IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 4, 2021, accepted April 6, 2021, date of publication April 9, 2021, date of current version April 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3072237

Information Mandala: Statistical Distance Matrix With Clustering

XIN LU, (Member, IEEE)
Faculty of Science and Engineering, Iwate University, 3-18-8 Ueda, Morioka 020-8550, Japan
e-mail: luxin@iwate-u.ac.jp

ABSTRACT In machine learning, observation features are measured in a metric space to obtain their distance function for optimization. Given similar features that are statistically sufficient as a population, a statistical distance between two probability distributions can be calculated for more precise learning. Provided the observed features are multi-valued, the statistical distance function is still efficient. However, due to its scalar output, it cannot be applied to represent detailed distances between feature elements. To resolve this problem, this paper extends the traditional statistical distance to a matrix form, called a statistical distance matrix. The proposed approach performs well in object recognition tasks and clearly and intuitively represents the dissimilarities between cat and dog images in the CIFAR dataset, even when directly calculated using the image pixels. By using the hierarchical clustering of the statistical distance matrix, the image pixels can be separated into several clusters that are geometrically arranged around a center like a Mandala pattern. The statistical distance matrix with clustering is called the Information Mandala.

INDEX TERMS Statistical distance matrix, hierarchical clustering, Mandala.

I. INTRODUCTION

Classification is a type of supervised learning in machine learning that identifies to which of a set of categories a new observation belongs, based on a training set of labeled observations. The corresponding procedure for unsupervised learning is called clustering, which groups observations into categories based on their inherent similarities. In both classification and clustering, observation features are measured in a metric space, and their dissimilarities or distances are calculated for optimization. For example, a support vector machine (SVM) [1] in classification needs to measure the distance between two observation categories using the most efficient kernel function. For clustering, the k-means [2] approach aims to divide observations into categories to minimize the within-cluster sum of squares metric of the features in Euclidean or Mahalanobis space.

Provided the observation features are considered random variables or the feature set is considered a random vector in a probability space, measuring the distance between observations can be interpreted as quantifying a statistical distance between two probability distributions. Statistical distances have special mathematical properties that not all distances have. These properties include making distance measurements not only more effective and appropriate but also more robust to small outliers. Some important statistical distances, such as the Mahalanobis distance [3], Bhattacharyya distance [4], Hellinger distance [5], Kullback-Leibler divergence [6], [7], and Chernoff distance [8], have been applied to artificial intelligence applications, such as image segmentation [9], [10], texture segmentation [11], color and texture matching [12], feature extraction [13], speech recognition [14], and action recognition [15]. However, a clear limitation of these general statistical distances is that they only provide a scalar output to represent a global feature distance between two observations, regardless of the size of the feature set or the dimensionality of the corresponding random vector. Thus, the local distances of all features in the set or the relationships of all elements in the random vector cannot be elaborated. Therefore, an important problem is how to refine the concept of statistical distance to move from a scalar to a matrix.

The concept of a distance matrix has been introduced in graph theory [16]. In a directed graph, a distance matrix is defined by a weighted adjacency matrix. Given that each edge is assigned a weight, the distance between two vertices can be defined as the minimum sum of the weights of the shortest paths connecting the two vertices. The

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

distance matrix is asymmetric and not metric because the paths are oriented. If there are enough samples of each vertex, the cross-correlation matrix or partial cross-correlation matrix is used to identify the weights and quantify the distance matrix. However, due to its information loss from the assumption that all the data is in a probability space, the correlation matrix is not yet delicate or precise enough to satisfy some machine learning requirements.

To solve the above problems, this paper transforms traditional statistical distances into their matrix forms through a simple de-trace operation, and experimentally demonstrates the results for complicated distance performances using the CIFAR-10 dataset, which is the most famous dataset in machine learning.

II. PRELIMINARIES

First, we provide definitions of probability theory to specify the statistical distance.

Definition 1: In probability theory, a measurable function from a probability space $(\Omega, \mathcal{F}, P)$ to a measurable space $(\Lambda, \mathcal{G})$ is called a $(\Lambda, \mathcal{G})$-valued random variable, and is denoted by one of X, Y, Z, ... Let $\mathcal{G} := \mathcal{B}(\Lambda)$ (Borel σ-field). If X is a measurable function from $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, it is called a real-valued random variable. If $X := [X_1, \ldots, X_d]^\intercal$ is a measurable function from $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, it is called a d-dimensional random vector, where X_i is the i-th component of X, and X_1, ... X_d are random variables on a common probability space.

Definition 2: Let X be a $(\Lambda, \mathcal{G})$-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$. Then, a probability measure $P \circ X^{-1}$ on a measurable space $(\Lambda, \mathcal{G})$ is defined as $P \circ X^{-1}(B) := P(X^{-1}(B)) = P(X \in B)$, $B \in \mathcal{G}$. Then $P \circ X^{-1}$ is called the distribution of X, and is denoted by P_X.

Definition 3: Consider a real-valued random variable or random vector X with distribution P_X on a measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If $F(x) := P(X \leq x) = P_X((-\infty, x])$, for $x \in \mathbb{R}$, then $F(x)$ is called the cumulative distribution function of X on $\mathbb{R}$. Moreover, if $F(x)$ is absolutely continuous on $\mathbb{R}$, then $f_X(x) := dF(x)/dx$ is called the probability density function of X.

Definition 4: Given a probability space $(\Omega, \mathcal{F}, P)$ for a d-dimensional random vector X_1 with distribution P_X on the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, the cumulative distribution function is defined as $F_1(\boldsymbol{x}) := P(X_1 \leq \boldsymbol{x}) = P_X((-\infty, \boldsymbol{x}])$, for $\boldsymbol{x} \in \mathbb{R}^d$. If $F_1(\boldsymbol{x})$ is absolutely continuous with respect to $\boldsymbol{x}$ on $\mathbb{R}^d$, then $p_X(\boldsymbol{x}) := dF_1(\boldsymbol{x})/d\boldsymbol{x}$ is called the probability density function of X_1. Given another probability space $(\Omega, \mathcal{F}, Q)$, for a d-dimensional random vector X_2 with distribution Q_X on the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we can similarly obtain the corresponding probability density function of X_2, i.e., $q_X(\boldsymbol{x}) := dF_2(\boldsymbol{x})/d\boldsymbol{x}$, based on the cumulative distribution function $F_2(\boldsymbol{x}) := Q(X \leq \boldsymbol{x}) = Q_X((-\infty, \boldsymbol{x}])$. Note that $p_X(\boldsymbol{x})$ and $q_X(\boldsymbol{x})$ are generally abbreviated to $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$.

Based on these definitions, several indices have been introduced in statistics to reflect the dissimilarity between two probability distributions, $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$. The Bhattacharyya distance D_B was first proposed by [4] as a metric for quantifying dissimilarity:

$$D_B := -\ln \int_{\mathbb{R}^d} p^{\frac{1}{2}}(\boldsymbol{x}) q^{\frac{1}{2}}(\boldsymbol{x}) d\boldsymbol{x}. \tag{1}$$

The Chernoff distance D_C, an extension of D_B, was introduced in [8]. Here, the square root operator is replaced with an exponent coefficient s.

$$D_C := -\ln \int_{\mathbb{R}^d} p^s(\boldsymbol{x}) q^{1-s}(\boldsymbol{x}) d\boldsymbol{x} \tag{2}$$

The Kullback-Leibler divergence D_{KL}, formulated as follows, was proposed in [6], [7]. Note that it is not a metric because it does not satisfy the metric axiom.

$$D_{KL} := \int_{\mathbb{R}^d} [p(\boldsymbol{x}) - q(\boldsymbol{x})] \ln \left[\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right] d\boldsymbol{x} \tag{3}$$

The Hellinger distance D_H, as introduced in [5], is defined as follows through the Hellinger integral:

$$D_H := \frac{1}{\sqrt{2}} \int_{\mathbb{R}^d} \left\| p^{\frac{1}{2}}(\boldsymbol{x}) - q^{\frac{1}{2}}(\boldsymbol{x}) \right\|_2 d\boldsymbol{x}. \tag{4}$$

These measures all give scalar dissimilarities between the two probability distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, regardless of the dimensionality of the corresponding random vector.

III. MAIN RESULTS

A. DE-TRACE OPERATION FOR DISTANCE MATRIX

We use the de-trace operation to convert the scalar-valued statistical distances into their matrices. We first focus on the Mahalanobis distance D_M introduced in [3] for easy understanding because it can be considered a particular case of the Bhattacharyya distance D_B. Given two populations with respective mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathbb{R}^{d \times d}$, the squared Mahalanobis distance D_M^2 is written in a quadratic form:

$$\begin{aligned} D_M^2 &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \operatorname{tr}\left\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}^{-1}\right\} = \operatorname{tr} \boldsymbol{D}_M, \end{aligned} \tag{5}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. This quadratic form can be transformed to a trace form, as noted in Eqn. (5). By removing the trace, we can obtain the Mahalanobis distance matrix $\boldsymbol{D}_M$ in a de-trace form, Eqn. (6).

$$\boldsymbol{D}_M = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}^{-1} \tag{6}$$

By contrast, the Bhattacharyya distance D_B with corresponding distance matrix $\boldsymbol{D}_B$ is defined in a continuous measurable space as per Definition 4. In this paper, we suppose that two d-dimensional random vectors X_1 and X_2 follow two normal distributions, $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively. Thus, D_B for X_1 and X_2 is defined as

$$\begin{aligned} D_B = &\frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &+ \frac{1}{2} \ln \left[\det \boldsymbol{\Sigma}_1^{-\frac{1}{2}} \det \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right) \det \boldsymbol{\Sigma}_2^{-\frac{1}{2}} \right] \end{aligned}$$

$$= \mathrm{tr}\left\{\frac{1}{4}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1} + \frac{1}{2}\left[\ln\left(\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right) - \ln(\boldsymbol{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_2^{\frac{1}{2}})\right]\right\} = \mathrm{tr}\,\boldsymbol{D}_B. \quad (7)$$

We find that the first term of Eqn. (7) is similar to Eqn. (5) and can be transformed to the same trace form. The second term can also be changed into a trace form based on the following equations.

$$\det\boldsymbol{A}\det\boldsymbol{B} = \det(\boldsymbol{AB}), \quad (8)$$

$$\ln(\det\boldsymbol{A}) = \mathrm{tr}(\ln\boldsymbol{A}), \quad (9)$$

$$\mathrm{tr}[\ln(\boldsymbol{AB})] = \mathrm{tr}(\ln\boldsymbol{A}) + \mathrm{tr}(\ln\boldsymbol{B}) \quad (10)$$

Here, all above equations hold, if and not only if $\boldsymbol{A}$ and $\boldsymbol{B}$ are two positive definite matrices in $\mathbb{R}^{d\times d}$. Note that Eqn. (9) holds by Jacobi's formula for any complex square matrix where $\ln(\boldsymbol{A})$ is defined. Then, by dissolving the trace, the Bhattacharyya distance matrix $\boldsymbol{D}_B$ can be written as Eqn. (11).

$$\boldsymbol{D}_B = \frac{1}{4}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1} + \frac{1}{2}\left[\ln\left(\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right) - \ln(\boldsymbol{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_2^{\frac{1}{2}})\right] \quad (11)$$

The Chernoff distance D_C between the two normal distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is defined as

$$\begin{aligned} D_C &= \frac{1}{2}s(1-s)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}[(1-s)\boldsymbol{\Sigma}_1+s\boldsymbol{\Sigma}_2]^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\ &\quad + \frac{1}{2}\ln\left\{\det\boldsymbol{\Sigma}_1^{s-1}\det[(1-s)\boldsymbol{\Sigma}_1+s\boldsymbol{\Sigma}_2]\det\boldsymbol{\Sigma}_2^{-s}\right\} \\ &= \mathrm{tr}\left\{\frac{1}{2}s(1-s)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}[(1-s)\boldsymbol{\Sigma}_1+s\boldsymbol{\Sigma}_2]^{-1}\right. \\ &\quad \left. + \frac{1}{2}\{\ln[(1-s)\boldsymbol{\Sigma}_1+s\boldsymbol{\Sigma}_2] - \ln\left(\boldsymbol{\Sigma}_1^{1-s}\boldsymbol{\Sigma}_2^{s}\right)\}\right\} = \mathrm{tr}\,\boldsymbol{D}_C, \end{aligned} \quad (12)$$

where $s \in [0, 1]$. After the transformation for obtaining the trace form, the corresponding matrix $\boldsymbol{D}_C$ is obtained as Eqn. (13).

$$\begin{aligned} \boldsymbol{D}_C &= \frac{1}{2}s(1-s)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}[(1-s)\boldsymbol{\Sigma}_1+s\boldsymbol{\Sigma}_2]^{-1} \\ &\quad + \frac{1}{2}\{\ln[(1-s)\boldsymbol{\Sigma}_1+s\boldsymbol{\Sigma}_2] - \ln\left(\boldsymbol{\Sigma}_1^{1-s}\boldsymbol{\Sigma}_2^{s}\right)\} \end{aligned} \quad (13)$$

Note that the Chernoff distance D_C with distance matrix $\boldsymbol{D}_C$ extends D_B with $\boldsymbol{D}_B$, and is more flexible and adaptive due to the exponent coefficient s being adjustable according to computation requirements.

The Kullback-Leibler divergence D_{KL} between $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is given by

$$\begin{aligned} D_{KL} &= \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\ &\quad + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1+2\boldsymbol{I}_d) \\ &= \mathrm{tr}\left\{\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}\right. \\ &\quad \left. + \frac{1}{2}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1+2\boldsymbol{I}_d)\right\} = \mathrm{tr}\,\boldsymbol{D}_{KL}, \end{aligned} \quad (14)$$

where $\boldsymbol{I}_d$ is a d-dimensional identity matrix. It is easy to write its trace form and obtain the corresponding distance matrix $\boldsymbol{D}_{KL}$ as in Eqn. (15), where there exists no logarithm operation in the second term.

$$\boldsymbol{D}_{KL} = \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1} + \frac{1}{2}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1+2\boldsymbol{I}_d) \quad (15)$$

The Hellinger distance D_H between $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ is written as

$$D_H = \left[1-\exp(-D_B)\right]^{\frac{1}{2}} = \left[1-\exp(-\mathrm{tr}\,\boldsymbol{D}_B)\right]^{\frac{1}{2}}. \quad (16)$$

It can be considered a function with respect to $\boldsymbol{D}_B$, but cannot be changed into a complete trace form. Thus, the Hellinger distance D_H has no distance matrix.

Here, under the assumption that the distance is non-negative real number, all the elements of the above distance matrices are set to the absolute value of the original one. Based on these statistical distance matrices, we apply an ordinary hierarchical clustering [17], [18] to cluster the elements of a random vector. Given a cut-off threshold, this clustering algorithm can provide stable and reliable clustering results for the elements of a random vector.

B. HIERARCHICAL CLUSTERING FOR DISTANCE MATRIX

The input to the hierarchical clustering algorithm is defined as a finite element set S of the random vector X with a distance function δ, which is the map $\delta : S \times S \to \mathbb{R}$. Here, $\delta(u, v)$ is assigned the element value in a distance matrix $\boldsymbol{D}$ of X at location (u, v) and may be zero, where $u, v \in S$ and $\delta(u, u)$ is set to 0. Given that the set S has d elements, there exist $\binom{d}{2}$ pairwise distances.

The output of the hierarchical clustering algorithm is defined by a dendrogram, which can be considered as a data structure and is expressed as a mathematical graph. A stepwise dendrogram is used in this paper. Given a finite set S_0 with cardinality $d = |S_0|$, a stepwise dendrogram is a list of triples $\langle u_i, v_i, \delta(u_i, v_i)\rangle, i = 0, \ldots, d-2$ with the corresponding node labels n_i, where $u_i, v_i \in S_i$. Set S_0 is the initial data point. Set S_{i+1} is recursively defined as $(S_i \backslash \{u_i, v_i\}) \cup n_i$. In each step, the new node labeled n_i is formed by joining nodes u_i and v_i at distance $\delta(u_i, v_i)$. The procedure contains $d-1$ steps, such that the final state is a single node containing all d initial nodes.

The proposed hierarchical clustering algorithm is given in Algorithm 1.

Here, the agglomerative formula for updating δ is defined as largest-distance function between $\delta(u_i, x)$ and $\delta(v_i, x)$:

$$f(\delta(u_i, x), \delta(v_i, x)) := \max(\delta(u_i, x), \delta(v_i, x)). \quad (17)$$

Algorithm 1 Hierarchical Clustering Algorithm for Distance Matrix

Inputs:
Node labels: S_0
Distance function: δ
Initialize:
Number of input nodes: $d \leftarrow |S|$
Stepwise dendrogram: $L \leftarrow \varnothing$
for $i = 0$ **to** $d - 2$ **do**
$(u_i, v_i) \leftarrow \arg\min_{S_i \times S_i \setminus \Delta_i} \delta$, Δ_i denotes diagonal elements in $S_i \times S_i$.
Append triple $\langle u_i, v_i, \delta(u_i, v_i)\rangle$ to L
$S_i \leftarrow S_i \setminus \{u_i, v_i\}$
Create a new node label $n_i \notin S_i$
Update δ for all $x \in S_i$ by

$$\delta(n_i, x) = \delta(x, n_i) := f(\delta(u_i, x), \delta(v_i, x))$$

$S_{i+1} \leftarrow S_i \cup \{n_i\}$
end for
Outputs:
Stepwise dendrogram: L

There exist other useful formulas like smallest-distance function:

$$f(\delta(u_i, x), \delta(v_i, x)) := \min(\delta(u_i, x), \delta(v_i, x)), \tag{18}$$

or average-distance function:

$$f(\delta(u_i, x), \delta(v_i, x)) := \frac{1}{|S_i|^2} \sum_{i \in S_i} \sum_{i \in S_i} (\delta(u_i, x), \delta(v_i, x)). \tag{19}$$

Given a cut-off threshold, this algorithm can provide stable and reliable clustering results for the elements of a random vector.

IV. COMPUTED EXAMPLES

We use the CIFAR-10 dataset [19] to test the effects of the statistical distance matrices. This dataset contains 60,000 32×32 color images in 10 different classes. To simplify the calculation and obtain distinguishable results, as shown in Figure 1, we calculated only the distance matrices between airplanes and dogs, birds and dogs, cats and dogs, such that the similarities between every two objects could range from weak to strong.

A. STATISTICAL DISTANCE MATRIX

Given a value space $U = [0, 1]$ for an image pixel, an image $\boldsymbol{A} := [a_{ij}] \in U^{m \times n}$, $i = 1, \ldots, m, j = 1, \ldots, n$ is re-formed as $\boldsymbol{a} := [\acute{a}_t] = \mathrm{vec}\boldsymbol{A} \in U^{d \times 1}$, $t = 1, \ldots, d, d = m \times n$ through matrix vectorization $\mathrm{vec}(\cdot)$. Let image sets $\{\boldsymbol{a}_k\}$ and $\{\boldsymbol{b}_k\}$, $k = 1, \ldots, N$ of two classes be regarded as two populations with d-dimensional random vectors X_1 and X_2, which respectively follow the two normal distributions

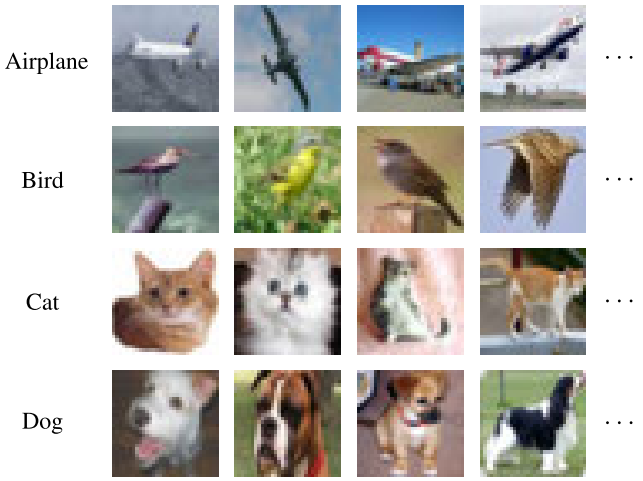

FIGURE 1. Examples of airplanes, birds, cats, and dogs in the CIFAR-10 dataset.

$\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where

$$\boldsymbol{\mu}_1 := \mathbb{E}[\boldsymbol{X}_1] = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{a}_k, \quad \boldsymbol{\mu}_2 := \mathbb{E}[\boldsymbol{X}_2] = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{b}_k,$$

$$\boldsymbol{\Sigma}_1 := \mathrm{var}[\boldsymbol{X}_1] = \frac{1}{N-1} \sum_{k=1}^{N} [(\boldsymbol{a}_k - \boldsymbol{\mu}_1)(\boldsymbol{a}_k - \boldsymbol{\mu}_1)^{\mathsf{T}}],$$

$$\boldsymbol{\Sigma}_2 := \mathrm{var}[\boldsymbol{X}_2] = \frac{1}{N-1} \sum_{k=1}^{N} [(\boldsymbol{b}_k - \boldsymbol{\mu}_2)(\boldsymbol{b}_k - \boldsymbol{\mu}_2)^{\mathsf{T}}]. \tag{20}$$

In the examples, $\{\boldsymbol{a}_k\}$ and $\{\boldsymbol{b}_k\}$ were set to the image sets of airplanes and dogs, birds and dogs, and cats and dogs in order, where m and n are set to 32, respectively. Note that $N = 5000$ because of only the training images are used for each class. By calculating and substituting their mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ with covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ into Eqn. (6)-(15), we can obtain the four statistical distance matrices $\boldsymbol{D}_M$, $\boldsymbol{D}_{KL}$, $\boldsymbol{D}_B$, and $\boldsymbol{D}_C$ for all of three cases, where each $\boldsymbol{D} := [\delta_{uv}] \in \mathbb{R}^{d \times d}$, $u, v = 1, \ldots, d$.

As shown in Figure 2, $\boldsymbol{D}_M$ and $\boldsymbol{D}_{KL}$ for all of three cases appear chaotic and uninformative. By contrast, the local distances with high values represent a grid-like pattern in the middle of $\boldsymbol{D}_B$ and $\boldsymbol{D}_C$, where the exponent coefficient s in $\boldsymbol{D}_C$ was set to 0.3. These high-valued local distances can effectively be used to distinguish the corresponding elements of the random vectors X_1 and X_2 in the measurable space, which are also regarded as the corresponding pixels of the images $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$. Considering that $\boldsymbol{D}_B$ is a particular case of $\boldsymbol{D}_C$ where s is set to $1/2$, the $\boldsymbol{D}_C$-like statistical distance matrices are confirmed to be valid.

Based on the definition equations of statistical distance matrices, the vertical and horizontal lines of elements in these matrices, called rows and columns, are related to $\{\boldsymbol{a}_k\}$ and $\{\boldsymbol{b}_k\}$, respectively. Note that $\{\boldsymbol{a}_k\}$ is regarded as the image sets of airplanes, birds, or cats. $\{\boldsymbol{b}_k\}$ is only assigned to the image set of dogs.

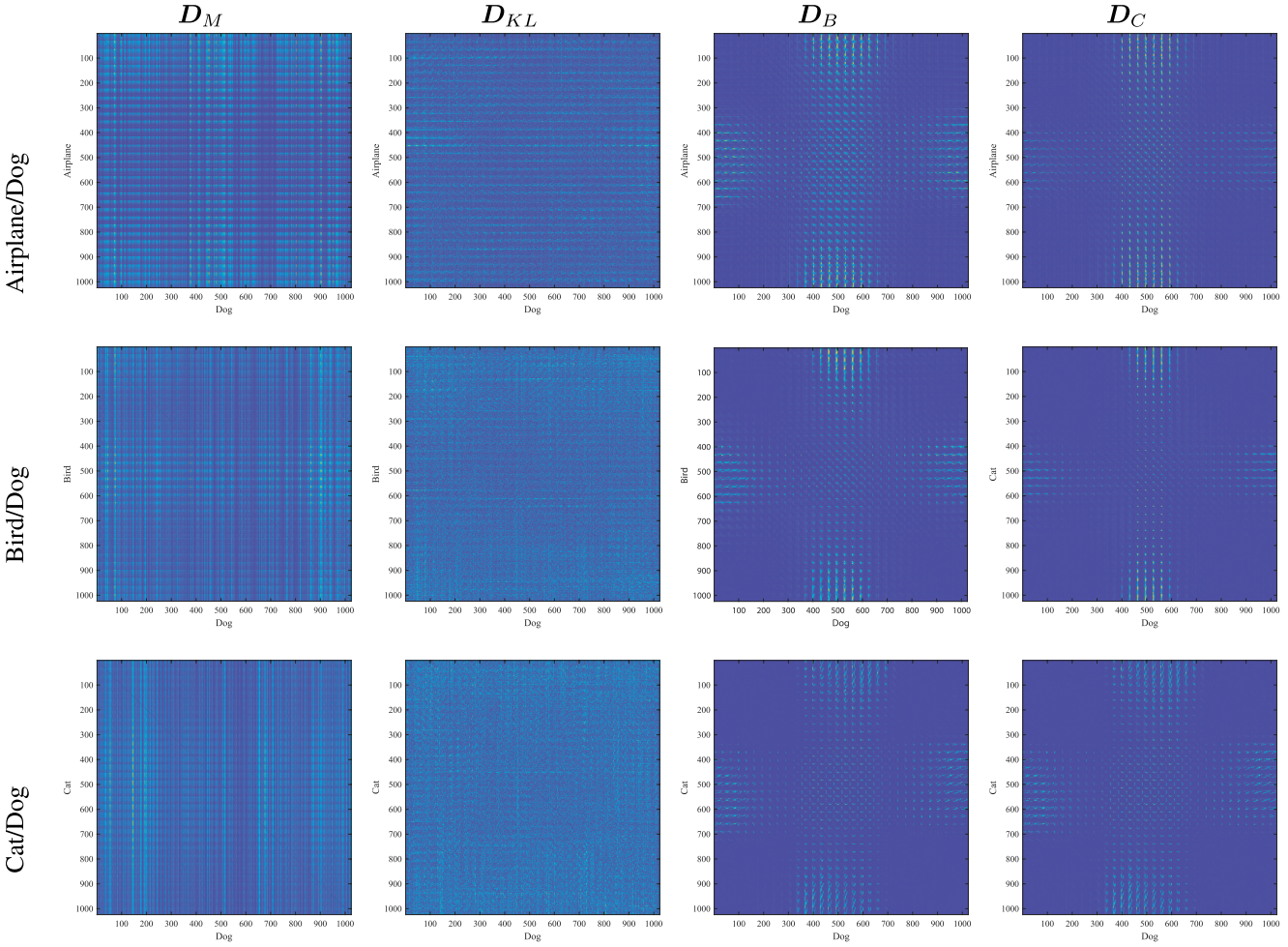

FIGURE 2. Statistical distance matrices D_M, D_{KL}, D_B, and D_C for the cases of airplanes and dogs, birds and dogs, cats and dogs, respectively. Every subfigure corresponds a distance matrix in the same from. Here, D_B and D_C perform better than D_M and D_{KL} due to their grid-like patterns.

B. DISTANCE-ACCUMULATION IMAGE

Furthermore, for each image pixel, we can accumulate all of its related local distances as a value and assign it to the current pixel to form a distance-accumulation vector $\boldsymbol{\phi} := [\acute{\phi}_t] \in \mathbb{R}^{d\times 1},\ t = 1, \ldots, d$ by using the following distance-matrix-imaging method:

$$\acute{\phi}_t = \sum_{u=1}^{d} \delta_{ut} + \sum_{v=1}^{d} \delta_{tv}. \tag{21}$$

Then, $\boldsymbol{\phi}$ is re-formed as a distance-accumulation image $\boldsymbol{\Phi} := [\phi_{ij}] = \widehat{\text{vec}}\boldsymbol{\phi},\ i = 1, \ldots, m, j = 1, \ldots, n$, where $\widehat{\text{vec}}(\cdot)$ denotes the reverse process of vec$(\cdot)$. Note that $\boldsymbol{\Phi}$ has the same form with the image $\boldsymbol{A}$. Unlike the statistical distance matrix $\boldsymbol{D}$, every axis of $\boldsymbol{\Phi}$ is related to the double classes owing to Eqn. (21).

The effects of the statistical distance matrices are more clearly reflected in their distance-accumulation images than themselves. As illustrated in Figure 3, the distance-accumulation images of $\boldsymbol{D}_M$ and $\boldsymbol{D}_D$ are disordered, and those of $\boldsymbol{D}_B$ and $\boldsymbol{D}_C$ are ordered. The main representation is that the high-value pixels are all concentrated in the center of the distance-accumulation image, showing a distribution similar to a circle or an ellipse. For objects with low similarities, such as airplanes and dogs, the number of high-value pixels is greater and their locations are more concentrated. By contrast, between objects with high similarities, such as cats and dogs, high-value pixels are fewer and concentrated in the center of the image more broadly along with middle-value pixels. Therefore, the distance-matrix-imaging method can simultaneously quantify the differences between the pixels of every two objects in degree and position.

C. TEST FOR CENTRALIZATION OF HIGH-VALUED ELEMENTS

Let the distance-accumulation image $\boldsymbol{\Phi}$ be normalized as a discrete distribution $\bar{\boldsymbol{\Phi}} := [\bar{\phi}_{ij}],\ i = 1, \ldots, m, j = 1, \ldots, n$ in two dimensions, where the summation of all the elements in $\bar{\boldsymbol{\Phi}}$ is set to 1 by $\bar{\phi}_{ij} := \phi_{ij} / \sum_{j=1}^{n} \sum_{i=1}^{m} \phi_{ij}$. As proposed in [20], this distribution shape can be fitted by a 2-dimensional normal distribution with a probability density

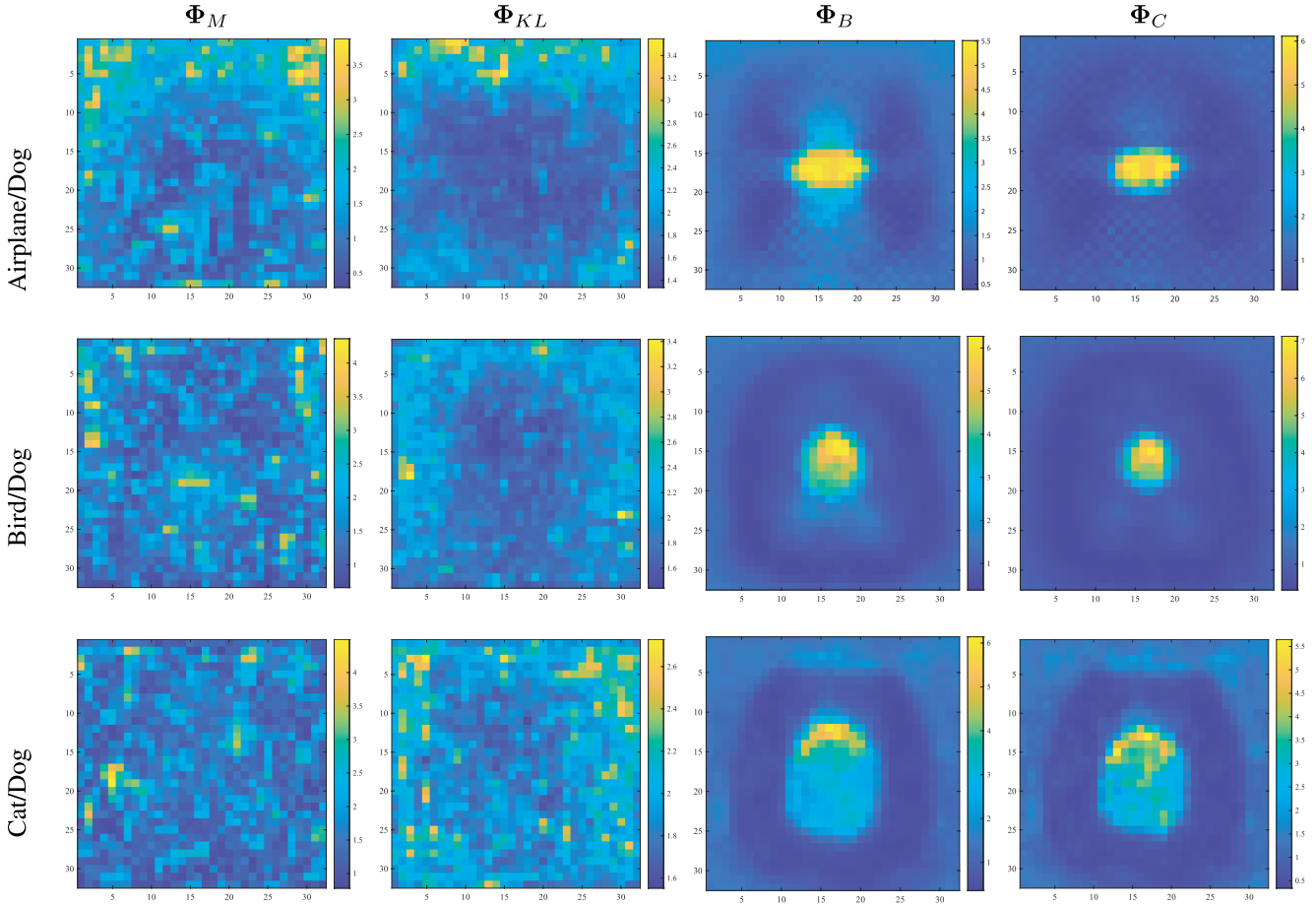

FIGURE 3. Distance-accumulation images Φ_M, Φ_{KL}, Φ_B, and Φ_C for the cases of airplanes and dogs, birds and dogs, cats and dogs, respectively. Every subfigure corresponds a distance-accumulation image in the same from. Here, Φ_B and Φ_C perform better than Φ_M and Φ_{KL} due to the centralization of their high-valued elements.

function $p(x, y)$. Then, the normal distribution is quantized as a discrete distribution $\boldsymbol{P} := [p_{ij}]$, $i = 1, \ldots, m, j = 1, \ldots, n$. The Hadamard-production summation of $\bar{\boldsymbol{\Phi}}$ and $\boldsymbol{P}$, defined as $\bar{\boldsymbol{\Phi}} \odot \boldsymbol{P} := \sum_{j=1}^{n} \sum_{i=1}^{m} \bar{\phi}_{ij} p_{ij}$, is considered as a measure to estimate whether the high-valued elements of $\bar{\boldsymbol{\Phi}}$ are centralized. As shown in Figures 4-6, the values of $\bar{\boldsymbol{\Phi}}_B \odot \boldsymbol{P}_B$ and $\bar{\boldsymbol{\Phi}}_C \odot \boldsymbol{P}_C$ are greater than those of $\bar{\boldsymbol{\Phi}}_M \odot \boldsymbol{P}_M$ and $\bar{\boldsymbol{\Phi}}_{KL} \odot \boldsymbol{P}_{KL}$ for every case; therefore, $\bar{\boldsymbol{\Phi}}_B$ and $\bar{\boldsymbol{\Phi}}_C$ have more high-valued elements centralized in image than $\bar{\boldsymbol{\Phi}}_M$ and $\bar{\boldsymbol{\Phi}}_{KL}$. It causes that their elements, which also means the pixels of object image, can be clustered easily and clearly at the following step.

D. TEST FOR SHAPE SIMILARITY TO MANDALAS

Finally, the hierarchical clustering algorithms respectively based on the three agglomerative formulas, which are the smallest-distance function (18), the average-distance function (19), and the largest-distance function (17), are used to cluster the pixels of object images based on the statistical distance matrix $\boldsymbol{D}_B$. Here, the pixels of object images are separated into three clusters, and labeled in the distance-accumulation images for the three cases, respectively. As shown in Figures 8, 9, and 10, the hierarchical clustering algorithm based on the largest-distance function (17) performs best with the most abundant patterns.

Moreover, in Figures 8, 9, and 10, the cluster patterns are all circular or square, symmetrical, and radiate from a center point. Therefore, the cluster patterns can be considered as Mandalas.[1] Thus, we establish the term "Information Mandala" to describe the statistical distance matrix with clustering.

We also compare our proposed hierarchical clustering with spectrum clustering [21] and k-medoids clustering [22] to prove its effectiveness for the statistical distance matrix. The corresponding experimental results, shown in Figures 11, 13, and 15, reflect that our proposed hierarchical clustering performs better with clearer and more ordered clustering results than the other two clusterings. Here, the pixels of object images are separated into ten clusters.

[1]The word Mandala is a Sanskrit term meaning "sacred circle." In various religious traditions, such as Hinduism, Buddhism, Jainism, and Shintoism, a mandala is used as a map to represent paradise, gods, or actual shrines. Mandalas are circular or square and designed with repeating colors, shapes, and patterns that radiate from a center point. Mandalas can be precise, carefully measured, geometric, and perfectly symmetrical.

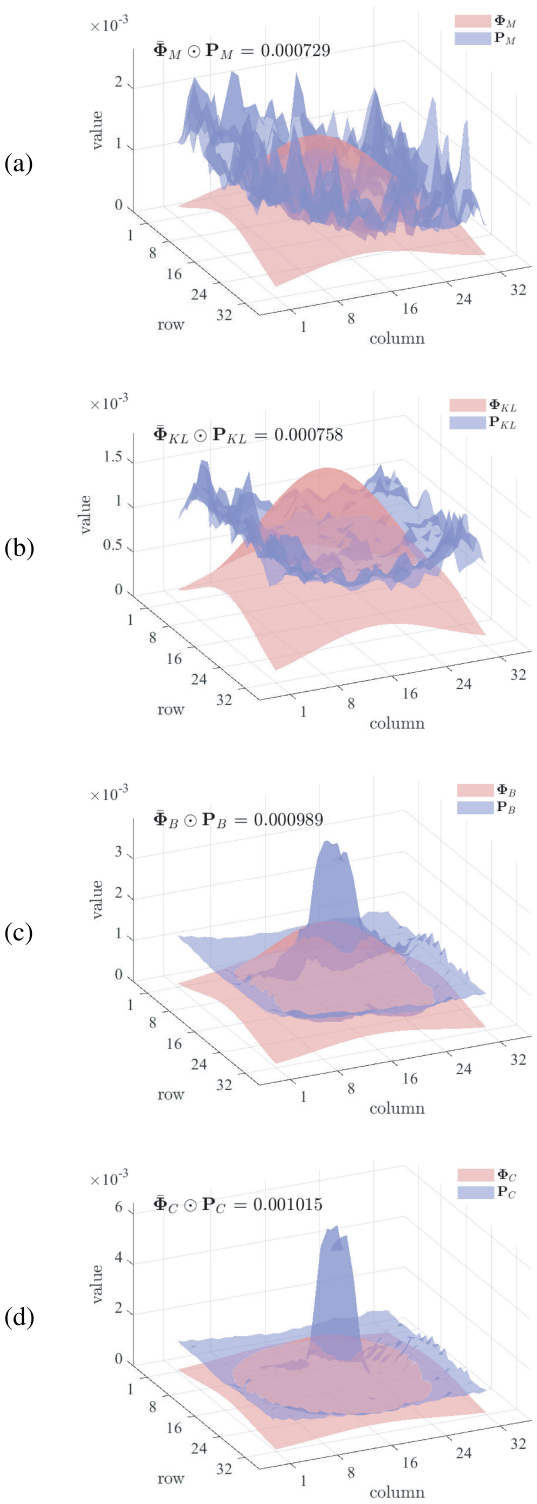

FIGURE 4. Hadamard-production summation of $\bar{\Phi}$ and P for the case of airplanes and dogs. (a): $\bar{\Phi}_M \odot P_M$; (b): $\bar{\Phi}_{KL} \odot P_{KL}$; (c): $\bar{\Phi}_B \odot P_B$; (d): $\bar{\Phi}_C \odot P_C$. Here, $\bar{\Phi}_B \odot P_B$ and $\bar{\Phi}_C \odot P_C$ are greater than $\bar{\Phi}_M \odot P_M$ and $\bar{\Phi}_{KL} \odot P_{KL}$, that is Φ_B and Φ_C have more high-valued elements centralized in image than Φ_M and Φ_{KL}.

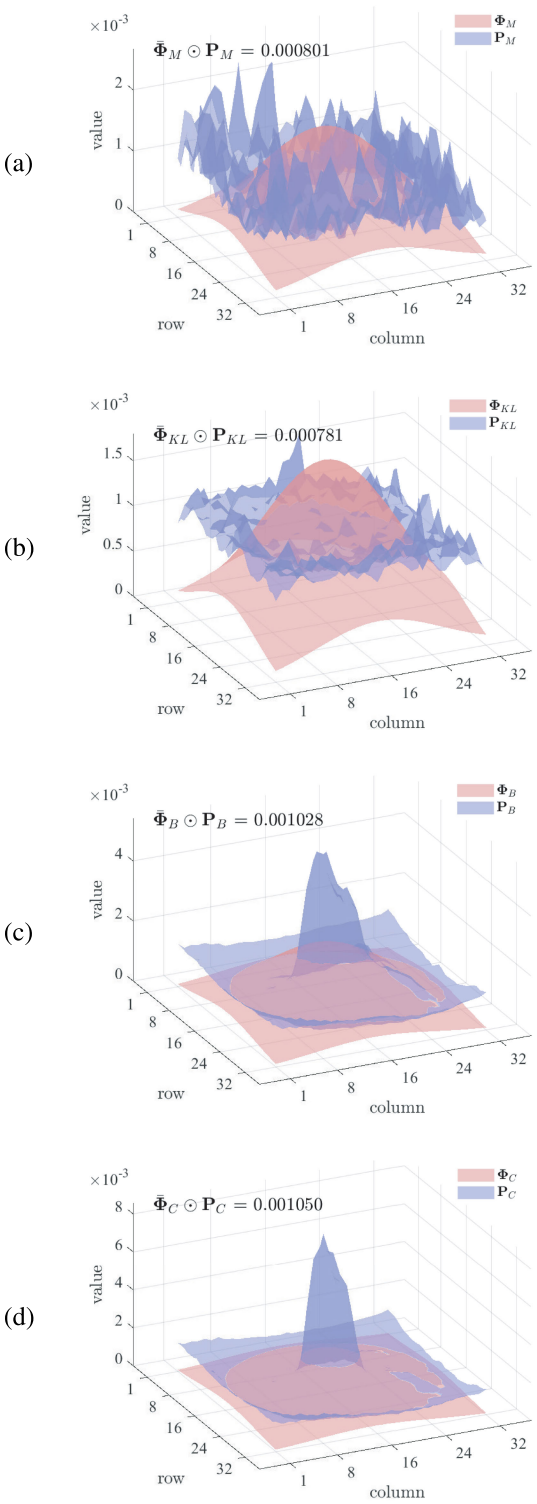

FIGURE 5. Hadamard-production summation of $\bar{\Phi}$ and P for the case of birds and dogs. (a): $\bar{\Phi}_M \odot P_M$; (b): $\bar{\Phi}_{KL} \odot P_{KL}$; (c): $\bar{\Phi}_B \odot P_B$; (d): $\bar{\Phi}_C \odot P_C$. Here, $\bar{\Phi}_B \odot P_B$ and $\bar{\Phi}_C \odot P_C$ are greater than $\bar{\Phi}_M \odot P_M$ and $\bar{\Phi}_{KL} \odot P_{KL}$, that is Φ_B and Φ_C have more high-valued elements centralized in image than Φ_M and Φ_{KL}.

Furthermore, we propose a novel measure, referred to as relative span summation, to decide whether the clustering results are similar to Mandalas. As shown in Figure 7, the elements of $\boldsymbol{A}$ are reordered as $\tilde{\boldsymbol{a}} := [\tilde{a}_t]$, $t = 1, \ldots, d$,

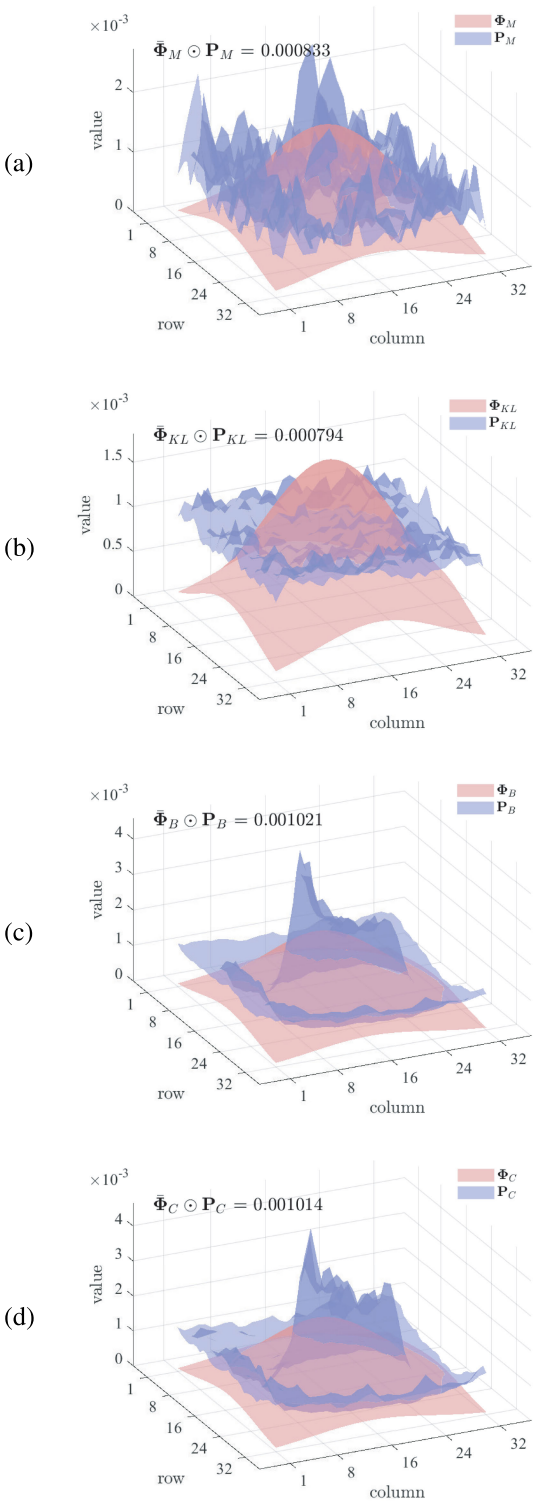

FIGURE 6. Hadamard-production summation of $\bar{\Phi}$ and P for the case of cats and dogs. (a): $\bar{\Phi}_M \odot P_M$; (b): $\bar{\Phi}_{KL} \odot P_{KL}$; (c): $\bar{\Phi}_B \odot P_B$; (d): $\bar{\Phi}_C \odot P_C$. Here, $\bar{\Phi}_B \odot P_B$ and $\bar{\Phi}_C \odot P_C$ are greater than $\bar{\Phi}_M \odot P_M$ and $\bar{\Phi}_{KL} \odot P_{KL}$, that is Φ_B and Φ_C have more high-valued elements centralized in image than Φ_M and Φ_{KL}.

$d = m \times n$ along an Archimedean spiral, which winds around the center point of $\boldsymbol{A}$. Then, as shown

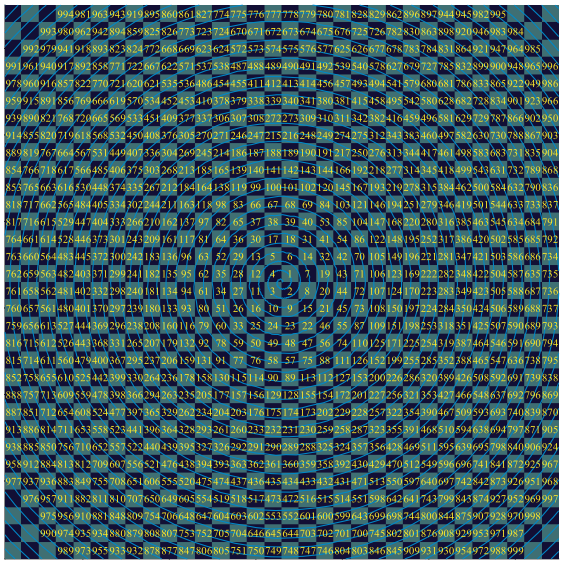

FIGURE 7. Image element reordering, where the elements of A are reordered as $\tilde{a} := [\tilde{a}_t]$, $t = 1, \ldots, d$, $d = m \times n$ along an Archimedean spiral winding around the center point of A.

in Figures 12, 14, and 16, the r-th cluster can be denoted as a set of the element subscripts $T^{(r)} := \{t^{(r)}(1), \ldots, t^{(r)}(N^{(r)})\}$ from $\tilde{\boldsymbol{a}}$, where $N^{(r)}$ is the amount of the elements included in the r-th cluster. Given $t^{(r)}(1) < \ldots < t^{(r)}(N^{(r)})$, the span of $T^{(r)}$ can be defined as $\text{span}(r) := t^{(r)}(N^{(r)}) - t^{(r)}(1) + 1$.

For any cluster $T^{(r)}$, suppose that its element subscripts are consecutive integers, which means these elements are arranged continuously on the Archimedean spiral, there exists a minimum span assigned as $\text{span}(r) = N^{(r)}$. Then, we can obtain a minimum span summation $\text{Span} := \sum_r \text{span}(r) = \sum_r N^{(r)} = d$ of all the clusters, that is all the clusters are arranged integrally on the Archimedean spiral and connected end-to-end. Otherwise, we have a general span summation $\text{Span} = \sum_r \text{span}(r) > \sum_r N^{(r)} = d$, which indicates some clusters overlap on the Archimedean spiral. Note that Span will increase more greatly, while the clusters overlap more richly and disordered. Therefore, the ratio of Span to d, also called as relative span summation Span/d, can be considered as the measure to decide whether the clustering results are circular and radiate from a center point like Mandalas.

As illustrated in Figures 12, 14, and 16, it obvious that the Span/ds of our proposed hierarchical clustering are smaller than those of the spectrum clustering and the k-medoids clustering for every case. Therefore, our hierarchical clustering is more effective to extract the clear and ordered clustering results than the other two clusterings. Here, the amount of cluster is set to 10.

V. DISCUSSIONS

We first discuss why the statistical distance matrices $\boldsymbol{D}_B$ and $\boldsymbol{D}_C$ are effective in the feature-distance measurement. There is a quadratic term with respect to the mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ in $\boldsymbol{D}_B$ and $\boldsymbol{D}_C$. This

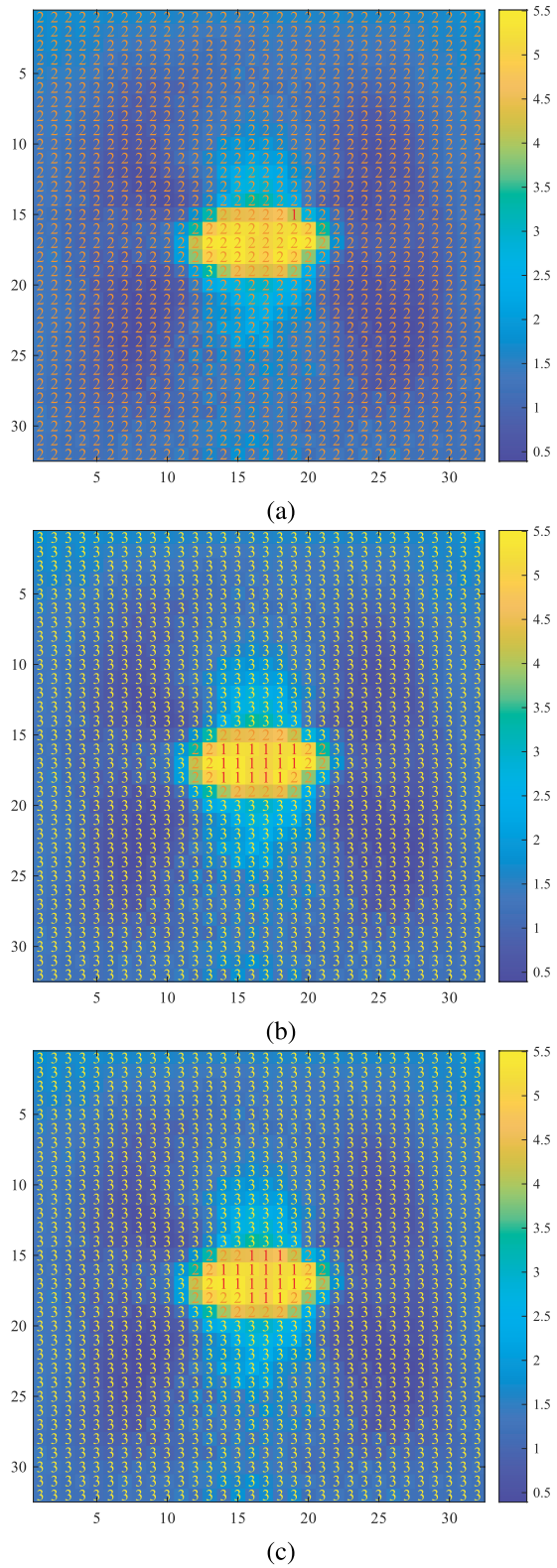

FIGURE 8. The labeled distance-accumulation image Φ_B to show the proposed hierarchical clustering results (three clusters) for the case of airplanes and dogs using the statistical distance matrix D_B. (a): smallest-distance function (18); (b): average-distance function (19); (c): largest-distance function (17). Here, (c) is best with the most abundant patterns.

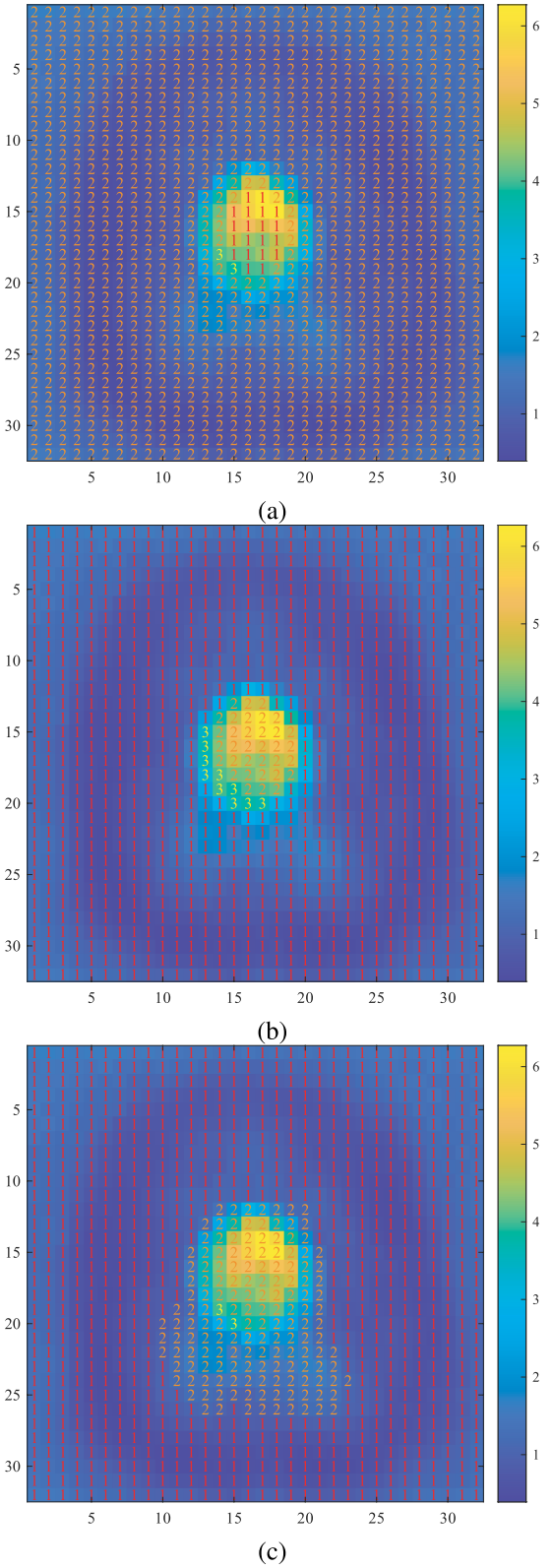

FIGURE 9. The labeled distance-accumulation image Φ_B to show the clustering results (three clusters) for the case of birds and dogs using the statistical distance matrix D_B. (a): smallest-distance function (18); (b): average-distance function (19); (c): largest-distance function (17). Here, (c) is best with the most abundant patterns.

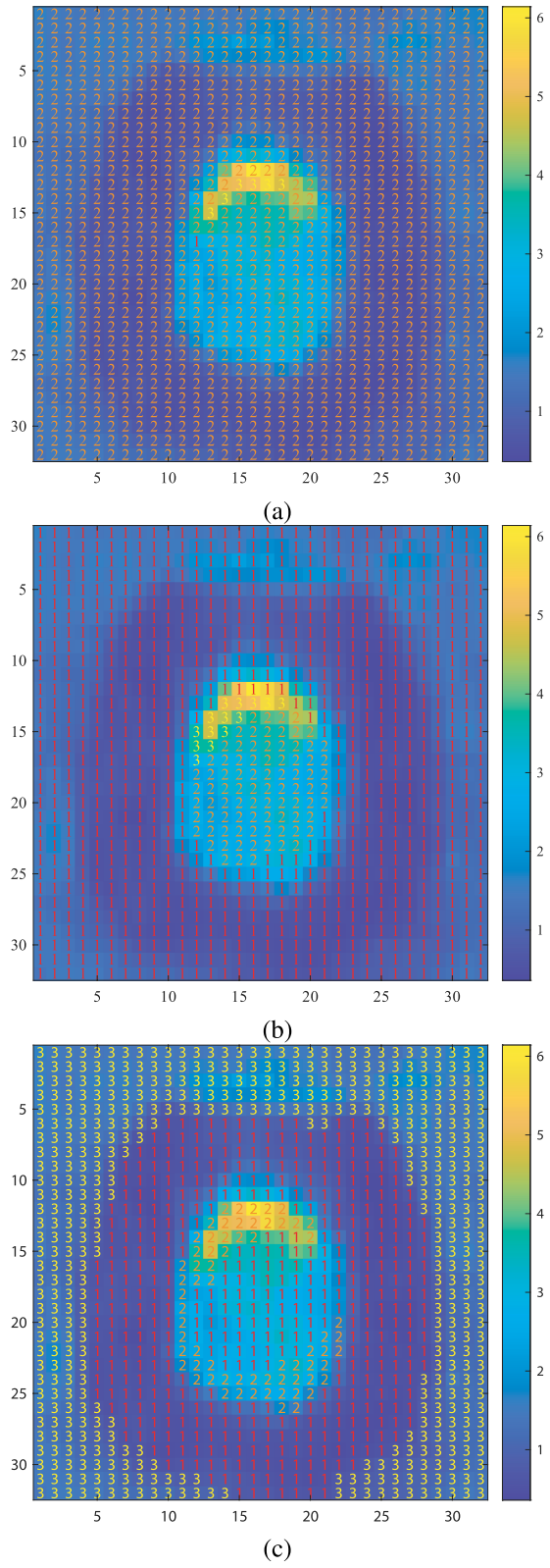

FIGURE 10. The labeled distance-accumulation image Φ_B to show the clustering results (three clusters) for the case of cats and dogs using the statistical distance matrix D_B. (a): smallest-distance function (18); (b): average-distance function (19); (c): largest-distance function (17). Here, (c) is best with the most abundant patterns.

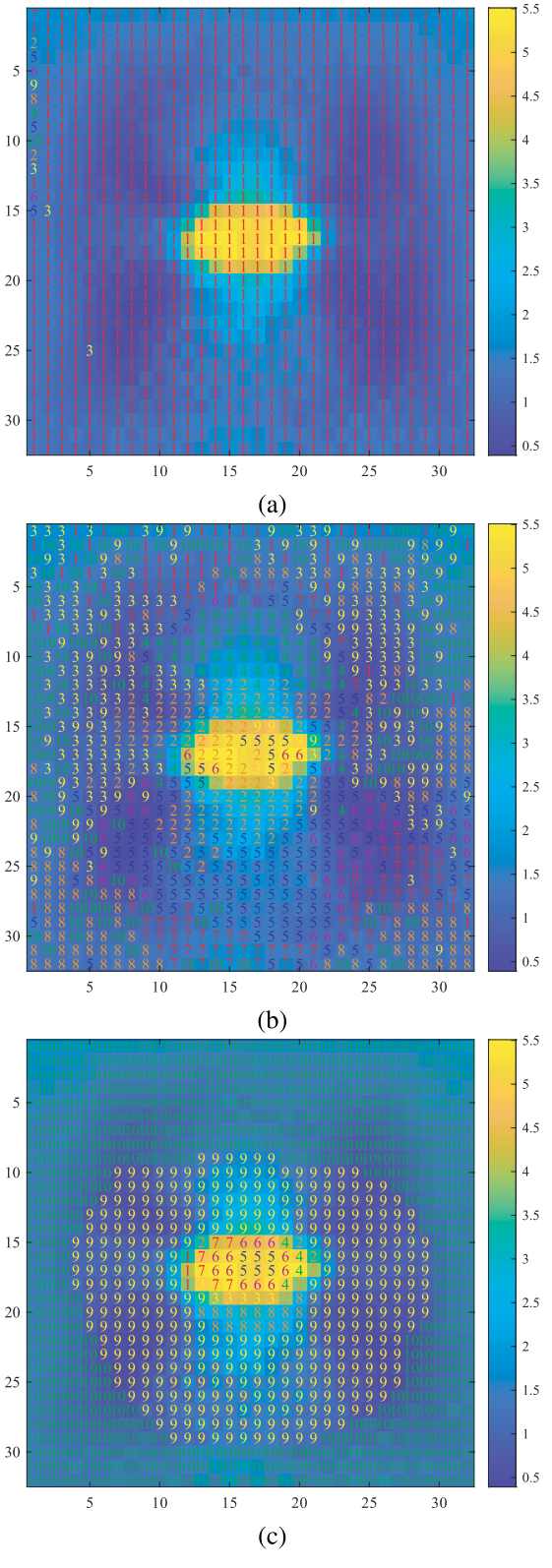

FIGURE 11. The labeled distance-accumulation image Φ_B to show the clustering results (ten clusters) for the case of airplanes and dogs using the statistical distance matrix D_B. (a): spectrum clustering; (b): k-medoids clustering; (c): proposed hierarchical clustering. Here, (c) is most similar to Mandalas.

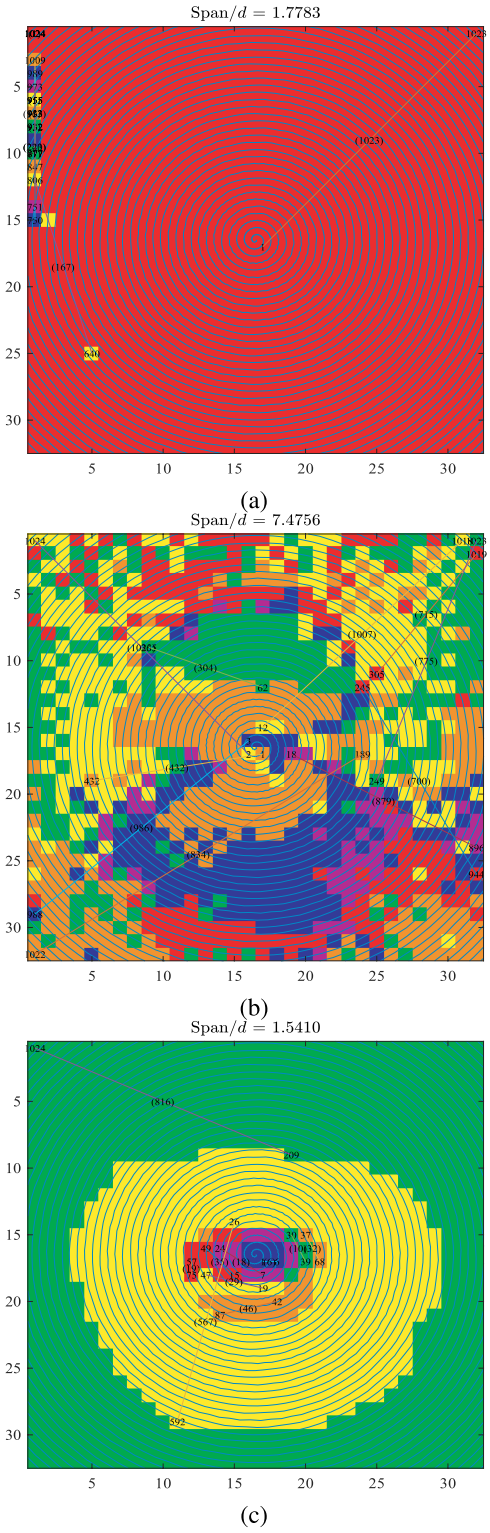

FIGURE 12. The colored clustering mask on the reordered image $\tilde{a}$ to show the clustering results (ten clusters) for the case of airplanes and dogs using the statistical distance matrix D_B. The head $t^{(r)}(1)$ and tail $t^{(r)}(N^{(r)})$ of the r-th cluster $T^{(r)}$ are connected by a line, whose center are labeled by its span $\text{span}(r) := t^{(r)}(N^{(r)}) - t^{(r)}(1) + 1$ in brackets. (a): spectrum clustering; (b): k-medoids clustering; (c): proposed hierarchical clustering. Here, (c) is most similar to Mandalas.

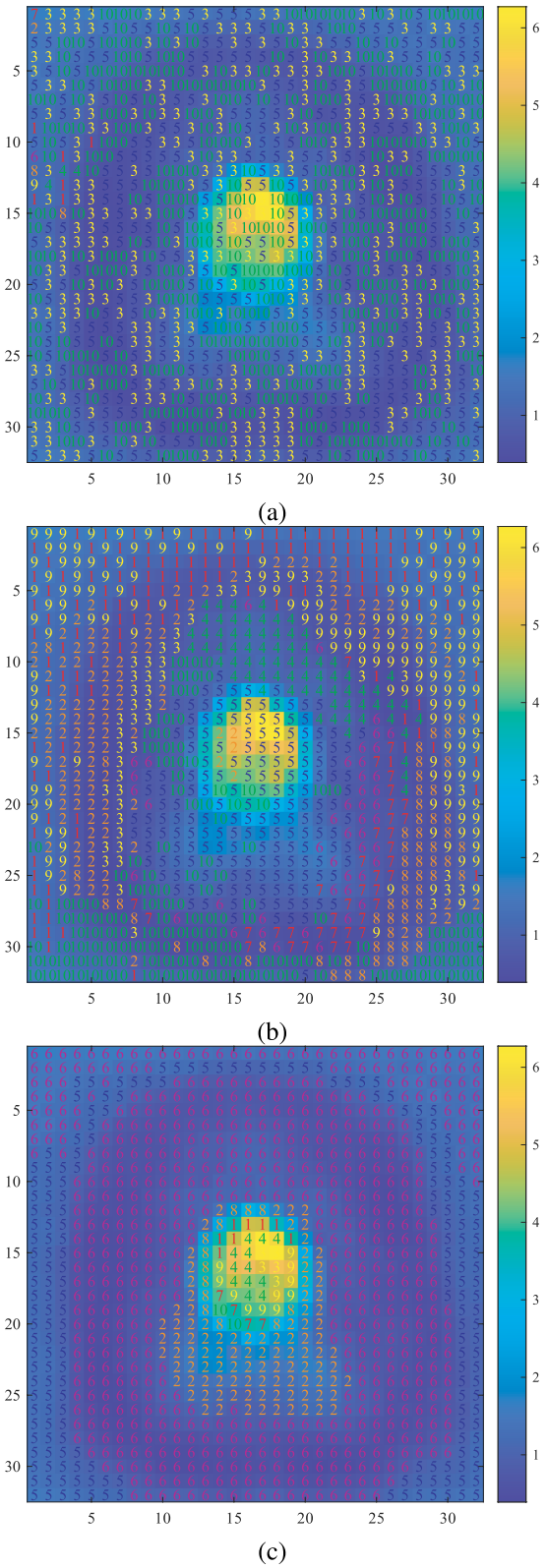

FIGURE 13. The labeled distance-accumulation image Φ_B to show the clustering results (ten clusters) for the case of birds and dogs using the statistical distance matrix D_B. (a): spectrum clustering; (b): k-medoids clustering; (c): proposed hierarchical clustering. Here, (c) is most similar to Mandalas.

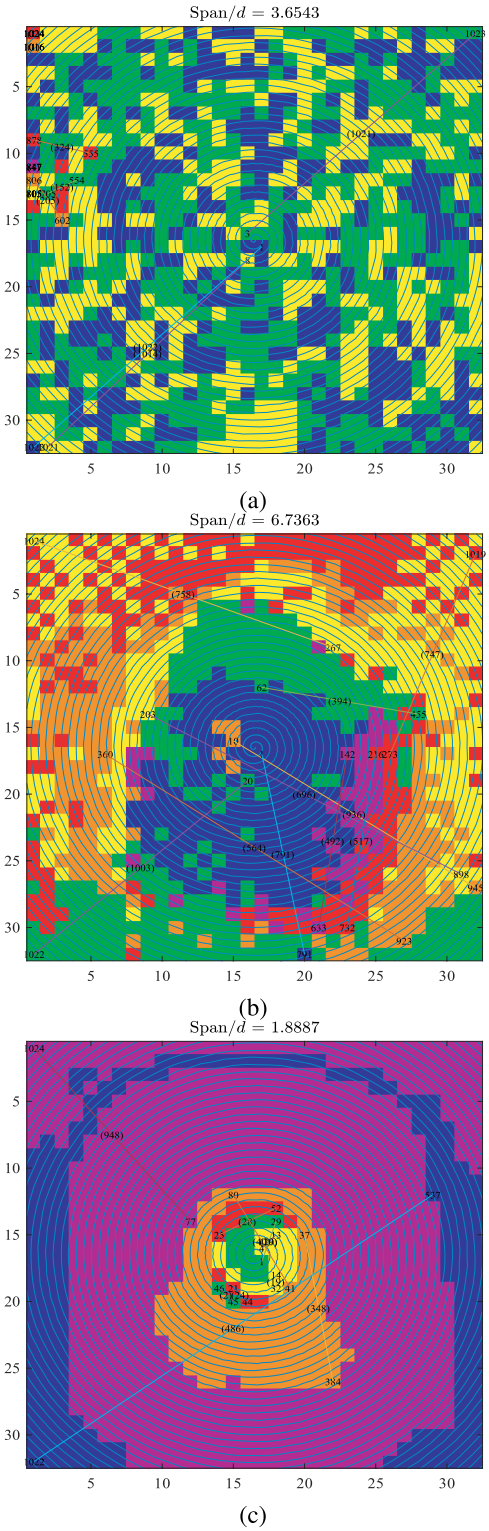

FIGURE 14. The colored clustering mask on the reordered image $\tilde{a}$ to show the clustering results (ten clusters) for the case of birds and dogs using the statistical distance matrix D_B. The head $t^{(r)}(1)$ and tail $t^{(r)}(N^{(r)})$ of the r-th cluster $T^{(r)}$ are connected by a line, whose center are labeled by its span $\text{span}(r) := t^{(r)}(N^{(r)}) - t^{(r)}(1) + 1$ in brackets. (a): spectrum clustering; (b): k-medoids clustering; (c): proposed hierarchical clustering. Here, (c) is most similar to Mandalas.

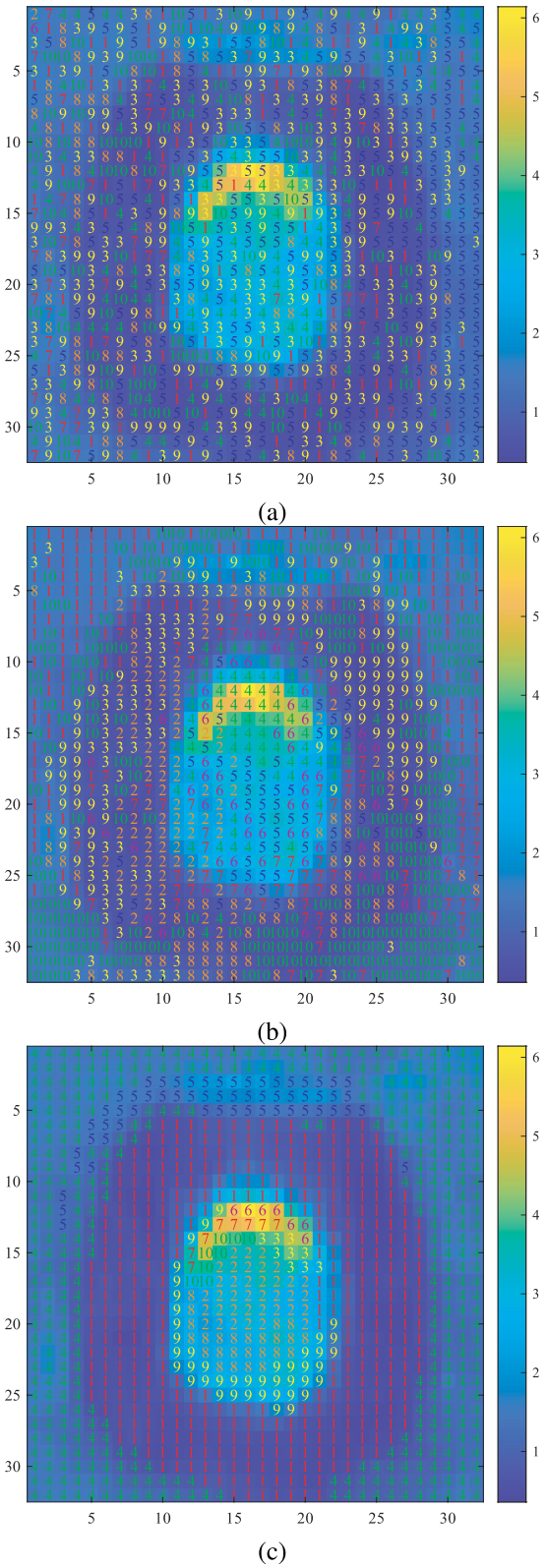

FIGURE 15. The labeled distance-accumulation image Φ_B to show the clustering results (ten clusters) for the case of cats and dogs using the statistical distance matrix D_B. (a): spectrum clustering; (b): k-medoids clustering; (c): proposed hierarchical clustering. Here, (c) is most similar to Mandalas.

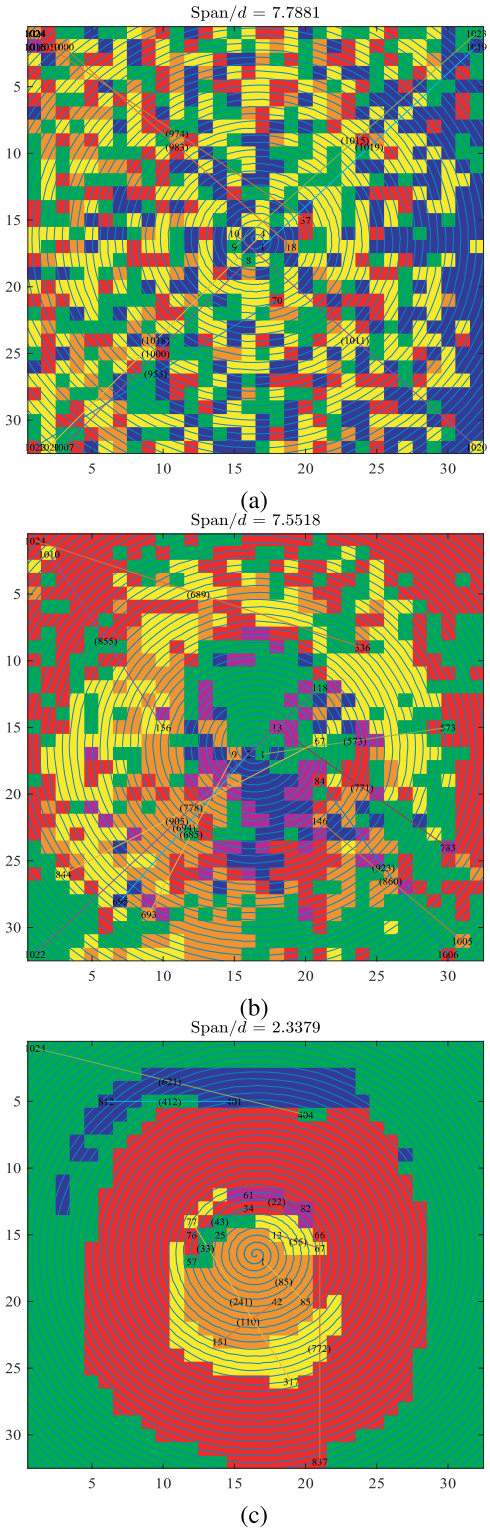

FIGURE 16. The colored clustering mask on the reordered image $\tilde{a}$ to show the clustering results (ten clusters) for the case of cats and dogs using the statistical distance matrix D_B. The head $t^{(r)}(1)$ and tail $t^{(r)}(N^{(r)})$ of the r-th cluster $T^{(r)}$ are connected by a line, whose center are labeled by its span $\text{span}(r) := t^{(r)}(N^{(r)}) - t^{(r)}(1) + 1$ in brackets. (a): spectrum clustering; (b): k-medoids clustering; (c): proposed hierarchical clustering. Here, (c) is most similar to Mandalas.

term also exists in $\boldsymbol{D}_M$ and $\boldsymbol{D}_{KL}$. There is also an item that contains only the logarithm of the covariance ratio in $\boldsymbol{D}_B$ and $\boldsymbol{D}_C$ and does not contain the mean vector. There is no such item containing only the covariance matrix in $\boldsymbol{D}_M$, and the only item containing the covariance ratio in $\boldsymbol{D}_{KL}$ has not been calculated logarithmically. When the two mean vectors are equal or approximate, the value of the quadratic term tends to zero, meaning that when the two probability distributions overlap heavily, the term of the covariance ratio plays a more important role than the quadratic term. This is the reason $\boldsymbol{D}_B$ and $\boldsymbol{D}_C$ are effective.

Next, we explain why cluster analysis is needed. Before cluster analysis, the statistical distance matrix $\boldsymbol{D}$ represents all the local distances between the elements of the random vector X. A large number of the local distances are so near zero that $\boldsymbol{D}$ becomes sparse. On the other hand, in graph theory, $\boldsymbol{D}$ can be mapped to a directed graph, whose vertices are defined as the elements of X, and whose edges are assigned to the local distances. However, in some applications when this graph is large, processing the edges with small values increases the computational complexity. Therefore, a tree structure is applied to reduce the number of unimportant edges in the graph and to arrange the vertices hierarchically according to their important edges with reassigned values. This tree structure can greatly accelerate the access speed and save memory space in the computer.

Then, we specify why the statistical distance matrices have further information than the cross-correlation matrix. The cross-correlation matrix $\boldsymbol{\Sigma}_{1,2}$ of X_1 and X_2 is defined as

$$\begin{aligned}\boldsymbol{\Sigma}_{1,2} &:= \text{cov}[X_1, X_2] = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])^{\mathsf{T}}\right] \\ &= \frac{1}{N-1}\sum_{k=1}^{N}[(\boldsymbol{a}_k - \boldsymbol{\mu}_1)(\boldsymbol{b}_k - \boldsymbol{\mu}_2)^{\mathsf{T}}]. \end{aligned} \tag{22}$$

It can be normalized as

$$\boldsymbol{R} := \text{corr}[X_1, X_2] = (\text{diag}\,\boldsymbol{\Sigma}_1)^{-\frac{1}{2}}\boldsymbol{\Sigma}_{1,2}(\text{diag}\,\boldsymbol{\Sigma}_2)^{-\frac{1}{2}}, \tag{23}$$

where diag $\boldsymbol{\Sigma}$ denotes the diagonal matrix of $\boldsymbol{\Sigma}$. Note that diag $\boldsymbol{\Sigma}_1$ and diag $\boldsymbol{\Sigma}_2$ only include the diagonal elements of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. It causes the cross-correlation matrix $\boldsymbol{R}$ to lose information included in the inverses of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, which is important on the assumption that all the data is in a probability space. Therefore, the cross-correlation matrix is not yet delicate or precise enough to satisfy some machine learning requirements.

Finally, we consider the relation between the statistical distance matrices and the Capsule Neural Network (CapsNet), which is a novel and useful model of neural networks proposed in [23], [24]. In this paper, the statistical distance matrices are similar to the matrix of weights in CapsNet. Thus, they are viewpoint-invariant and can be used to distinguish an object no matter how much its pose has changed in the image. However, compared to CapsNet, the proposed

statistical distance matrices are more intuitive based on the distance-accumulation images and more specifiable by using the hierarchical clustering method. Therefore, the statistical distance matrices, represented as Information Mandalas, can be considered an extended version of the matrix of weights.

VI. SUMMARY

Through the experimental comparisons of object images whose pixels are considered features, we confirmed that $\boldsymbol{D}_C$-like statistical distance matrices are more effective in distinguishing objects than other distance matrices. Their distance-accumulation images showed that high-valued pixels were concentrated in the middle of the image. Moreover, we found that after the hierarchical clustering of the distance matrix, all the pixel clusters basically surround the center of the image and are arranged radially from inside to outside according to the distance value. Since these patterns are very similar to Mandalas, we refer to the statistical distance matrix with clustering as the Information Mandala. The Information Mandala is a new form of entropy. We will use it to understand convolutional neural networks in our future work.

A. DERIVATION FOR THE BHATTACHARYYA DISTANCE D_B (EQN. (7))

Let two d-dimensional random vectors X_1 and X_2 follow two normal distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively. Their corresponding probability density functions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are defined as

$$\begin{aligned} p(\boldsymbol{x}) &:= \det(2\pi\boldsymbol{\Sigma}_1)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\intercal \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right], \\ q(\boldsymbol{x}) &:= \det(2\pi\boldsymbol{\Sigma}_2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right]. \end{aligned} \tag{24}$$

The product of their square roots is written as

$$\begin{aligned} p^{\frac{1}{2}}(\boldsymbol{x})q^{\frac{1}{2}}(\boldsymbol{x}) = {} & \det(2\pi\boldsymbol{\Sigma}_1)^{-\frac{1}{4}} \det(2\pi\boldsymbol{\Sigma}_2)^{-\frac{1}{4}} \\ & \times \exp\left[-\frac{1}{4}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\intercal \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right. \\ & \left. -\frac{1}{4}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right]. \end{aligned} \tag{25}$$

By integrating Eqn. (25) in $\mathbb{R}^d$ with respect to $\boldsymbol{x}$, we obtain

$$\begin{aligned} & \int_{\mathbb{R}^d} p_X^{\frac{1}{2}}(\boldsymbol{x}) q_X^{\frac{1}{2}}(\boldsymbol{x}) d\boldsymbol{x} \\ & = \det(2\pi\boldsymbol{\Sigma}_1)^{-\frac{1}{4}} \det(2\pi\boldsymbol{\Sigma}_2)^{-\frac{1}{4}} && (26) \\ & \quad \times \exp\left[-\frac{1}{4}(\boldsymbol{\mu}_1^\intercal \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^\intercal \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\right] && (27) \\ & \quad \times \int_{\mathbb{R}^d} \exp\left[-\frac{1}{4}\boldsymbol{x}^\intercal(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}\right. \\ & \quad \left. + \frac{1}{2}(\boldsymbol{\mu}_1^\intercal \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\mu}_2^\intercal \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}\right] d\boldsymbol{x}. && (28) \end{aligned}$$

Eqns. (26), (27), and (28) can be transformed as follows. First, we define the following term:

This transformation is based on

$$\begin{aligned} (\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1} &= \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\Sigma}_2^{-1}+\boldsymbol{\Sigma}_1^{-1})^{-1}\boldsymbol{\Sigma}_1^{-1} \\ &= \boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}\boldsymbol{\Sigma}_2^{-1}, \end{aligned} \tag{30}$$

with

$$\begin{aligned} (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}) &= \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^{-1} + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^{-1} \\ &= \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)\boldsymbol{\Sigma}_2^{-1} \\ &= \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)\boldsymbol{\Sigma}_1^{-1}. \end{aligned} \tag{31}$$

Here, Eqn. (30) holds by

$$\begin{aligned} (\boldsymbol{A}+\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^\intercal)^{-1} &= \boldsymbol{A}^{-1} \\ & \quad -\boldsymbol{A}^{-1}\boldsymbol{C}(\boldsymbol{B}^{-1}+\boldsymbol{C}^\intercal\boldsymbol{A}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^\intercal\boldsymbol{A}^{-1}, \end{aligned} \tag{32}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ are all positive-definite matrices. $\boldsymbol{\Sigma}$ is a mean of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ as

$$\boldsymbol{\Sigma} := \frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}. \tag{33}$$

Then let Eqn. (28) be

$$\begin{aligned} & \int_{\mathbb{R}^d} \exp\left[-\frac{1}{2}\boldsymbol{x}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \frac{1}{2}(\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\mu}_2^\intercal\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}\right] d\boldsymbol{x} = \int_{\mathbb{R}^d} \\ & \quad \times \exp\left[-\frac{1}{2}\boldsymbol{y}^\intercal\boldsymbol{y} + \frac{1}{2}(\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\mu}_2^\intercal\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2^{-1})^{-\frac{1}{2}}\boldsymbol{y}\right] \\ & \quad \times d\left[(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2^{-1})^{-\frac{1}{2}}\boldsymbol{y}\right], \\ & = \underbrace{\text{Eqn. (29)}}_{\text{first factor}} \times \underbrace{\text{Eqn. (28)/Eqn. (29)}}_{\text{second factor}}, \end{aligned} \tag{34}$$

where

$$\boldsymbol{y} := (\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2^{-1})^{\frac{1}{2}}\boldsymbol{x}. \tag{35}$$

By multiplying Eqn. (27) and the first factor of Eqn. (34) together, we obtain

$$\begin{aligned} & \text{Eqn. (27)} \times \text{Eqn. (29)} \\ & = \exp\left[-\frac{1}{8}(\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2)\right] \\ & = \exp\left[-\frac{1}{8}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2))\right]. \end{aligned} \tag{36}$$

The second factor of Eqn. (34) is transformed as

$$\begin{aligned} & \det\left[(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2{}^{-1})^{-\frac{1}{2}}\right] \int_{\mathbb{R}^d} \\ & \quad \times \exp\left\{-\frac{1}{2}\left[\boldsymbol{y} - \frac{1}{2}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2{}^{-1})^{-\frac{1}{2}}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\right]^\intercal\right. \\ & \quad \left. \times \left[\boldsymbol{y} - \frac{1}{2}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2{}^{-1})^{-\frac{1}{2}}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\right]\right\} d\boldsymbol{y} \\ & = (2\pi)^{\frac{k}{2}} \det(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2{}^{-1})^{-\frac{1}{2}} \end{aligned} \tag{37}$$

using change-of-variable technique. Given

$$\begin{aligned} & \text{Eqn. (26)} \times \text{Eqn. (37)} \\ & = \det(2\pi\boldsymbol{\Sigma}_1)^{-\frac{1}{4}} \det(2\pi\boldsymbol{\Sigma}_2)^{-\frac{1}{4}} (2\pi)^{\frac{k}{2}} \det(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2{}^{-1})^{-\frac{1}{2}} \\ & = (2\pi)^{-\frac{k}{4}} \det\boldsymbol{\Sigma}_1^{-\frac{1}{4}} (2\pi)^{-\frac{k}{4}} \end{aligned}$$

$$\begin{aligned}
&\exp\left[\frac{1}{8}(\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\mu}_2^\intercal\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\right]\\
&\quad=\exp\left[\frac{1}{8}(\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+2\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\boldsymbol{\mu}_2^\intercal\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2)\right]\\
&\quad=\exp\left[\frac{1}{8}(2\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{I}-\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1})\boldsymbol{\mu}_1+2\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+2\boldsymbol{\mu}_2^\intercal\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{I}-\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1})\boldsymbol{\mu}_2)\right]\\
&\quad=\exp\left[\frac{1}{8}(2\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+2\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+2\boldsymbol{\mu}_2^\intercal\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2-\boldsymbol{\mu}_2^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2)\right]. \qquad (29)
\end{aligned}$$

$$\begin{aligned}
&\times\det\boldsymbol{\Sigma}_2^{-\frac{1}{4}}(2\pi)^{\frac{k}{2}}\det(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2{}^{-1})^{-\frac{1}{2}}\\
&=\det\boldsymbol{\Sigma}_1^{\frac{1}{4}}\det\boldsymbol{\Sigma}^{-\frac{1}{2}}\det\boldsymbol{\Sigma}_2^{\frac{1}{4}}, \qquad (38)
\end{aligned}$$

the Bhattacharyya distance D_B is achieved by

$$\begin{aligned}
&-\ln(\text{Eqn. (36)}\times\text{Eqn. (38)})\\
&\quad=\frac{1}{8}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\
&\qquad+\frac{1}{2}\ln\left[\det\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\det\boldsymbol{\Sigma}\det\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\right]. \qquad (39)
\end{aligned}$$

REFERENCES

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[2] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, L. M. L. Cam and J. Neyman, Eds., 1967, pp. 281–297.

[3] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. Nat. Inst. Sci., Calcutta*, vol. 2, no. 1, pp. 49–55, 1936.

[4] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, no. 1, pp. 99–109, 1943.

[5] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen," *J. für die reine und angewandte Mathematik*, vol. 1909, no. 136, pp. 210–271, 1909.

[6] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[7] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Wiley, 1959.

[8] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sums of observations," *Ann. Math. Statist.*, vol. 23, no. 4, pp. 409–507, 1952.

[9] I. B. Ayed, H.-M. Chen, K. Punithakumar, I. Ross, and S. Li, "Graph cut segmentation with a global constraint: Recovering region distribution via a bound of the Bhattacharyya measure," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3288–3295.

[10] S. M. Kang and J. W. L. Wan, "A multiscale graph cut approach to bright-field multiple cell image segmentation using a Bhattacharyya measure," *Proc. SPIE*, vol. 8669, Mar. 2013, Art. no. 86693S.

[11] C. C. Reyes-Aldasoro and A. Bhalerao, "The Bhattacharyya space for feature selection and its application to texture segmentation," *Pattern Recognit.*, vol. 39, no. 5, pp. 812–826, May 2006.

[12] K. G. P. Derpanis and R. Wildes, "Spacetime texture representation and recognition based on a spatiotemporal orientation analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1193–1205, Jun. 2012.

[13] D. Shirosawa, X. Lu, and A. Kimura, "A performance evaluation of variation-HOG descriptor for human face detection," in *Proc. Int. Workshop Adv. Image Texhnology*, 2017, p. 4.

[14] C. H. You, K. A. Lee, and H. Li, "An SVM Kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 49–52, Jan. 2009.

[15] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes, "Action spotting and recognition based on a spatiotemporal orientation analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 527–540, Mar. 2013.

[16] F. Harary, R. Z. Norman, and D. Cartwright, *Structural Models: An Introduction to the Theory of Directed Graphs*. New York, NY, USA: Wiley, 1965.

[17] O. Maimon and L. Rokach, Eds., *Data Mining and Knowledge Discovery Handbook*. New York, NY, USA: Springer, 2010.

[18] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," Sep. 2011, *arXiv:1109.2378*. [Online]. Available: http://arxiv.org/abs/1109.2378

[19] A. Krizhevsky, V. Nair, and G. E. Hinton. *CIFAR-10 (Canadian Institute for Advanced Research)*. Accessed: Feb. 2020. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[20] K. Nishiyama and X. Lu, "A novel view of color-based visual tracker using principal component analysis," *IEICE Trans. Fundam.*, vol. 19, no. 12, pp. 3843–3848, 2008.

[21] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[22] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 2009.

[23] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3856–3866.

[24] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15. [Online]. Available: https://openreview.net/forum?id=HJWLfGWRb

XIN LU (Member, IEEE) received the B.E. degree in computer science and information technology from Hohai University, China, in 2000, and the M.E. and Dr.Eng. degrees in information science from the University of Tokushima, in 2003 and 2006, respectively. Since 2006, he has been an Assistant Professor with the Faculty of Science and Engineering, Iwate University, Japan. His research interests include statistics, machine learning, and signal processing.