Spherical Space Domain Adaptation with Robust Pseudo-label Loss
Supplementary Material

Xiang Gu, Jian Sun (✉) and Zongben Xu
Xi'an Jiaotong University, Xi'an, 710049, China
xianggu@stu.xjtu.edu.cn, {jiansun, zbxu}@xjtu.edu.cn

1. Appendix A : Additional Details

1.1. Center of Samples on Sphere

This section computes the center of spherical samples shown in Sect. 3.2 and Sect. 4.1 of paper.

Let $f_1, f_2, \cdots, f_m$ be samples on sphere $\mathbb{S}_r^{n-1} = \{f \in \mathbb{R}^n : ||f|| = r\}$, the center $\mathcal{C}$ of the samples on sphere is the point closest to all samples, i.e., the solution of the following optimizing problem,

$$\min_{f \in \mathbb{S}_r^{n-1}} \frac{1}{m} \sum_{i=1}^{m} \text{dist}(f, f_i), \tag{1}$$

where $\text{dist}(u, v) = 1 - \frac{u^T v}{||u||||v||}$ is the cosine distance. Since $||f|| = r, \forall f \in \mathbb{S}_r^{n-1}$, problem in Eq. (1) can be written as

$$\begin{aligned} &\max_f f^T \left(\sum_{i=1}^{m} f_i \right) \\ &s.t. \quad ||f|| = r. \end{aligned} \tag{2}$$

With the method of Lagrange multipliers, the center can be obtained as

$$\mathcal{C} = \frac{r}{||\tilde{f}||} \tilde{f}, \tag{3}$$

where $\tilde{f} = \sum_{i=1}^{m} f_i$.

1.2. Spherical Linear Transform

This section describes details of spherical linear transform shown in Sect. 5 of paper.

Spherical exponential and logarithmic maps. The exponential and logarithmic maps connect the tangent space and the sphere [6]. Let $N = (0, 0, \cdots, r) \in \mathbb{R}^n$ be the north pole of sphere $\mathbb{S}_r^{n-1} = \{x \in \mathbb{R}^n : ||x|| = r\}$, then the tangent space $T_N \mathbb{S}_r^{n-1}$ at N becomes the hyperplane $e_n^T z - r = 0, \forall z \in \mathbb{R}^n$, where $e_n = (0, \cdots, 0, 1) \in \mathbb{R}^n$. Thus, any vector $\tilde{v}$ in $T_N \mathbb{S}_r^{n-1}$ can be expressed as $\tilde{v} = (v, r)$, where $v \in \mathbb{R}^{n-1}$. The exponential map $\exp_N : T_N \mathbb{S}_r^{n-1} \to \mathbb{S}_r^{n-1}$ is given by

$$\exp_N(\tilde{v}) = N \cos\theta + \tilde{v} \frac{\sin\theta}{\theta}, \tag{4}$$

$\forall \tilde{v} = (v, r) \in T_N \mathbb{S}_r^{n-1}$, where $\theta = \frac{||\tilde{v}||}{r}$. The logarithmic map $\log_N : \mathbb{S}_r^{n-1} \to T_N \mathbb{S}_r^{n-1}$ is given by

$$\log_N(x) = \frac{\varphi}{\sin\varphi}(x - N \cos\varphi), \tag{5}$$

$\forall x \in \mathbb{S}_r^{n-1}$,where $\varphi = \arccos(N^T x / r^2)$.

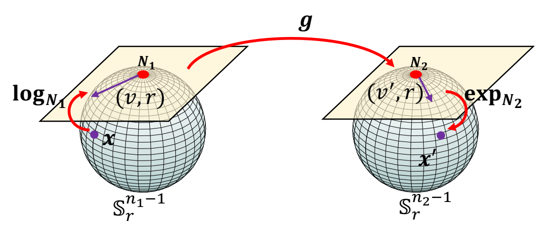


Figure 1. Spherical linear transform.

Definition of spherical linear transform. As illustrated in Fig. 1, to define the spherical linear transform, we project spherical features on $\mathbb{S}_r^{n_1-1}$ to the tangent space $T_{N_1} \mathbb{S}_r^{n_1-1}$ by logarithmic map $\log_{N_1}$, then transfer the projected features into $T_{N_2} \mathbb{S}_r^{n_2-1}$ by linear transform g, finally project the transferred features to the sphere $\mathbb{S}_r^{n_2-1}$ by exponential map $\exp_{N_2}$, where $N_i = (0, \cdots, 0, r) \in \mathbb{R}^{n_i}$ is the north pole of $\mathbb{S}_r^{n_i-1}$, for $i = 1, 2$. Since any vector $\tilde{v}$ in $T_{N_i} \mathbb{S}_r^{n_i-1}$ can be expressed as $\tilde{v} = (v, r), v \in \mathbb{R}^{n_i-1}$, the linear transform $g : T_{N_1} \mathbb{S}_r^{n_1-1} \to T_{N_2} \mathbb{S}_r^{n_2-1}$ can be expressed as

$$g(\tilde{v}) = g((v, r)) = (Wv + b, r) \tag{6}$$

$\forall \tilde{v} = (v, r) \in T_{N_1} \mathbb{S}_r^{n_1-1}$, where $W \in \mathbb{R}^{(n_2-1)*(n_1-1)}$ and $b \in \mathbb{R}^{n_2-1}$ are parameters. Therefore, the spherical linear transform from $\mathbb{S}_r^{n_1-1}$ to $\mathbb{S}_r^{n_2-1}$ can be defined by

$$g_s(x) = \exp_{N_2}(g(\log_{N_1}(x))), \quad \forall x \in \mathbb{S}_r^{n_1-1}. \tag{7}$$

2. Appendix B : Proofs

2.1. Bound of Spherical Radius

This section proves the bound of spherical radius shown in Sect. 5 of paper. Suppose the learned class center of spherical features of the last perceptron layer on $\mathbb{S}_r^{n-1}$ is in direction of the corresponding weight vector of spherical logistic regression, *i.e.*, the class center $\bar{x}_i = rw_i$, where $||w_i|| = 1, i = 1, \cdots, K$. Suppose the number of classes $K > 1$. Let P_w denote expected minimum classification probability of class center. Then the lower bound of r is formulated as

$$r \geq \frac{K-1}{K} \ln \frac{(K-1)P_w}{1-P_w}. \tag{8}$$

Proof:

This proof is inspired by [5]. $\forall i$, we have

$$P(y=i|\bar{x}_i) = \frac{e^{w_i^T \bar{x}_i + b_i}}{e^{w_i^T \bar{x}_i + b_i} + \sum_{j,j\neq i} e^{w_j^T \bar{x}_i + b_j}} \geq P_w, \tag{9}$$

$$\frac{e^{r+b_i}}{e^{r+b_i} + \sum_{j,j\neq i} e^{r w_j^T w_i + b_j}} \geq P_w, \tag{10}$$

$$1 + e^{-r} \sum_{j,j\neq i} e^{r w_j^T w_i + b_j - b_i} \leq \frac{1}{P_w}, \tag{11}$$

$$\sum_{i=1}^{K} \left(1 + e^{-r} \sum_{j,j\neq i} e^{r w_j^T w_i + b_j - b_i} \right) \leq \frac{K}{P_w}, \tag{12}$$

$$1 + \frac{e^{-r}}{K} \sum_{i,j,j\neq i} e^{r(w_j^T w_i + (b_j - b_i)/r)} \leq \frac{1}{P_w}. \tag{13}$$

Since $f(x) = e^{rx}$ is a convex function, according to Jensen's inequality, we have

$$\frac{1}{K(K-1)} \sum_{i,j,j\neq i} e^{r(w_j^T w_i + \frac{b_j - b_i}{r})} \geq e^{\frac{r \sum_{i,j,j\neq i} (w_j^T w_i + \frac{b_j - b_i}{r})}{K(K-1)}}. \tag{14}$$

Since

$$\begin{aligned} \sum_{i,j,j\neq i} w_j^T w_i &= \left(\sum_i w_i \right)^T \left(\sum_i w_i \right) - \sum_i (w_i^T w_i) \\ &\geq -K, \end{aligned} \tag{15}$$

$$\begin{aligned} \sum_{i,j,j\neq i} (b_j - b_i) &= \sum_{i,j} (b_j - b_i) - \sum_i (b_i - b_i) \\ &= \sum_i b_i - \sum_j b_j = 0, \end{aligned} \tag{16}$$

we have

$$\sum_{i,j,j\neq i} e^{r(w_j^T w_i + \frac{b_j - b_i}{r})} \geq K(K-1) e^{-\frac{r}{K-1}}. \tag{17}$$

Combining Eqs. (13) and (17), we have

$$1 + (K-1) e^{-\frac{rK}{K-1}} \leq \frac{1}{P_w}. \tag{18}$$

Thus, we can obtain the bound

$$r \geq \frac{K-1}{K} \ln \frac{(K-1)P_w}{1-P_w}. \tag{19}$$

2.2. Deduction of EM for Estimating ϕ

This section deduces EM algorithm for estimating ϕ shown in Sect. 6 of paper. We need to estimate parameters $\phi = \{\pi_k, \sigma_k, \delta_k\}_{k=1}^K$ of the following mixed model

$$p(d_j^t|\tilde{y}_j^t) = \pi_{\tilde{y}_j^t} \mathcal{N}^+(d_j^t|0, \sigma_{\tilde{y}_j^t}) + (1-\pi_{\tilde{y}_j^t}) \mathcal{U}(0, \delta_{\tilde{y}_j^t}), \tag{20}$$

where

$$\mathcal{N}^+(x|0,\sigma) = \begin{cases} 2\mathcal{N}(x|0,\sigma), & \text{if } x \geq 0. \\ 0, & \text{if } x < 0. \end{cases} \tag{21}$$

Let $\tilde{d}_j^t = (-1)^{m_j} d_j^t$, where m_j is sampled from Bernoulli distribution $B(1, 0.5)$, then $\tilde{d}_j^t$ follows the following mixed model

$$p(\tilde{d}_j^t|\tilde{y}_j^t) = \pi_{\tilde{y}_j^t} \mathcal{N}(\tilde{d}_j^t|0, \sigma_{\tilde{y}_j^t}) + (1-\pi_{\tilde{y}_j^t}) \mathcal{U}(\delta_{\tilde{y}_j^t}, \delta_{\tilde{y}_j^t}). \tag{22}$$

The proof is given later. Eq. (22) is exactly the model in [3, 4], of which the corresponding maximum likelihood model becomes

$$\max_{\sigma_k, \delta_k, \pi_k} \prod_{j=1}^{N_t} p(\tilde{d}_j^t|\tilde{y}_j^t). \tag{23}$$

Solving problem Eq. (23) with EM algorithm, as in [3], we have the following updating equations

$$\begin{aligned} \gamma_j^{(l+1)} &= \frac{\pi_{\tilde{y}_j^t}^{(l)} \mathcal{N}(\tilde{d}_j^t|0, \sigma_{\tilde{y}_j^t}^{(l)})}{\pi_{\tilde{y}_j^t}^{(l)} \mathcal{N}(\tilde{d}_j^t|0, \sigma_{\tilde{y}_j^t}^{(l)}) + (1-\pi_{\tilde{y}_j^t}^{(l)}) \mathcal{U}(-\delta_{\tilde{y}_j^t}^{(l)}, \delta_{\tilde{y}_j^t}^{(l)})}, \\ \pi_k^{(l+1)} &= \frac{1}{\sum_{j=1}^{N_t} I_{\{\tilde{y}_j^t = k\}}} \sum_{j=1}^{N_t} I_{\{\tilde{y}_j^t = k\}} \gamma_j^{(l+1)} \\ \sigma_k^{(l+1)} &= \frac{\sum_{j=1}^{N_t} I_{\{\tilde{y}_j^t = k\}} \gamma_j^{(l+1)} (\tilde{d}_j^t)^2}{\sum_{j=1}^{N_t} I_{\{\tilde{y}_j^t = k\}} \gamma_j^{(l+1)}}, \\ \delta_k^{(l+1)} &= \sqrt{3(q_2 - q_1^2)}, \end{aligned} \tag{24}$$

where

$$\begin{aligned} q_1 &= \frac{1}{\sum_{j=1}^{N_t} I_{\{\tilde{y}_j^t = k\}} \gamma_j^{(l+1)}} \sum_{j=1}^{N_t} \frac{1-\gamma_j^{(l+1)}}{1-\pi_k^{(l+1)}} I_{\{\tilde{y}_j^t = k\}} \tilde{d}_j^t, \\ q_2 &= \frac{1}{\sum_{j=1}^{N_t} I_{\{\tilde{y}_j^t = k\}} \gamma_j^{(l+1)}} \sum_{j=1}^{N_t} \frac{1-\gamma_j^{(l+1)}}{1-\pi_k^{(l+1)}} I_{\{\tilde{y}_j^t = k\}} (\tilde{d}_j^t)^2. \end{aligned}$$

Proof of Eq. (22) **:** To prove Eq. (22), we prove the following proposition.

Suppose random variable x follows

$$p(x) = \pi\mathcal{N}^{+}(x|0,\sigma) + (1-\pi)\mathcal{U}(0,\delta), \tag{25}$$

where

$$\mathcal{N}^{+}(x|0,\sigma) = \begin{cases} 2\mathcal{N}(x|0,\sigma), & \text{if } x \geq 0. \\ 0, & \text{if } x < 0. \end{cases} \tag{26}$$

Let $\tilde{x} = (-1)^m x, m \sim B(1, 0.5)$, then $\tilde{x}$ follows

$$\tilde{p}(\tilde{x}) = \pi\mathcal{N}(\tilde{x}|0,\sigma) + (1-\pi)\mathcal{U}(-\delta,\delta). \tag{27}$$

Proof:

The probability

$$\begin{aligned} & P(\tilde{x} < s) \\ =& P((-1)^m x < s) \\ =& P((-1)^m x < s|m=0)P(m=0) \\ & + P((-1)^m x < s|m=1)P(m=1) \\ =& 0.5P(x < s) + 0.5P(x > -s). \end{aligned}$$

If $s \geq 0$, then

$$\begin{aligned} & P(\tilde{x} < s) = 0.5P(x < s) + 0.5 \\ =& 0.5\int_{-\infty}^{s} p(x)dx + 0.5 \\ =& 0.5\int_{0}^{s} \left(\pi\mathcal{N}^{+}(x|0,\sigma) + (1-\pi)\mathcal{U}(0,\delta)\right)dx + 0.5 \\ =& \pi\left(0.5 + \int_{0}^{s}\mathcal{N}(x|0,\sigma)dx\right) \\ & + (1-\pi)\left(0.5 + 0.5\int_{0}^{s}\frac{1}{\delta}dx\right) \\ =& \pi\int_{-\infty}^{s}\mathcal{N}(x|0,\sigma)dx + (1-\pi)\int_{-\infty}^{s}\mathcal{U}(-\delta,\delta)dx \\ =& \int_{-\infty}^{s}\left(\pi\mathcal{N}(\tilde{x}|0,\sigma) + (1-\pi)\mathcal{U}(-\delta,\delta)\right)d\tilde{x}. \end{aligned}$$

If $s < 0$, similarly, we have the same equation. Thus, the density of $\tilde{x}$ is

$$\tilde{p}(\tilde{x}) = \pi\mathcal{N}(\tilde{x}|0,\sigma) + (1-\pi)\mathcal{U}(-\delta,\delta). \tag{28}$$

The proof is completed.

2.3. Proof of Lemma 1

This section proves Lemma 1 shown in Sect. 7 of paper.

Lemma 1. *Let $h \in \mathcal{H}$ be a hypothesis, f_S and f_T be the true labeling function for source and target respectively, f'_T be the pseudo-labeling function for target domain, then*

$$\begin{aligned} \varepsilon_T(h) \leq & \frac{1}{2}(\varepsilon_S(h) + \varepsilon_T(h, f'_T) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(P_S, P_T)) \\ & + \varepsilon_T(f'_T, f_T) + \frac{1}{2}\beta, \end{aligned} \tag{29}$$

where $\varepsilon_T(h, h') = \mathbb{E}_{x\sim P_T}[h(x) \neq h'(x)]$, $\beta = \min_{h'\in\mathcal{H}}\{\varepsilon_S(h') + \varepsilon_T(h', f'_T)\}$ *is a constant to* h*.*

Proof:

Recall the triangle inequality for classification error [2], which implies that for any hypothesis f_1, f_2andf_3, we have $\varepsilon(f_1, f_2) \leq \varepsilon(f_1, f_3) + \varepsilon(f_2, f_3)$. Then

$$\varepsilon_T(h) = \varepsilon_T(h, f_T) \leq \varepsilon_T(h, f'_T) + \varepsilon_T(f'_T, f_T). \tag{30}$$

According to Theorem 2 in [1], we have

$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(P_S, P_T) + \lambda^*, \tag{31}$$

where $\lambda^* = \min_{h'\in\mathcal{H}}\{\varepsilon_S(h') + \varepsilon_T(h')\}$. Recall Eq. (30), we have

$$\begin{aligned} \varepsilon_T(h) \leq & \varepsilon_S(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(P_S, P_T) \\ & + \min_{h'\in\mathcal{H}}\{\varepsilon_S(h') + \varepsilon_T(h', f'_T)\} + \varepsilon_T(f'_T, f_T). \end{aligned} \tag{32}$$

Combining Eq. (30) and Eq. (32), we have

$$\begin{aligned} \varepsilon_T(h) \leq & \frac{1}{2}(\varepsilon_S(h) + \varepsilon_T(h, f'_T) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(P_S, P_T)) \\ & + \varepsilon_T(f'_T, f_T) + \frac{1}{2}\beta, \end{aligned} \tag{33}$$

where $\beta = \min_{h'\in\mathcal{H}}\{\varepsilon_S(h') + \varepsilon_T(h', f'_T)\}$.

3. Appendix C : Experiments

3.1. Full Results of Ablation Experiments

This section reports full results of ablation study shown in Sect. 8.2 of paper. The full results of ablation experiments on Office-31 and ImageCLEF-DA are given in Table 1 of this document. The full results of ablation experiments on Office-Home are given in Table 2 of this document.

3.2. Stability of Losses

This section testifies the stability of our losses. Considering the objective function Eq.(1) in paper that we want to minimize, we design an iterative optimization algorithm by alternately optimizing networks and estimating parameters of Gaussian-uniform mixture model using EM algorithm.

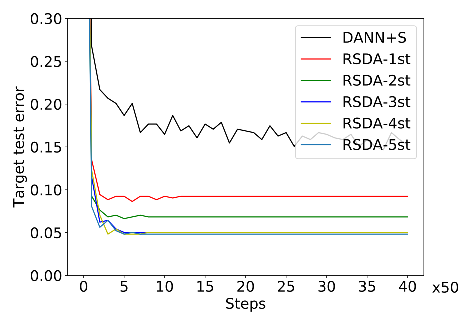
Table 1. Accuracy(%) of ablation experiments on Office-31 and ImageCLEF-DA.

Method	Office-31							ImageCLEF-DA						
	A→W	W→A	A→D	D→A	D→W	W→D	Avg	I→P	P→I	I→C	C→I	C→P	P→C	Avg
DANN	82.0	67.4	79.7	68.2	96.9	99.1	82.2	75.0	86.0	96.2	87.0	74.3	91.5	85.0
DANN+S	93.2	71.0	87.5	70.3	98.0	100.0	86.7	78.3	91.0	96.8	91.8	77.7	95.2	88.5
DANN+R	93.7	74.0	91.8	74.1	98.6	100.0	88.7	78.7	92.6	96.7	93.3	78.4	95.0	89.1
DANN+S+R	94.2	75.4	92.5	73.7	99.1	100.0	89.2	78.5	91.8	97.8	93.5	**78.7**	96.3	89.4
DANN+S+E	91.7	68.7	92.2	73.0	98.5	100.0	87.4	78.5	**93.7**	96.5	91.7	76.8	94.8	88.7
DANN+R+E	**95.5**	74.8	**95.2**	74.5	98.6	100.0	89.8	78.8	93.0	97.2	**93.7**	78.5	96.3	89.4
DANN+S+R+E (RSDA)	95.3	**76.0**	**95.2**	**75.5**	**99.3**	**100.0**	**90.2**	**79.2**	93.0	**98.3**	93.6	78.5	**98.2**	**90.1**

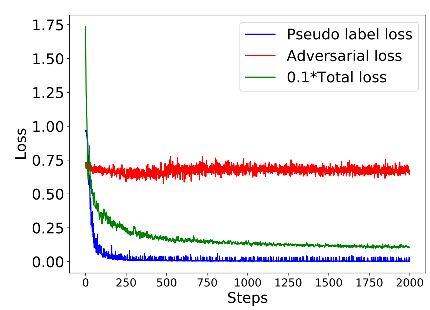
Table 2. Accuracy(%) of ablation experiments on Office-Home.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
DANN	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
DANN+S	45.5	61.9	72.2	54.6	59.2	62.8	52.0	43.9	71.8	66.3	51.5	76.5	59.8
DANN+R	50.5	75.1	79.2	62.2	72.1	73.8	61.6	47.3	79.8	70.2	54.6	81.1	67.3
DANN+S+R	49.5	74.0	79.2	64.3	72.3	75.2	63.5	**51.5**	80.2	72.5	55.2	83.1	68.4
DANN+S+E	48.5	73.3	78.4	65.2	72.4	71.5	**66.5**	49.8	79.8	**75.2**	53.5	82.5	68.0
DANN+R+E	51.1	75.1	79.7	65.4	**74.2**	75.7	63.1	50.3	80.5	71.7	55.7	83.1	68.8
DANN+S+R+E (RSDA)	**51.5**	**76.8**	**81.1**	**67.1**	72.1	**77.0**	64.2	51.1	**81.8**	74.9	**55.9**	**84.5**	**69.8**

As an iterative optimization algorithm, our training method can stably decreases the loss and converges in all our training experiments. As an example, we show target test errors of the first five iterations and loss functions of the first iteration (since curves of losses in each iteration are similar) on task **A**→**D** in Fig. 2. Figure 2(a) shows that the pseudo-label loss can gradually calibrate model. All training losses decrease stably in network optimization, as shown in Fig. 2(b).



(a) Target test errors



(b) Training losses

Figure 2. (a) Target test errors during alternative optimization. (b) Training errors during network optimization.

3.3. Effectiveness of Gaussian-uniform Model

This section evaluates effectiveness of Gaussian-uniform model on real data, which is complementary to Sect. 4 and Sect. 8.2 of paper. To further verify effectiveness of our Gaussian-uniform model on real data, we show in Fig. 3 the estimated Gaussian density of target feature distances of several classes in task **W**→**A** on Office-31 dataset. Figure 3 illustrates that distances of wrong labeled samples (red circles) have low Gaussian density, thus can be detected.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *ML*, 79(1-2):151–175, 2010. 3

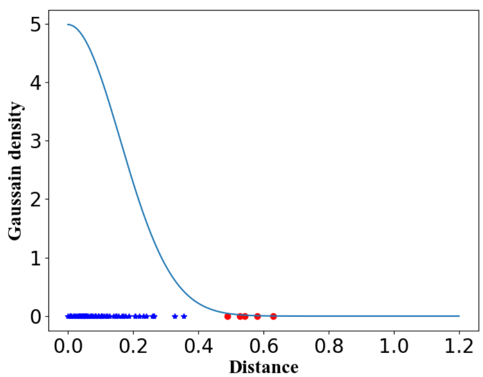
[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, 2007. 3

[3] Pietro Coretto and Christian Hennig. Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust gaussian clustering. *JASA*, 111(516):1648–1659, 2016. 2

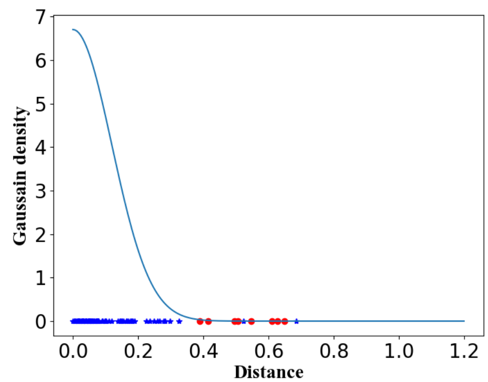
[4] Stéphane Lathuilière, Pablo Mesejo, Xavier Alameda-Pineda, and Radu Horaud. Deepgum: Learning deep robust regression with a gaussian-uniform mixture model. In *ECCV*, 2018. 2

[5] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 2

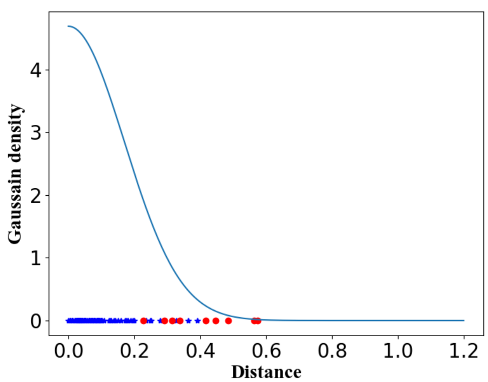
[6] Richard C Wilson, Edwin R Hancock, Elżbieta Pekalska, and Robert PW Duin. Spherical and hyperbolic embeddings of data. *IEEE TPAMI*, 36(11):2255–2269, 2014. 1



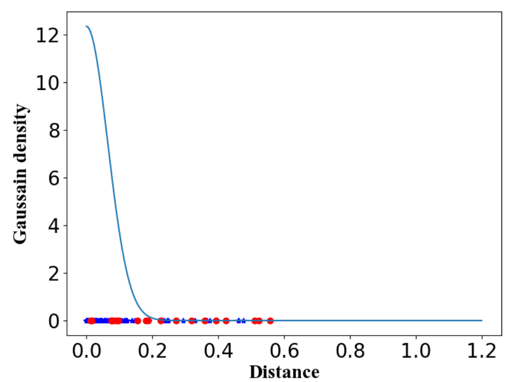
(a) bike helmet



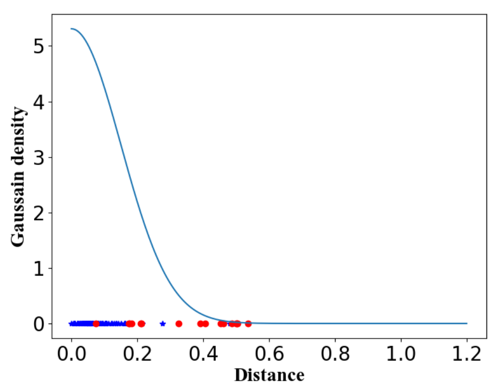
(b) desk chair



(c) keyboard



(d) laptop computer



(e) projector

Figure 3. The estimated Gaussian density w.r.t. feature distances to corresponding predicted class centers. The features are from several classes, *e.g.*, (a) bike helmet, (b) desk chair, (c) keyboard, (d) laptop computer and (e) projector, in task **W**→**A** on Office-31 dataset. Blue stars denote distances of correctly labeled samples and red circles denote distances of wrongly labeled samples.