# Nonlinear Ranking Loss on Riemannian Potato Embedding

Byung Hyung Kim
School of Computing
KAIST
bhyung@kaist.ac.kr

Yoon-Je Suh
School of Electrical Engineering
KAIST
yoonje@kaist.ac.kr

Honggu Lee
Looxid Labs
honggu.lee@looxidlabs.com

Sungho Jo
School of Computing
KAIST
shjo@kaist.ac.kr

***Abstract*—We propose a rank-based metric learning method by leveraging a concept of the *Riemannian Potato* for better separating non-linear data. By exploring the geometric properties of Riemannian manifolds, the proposed loss function optimizes the measure of dispersion using the distribution of Riemannian distances between a reference sample and neighbors and builds a ranked list according to the similarities. We show the proposed function can learn a hypersphere for each class, preserving the similarity structure inside it on Riemannian manifold. As a result, compared with Euclidean distance-based metric, our method can further jointly reduce the intra-class distances and enlarge the inter-class distances for learned features, consistently outperforming state-of-the-art methods on three widely used non-linear datasets.**

## I. Introduction

Recent research on deep metric learning(DML) has shown significant benefits by learning semantic distance on a non-linear embedding space using deep neural networks. With the development of these loss functions, a large corpus of literature has been presented in various areas such as face recognition, image clustering, and retrieval [1]. Loss function in DML plays a crucial role in learning similarity on a manifold, and many methods such as contrastive loss [2], triplet loss [3], lifted structured embedding [4], N-pair Loss method [5] have been proposed in the literature. All these learning methods take a Euclidean distance metric to measure the distance between paired examples in $n$-dimensional feature vector space. However, many scientific fields study data with an underlying structure that is a non-Euclidean space. Some real-world applications include social networks in computational social sciences, sensor networks in communications, functional networks in brain imaging, regulatory networks in genetics, and meshed surfaces in computer graphics. Hence directly applying the Euclidean-based state-of-the-art approaches often results in poor or less informative performance. Furthermore, Euclidean distance cannot preserve the correlation and the drawback limits to understand non-stationary data. To overcome this problem, we focus on developing a novel method on non-Euclidean space.

Interestingly, relatively simple machine learning techniques can produce state-of-the-art results as soon as the particular Riemannian geometry is taken into account [6], [7]. Although the overall structure of the metric learning is preserved in the context of neural networks, its generalization to Riemannian manifold requires geometric tools on the manifold. In this work, we further assess the particular interest of metric learning for Riemannian manifold in the context of learning on scarce non-linear data with lightweight models. We focus on the original architecture proposed in [8].

In this paper, we devise a new non-linear rank loss by leveraging a concept of the *Riemannian Potato* (RP) defined in [9]. The RP provides a measure of dispersion using the distribution of distances between covariance matrices and a reference matrix and rejects epochs whose covariance matrices lie out of a region of acceptability defined by a $z$-score threshold. Inspired by the principle of the RP, our loss function aims to pull positive points closer than the potato-shaped region of acceptability ($z$-score) and push negative points out of the boundary. We find that our loss on a Riemannian neural network [8] with rank perspective has better performance than current Euclidean-based DML approaches for learning discriminative non-linear embeddings. Our approach can further jointly reduce the intra-class distances and enlarge the inter-class distances for the learned features, and preserve the correlations of the non-linear features. As such, our contributions are the following:

- We propose a new ranking loss to learn discriminative embeddings on non-Euclidean spaces. By leveraging the concept of *Riemannian Potato*, we exploit the structure of non-linear embedding spaces.
- We achieve new state-of-the-art performance on three popular benchmarks, reducing the intra-class distances and enlarging the inter-class distances for the learned features.

## II. Related Work

### A. Metric Learning Methods

Metric learning aims to learn an embedding space, where the similar samples are encouraged to be closer, while dissimilar ones are pushed apart from each other. We give a brief review of the-state-of-the-art methods below, and we compare our proposed loss function to them experimentally in Section 4.

Contrastive Loss [2] aims to minimize the distance between two samples $f(x_i)$ and $f(x_j)$ if they are categorized in the same label and to maximize otherwise. Triplet Loss [3] takes

a set of triplets, $i$, $j$, and $k$ are the indexes of anchor, positive, and negative points, respectively. The loss function aims to pull the anchor point closer to the positive than to the negative point by a fixed margin $m$.

$$L(X, y) = \frac{1}{\mathcal{T}} \sum_{(i,j,k)\in\mathcal{T}} [d^2_{(i,j)} - d^2_{(i,k)} + \alpha]_+, \quad (1)$$

where $\mathcal{T}$ is the set of triplets, $d_{(i,j)}$ is the Euclidean distance between the two points. The operation $[\cdot]_+$ denotes the hinge function, and $m$ denotes a fixed margin constant.

N-pair-mc [5] interacts with more negative examples. The loss function aims to identify one positive point from $N-1$ negative points of $N-1$ classes.

$$L(X, y) = \frac{\lambda}{N} \sum_{i}^{N} \|f(X_i)\|_2 - \frac{1}{\mathcal{P}} \sum_{(i,j)\in\mathcal{P}} \log \frac{\Xi(X_i)}{\Xi(X_i) + \sum_{k:y[k]\neq y[i]} \Xi(X_i)}, \quad (2)$$

where $\Xi(X_i) = \exp\{f(X_i)^\intercal f(X_j)\}$, $N$ is the number of the samples, and $\lambda$ is the regularization constant on the embedding vectors.

Lifted Struct [4] argued that the contrastive and triplet loss function are challenging to explore full pair-wise relations between samples in a mini-batches. The objective of Lifted Struct is to pull one positive pair$(x_i, x_j)$ as close as possible and push associated negative points farther than a margin $m$

$$L(X, y) = \frac{1}{2|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} [\log( \sum_{(i,k)\in\mathcal{N}} \exp\{\alpha - d_{(i,k)}\} + \sum_{(j,l)\in\mathcal{N}} \exp\{\alpha - d_{(j,l)}\}) + d_{(i,j)}]^2_+, \quad (3)$$

where $\mathcal{N}$ denotes the set of pair-wise examples with different labels.

Recently, many efforts have been devoted to devising new loss functions to learn a non-linear embedding of data. Schall *et al.* [10] proposed a rank-based approximation with a non-linear transfer function. Xu *et al.* [11] presented a kernel-based approximation approach to capture non-linear relationships between samples. Another line of DML which is related to our approach is to use statistics of data. Yuan *et al.* [12] proposed a distance metric by leveraging a concept of the signal-to-noise ratio (DSML), denoting anchor features as signals and other features as noisy signals.

### *B. Riemannian Geometry in Metric Learning*

A Riemannian manifold is a differential manifold equipped with a varying inner product smoothly on each tangent space. Riemannian metrics of the manifold are the family of inner products on all tangent spaces. Given two points in a curved space, the shortest path can be defined by minimizing the length of a curve between the points, geodesic distance, which is analogous to straight lines in Euclidean space, but a more natural measure between the two points than the Euclidean distance. Several metrics have been presented to capture its non-linearity such as the affine-invariant metric (AIM) [13], log-Euclidean metric (LEM) [14], Stein divergence [15], Burg matrix divergence [16], and alpha-beta divergence [17]. The two most widely used distance measures as true geodesic distances induced by Riemannian metrics are the log-Euclidean distance

$$\delta_L(\Sigma_1, \Sigma_2) = \|\log(\Sigma_1) - \log(\Sigma_2)\|_F \quad (4)$$

and the affine-invariant distance.

$$\delta_R(\Sigma_1, \Sigma_2) = \|\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})\|_F = (\sum_{c=1}^{C} \log^2\lambda_c)^{1/2}, \quad (5)$$

where $\lambda_c$, $c = 1, \ldots, C$ are the eigenvalues of $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$.

Most existing methods learn the SPD distance measure in a more discriminative space such as the Euclidean tangent space [18], [19], [20] and the reproducing kernel Hilbert space (RKHS) [21]. Vemulapalli and Jacobs [18] proposed a Mahalanobis-based metric learning method on the tangent space. Huang *et al.* [19] introduced the LEM learning (LEML) approach to transform the matrix on the tangent space to other tangent spaces. Zhou *et al.* [20] presented a sample-specific version of LEML named $\alpha$-based covariance-like metric learning ($\alpha$-CML) that learns to adjust eigenvalues of the SPD matrix for more discriminative power. Besides, several metric learning methods [22], [23] combine the discriminative abilities of multiple types of manifold representations into RKHS. Quang *et al.* [21] generalized the LEM between two finite-dimensional SPD matrices to infinite-dimensional covariance matrices in RKHS by Hilbert–Schmidt operators. To overcome the inaccurate approximation of the Euclidean space and preserve the SPD manifold structure, Harandi *et al.* [24] proposed projecting the high-dimensional SPD matrix into a low-dimensional manifold and learning a metric in the new manifold.

Although various metric learning approaches have been studied on non-linear manifolds, most existing methods are notoriously computation-heavy on optimization problems. To overcome this problem, we devise a new rank-based metric learning method in the context of deep neural networks, learning on scarce data with lightweight models.

## III. METHODOLOGY

### *A. Principle of the Riemannian Potato*

The *Riemannian Potato* (RP) is a multivariate adaptive method for identifying artifacts in continuous data [9]. A potato-shaped region of acceptability is induced by the non-linearity of Riemannian manifold. Since the geometry of covariance matrices captures multivariate second-order statistics of data, the RP is based on covariance matrices that are symmetric positive definite (SPD) processed in a Riemannian manifold. The principle of the RP is to represent clean signals

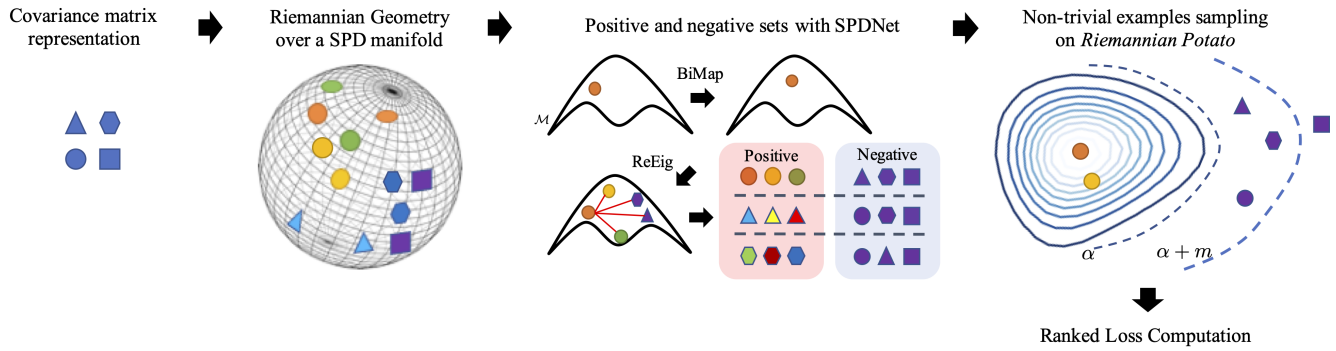


Fig. 1. The overview deep learning framework of our proposed ranking loss with SPDNet. For each input mini-batch, every SPD matrix acts as a query and obtains a list ranked by the $z$-scores on RP embeddings. For each ranked list, we sample and group less trivial samples into the positive and negative sets based on their margin losses with respect to the query. At last, our proposed loss is computed with the joint supervision of positive and negative sets for every query.

by estimating a reference covariance matrix and a measure of dispersion ($z$-score) for a epoch $t$,

$$z_t = \frac{\log(d_t/\mu_t)}{\log(\sigma_t)}, \tag{6}$$

where $d_t = \delta_R(\Sigma_t, \bar{\Sigma}_{t-1})$ is the Riemannian distance between the current covariance matrix $\Sigma_t$ and the reference matrix $\bar{\Sigma}_{t-1}$, which is the geometric mean of $i \in [0, I]$ numbers of $\Sigma_i$. The reference matrix $\bar{\Sigma}$ can be defined by minimizing the dispersion of $\Sigma$ on the manifold $\mathcal{M}$ such as the sum of squared distances:

$$\bar{\Sigma} = \arg\min_{\Sigma \in \mathcal{M}} \sum_{i=1}^{N_I} \delta_R^2(\Sigma_i, \Sigma), \tag{7}$$

where $N_I$ can be the maximum number of iterations. The mean $\mu$ and the standard deviation $\sigma$ are

$$\mu_t = \exp(\frac{1}{t}\sum_{i=1}^{t}\log(d_i)), \tag{8}$$

$$\sigma_t = \exp(\sqrt{\frac{1}{t}\sum_{i=1}^{t}(\log(d_i/\mu_i))^2}. \tag{9}$$

Then, the RP rejects all artifacts whose covariance matrices are too far from the reference $\bar{\Sigma}$ according to an objective statistical criterion based on $z$-score threshold $z_{th}$. Typically it is 2.5 to define the hull of acceptability, rejecting around 0.6% of data under Gaussian assumption. We should note that arithmetic definitions in the original RP [9] are not optimal because Riemannian distances are not normally distributed. Since distances empirically follow a non-negative highly right-skewed distribution, we modeled them in Eq. (6), (8), and (9) by a log-normal or chi-squared distribution [25].

### *B. Riemannian Potato-based Ranking Loss (RPL)*

Inspired by the principle of the RP, we define the hull of acceptability by estimating a reference matrix $\bar{\Sigma}$ with positive pairs which are the same classes as a anchor covariance matrix. Hence, adding a superscript $c$ to index the $C$ classes in Eq. (6) $\sim$ (9), our loss function, RPL, aims to pull positive samples from the same class with $c$ closer than a predefined RP threshold $z_{th}$ and push negative samples out of the boundary, separating the positive and negative sets by a margin $m$. Given a SPD matrix $\Sigma_i$ and the associated label $c = y_i \in C$, $m$ is the margin between the two boundaries as follows:

$$\begin{aligned} L(\Sigma_i, \Sigma_j, y_i; f) = (1 - y_{ij})[z_{th} - z_j^c)] \\ + y_{ij}[z_j^c - (z_{th} + m)]_+, \end{aligned} \tag{10}$$

where $y_{ij} = 1$ if $c = y_i = y_j$ and $y_{ij} = 0$ otherwise. $z_j^c$ is the $z$-score of $\Sigma_j$ for a class $c$ with the AIR distance between two points $\delta_R(\Sigma_j, \bar{\Sigma}_{i-1}^c)$ in Eq. (5).

*1) Mining Strategy on the Riemannian Potato Embedding:* High retrieval quality does not depend on the actual distances, but rather on the ranking order of the features from similar examples. Hence, given a query $\Sigma_i$, we rank all other sample points according to their similarities to the query (See Fig. 1). In each class $c \in C$, positive samples in the positive set $\mathcal{P}_i^c$ and negative samples in the negative set $\mathcal{N}_i^c$ are given by

$$\mathcal{P}_i^c = \{\forall \Sigma_j | j \neq i \wedge c = y_i = y_j\}, \tag{11}$$

$$\mathcal{N}_i^c = \{\forall \Sigma_j | c = y_i, y_i \neq y_j\}. \tag{12}$$

Metric learning methods have adopted sampling strategies for pairs and triplets in neighborhood relationships from class labels. Similar as in [26], we focus on less trivial samples which have non-zero losses in violation of the pairwise similarity for the retrieval problem. Furthermore, our strategy retrieves samples on the class level since instance-based sampling cannot guarantee that each example has at least one neighbor in the same minibatch. We denote the sets of non-trivial positive $\hat{\mathcal{P}}_i^c$ and negative $\hat{\mathcal{N}}_i^c$ samples with respect to a query $\Sigma_i$ as

$$\hat{\mathcal{P}}_i^c = \{\forall \Sigma_j | j \neq i \wedge c = y_i = y_j, z_j^c > z_{th}\}, \tag{13}$$

$$\hat{\mathcal{N}}_i^c = \{\forall \Sigma_j | c = y_i, y_i \neq y_j, z_j^c < z_{th} + m\}. \tag{14}$$

*2) Joint Loss:* A perfect clustering can be achieved if and only if all distance to negative examples are larger than a boundary $z_{th}$. Consequently, all samples from the same class are grouped into a hypersphere with $z_{th}$ diameter. To pull all non-trivial positive points in $\hat{\mathcal{P}}$ together and learn a class hypersphere, we minimize:

$$L_P(\Sigma_i, y_i; f) = \frac{1}{|\hat{\mathcal{P}}_i^c|} \sum_{\Sigma_j \in \hat{\mathcal{P}}_i^c} L(\Sigma_i, \Sigma_j, y_i; f) \quad (15)$$

To push the non-trivial negative points in $\hat{\mathcal{N}}$, beyond the boundary $z_{th} + m$, we minimize:

$$L_N(\Sigma_i, y_i; f) = \frac{1}{|\hat{\mathcal{N}}_i^c|} \sum_{\Sigma_j \in \hat{\mathcal{N}}_i^c} L(\Sigma_i, \Sigma_j, y_i; f) \quad (16)$$

In RPL, we adopt the joint supervision of the two objective functions to enhance the discriminative power of deep features as follows:

$$L_{RP}(\Sigma_i, y_i; f) = L_P(\Sigma_i, y_i; f) + \lambda L_N(\Sigma_i, y_i; f), \quad (17)$$

where $\lambda$ controls the balance between positive and negative sets. With the joint supervision, not only the inter-class features differences are enlarged, but also the variations of the intra-class feature are reduced.

### C. Stochastic RP Optimization on Mini-Batches

We optimize the proposed RPL using stochastic gradient descent (SGD) with mini-batches. Each mini-batch is randomly sampled from the whole training classes, emits one RPL value, and the overall objective is the average of the RPL values as follows:

$$\bar{L}_{RP} = \frac{1}{N} \sum L_{RP}(\Sigma_i, y_i; f), \quad (18)$$

where $N$ is the batch size and geometric statistics of RP are updated as

$$\mu_i = \exp((1 - \beta_\delta)\log(\mu_{i-1}) + \beta_\delta \log(d_i)), \quad (19)$$

$$\sigma_i = \exp(\sqrt{(1 - \beta_\delta)(\log(\sigma_{i-1})^2 + \beta_\delta(\log(d_i/\mu_i))^2}), \quad (20)$$

$$\bar{\Sigma}_i = \bar{\Sigma}_{i-1}^{\frac{1}{2}} (\bar{\Sigma}_{i-1}^{-\frac{1}{2}} \Sigma_i \bar{\Sigma}_{i-1}^{-\frac{1}{2}})^{\beta_\delta} \bar{\Sigma}_{i-1}^{\frac{1}{2}}, \quad (21)$$

where $\beta_\delta \in [0, 1]$ defines the learning rate for adaptation in online implementations. A hyper-parameter $N_I$ in Eq. (7) defines the number of positive covariance matrices used for initializing the region of acceptability to model an accurate estimate for the mean and the distribution of distances to it, which will significantly influence the retrieval performance. Otherwise, the RP will be inefficient to separate negative examples. For the calibration method to initialize the region of acceptability, we uniformly sample the $N_I$ numbers of samples for each class $c$ and estimate the geometric statistics $\mu$, $\sigma$, and $\bar{\Sigma}$ Eq. (7) $\sim$ (9) before training.

Since the geometric mean has no closed form and is negatively affected by ill-conditioned input matrices [27], we use a gradient descent algorithm [28] in stochastic optimization on mini-batches for solving the minimization of sum-of-squared distance function. The proposed loss function on one mini-batch is summarized in Algorithm 1, and the overall pipeline for our loss on SPDnet is shown in Fig. 1.

**Algorithm 1** Riemannian Potato-based Ranking Loss

**Input:** $\{\{z_i^c\}_{i=1}^{N_c}\}_{c=1}^{C} = \{(\Sigma_i, y_i)\}_{i=1}^{N}$, the embedding function $f(\cdot)$, the learning rate $\beta_L$ and $\beta_\delta$

**Output:** Updated $f(\cdot)$

1: **for** all embeddings $f(z_i^c) \in \{\{f(z_i^c)\}_{i=1}^{N_c}\}_{c=1}^{C}$ **do**
2: Sample less trivial positive points in $\mathcal{P}_{c,i}^*$.
3: Sample less trivial negative points in $\mathcal{N}_{c,i}^*$.
4: Compute the joint loss in Eq. (17).
Update geometric statistics of the RP in Eq.(19) $\sim$ (21).
5: **end for**
Compute the averaged loss in Eq. (18).
Compute the gradient $\nabla f = \partial \bar{L}_{RP}(\Sigma_i; f)/\partial f$.
6: **return** $f(\cdot) = f(\cdot) - \beta_\delta \cdot \nabla f$

## IV. Experiment

We evaluate the proposed loss function on different tasks: emotion classification from EEG [29] and facial images [30], and skeleton-based human action recognition [31] from images using three popular datasets, where SPD matrix representation has achieved great success.

- DEAP [29]: Database for Emotion Analysis Using Physiological Signals (DEAP) is a large-scale dataset for 32-channel EEG-based emotion recognition. This dataset contains EEG signals of 32 participants; each participant watched 40 one-minute-long excerpts of music videos. The dataset contained continuous valence ratings on scales from 1 to 9 rated directly after each trial. We grouped the continuous valence labels into $k$ discrete states, denoting as DEAP-k. For instance, DEAP-3 comprises negative ($1 \sim 3$), neutral ($4 \sim 6$), and positive ($7 \sim 9$) valence ratings. As similar in [32], for each channel, all EEG signals are first band-pass filtered with a bandwidth of $4 \sim 47$Hz, and electrode-wise exponential moving standardization is then performed to compute exponential moving means and variances, both of which are used to standardize the continuous data. As a result, each EEG signal is represented by a $32 \times 32$ SPD matrix.
- AFEW [30]: Acted Facial Expression in the Wild (AFEW) dataset contains 1,345 video sequences of seven facial expressions acted by 330 actors in movies. To evaluate the performance on the validation set, we follow the setting in [8]. Each facial frame is normalized to an image of size $20 \times 20$ for extracting the covariance matrix feature of size $64 \times 64$.
- HDM05 [31]: Hochschule der Medien (HDM05) is a large-scale dataset for the problem of skeleton-based human action recognition. The dataset contains 2,337 sequences of 130 action classes, which provides 3D locations of 31 joints of the subjects. As similar in [8], we divide the training sequences set to around 18,000 small sub-sequences in each random evaluation and represent

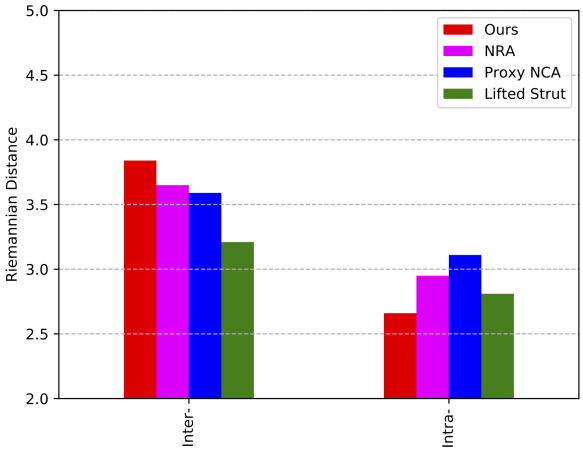


Fig. 2. The distribution of deeply learned SPD features (Inter- and Intra-distances) on DEAP dataset.

each sequence by a joint covariance descriptor(SPD matrix) of size $93\times93$, which is calculated by a second-order statistics of the 3D coordinates for the 31 joints in each frame.

### A. Experimental Setup

We compare our RPL to a series of state-of-the-art methods which are tested under the same setting: Triplet-Random, Triplet-Semihard, Lifted Struct, N-Pair, NRA, and DSML-Triplet. All methods have been described in Section 2. For a fair comparison, learning rate $\beta_L$ was set to $1 \times 10^{-3}$ with $5\times10^{-4}$ weight decay. $\beta_\delta$ was set to $1/t$ with the initialization of $N_I = 100$, giving a cumulative moving average. The batch size was set to 150, the weights were initialized as random semi-orthogonal matrices, and the rectification threshold $\epsilon$ was set to $10^{-4}$. Early-stopping during validation with a fixed patience size was adopted to prevent an overfitting in learning the deep features. We report the retrieval performance and the clustering quality in terms of F1 score, Recall@K, and NMI. SPDNet was used as our backbone network [8]. For AFEW and HDM05, three BiMap layers and two ReEig layers are configured. The parameters on AFEW are set to $400 \times 200$, $200 \times 100$, and $100 \times 50$, respectively. The parameters on HDM05 are set to $93 \times 80$, $80 \times 40$, and $40 \times 20$, respectively. For the DEAP dataset, we configured one BiMap Layer and one ReEig Layer as the feature extractor, whose parameters are set to $64 \times 32$. For other parameters of the compared methods, we empirically set the best parameters with the highest accuracy based on the original study.

### B. Experimental Results

Table I, II, and Fig. 2 report the performance on the three different datasets. We have the following observations.

- Overall, our method outperformed all the compared approaches, which validates the effectiveness of our proposed loss function. Particularly, the F1 and R@1 on the three datasets were higher than previous state-of-the-art methods. Among the baselines, the triplet loss with uniform sampling always performed the worst. These results support the significance of mining positive and negative examples on non-Euclidean space. Furthermore, when triplet loss with hard negatives, the results became poor severely in R@1, R@3, and NMI. This bad performance on triplet loss implies the loss may waste gradient update on SPD matrices far away from the decision boundary.
- Discriminative power of our loss function can be found in Fig. 2, which showed the pairwise distances between the Riemannian centers of each class and matrices within a class (intra-class distances), and with different classes (inter-class distances). The result indicates that the features learned by our loss exhibit more clear discriminative structures, while the other loss presents relatively vague structures. Although the NMI result from our method slightly underperforms with comparison to the Proxy NCA on the DEAP dataset, the encouraging performances of our loss in Table II and Fig. 2 show that our Riemannian distance-based metric learning approach has more power to enlarge the inter-class distances and reduce the intra-class distances than the traditional Euclidean distance metric such as Lifted Struct and NRA. While the Proxy NCA had a good performance in inter-class distance, the softmax-based computation was not discriminative enough to reduce intra-class variations. The Lifted Struct method had similar performance to reduce the intra-class distance, but their Euclidean-based distance metric was not accurate enough to enlarge inter-class distance. The proposed method on SPD matrices led to better performance when the study on reducing the semantic distances between continuous labels was conducted.
- Different modalities on the three datasets had different effects on learning their characteristics by discriminating each of $f$ functions. Except for the proposed loss, none of the other methods always wins due to the difference in dataset properties, including class imbalance, noisy labels, etc. Most loss functions had significant difficulty in learning non-stationary EEG signals on the DEAP dataset. This result indicates applying non-linear and non-stationary data to Euclidean-based metric learning leads to incorrect measurement of semantic distances. On the other hand, our method consistently shows the superiority to Euclidean-based metric learning methods.

### C. Ablation Studies

*1) The Effect of Batch Size:* The batch size is critical element for the model performance in deep metric learning since it determines the size of retrieval problems during training. In our loss, the batch size decides the number of negative class labels and influences to shape the RP in Eq. (8) and (9). In order to study the influence of batch size in our approach, we fix the number of SPD matrices per class and only vary the size of batch size and measure the R@1

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON DEAP, AFEW, AND HDM05 IN TERMS OF F1 (%), RECALL@K (%) AND NMI (%). ALL THE COMPARED METHODS USE SPDNET AS THE BACKBONE ARCHITECTURE.

| Method | DEAP-4 | | | | AFEW | | | | HDM05 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | R@1 | R@3 | NMI | F1 | R@1 | R@3 | NMI | F1 | R@1 | R@3 | NMI |
| DSML-Triplet | 38.7 | 35.5 | 37.8 | 36.1 | 29.3 | 31.3 | 34.2 | 31.9 | 53 | 57.3 | 58.5 | 52.4 |
| Triplet-Random | 33.5 | 31.4 | 32.5 | 28.7 | 25.8 | 25.1 | 25.4 | 27.4 | 48.3 | 44.5 | 47.5 | 45.6 |
| Triplet-Semihard | 35.5 | 30.1 | 31.4 | 27.3 | 27.4 | 24.4 | 30.5 | 28 | 51.3 | 50.4 | 55.3 | 47.5 |
| Lifted Struct | 35.3 | 35.2 | 35.8 | 33.4 | 32.5 | 35.4 | 38.4 | 39.4 | 55.5 | 59.3 | 59.2 | 53.4 |
| N-pair-mc | 41.5 | 38.4 | 39.5 | 34.8 | 34.4 | 33.5 | 36.4 | 35.1 | 59.8 | 60 | 61 | 59.4 |
| Proxy NCA | 39.8 | 41.3 | 41.4 | **38.1** | 34.5 | 34.2 | 36 | 36.3 | 59 | 63.3 | 64.5 | 62 |
| NRA | 42.2 | 44.4 | **46.2** | 37.2 | 35.2 | **36.5** | 38.6 | **36.8** | 59.2 | 64.3 | 65.2 | 64.1 |
| **RPL** | **43.3** | **44.7** | **46.2** | 37.5 | **36.4** | **36.5** | **39.4** | 36.4 | **59.4** | **66.7** | **68.8** | **65.4** |

TABLE II
COMPARISON WITH THE STATE-THE-OF-ART METHODS ON DIFFERENT DEAP DATASET SETTINGS. THE EVALUATION SETTINGS FOLLOW TABLE I. DEAP-K GROUPED THE CONTINUOUS VALENCE LABELS INTO $k$ DISCRETE STATES.

| Method | DEAP-2 | | DEAP-3 | | DEAP-9 | |
|---|---|---|---|---|---|---|
| | F1 | R@1 | F1 | R@1 | F1 | R@1 |
| DSML-Triplet | 54.3 | 57.9 | 40.4 | 39.1 | 20.4 | 19.7 |
| Triplet-Random | 46.2 | 52.7 | 36.8 | 35.4 | 14.3 | 13.6 |
| Triplet-Semihard | 53.1 | 55 | 38.4 | 37.3 | 17.4 | 18.4 |
| Lifted Struct | 56.4 | 59.2 | 40.8 | 39.4 | 21.5 | 19.4 |
| N-pair-mc | 59.6 | 58.4 | 43.5 | 42.8 | 20.4 | 21.2 |
| Proxy NCA | 58.9 | 61.3 | 43.4 | 43.1 | 20.5 | 19.2 |
| NRA | 60.5 | 63.4 | 44.6 | 47.1 | 21.1 | 20.9 |
| **RPL** | **63.8** | **66.9** | **45.2** | **47.5** | **23.4** | **23.5** |

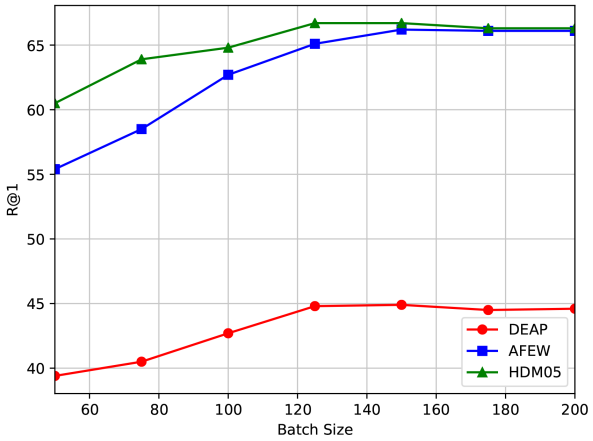


Fig. 3. Recall@1 results of different batch size on the three datasets (DEAP, AFEW, and HDM05).

on all three datasets. The results are reported in Fig. 3. We observe R@1 monotonically improves with larger batch size. This may resonate with the fact that large batches reduce the variance of the stochastic gradients and statistic values in the RP.

*2) The Effect of the Parameter $\lambda$:* The $\lambda$ is a hyper-parameter of RPL to dominate negative sets. It determines the balance of the two sets in discriminative feature learning. To investigate the sensitiveness of the parameter, we vary $\lambda$ from 0.1 to 1, keeping other parameters fixed. Fig. 4 shows the impact on test set performance in terms of R@1. Intuitively, increasing the value of $\lambda$ during training would improve the classification performance of the deeply learned features. We also observe that the performance remains largely stable across a wide range of $\lambda$.

*3) Geometric Statistics on Online RP:* Another hyper-parameter of RPL is the maximum number of iteration $N_I$ for initializing the RP to reach convergence on geometric statistics in Eq. (7). It determines the robustness of representing positive sets $\mathcal{P}$ as RP. To study its effect, we also vary numbers of $N_I$, keeping other parameters fixed. Fig. 5 shows the impact on test set performance in terms of R@1 and the convergence curve on the DEAP dataset. Intuitively, increasing the number of $N_I$ during training would result in a robust approximation of RP, converging quickly. However, we observe that retrieval performance and the convergence have not necessarily improved with more epochs.

## V. CONCLUSION

In this paper, we proposed a rank-based metric learning method for learning discriminative embeddings and showed the efficacy on classifying non-linear data. Given a query covariance matrix, our RPL splits its positive and negative sets and forces a margin between them on a SPD manifold. In addition, non-trivial samples mining and negative examples weighting are exploited to make better use of all informative data points. The proposed method achieves state-of-the-art performance, reducing the intra-class distances and enlarging the inter-class distances for learned features. Our next work will study the non-stationary nature of brain activity as revealed by EEG, which has been subject to noises from various artifacts,

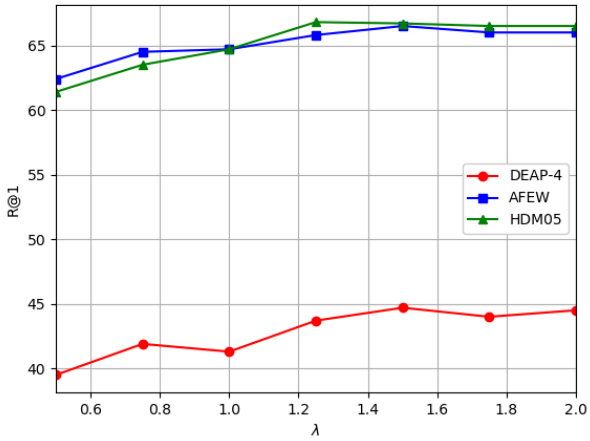


Fig. 4. Recall@1 results of different $\lambda$ on the three datasets (DEAP, AFEW, and HDM05).

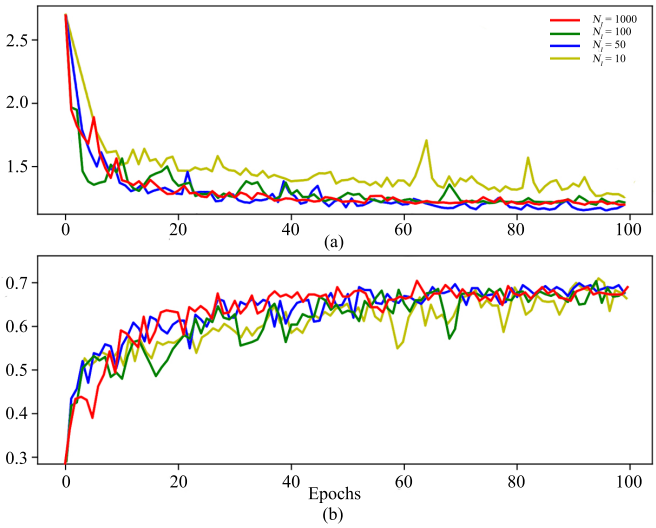


Fig. 5. The (a) convergence curves (loss) and (b) the Recall@1 performance of different initialization $N_I$ on the DEAP-2 dataset during training

low signal-to-noise ratio (SNR) of sensors, and inter- and intra-subject variability. Hence, we will investigate the efficacy of RP-based metric learning for discriminating EEG signals and show the progression patterns of the classification in training on additional datasets such as EEGBCI [33].

## ACKNOWLEDGMENT

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00432)

## REFERENCES

[1] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. P. Cohen, "Revisiting training strategies and generalization performance in deep metric learning," in *International Conference on Machine Learning*, 2020.

[2] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 1735–1742.

[3] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[4] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.

[5] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, 2016, pp. 1857–1865.

[6] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2011.

[7] H. Liu, J. Li, Y. Wu, and R. Ji, "Learning neural bag-of-matrix-summarization with riemannian network," in *AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8746–8753.

[8] Z. Huang and L. Van Gool, "A riemannian network for spd matrix learning," in *AAAI Conference on Artificial Intelligence*, 2017.

[9] A. Barachant, A. Andreev, and M. Congedo, "The riemannian potato: an automatic and adaptive artifact detection method for online experiments using riemannian geometry," in *TOBI Workshop lV*, 2013, pp. 19–20.

[10] K. Schall, K. U. Barthel, N. Hezel, and K. Jung, "Deep metric learning using similarities from nonlinear rank approximations," in *IEEE International Workshop on Multimedia Signal Processing*. IEEE, 2019, pp. 1–6.

[11] Y. Xu, C. Miao, Y. Liu, H. Song, Y. Hu, and H. Min, "Kernel-target alignment based non-linear metric learning," *Neurocomputing*, 2020.

[12] T. Yuan, W. Deng, J. Tang, Y. Tang, and B. Chen, "Signal-to-noise ratio: A robust distance metric for deep metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4815–4824.

[13] X. Pennec, P. Fillard, and N. Ayache, "A riemannian framework for tensor computing," *International Journal of computer vision*, vol. 66, no. 1, pp. 41–66, 2006.

[14] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 56, no. 2, pp. 411–421, 2006.

[15] S. Sra, "A new metric on the manifold of kernel matrices with application to matrix geometric means," in *Advances in Neural Information Processing Systems*, 2012, pp. 144–152.

[16] B. Kulis, M. Sustik, and I. Dhillon, "Learning low-rank kernel matrices," in *International Conference on Machine Learning*, 2006, pp. 505–512.

[17] A. Cichocki, S. Cruces, and S.-i. Amari, "Log-determinant divergences revisited: Alpha-beta and gamma log-det divergences," *Entropy*, vol. 17, no. 5, pp. 2988–3034, 2015.

[18] R. Vemulapalli and D. W. Jacobs, "Riemannian metric learning for symmetric positive definite matrices," *arXiv preprint arXiv:1501.02393*, 2015.

[19] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen, "Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification," in *International Conference on Machine Learning*, 2015, pp. 720–729.

[20] L. Zhou, L. Wang, J. Zhang, Y. Shi, and Y. Gao, "Revisiting metric learning for spd matrix based visual representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3241–3249.

[21] M. H. Quang, M. San Biagio, and V. Murino, "Log-hilbert-schmidt metric between positive definite operators on hilbert spaces," in *Advances in Neural Information Processing Systems*, 2014, pp. 388–396.

[22] Z. Huang, R. Wang, S. Shan, L. Van Gool, and X. Chen, "Cross euclidean-to-riemannian metric learning with application to face recognition from video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2827–2840, 2017.

[23] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *International Conference on Multimodal Interaction*, 2014, pp. 494–501.

[24] M. Harandi, M. Salzmann, and R. Hartley, "Dimensionality reduction on spd manifolds: The emergence of geometry-aware methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 48–62, 2017.

[25] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 155–174, 2017.
[26] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5207–5216.
[27] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.
[28] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Transactions on Medical Imaging*, vol. 23, no. 8, pp. 995–1005, 2004.
[29] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.
[30] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *International Conference on Multimodal Interaction*, 2014, pp. 461–466.
[31] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," Universität Bonn, Tech. Rep. CG-2007-2, June 2007.
[32] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
[33] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.