.08**	**3.16**	**3.01**	**3.08**	**4.12**	**4.16**	**4.02**	**4.10**

4.2 Implementation Detail

There are two hyper-parameters in our approach. One is the coefficient α for the regression loss term $\mathcal{L}_{reg}$. Another one is the number of sampled points M. Figure 5 shows the relationship between number of sampled points M and angle between adjacent points. We set $\alpha = 0.2$ and $M = 600$ in our experiments.

We implement our proposed approach based on PyTorch and adopt ResNet-18 [19] as the backbone. In training, we adopt Adam optimizer with the initial learning rate of 0.0001. The total training epoch is set to be 50 with the decay rate of 0.95 for every epoch. Batch size is set to be 64 and every image is resized to 224×224. All the experiments are conducted on a RTX 2080 Ti GPU.

We augment training images with random crop, noise and random zoom with scale from 0.8 to 1.2.

Table 2. Comparison with state-of-the-art methods on the AFLW2000 and BIWI datasets. All methods are trained on 300W-LP. We highlight the **best** results and **our** results.

Method	Backbone	AFLW2000				BIWI (full)			
		Pitch	Yaw	Roll	MAE	Pitch	Yaw	Roll	MAE
3DDFA [51]	Two-stream	27.09	4.71	28.43	20.08	41.90	5.50	13.22	20.21
Dlib [22]	-	11.25	8.49	22.83	14.19	13.00	11.86	19.56	14.81
HPE [21]	ResNet-50	6.18	4.87	4.80	5.28	5.18	4.57	3.12	4.29
Hopenet [35]	ResNet-50	7.12	5.31	6.13	6.19	5.89	6.01	3.72	5.20
Quatnet [20]	GoogLeNet	5.62	3.97	3.92	4.50	5.49	4.01	2.94	4.15
Liu *et al.* [28]	ResNet-50	5.06	**3.03**	3.68	3.93	5.61	4.12	3.15	4.29
FSA-Net [46]	SSR-Net	6.34	4.96	4.78	5.36	5.21	4.56	3.07	4.28
TriNet [2]	ResNet-50	5.77	4.20	4.04	4.67	4.75	**3.05**	4.11	3.97
MNN [38]	Encoder-Decoder	**4.69**	3.34	3.48	3.83	4.61	3.98	**2.39**	3.66
img2pose [1]	ResNet-18	5.03	3.43	3.28	3.91	3.55	4.57	3.24	3.79
Ours	ResNet-18	**4.74**	**3.08**	**3.11**	**3.64**	**3.52**	**4.21**	**3.10**	**3.61**

4.3 Analysis of Rotation Representations

Even though there are multiple ways to describe a rotation and the most commonly used ones include Euler angles, quaternion, Lie algebra and rotation matrices, it remains under-studied that which representation is the best option for the task of facial pose estimation. [2] briefly discussed Euler Angle and quaternions. However, they omitted Lie algebra and did not provide any experimental support. We implement a thorough comparison between the performances of different representations using the same backbone of ResNet-18 (see Table 1). Since MAE is not an accurate measure for profile faces, we also adopt mean absolute error of vectors (MAEV) to make a comprehensive comparison. Experiments show that rotation matrices achieve the best result among all representations under both scenarios.

The experimental results accord with the continuity properties of each representation. As shown by the work [40,50], it needs at least 5 dimensions to describe the rotation continuously, otherwise it incurs discontinuity issue similar to Euler angles. Both Euler angle and Lie algebra $\in \mathbb{R}^3$ and quaternion $\in \mathbb{R}^4$. Therefore, none of them can describe the rotation continuously. Here we include some cases when the phenomenons of discontinuity occur. For a unit quaternion $\boldsymbol{q} = w + x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k}$, where $w^2 + x^2 + y^2 + z^2 = 1$. Then $(1, 0, 0, 0)$ and $(-1, 0, 0, 0)$ represents the same rotation. For Lie algebra $\mathfrak{so}(3)$ which is denoted by an anti-symmetric matrix $\boldsymbol{\phi}^\wedge$ where $\boldsymbol{\phi} = \theta\boldsymbol{a}$ and $\boldsymbol{a} \in \mathbb{R}^3$, $||\boldsymbol{a}||_2 = 1, \theta \in [-\pi, +\pi]$. For any $\boldsymbol{a}$, faces have similar appearances when θ approaches π and $-\pi$. Therefore, the rotation matrix is the best representation in terms of the performance and continuity property.

Table 3. Comparison between **direct regression** and **distribution learning**. Results are obtained on the AFLW2000 and BIWI benchmarks.

Training set	300W-LP		BIWI (70%)
Testing set	ALFW2000	BIWI (full)	BIWI (30%)
Direct regression	4.31	4.19	3.08
SG learning	3.79	3.71	2.93
ASG learning	**3.64**	**3.61**	**2.77**

4.4 Comparison with State-of-the-Arts

We compare the performance of our method with other state-of-the-art methods (see Table 2) under scenario 1. Since the training/test set division in scenario 2 is arbitrary and thus is not adopted by methods [1,38], we choose only scenario 1 for comparison. The results of the compared methods are directly cited from their original papers. Liu *et al*. [28] is the first work that follows the distribution learning paradigm for wild pose estimation. Different from our work, they convert the Euler angles to 3 Gaussian distributions with the same variance. Even though they use the ResNet-50 as the backbone which is deeper than our ResNet-18, our ASG-based distribution learning surpasses their performance. FSA-Net [46] and TriNet [2] take advantage of the combination of attention module and capsule network and append them to the backbone network to improve the learning ability of the network. Even though both have more complex structures and more parameters than our network, their performance is inferior to ours.

It is worth mentioning that some of the methods such as MNN [38] and img2pose [1] also use face landmarks in a weakly supervised manner to help improve the performance of network. To highlight the effectiveness of our distri-

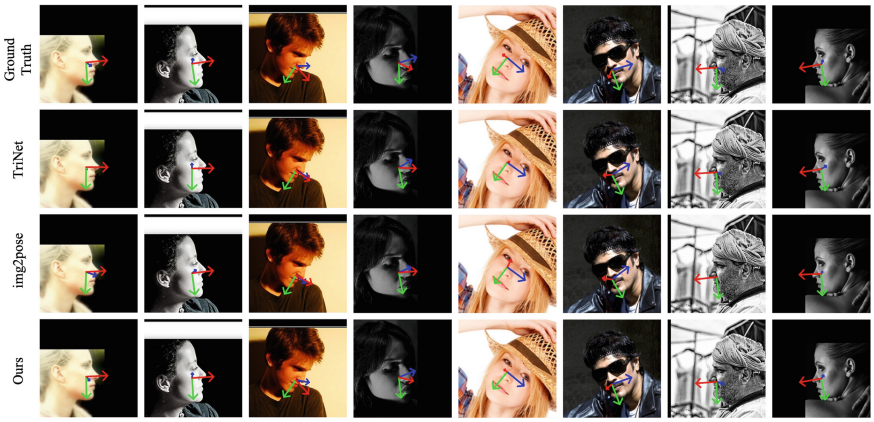


Fig. 6. Qualitative comparison of different methods. Trained on 300W-LP and tested on AFLW2000.

bution learning strategy, we make the network learn the pose estimation without the landmark labels. Experiments show that even though less information is provided, our network still achieves better performance. The qualitative results are demonstrated in Fig. 6. It can be seen that our approach makes more accurate predictions when faces are partially occluded.

Table 4. Comparison between adaptive **parameters learning** and **fixed** ASG parameters. Results are obtained on the AFLW2000 and BIWI benchmarks.

Training set	300W-LP		BIWI (70%)
Testing set	AFLW2000	BIWI (full)	BIWI (30%)
$\lambda = 1$, $\eta = 1$	3.79	3.84	2.92
$\lambda = 5$, $\eta = 5$	3.86	3.97	3.03
$\lambda = 1$, $\eta = 5$	3.92	4.03	2.97
Adaptive parameters	**3.64**	**3.61**	**2.77**

Table 5. Comparison of the effects of **different loss terms**. Results are obtained on the AFLW2000 and BIWI benchmarks.

Training set	300W-LP		BIWI (70%)
Testing set	ALFW2000	BIWI (full)	BIWI (30%)
$\mathcal{L}_{cls}$	3.67	3.68	2.81
$\mathcal{L}_{reg}$	4.31	4.19	3.08
$\mathcal{L}_{cls} + \mathcal{L}_{reg}$	**3.64**	**3.61**	**2.77**

4.5 Ablation Study

In this section, we investigate the effectiveness of our method by carrying out ablation experiments on the adaptive ASG label distribution learning and different loss components.

Distribution Learning *vs.* Regression. We examine the advantages of the ASG distribution to isotropic SG distribution and use the direct regression of the rotation matrix as baseline. Results are shown in Table 3. Our ASG distribution can effectively improve the performance compared with other two baseline methods.

Adaptive Parameters *vs.* Fixed Parameters. We conduct experiments to compare the performance of methods with adaptive parameters and fixed parameters (see Table 4). While achieving superior performance over the fixed parameters, our adaptive parameter learning is computationally efficient as it avoids the exhaustive search of parameters. All the ASG parameters λ and η learned by the samples in the 300W-LP dataset are demonstrated in Fig. 7. We

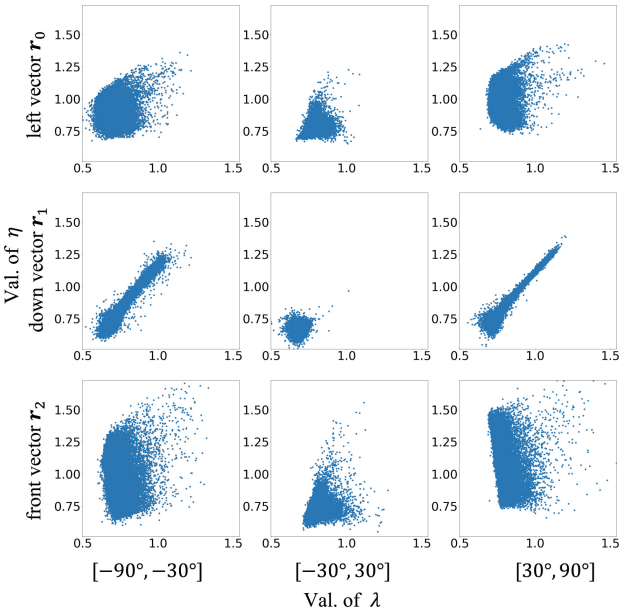


Fig. 7. Visualization of η and λ distribution for three pose vectors $\boldsymbol{r}_0, \boldsymbol{r}_1$ and $\boldsymbol{r}_2$ of all the samples in different ranges of Yaw. All trained on the 300W-LP dataset.

divide the angle of yaw into three equal ranges. It is worth noting that our learning behavior follows a clear pattern. For instance, the parameter distributions in the first and third columns resemble each other. Because turning faces to left and right results in symmetric images, the parameters of ASG should be similar. This is reflected by the distribution of parameters.

Loss Functions. We examine the effectiveness of each loss term (see Table 5). Notice when only $\mathcal{L}_{reg}$ is applied, the network is supervised by the arbitrary distribution with expectation of the same as the ground truth. The above results confirm that the classification term and regression term can work collaboratively to operate effective label distribution learning for the facial pose estimation.

5 Conclusion

In this paper, we introduce a novel ASG-based Label Distribution Learning method for estimating facial pose. This is the first attempt to include directional statistics in the estimation of pose. We anticipate that this work will illustrate potential future directions for the community to investigate.

References

1. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: face alignment and detection via 6DoF, face pose estimation. In: CVPR (2021)

2. Cao, Z., Chu, Z., Liu, D., Chen, Y.: A vector-based representation to enhance head pose estimation. In: WACV (2021)
3. Chang, F.J., Tuan Tran, A., Hassner, T., Masi, I., Nevatia, R., Medioni, G.: FacePoseNet: making a case for landmark-free face alignment. In: ICCV Workshops (2017)
4. Chen, Z., Liu, Z., Hu, H., Bai, J., Lian, S., Shi, F., Wang, K.: A realistic face-to-face conversation system based on deep neural networks. In: ICCV (2019)
5. Cheng, Z., et al.: Physical attack on monocular depth estimation with optimal adversarial patches. In: ECCV (2022)
6. Cui, Y., Yan, L., Cao, Z., Liu, D.: TF-blender: temporal feature blender for video object detection. In: ICCV (2021)
7. De Rousiers, C., Bousseau, A., Subr, K., Holzschuch, N., Ramamoorthi, R.: Real-time rough refraction. In: Symposium on Interactive 3D Graphics and Games, pp. 111–118 (2011)
8. Diaz, R., Marathe, A.: Soft labels for ordinal regression. In: CVPR (2019)
9. Fan, Y.Y., et al.: Label distribution-based facial attractiveness computation by deep residual learning. IEEE Trans. Multimedia **20**(8), 2196–2208 (2017)
10. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3D face analysis. Int. J. Comput. Vis. **101**(3), 437–458 (2013)
11. Fisher, R.A.: Dispersion on a sphere. Proc. R. Soc. London Ser. A Math. Phys. Sci. **217**(1130), 295–305 (1953)
12. Gao, G., Lauri, M., Zhang, J., Frintrop, S.: Occlusion resistant object rotation regression from point cloud segments. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11129, pp. 716–729. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11009-3_44
13. Geng, X., Hou, P.: Pre-release prediction of crowd opinion on movies by label distribution learning. In: IJCAI (2015)
14. Geng, X., Xia, Y.: Head pose estimation based on multivariate label distribution. In: CVPR (2014)
15. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE Trans. Pattern Anal. Mach. Intell. **35**(10), 2401–2412 (2013). https://doi.org/10.1109/TPAMI.2013.51
16. Geronimo, D., Lopez, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1239–1258 (2009)
17. González, Á.: Measurement of areas on a sphere using Fibonacci and latitude-longitude lattices. Math. Geosci. **42**(1), 49–64 (2010)
18. Hara, K., Nishino, K., Ikeuchi, K.: Multiple light sources and reflectance property estimation based on a mixture of spherical distributions. In: ICCV (2005)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
20. Hsu, H.W., Wu, T.Y., Wan, S., Wong, W.H., Lee, C.Y.: QuatNet: quaternion-based head pose estimation with multiregression loss. IEEE Trans. Multimedia **21**(4), 1035–1046 (2018)
21. Huang, B., Chen, R., Xu, W., Zhou, Q.: Improving head pose estimation using two-stage ensembles with top-k regression. Image Vis. Comput. **93**, 103827 (2020)
22. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: CVPR (2014)
23. Kent, J.T.: The Fisher-Bingham distribution on the sphere. J. R. Stat. Soc. Ser. B (Methodol.) **44**(1), 71–80 (1982)

24. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: an accurate O(n) solution to the PnP problem. Int. J. Comput. Vis. **81**(2), 155 (2009)
25. Liu, D., Cui, Y., Tan, W., Chen, Y.: SG-Net: spatial granularity network for one-stage video instance segmentation. In: CVPR (2021)
26. Liu, et al..: DenserNet: weakly supervised visual localization using multi-scale feature aggregation. In: AAAI (2021)
27. Liu, X., et al.: AgeNet: deeply learned regressor and classifier for robust apparent age estimation. In: ICCVW (2015)
28. Liu, Z., Chen, Z., Bai, J., Li, S., Lian, S.: Facial pose estimation by deep learning from label distributions. In: CVPR Workshops (2019)
29. Liu, Z., Hu, H., Wang, Z., Wang, K., Bai, J., Lian, S.: Video synthesis of human upper body with realistic face. In: 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp. 200–202. IEEE (2019)
30. Liu, Z., et al.: Unveiling the power of mixup for stronger classifiers. arXiv preprint arXiv:2103.13027 (2021)
31. Mahendran, S., Ali, H., Vidal, R.: 3D pose regression using convolutional neural networks. In: ICCV Workshops (2017)
32. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: Proceedings of the First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (2011)
33. Murphy-Chutorian, E., Doshi, A., Trivedi, M.M.: Head pose estimation for driver assistance systems: a robust algorithm and experimental evaluation. In: 2007 IEEE Intelligent Transportation Systems Conference, pp. 709–714. IEEE (2007)
34. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: PVNet: pixel-wise voting network for 6DoF pose estimation. In: CVPR (2019)
35. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: CVPR Workshops (2018)
36. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: ICCV Workshops (2013)
37. Song, C., Song, J., Huang, Q.: HybridPose: 6D object pose estimation under hybrid representations. In: CVPR (2020)
38. Valle, R., Buenaposada, J.M., Baumela, L.: Multi-task head pose estimation in-the-wild. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 2874–2881 (2020)
39. Wang, J., Ren, P., Gong, M., Snyder, J., Guo, B.: All-frequency rendering of dynamic, spatially-varying reflectance. In: ACM SIGGRAPH Asia 2009 papers, pp. 1–10 (2009)
40. Xiang, S.: Eliminating topological errors in neural network rotation estimation using self-selecting ensembles. ACM Trans. Graph. (TOG) **40**(4), 1–21 (2021)
41. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017)
42. Xu, K., Sun, W.L., Dong, Z., Zhao, D.Y., Wu, R.D., Hu, S.M.: Anisotropic spherical gaussians. ACM Trans. Graph. (TOG) **32**(6), 1–11 (2013)
43. Yan, L., et al.: GL-RG: global-local representation granularity for video captioning. In: IJCAI (2022)
44. Yang, H., Mou, W., Zhang, Y., Patras, I., Gunes, H., Robinson, P.: Face alignment assisted by head pose estimation. arXiv preprint arXiv:1507.03148 (2015)
45. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: a face detection benchmark. In: CVPR (2016)

46. Yang, T.Y., Chen, Y.T., Lin, Y.Y., Chuang, Y.Y.: FSA-net: learning fine-grained structure aggregation for head pose estimation from a single image. In: CVPR (2019)
47. Zakharov, S., Shugurov, I., Ilic, S.: DPOD: 6D pose object detector and refiner. In: ICCV (2019)
48. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
49. Zhang, Z., Wang, M., Geng, X.: Crowd counting in public video surveillance by label distribution learning. Neurocomputing **166**, 151–163 (2015)
50. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019)
51. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: CVPR (2016)
52. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: CVPR (2015)