Deep Metric Learning on the SPD Manifold for Image Set Classification

Rui Wang, Xiao-Jun Wu*, Tianyang Xu, Cong Hu, and Josef Kittler, *Life Member, IEEE*

***Abstract*—Thanks to the efficacy of Symmetric Positive Definite (SPD) manifold in characterizing video sequences (image sets), image set-based visual classification has made remarkable progress. However, the issue of large intra-class diversity and inter-class similarity is still an open challenge for the research community. Although several recent studies have alleviated the above issue by constructing Riemannian neural networks for SPD matrix nonlinear processing, the degradation of structural information during multi-stage feature transformation impedes them from going deeper. Besides, a single cross-entropy loss is insufficient for discriminative learning as it neglects the peculiarities of data distribution. To this end, this paper develops a novel framework for image set classification. Specifically, we first choose a mainstream neural network built on the SPD manifold (SPDNet) [25] as the backbone with a stacked SPD manifold autoencoder (SSMAE) built on the tail to enrich the structured representations. Due to the associated reconstruction error terms, the embedding mechanism of both SSMAE and each SPD manifold autoencoder (SMAE) forms an approximate identity mapping, simplifying the training of the suggested deeper network. Then, the ReCov layer is introduced with a nonlinear function for the constructed architecture to narrow the discrepancy of the intra-class distributions from the perspective of regularizing the local statistical information of the SPD data. Afterward, two progressive metric learning stages are coupled with the proposed SSMAE to explicitly capture, encode, and analyze the geometric distributions of the generated deep representations during training. In consequence, not only a more powerful Riemannian network embedding but also effective classifiers can be obtained. Finally, a simple maximum voting strategy is applied to the outputs of the learned multiple classifiers for classification. The proposed model is evaluated on three typical visual classification tasks using widely adopted benchmarking datasets. Extensive experiments show its superiority over the state of the arts.**

***Index Terms*—SPD Manifold, Image Set Classification, Riemannian Neural Network, Stacked SPD Manifold Autoencoder (SSMAE), Metric Learning.**

I. Introduction

THE covariance matrices are ubiquitous in any statistical-related field, but it is less common for them to be used as a representation of the data for computer vision and pattern recognition (CV&PR). In spite of this, their usefulness has been proved in a variety of applications. In medical imaging, covariance matrices are exploited to analyze magnetic resonance imaging (MRI) [16], [34] and classify time-series for Brain-Computer Interfaces (BCI) [17]. In visual classification, since they have the capacity to globally capture and characterize the spatiotemporal fluctuations of image sets (video clips) of different lengths as fixed-dimensional second-order representations, covariance features have gained great popularity in the tasks of face recognition [11], [20], [27], dynamic scene classification [8], [9], [44], and action recognition [25], [26], [31], *etc*.

Given the convenience of covariance matrix, we select it as the feature descriptor for image set data in this paper. However, the essential difficulty of processing and classifying these matrices, which are actually SPD, is that they can not be viewed as Euclidean points, as the topological space spanned by a family of SPD matrices of the same dimensionality is not a vector space, but a curved Riemannian manifold, *i.e.*, SPD manifold [46]. Therefore, it is inappropriate to perform a direct Euclidean computation on the SPD manifold-valued data. To overcome this limitation, [34], [46] advocate the usage of Riemannian metrics to capture and encode the Riemannian geometry of SPD matrices, including Log-Euclidean metric (LEM) [46] and Affine-Invariant Riemannian metric (AIRM) [34]. By utilizing these well-studied Riemannian metrics, the Euclidean tools can be generalized to the SPD manifold by either mapping it into an associated flat space via tangent approximation [2]–[4] or embedding it into the Reproducing Kernel Hilbert space (RKHS) via Riemannian kernel functions [5]–[7], [9], [23], [36]. However, the representation learning and classification process of these two types of approaches is basically carried out in the Euclidean vector space, which will inevitably distort the geometrical structure of the original data manifold. To tackle this issue, several SPD matrix discriminant analysis algorithms [8], [11], [12], [20] concerned with dimensionality reduction (DR) have recently been suggested as a means of geometry-aware feature selection. The philosophy of this type of approach is to generate a lower-dimensional feature manifold with maximum discriminatory power by jointly learning a manifold-to-manifold embedding mapping and a similarity metric on the original SPD manifold. As a consequence, the Riemannian geometry of the input SPD data points could be well preserved in the resulting space.

However, image set data usually cover a wide range of intra-class diversity and inter-class ambiguity, caused by changes in several natural conditions along with the data capturing pro-

R. Wang, X.-J. Wu (*Corresponding author*), T. Xu, and C. Hu are with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China. R. Wang, X.-J. Wu, T. Xu, and C. Hu are also with Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University. e-mail: ({cs_wr, conghu}@jiangnan.edu.cn; {xiaojun_wu_jnu, tianyang_xu}@163.com).

J. Kittler is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. J. Kittler is also with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China. e-mail: j.kittler@surrey.ac.uk.

cess, such as pose, morphology, illumination, and background characterizing each instance, *etc*. This characteristic increases the difficulty of image set classification, which aims to identify an object of interest from a group of image instances of the same visual content, rather than from a single instance [5], [8]–[10], [13], [19]–[21], [28], [29], [31], [38]. Regrettably, although the above-mentioned image set classification methods based on SPD manifold learning are fruitful, the intrinsic shallow learning mechanism impedes them from extracting powerful geometric features for improved classification, especially for complicated data scenarios. In such a case, how to effectively mine discriminative patterns from encoded SPD matrix-based representations still remains a key challenge.

It is widely acknowledged that deep neural networks [50], [51], [69] have achieved remarkable progress in the CV&PR community. Their advantages stem both from the ability to learn powerful semantic information, and from the simplicity and scalability of the gradient-descent training procedure used in backpropagation. Accordingly, some researchers embarked on generalizing the paradigm of Euclidean deep learning to the context of Riemannian manifolds to inject new vitality into image set classification. Recently, several Riemannian neural networks (RiemNets) [25]–[27], [42] comprised of a stack of feature transformation and activation layers have been put forward for SPD matrix learning. The fundamental reasons for their successful applications in some computer vision tasks lie in two factors: 1) the Riemannian geometry of the input data manifold can be preserved through all the layers during training; 2) the deep and nonlinear data embedding mechanism. Therefore, this kind of approach is qualified to learn fine-grained geometric representations for improved classification in comparison with the methods introduced above.

However, exploring the potential of deep learning techniques in the field of Riemannian manifolds is still in the primary stage. One of the main limitations confronted by RiemNets is the degradation problem, *i.e.*, with the increase of network depth, classification accuracy will be degraded, as illustrated in Fig. 1. This phenomenon demonstrates that simply stacking more layers on top of each other does not mean that a better RiemNet can be delivered. Besides, most of the existing studies directly utilize a single cross-entropy loss to supervise the whole network, while neglecting the peculiarities of the data distribution. Consequently, the intra- and inter-class variability information conveyed by the generated deep representations can not be explicitly encoded and learned during training, suppressing the capacity of the obtained geometric features. In this context, learning a discriminative deep Riemannian network embedding for the original image set data to achieve better inter-class separability and smaller intra-class diversity is an interesting proposition.

To address the issues mentioned above, this paper presents a novel SPD matrix learning method for image set classification. Its overall framework is illustrated in Fig. 2. The main purpose of designing such architecture is to probe a deeper manifold-to-manifold embedding mapping to transform the input SPD matrices from the original Riemannian manifold to a new space that not only has the same Riemannian geometry but also possess better discriminatory power. To achieve this objective,

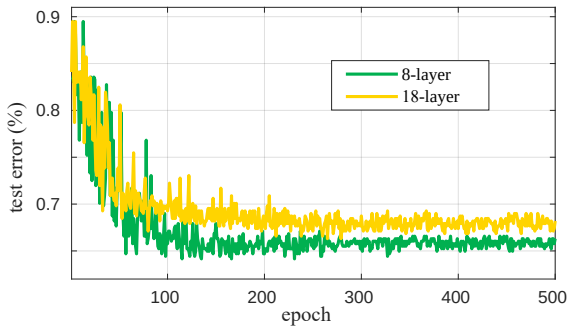

Fig. 1. The classification error of 8-layer and 18-layer SPDNets versus the number of epochs on the AFEW dataset. The deeper Riemannian network exhibits higher test error than that of the shallower one.

there are two pivotal challenges need to be tackled: 1) how to design a deeper SPD manifold neural network to effectively deal with the degradation of structural information during multi-stage data compressed sensing; 2) how to boost the discriminability of the learned geometric representations. For the first challenge, we first select SPDNet [25] as the backbone of our model, considering its merits in nonlinear feature extraction and Riemannian geometry maintaining of SPD matrices. Motivated by the fact that the depth of representations is of crucial importance for many computer vision tasks, we then construct a stacked SPD manifold autoencoder (SSMAE) at the end of the backbone to enrich the 'levels' of structured features. With the aid of the associated reconstruction error terms, the network embedding scheme of both SSMAE and each SPD manifold autoencoder (SMAE) will gradually approximate an identity mapping, being qualified to simplify the training of the built deeper network. Furthermore, with the identity mapping, the norm of the Riemannian gradient will not approach zero, thus avoiding vanishing gradients. In theory, this solution allows the suggested deeper architecture to not incur higher classification error than that of the shallower backbone.

Regarding the second challenge, considering the potential negative impact of the variations of deep representations on the model capacity and the used single cross-entropy loss is incapable of explicitly capturing and encoding the geometric distributions of such data, we develop two progressive metric learning stages for the SSMAE module to facilitate the learning of a discriminative Riemannian network embedding. In the first stage, we append a novel metric learning regularization term to the hidden layer of each SMAE for the sake of learning discriminative SPD representations with higher intra-class similarity and lower inter-class ambiguity. Since the reconstruction error term of each SMAE enables the network to maintain a higher sensitivity against the data variations in the generated new feature space, the metric learning terms could be able to capture, encode, and process such relevant information in an efficient way during training. We can also obtain a series of effective classifiers in parallel by minimizing the classification terms. Due to the high-level structured features contain richer semantic information, at the second stage, we optimize another objective function consisting of a metric learning regularizer and a classification term on top of this network, which can not

only train a new classifier but also fine-tune the whole model.

As the correlation values of the SPD matrix represent the intra-subject data variations, they are crucial for pattern analysis in image set classification. However, most of the existing Riemannian learning approaches reduce the discrepancy of the intra-class data distributions by treating each SPD matrix as a Riemannian element and utilizing the metric learning- or deep learning-related holistic computing tactics, ignoring the importance of the local statistical information within each element. Therefore, in this article, we endow the network with the ReCov operation [27] to achieve local feature nodes regularization by magnifying the selected negative values of the SPD matrix in the negative orientation using a nonlinear activation function. As a result, the statistical relevance of the local feature regions in the original image set data can be enhanced, enabling the learning system to parse the geometry of within-subject data variations better.

To improve the visual classification performance, traditional single-still image-based methods using discriminative learning- or deep learning-based philosophy to relieve the problem of intra-class ambiguity and inter-class similarity. To name a few, [14] proposes a novel discriminative projection learning method to make the features in the generated low-dimensional subspace can not only fit SRC [18] well, but also capture the pivotal structural information embedded in the original data points. To adapt the learning systems to the complex visual scenarios better, [15] comes up with a novel framework for cross-resolution person re-ID task. In this architecture, with the extracted fine-grained details from cross-resolution person images using the modules of VDSR-CA and HRNet-ReID, a multi-task learning-based pseudo-siamese framework is then designed to narrow the discrepancy of the distributions between low-resolution and high-resolution images. It is evident that compared with [14], [15], the SPD manifold learning-based image set classification approach represents the intra-class data variations using structured SPD matrices, and conducts manifold-to-manifold discriminative learning to mitigate the inter-class similarity. Besides, the parameter optimization is also implemented on the Riemannian manifolds, rather than the Euclidean space. These operations are beneficial to preserve and characterize the Riemannian geometry of SPD data points.

In summary, the main contributions of the proposed method contain the following four aspects:

- To cope with the degradation problem, a stacked SPD manifold autoencoder (SSMAE) is constructed on the tail of the backbone network [25], with a series of reconstruction error terms to train. As this design could make the suggested SSMAE architecture approach an identity mapping, it can theoretically render an effective deeper model with lower classification error compared with the shallower counterpart.
- To boost the discriminability of the designed SPD network, two progressive metric learning stages are developed for the SSMAE module to explicitly inject the encoding and learning of the intra- and inter-class data variability information into the network training process. In this scenario, we could expect that a more powerful manifold-to-manifold Riemannian network embedding can be delivered.
- The ReCov layer [27] is generalized from the lightweight feedforward SPD network to the proposed deeper model to mitigate the intra-class diversity by performing local statistics regularization within SPD matrix. We demonstrate through experiments that the ReCov operation is still effective for end-to-end SPD matrix learning.
- Based on the learned geometric features and classifiers, the maximum voting method is applied for the final decision. Extensive experiments show that our method achieves the state-of-the-art accuracy on four benchmarking datasets.

II. Background Theory

In this section, we first present a brief introduction to the Riemannian manifold of SPD matrices and the Log-Euclidean metric on the SPD manifold, which provides the fundamental theory for the proposed approach. Then, the relationship between our method and some previous works is discussed.

A. Riemannian Manifold of SPD Matrices and the Corresponding Log-Euclidean Metric

For all non-zero $v \in \mathbb{R}^d$, a real valued SPD matrix $C \in \mathbb{R}^{d\times d}$ has an intrinsic property, which is $v^T Cv > 0$. The space spanned by a set of d-by-d SPD matrices is the interior of a convex cone in the $d(d+1)/2-$dimensional Euclidean space, denoted as Sym_d^+. As well-studied in [34], [46], a specific Riemannian manifold can be generated, *i.e.*, SPD manifold, when endowing Sym_d^+ with an appropriate Riemannian metric. Due to the topological space of SPD manifold locally conforming to the Euclidean properties and with globally defined differential structure, the derivatives of the curves at point C_i ($C_i \in Sym_d^+$) on the SPD manifold can be expressed under the matrix logarithm map: $\log_{C_i} : Sym_d^+ \rightarrow T_{C_i}Sym_d^+$, where $T_{C_i}Sym_d^+$ represents the tangent space of Sym_d^+ at point C_i. The group of inner products $\langle , \rangle_{C_i}$ on all the tangent spaces is known as the Riemannian metric.

As studied in [46], the space of SPD matrices is a commutative Lie group structure. Since any bi-invariant metric $\langle , \rangle$ on the Lie group of SPD matrices corresponds to an Euclidean metric in the SPD matrix logarithmic domain (the tangent space at identity matrix, $T_I Sym_d^+$), it is also called the Log-Euclidean Metric. To be specific, for any two tangent elements T_i, T_j, their scalar product in $T_C Sym_d^+$ is given as:

$$\langle T_i, T_j \rangle_C = \langle D_C \log_. T_i, D_C \log_. T_j \rangle_I, \tag{1}$$

where $D_C \log_. T$ is the directional derivative of the matrix logarithm at C along T. The logarithm map associated to the Riemannian metric is defined in terms of matrix logarithm:

$$\log_{C_i}(C_j) = D_{\log(C_i)} \exp_.(\log(C_j) - \log(C_i)). \tag{2}$$

Due to the differentiation of the equality $\log \circ \exp = I$, $D_{\log(C)}\exp_. = (D_C \log_.)^{-1}$. Similarly, the matrix exponential map can be expressed as:

$$\exp_{C_i}(T_j) = \exp(\log(C_i) + D_{C_i} \log_. T_j). \tag{3}$$

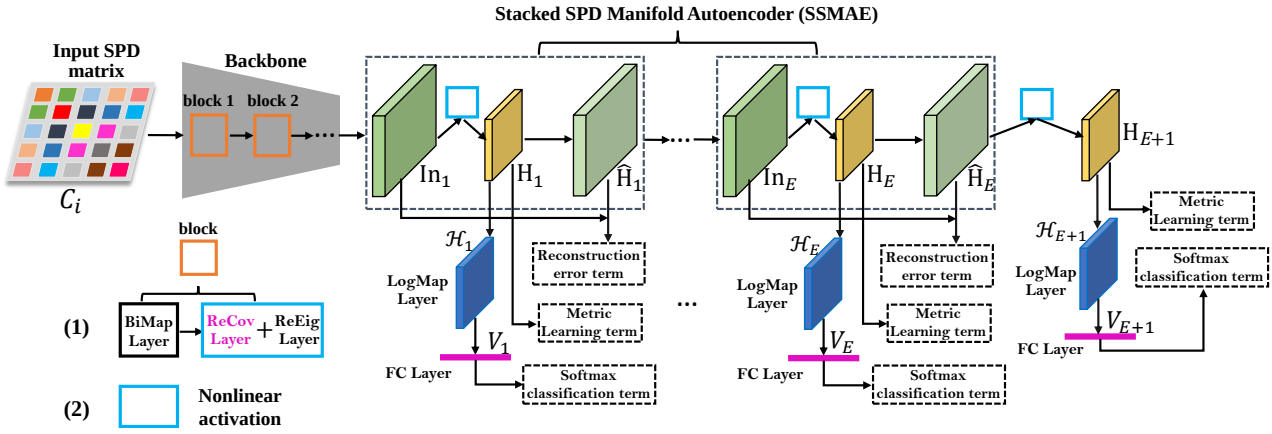

Fig. 2. A schematic diagram of the proposed SPD manifold deep metric learning framework. This framework consists of two parts. The first part is the backbone network for SPD matrix encoding, which has the same structure as SPDNet [25]. The second subsystem is the stacked SPD manifold autoencoder (SSMAE), designed to generate deep representations without structural information degradation. The e-th ($e = 1 \rightarrow E$) SPD manifold autoencoder (SMAE) is comprised of two branches: 1) the first branch contains the layers of input, nonlinear activation, hidden, and reconstruction for feature transformation and reconstruction; 2) the second branch is the classification module, consisting of the LogMap, FC, and softmax layers to generate Euclidean representations for image set classificaition. To make better use of the variability information conveyed by the network inputs, the SSMAE architecture is endowed with two-stage metric learning to supervise a powerful Riemannian network embedding. Besides, we introduce the ReCov layer for the proposed SPD network to inject nonlinearity, chiefly to perform local statistical information regularization of SPD data.

According to Eq. (1), Eq. (2), and Eq. (3), the LEM on the SPD manifold can be formulated as:

$$\mathcal{D} = \langle \log_{C_i}(C_j), \log_{C_i}(C_j) \rangle_{C_i} = ||\log(C_i) - \log(C_j)||_{\mathrm{F}}^2. \quad (4)$$

Compared to AIRM, LEM works directly in the domain of SPD matrices logarithms, resulting in higher computational efficiency. Accordingly, it is selected as the similarity measurement in this paper. For more detailed information, please kindly refer to [40], [46].

B. Relation with the Previous Works

In this paper, we propose modifications over the original SPD manifold neural network [25] in terms of deeper architecture and comprehensive loss. As mentioned above, SPDNet [25] serves as the backbone of our model, which can yield compact and efficient data representations for the input SPD matrices. Then, a stacked SPD manifold autoencoder (SSMAE), trained by a series of reconstruction error terms, is incorporated at the end of the backbone network. Since this design is qualified to enable the embedding mechanism of the SSMAE network to be an approximate identity mapping, more informative deep features could be obtained. In addition, as far as we know, our SSMAE is the first architecture to generalize the paradigm of Euclidean autoencoder to the domain of SPD manifolds, while the reconstruction error terms can make the suggested network remain sensitive to the data variations in the generated new feature spaces. Thereby, the designed two successive metric learning stages are able to capture, encode, and learn more comprehensive data structures with complex variability information. In this way, a discriminative manifold-to-manifold transforming network could be trained, so that the produced geometric features have lower intra-class diversity and better inter-class separability.

Considering the correlation values of SPD matrix characterize the intra-subject variations of the original image set data, in this article, we further inject the ReCov module into the designed SPD network to narrow the discrepancy of the intra-class feature distributions by performing local statistics regularization within SPD data points. Since the ReEig layer can be regarded as a holistic manifold-valued regularization strategy, their integration is eligible to improve the discriminability of the learned representations. Although the ReCov layer was originally designed by [27], the main contributions of this paper on the ReCov operation can be summarized into the following three aspects: 1) the validity of the ReCov layer shown in [27] is only confined to the lightweight Riemannian networks without backpropagation. This article demonstrates through experiments that the ReCov regularization is still effective for end-to-end Riemannian network; 2) this article utilizes Lemma 1 and Remark 1 to prove that when the value of ϵ is small enough, the impact of the ReCov operation on the eigenvalues of the input SPD matrices is negligible. In addition, Theorem 1 also serves as a theoretical guide for the selection of ϵ. However, [27] does not make such discussions; 3) in addition to the aforementioned Eq. (7), another activation function, *i.e.,* the following Eq. (29), has also been studied for the ReCov layer in this paper. The classification results further confirm the effectiveness of Eq. (8)-based ReCov regularization.

To the best of our knowledge, solving the problem of image set classification using the ideology of deep metric learning in the context of SPD manifolds is the first attempt in the research community. Compared with the existing SPD manifold neural networks [25], [27], the advantage of our method stems from the integration of deeper RiemNet and Riemannian metric learning, which can not only overcome the degradation problem when increasing the network depth, but also mitigate the influence of the intra- and inter-class data variations on the model capacity. Besides, compared with the conventional image set classification methods [11],

[20], [28], [29], [31], the innovation of our work is reflected in the methodology. The methods of [28], [29] dedicate to learn a discriminative and robust distance measure using a shallow metric learning framework. However, the image set (video) data is usually distributed on a nonlinear manifold, resulting in that the learned Euclidean metric may be sub-optimal. Since the Riemannian geometry can effectively and naturally characterize the global spatiotemporal fluctuations of a sequence of data, the works of [11], [20], [31] try to improve the accuracy of correct matching between image sets by pursuing an effective manifold-to-manifold embedding mapping. However, the data transformation scheme realized on the nonlinear SPD manifolds in [11], [20], [31] is shallow and linear, which may impact the capacity of the generated features. The strength of the proposed method lies in the generalization of image set encoding and learning to the context of RiemNets, which is instrumental to capturing fine-grained geometric representations.

Inspired by the effectiveness of Euclidean deep metric learning in fully exploiting the nonlinearity of samples to supervise a powerful embedding space [38], the metric learning mechanism is introduced to the proposed SPD network to better address the issue of large intra-class ambiguity and inter-class similarity of deep representations. It is clear that our method mainly focuses on learning a holistic, discriminative deep Riemannian metric from the input SPD matrices, while the methods in [38], [51] aim to mine local discriminative patterns from different video frames in the Euclidean space. Besides, the parameter optimization of the our network is realized by exploiting the stochastic gradient descent (SGD) setting on the Stiefel manifolds with the Riemannian matrix backpropagation for characterizing and preserving the Riemannian geometry of SPD data, while the conventional SGD-based Euclidean backpropagation is used in [38], [51].

III. Proposed Algorithm

As stated above, compared to the existing SPD manifold DR methods, updating the shallow linear feature embedding mechanism to the deep nonlinear function is the central merit of designing neural networks in the context of SPD manifolds, as more effective geometric features can be extracted, especifically for the complex data scenarios. However, the following two issues restrict the learning capacity of existing SPD networks: 1) the degradation (of structural information) problem prevents them from acquiring affluent semantic information by going deeper; 2) the used single cross-entropy loss cannot encode and analyze the intra- and inter-class geometric distributions conveyed by the input data, explicitly and adequately.

We address these two problems using the proposed SPD matrix deep learning architecture, which is detailedly introduced in this section. In Section III-A, we review how to perform image set modeling with the second-order statistics. Section III-B discusses the process of SPD matrix nonlinear encoding. The two progressive metric learning stages incorporated with the suggested SSMAE module are elaborated in Section III-C and Section III-D, respectively. This is followed by the presentation of image set classification based on the maximum voting strategy in Section III-E.

TABLE I
Comparison (%) on the AFEW, MDSD, and FPHA datasets.

Methods	AFEW	MDSD	FPHA
backbone, *i.e.*, SPDNet [25]	**34.23**	**32.05**	**84.23**
backbone-Eq. (7)	27.22	23.08	74.09

A. Set Modeling with Second-Order Statistics

Let $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d\times n}$ be a given image set (video sequence) with n instances, where $x_k \in \mathbb{R}^{d\times 1}$ denotes the k-th vectorized video frame of dimension d. For the image set X, its corresponding covariance matrix can be computed by:

$$C = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - m)(x_k - m)^T, \tag{5}$$

where m is the mean of X, expressed as: $m = \frac{1}{n}\sum_{k=1}^{n} x_k$. Due to the SPD manifold is spanned by a set of nonsingular covariance matrices, we maintain the positive definite property of C by exploiting the following trick: $C \leftarrow C + uI_d$, where I_d is an identity matrix of size $d \times d$, and the perturbation parameter u is configured as $trace(C) \times 10^{-3}$ in this paper.

B. SPD Matrix Nonlinear Encoding

To generate more appropriate SPD manifold-valued feature representations for the original image set data, as mentioned above, a prevailing neural network for SPD matrix nonlinear learning (SPDNet) [25] is chosen as the backbone of our model. As illustrated in Fig. 2, this backbone is comprised of a heap of BiMap and ReEig layers for data transformation and nonlinear activation. Besides, we also inject the ReCov module into each block to enhance the nonlinear learning capacity of the backbone network. These three basic layers are introduced as below.

BiMap Layer: This layer corresponds to the usual dense layer, exploited to transform the input SPD matrices into new ones with lower dimensionality via a bilinear mapping operator f_b, expressed as [25]:

$$C_k = f_b^{(k)}(C_{k-1}; W_k) = W_k^T C_{k-1} W_k, \tag{6}$$

where C_{k-1} is the input SPD matrix of the k-th layer, $C_k \in \mathbb{R}^{d_k\times d_k}$ is the corresponding output, and $W_k \in \mathbb{R}^{d_{k-1}\times d_k}(d_k \leq d_{k-1})$ represents the projection matrix (connection weights) to be learnt. To ensure that C_k resides on the SPD manifold, W_k is required to be a column full-rank matrix. In addition, W_k is also assumed to be a semi-orthogonal matrix [25], [27], such that optimizing W_k over the compact Stiefel manifold [40], [41] prompts the potential to obtain optimal solutions.

ReCov Layer: In the context of Euclidean deep learning, rectified linear units (ReLUs) are verified to be effective activation function in improving the nonlinear learning ability of the models. In this article, we propose to popularize this philosophy to the domain of Riemannian neural networks by directly imposing nonlinear sparseness on each input SPD matrix via the ReLU operation f_{rl}, expressed

as: $\hat{C}_k = f_{rl}^{(k)}(\hat{C}_{k-1}) = \max(\mathbf{0}, \hat{C}_{k-1})$. Wherein, $\hat{C}_{k-1}$ and $\hat{C}_k$ are the input and the corresponding output of this layer, $\mathbf{0} \in \mathbb{R}^{d_{k-1} \times d_{k-1}}$ is a matrix of all zeros, and $\max(\mathbf{0}, \hat{C}_{k-1})$ is defined as:

$$\max(\mathbf{0}, \hat{C}_{k-1})_{ij} = \begin{cases} 0, & \text{if } i \neq j \text{ and } \hat{C}_{k-1}(i,j) \leq 0, \\ \hat{C}_{k-1}(i,j), & \text{otherwise.} \end{cases} \quad (7)$$

However, as the correlation values of SPD matrix reflect the intra-subject data distributions, the above Eq. (7), which sets all the negative elements of $\hat{C}_{k-1}$ to zero, may cause some useful statistical information of the input data to be lost. From Table I, it is evident that backbone-Eq. (7) obtains significantly lower classification scores on the AFEW, FPHA, and MDSD datasets (these three datasets will be detailedly described in Section IV-C), demonstrating that utilizing Eq. (7) for feature regularization is counterproductive. Accordingly, we introduce the ReCov layer [27] to just regularize the points of each input SPD matrix in the $(-\epsilon, 0]$ interval using a nonlinear function f_{rc}, *i.e.*, $\hat{C}_k = f_{rc}^{(k)}(\hat{C}_{k-1}, -\epsilon\mathbf{1})$. Here, ϵ is a small activation threshold determined by cross-validation and $\mathbf{1} \in \mathbb{R}^{d_{k-1} \times d_{k-1}}$ is a matrix of all elements being ones. The mathematical form of this ReCov operation for an SPD matrix in the k-th layer is formulated as:

$$\hat{C}_k(i,j) = \begin{cases} -\epsilon, & \text{if } i \neq j \text{ and } \hat{C}_{k-1}(i,j) \in (-\epsilon, 0], \\ \hat{C}_{k-1}(i,j), & \text{otherwise.} \end{cases} \quad (8)$$

It can be intuitively found that Eq. (8) amplifies the elements in the scope of $(-\epsilon, 0]$ towards the negative direction, enhancing the statistical negative correlation of the local feature regions in the corresponding original video scenario. In consequence, the inconspicuous yet useful data variability information distributed in the $(-\epsilon, 0]$ scope can be intensified, enabling the subsequent metric learning modules to capture, encode, and process the intra-subject feature variations better.

To demonstrate that the ReCov operation will only bring about minor distortion to the main structural information of the input feature matrix, an implicit mapping $\phi(\cdot) : \mathcal{M} \mapsto \mu, \forall \mathcal{M} \in Sym_d^+$ is defined, where μ is an eigenvalue of $\mathcal{M}$. Then, the following conclusion can be made.

Theorem 1: Given any $\mathcal{M}_0 \in Sym_d^+$, $\phi(f_{rc}(\mathcal{M}_0, \epsilon))$ is continuous on $\epsilon \in [0, +\infty)$.

Proof 1: According to Eq. (8), the following inequation can be derived: $\|f_{rc}(\mathcal{M}_0, \epsilon') - f_{rc}(\mathcal{M}_0, \epsilon)\|_{\mathrm{F}} \leq [\sum_{i,j=1, i\neq j}^{d} (\epsilon' - \epsilon)^2]^{\frac{1}{2}} < d|\epsilon' - \epsilon|$. Then, given $\mathcal{M}_0 \in Sym_d^+$, $\forall \epsilon \in [0, +\infty)$, $\forall \varepsilon > 0$, there exists an upper-bound of $\delta \in \mathbb{R}^+$, *i.e.*, $\delta < \frac{\varepsilon}{d}$, such that $\forall \epsilon' \in [0, +\infty) : |\epsilon' - \epsilon| < \delta \Rightarrow \|f_{rc}(\mathcal{M}_0, \epsilon') - f_{rc}(\mathcal{M}_0, \epsilon)\|_{\mathrm{F}} < \varepsilon$. According to the algebra and complex analysis theories, the mapping $\phi(\cdot)$ is continuous. As the continuity of $f_{rc}(\mathcal{M}_0, \epsilon)$ is proved above, the composition function $\phi(f_{rc}(\mathcal{M}_0, \epsilon))$ is continuous accordingly. □

Remark 1: Based on Theorem 1, in the case of $\epsilon = 0$, the following statement can be obtained: given $\mathcal{M}_0 \in Sym_d^+$, $\forall \varepsilon > 0, \exists \delta > 0, \forall \epsilon' \in [0, +\infty) : |\epsilon' - \epsilon| < \delta \Rightarrow |\mu' - \mu^0| < \varepsilon$, where $\mu' = \phi(f_{rc}(\mathcal{M}_0, \epsilon'))$ and $\mu^0 = \phi(f_{rc}(\mathcal{M}_0, 0))$. Here, $\epsilon = 0$ *means that the ReCov operation does not work*. In this scenario, we could expect that with a sufficiently small ϵ', the ReCov layer will perturb the eigenvalue space of each input SPD matrix only in a minor way. Thereby, the main structural information of the data could be well-maintained. The following experimental results indicate that the ReCov operation can improve the classification performance of the designed network, thus verifying the previous expectation experimentally. This in turn reveals that Eq. (7) distorts the geometry of the input feature manifold severely, thus delivering poor performance of backbone-Eq. (7). Besides, the above theorem also provides theoretical guidance for the selection of ϵ. More discussions about this layer will be reported later.

ReEig Layer: This layer is analogous to the ReLU activation, which plays the role of eigenvalue regularization. Specifically, it is designed to adjust the small positive eigenvalues of each input matrix $\bar{C}_{k-1}$ to the proper ones with a nonlinear rectification function f_{re}, denoted as $\bar{C}_k = f_{re}^{(k)}(\bar{C}_{k-1}) = U \max(\zeta I, \Sigma) U^T$. Therein, $\bar{C}_{k-1} = U\Sigma U^T$ denotes the eigenvalue decomposition, and ζ is a small rectification threshold. Obviously, the ReEig operation can not only introduce nonlinearity but also preserve the SPD data from degeneracy.

Since the outputs of the (k-1)-th layer are the exact inputs of the k-th layer, we can obtain that $\hat{C}_{k-1} = C_k$, $\bar{C}_{k-1} = \hat{C}_k$, and $C_{k-1} = \bar{C}_k$. The operations introduced in the ReCov and ReEig layers convey the core nonlinear embedding mechanisms of the proposed model. Hence, we bundle these two consecutive layers together as a unit (shown in Fig. 2) to perform nonlinear activation in this work, which ensures that the generated new feature matrices are faithful to the Riemannian geometry of SPD manifolds.

C. The First Stage Metric Learning

As shown in Fig. 2, the proposed SSMAE module is constructed by multiple SMAEs, in which the outputs of each SMAE are used as the inputs of the subsequent SMAE. The network structure of each SMAE is made up of the input, nonlinear activation (*i.e.*, ReCov+ReEig), hidden, and reconstruction layers, respectively. Moreover, each hidden layer also connects to a head branch, consisting of the layers of LogMap, FC, and softmax, to produce the Euclidean representations for classification.

Let $\mathbb{X} = [X_1, X_2, ..., X_N] \in \mathbb{R}^{d \times \mathcal{N}}$ ($\mathcal{N} = \sum_{i=1}^{N} n_i$, where n_i denotes the number of instances contained in X_i) and $L = [l_1, l_2, ..., l_N] \in \mathbb{R}^{1 \times N}$ be the training data and its corresponding label vector, respectively. Here, N represents the total number of training samples. In this article, the modeled SPD manifold-valued data of $\mathbb{X}$ is expressed as $\mathbb{C} = [C_1, C_2, ..., C_N] \in \mathbb{R}^{d \times dN}$. For a given input SPD matrix C_i ($i = 1 \rightarrow N$), the corresponding low-dimensional and compact output of the backbone network is described as $\mathcal{T}_i = \phi_{\theta_1}(C_i)$, where ϕ_{θ_1} represents the nonlinear embedding from the original SPD manifold to the target one, implemented by a stack of BiMap, ReCov, and ReEig layers, and θ_1 signifies the set of to-be-learnt parameters of the backbone. For the e-th ($e = 1 \rightarrow E$) SMAE, we use In_e ($\mathrm{In}_e(\mathcal{T}_i) = \mathcal{T}_i$, when $e = 1$), $\mathrm{H}_e(\mathcal{T}_i)$, and $\hat{\mathrm{H}}_e(\mathcal{T}_i)$ to represent its input, output of the hidden layer, and reconstruction of the input, respectivety. Actually, $\mathrm{In}_e(\mathcal{T}_i)$ is equivalent to $\hat{\mathrm{H}}_{e-1}(\mathcal{T}_i)$. In the following, we replace

$\text{In}_e(\mathcal{T}_i)$ with $\hat{\text{H}}_{e-1}(\mathcal{T}_i)$ for clarity. Therefore, $\text{H}_e(\mathcal{T}_i)$ and $\hat{\text{H}}_e(\mathcal{T}_i)$ can be computed by:

$$\text{H}_e(\mathcal{T}_i) = f_{b_e}(W_{e1}, \hat{\text{H}}_{e-1}(\mathcal{T}_i)) = W_{e1}^T \pi(\hat{\text{H}}_{e-1}(\mathcal{T}_i))W_{e1}, \quad (9)$$

$$\hat{\text{H}}_e(\mathcal{T}_i) = f_{b_e}(W_{e2}, \text{H}_e(\mathcal{T}_i)) = W_{e2}^T \text{H}_e(\mathcal{T}_i)W_{e2}, \quad (10)$$

where f_{b_e}, π, and $W_{e1} \in \mathbb{R}^{s_{e-1} \times s_e}$, $W_{e2} \in \mathbb{R}^{s_e \times s_{e-1}}$ represent the bilinear mapping function, nonlinear activation operation, and to-be-learnt transformation matrices of the e-th SMAE, respectively.

Briefly speaking, our goal is to learn deep representations with diminished intra-class ambiguity and enlarged inter-class separability for the input data during training. Considering the degradation problem caused by increasing the network depth, we build a stack of SMAEs on the tail of the backbone with each trained by a reconstruction error term. This design can guide the network embedding mechanism of both SSMAE and each SMAE to form an approximate identity mapping, producing less classification error than the shallower backbone in theory. Besides, minimizing the reconstruction error terms enables SSMAE to be as sensitive as possible to the data variations in the yielded new feature manifolds, rendering the metric learning regularization term, imposed on the hidden layer of each SMAE, to be more effective to encode and learn the feature distributions. With these preparations, a discriminative manifold-to-manifold Riemannian network embedding can be supervised via the following objective function of the e-th SMAE:

$$\begin{aligned}\mathcal{L}(\theta_2, \mathbb{P}_e, \phi; \mathbb{C}) = &\min[\lambda_1 \mathcal{L}_1(\text{H}_e) + \\ &\lambda_2 \mathcal{L}_2(\hat{\text{H}}_{e-1}, \hat{\text{H}}_e) + \lambda_3 \mathcal{L}_3(C_i, l_i)],\end{aligned} \quad (11)$$

where λ_1, λ_2, and λ_3 are three trade-off parameters, $\theta_2 = \{\theta_1, W_{e1}, W_{e2}\}$, and $\mathbb{P}_e$ represents the projection matrix of the FC layer of the e-th SMAE, which will be introduced later.

The first term of Eq. (11) is a metric learning regularizer designed to explore an efficient metric space by encoding and learning the within- and the between-class geometric distributions of the generated representations. In such a space, similar samples are expected to be mapped closely to each other, while dissimilar samples could be separated by an appropriate manifold margin. This requirement is formulated as the following generalized logistic loss function, which decays smoothly instead of having a hard cut-off:

$$\mathcal{L}_1(\text{H}_e) = \log\left[1 + \exp\left(\max(\mathcal{S}_{w_e} - \mathcal{S}_{b_e}, \rho_1)\right)\right], \quad (12)$$

where ρ_1 is a pre-difined threshold used to restrain the manifold margin between the intra-class scatter $\mathcal{S}_{w_e}$ and the inter-class scatter $\mathcal{S}_{b_e}$. Their specific forms are respectively defined as:

$$\mathcal{S}_{w_e} = \frac{1}{N_w} \sum_{i=1}^{N} \sum_{j \neq i, l_i = l_j}^{N_w} ||\log(\text{H}_e(\mathcal{T}_i)) - \log(\text{H}_e(\mathcal{T}_j))||_{\text{F}}^2, \quad (13)$$

$$\mathcal{S}_{b_e} = \frac{1}{N_b} \sum_{i=1}^{N} \sum_{j=1, l_i \neq l_j}^{N_b} ||\log(\text{H}_e(\mathcal{T}_i)) - \log(\text{H}_e(\mathcal{T}_j))||_{\text{F}}^2, \quad (14)$$

where N_w and N_b denote the number of intra- and inter-class nearest neighbors of $\text{H}_e(\mathcal{T}_i)$, respectively.

However, the term $\max(\mathcal{S}_{w_e} - \mathcal{S}_{b_e}, \rho_1)$ presented in Eq. (12) focuses on maximizing the manifold margin between different classes, which does not stipulate how small $\mathcal{S}_{w_e}$ should be in the training process. As a consequence, there will be cases where the instances belong to the same category may form a large cluster with a relatively large $\mathcal{S}_{w_e}$ in the learnt feature space. To eliminate its potential negative impact on the image set classification performance of the proposed model, a new constraint term is added to the original metric loss to further require that the intra-class scatter $\mathcal{S}_{w_e}$ be less than a new margin ρ_2. It is suggested that ρ_2 be smaller than $|\rho_1|$.

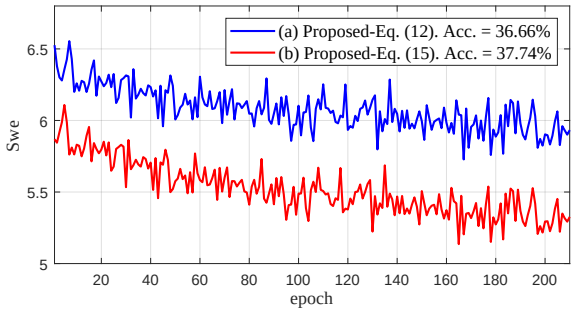

Fig. 3. The average intra-class scatter $\mathcal{S}_{w_e}$ versus the number of training epochs on the AFEW dataset under different metric learning functions.

Hence, Eq. (12) can be rewritten as:

$$\begin{aligned}\mathcal{L}_1(\text{H}_e) = \log[1 + \exp(\max(\mathcal{S}_{w_e} - \mathcal{S}_{b_e}, \rho_1) \\ + \beta \max(\mathcal{S}_{w_e}, \rho_2))],\end{aligned} \quad (15)$$

where β is a trade-off coefficient. Fig. 3 shows the trend of the average intra-class scatter $\mathcal{S}_{w_e}$ of the proposed model as a function of training epochs under different metric learning terms, choosing the AFEW dataset as an example. From this figure, it is evident that both curves show a trend of decreasing first and then leveling off, but the red curve is distributed below the blue curve in all the cases. From Fig. 3, we can also observe that the blue curve has a holistic lower rate of decline than that of the red curve. On one hand, these experimental observations manifest that both Eq. (12) and Eq. (15) can be used to achieve the purpose of reducing the intra-class diversity. On the other hand, it also points out that Eq. (12) is somewhat insufficient for learning the intra-class discriminative information. Besides, the classification accuracy of Proposed-Eq. (15) is 1.08% higher than that of Proposed-Eq. (12) on the AFEW dataset (shown in Fig. 3), justifying that placing additional constraint on $\mathcal{S}_{w_e}$ is effective.

The second term of Eq. (11) is defined to measure the reconstruction error between the input data and the corresponding reconstructed data, expressed as:

$$\mathcal{L}_2(\hat{\text{H}}_{e-1}, \hat{\text{H}}_e) = \sum_{i=1}^{N} ||\hat{\text{H}}_{e-1}(\mathcal{T}_i) - \hat{\text{H}}_e(\mathcal{T}_i)||_{\text{F}}^2. \quad (16)$$

We want to emphasize that the fundamental reasons for using the Euclidean distance (ED) to replace LEM for similarity measurement in Eq. (16) are two-fold: 1) it can verify the 'pixel-level' similarity between the input and the reconstructed samples intuitively; 2) the computation of the inverse of $\hat{\text{H}}_{e-1}(\mathcal{T}_i)$ can be avoided during optimization.

The third term of Eq. (11) is a softmax loss function for image set classification. It is implemented with the assistance

of the LogMap and the FC layers, as illustrated in Fig. 2. The LogMap layer [25] is mainly exploited to carry out Riemannian computing on each resulting SPD matrix via the logarithmic mapping function f_{l_e}, expressed as:

$$\mathcal{H}_e(\mathcal{T}_i) = f_{l_e}(\mathrm{H}_e(\mathcal{T}_i)) = \log(\mathrm{H}_e(\mathcal{T}_i)) = U\log(\Sigma)U^T, \quad (17)$$

such that a flat space for Euclidean computations can be generated. In Eq. (17), $\mathrm{H}_e(\mathcal{T}_i) = U\Sigma U^T$ represents the eigenvalue decomposition and $\log(\Sigma)$ is a diagonal matrix composed of the logarithm of the eigenvalues.

Now, the loss function $\mathcal{L}_3$ is given below:

$$\mathcal{L}_3(C_i, l_i) = -\sum_{i=1}^{N}\sum_{r=1}^{c}\tau(l_i, r) \times \log\frac{e^{\mathbb{P}_e^r V_e(\mathcal{T}_i)}}{\sum_o e^{\mathbb{P}_e^o V_e(\mathcal{T}_i)}}, \quad (18)$$

where $\mathbb{P}_e^r$ denotes the r-th row of the projection matrix $\mathbb{P}_e$, $V_e(\mathcal{T}_i)$ represents the vectorized form of $\mathcal{H}_e(\mathcal{T}_i)$, and $\tau(l_i, r)$ is an indicator function, where $\tau(l_i, r) = 1$ if $l_i = r$, and 0 otherwise.

D. The Second Stage Metric Learning

After the first metric learning stage, the discriminatory power of the generated deep representations will be improved. Besides, we could also obtain multiple effective classifiers for image set classification. Due to the approximate identity mapping scheme of the designed SSMAE, a certain amount of pivotal data variations will be contained in the resulting feature maps of the final reconstruction layer. To make better use of such latent information for enhancing the discrimination of the network further, as an exploration, we enable the network to be supervised by the second metric learning stage. As shown in Fig. 2, we first incorporate a nonlinear activation layer onto the top of SSMAE, followed by a BiMap layer to reduce the dimension of the input SPD matrices. Then, a metric learning regularizer and a classification module (presented in Section III-C) are coupled with the added BiMap layer to fine-tune the whole network and train a new classifier at the same time.

We denote the i-th output of SSMAE corresponding to its i-th input $\mathcal{T}_i$ as $\hat{\mathrm{H}}_E(\mathcal{T}_i)$ ($i = 1 \rightarrow N$). Since the last BiMap layer can be regarded as the hidden layer of the $(E+1)$-th SMAE, and $\hat{\mathrm{H}}_E(\mathcal{T}_i)$ is equivalent to its i-th input, the i-th output of the proposed SPD network is given by:

$$\mathrm{H}_{E+1}(\mathcal{T}_i) = f_{b_{E+1}}(W_{E+1}, \hat{\mathrm{H}}_E(\mathcal{T}_i)) = W_{E+1}^T\pi(\hat{\mathrm{H}}_E(\mathcal{T}_i))W_{E+1}, \quad (19)$$

where W_{E+1} represents the to-be-learnt transformation matrix of the final BiMap layer.

With these definitions, the objective function of this metric learning term can be defined as:

$$\mathcal{J}(\theta_W, \mathbb{P}_{E+1}, \phi; \mathbb{C}) = \lambda_4\mathcal{J}_1(\mathrm{H}_{E+1}) + \lambda_5\mathcal{J}_2(C_i, l_i), \quad (20)$$

where λ_4 and λ_5 are two trade-off parameters, $\theta_W = \{\theta_2, W_{E+1}\}$, and $\mathbb{P}_{E+1}$ denotes the projection matrix of the FC layer of the $(E+1)$-th SMAE.

The first term of Eq. (20) is a metric learning regularizer to further alleviate the intra-class diversity as well as the inter-class ambiguity by characterizing and analyzing the geometric distribution of the produced high-level features. Similar to Eq. (15), this term can be formulated as:

$$\begin{aligned}\mathcal{J}_1(\mathrm{H}_{E+1}) = \log[1 + \exp(\max(\mathcal{S}_{w_{E+1}} - \mathcal{S}_{b_{E+1}}, \rho_3) \\ + \eta\max(\mathcal{S}_{w_{E+1}}, \rho_4))],\end{aligned} \quad (21)$$

where ρ_3, ρ_4, and η play the same role as that of ρ_1, ρ_2, and β introduced in Eq. (15). The mathematical forms of $\mathcal{S}_{w_{E+1}}$ and $\mathcal{S}_{b_{E+1}}$ are analogous to Eq. (13) and Eq. (14), given as follows:

$$\mathcal{S}_{w_{E+1}} = \frac{1}{N_w}\sum_{i=1}^{N}\sum_{\substack{j\neq i \\ l_i=l_j}}^{N_w}||\log(\mathrm{H}_{E+1}(\mathcal{T}_i)) - \log(\mathrm{H}_{E+1}(\mathcal{T}_j))||_{\mathrm{F}}^2, \quad (22)$$

$$\mathcal{S}_{b_{E+1}} = \frac{1}{N_b}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ l_i\neq l_j}}^{N_b}||\log(\mathrm{H}_{E+1}(\mathcal{T}_i)) - \log(\mathrm{H}_{E+1}(\mathcal{T}_j))||_{\mathrm{F}}^2, \quad (23)$$

The second term of Eq. (20) is the softmax loss function to minimize the classification error with the input-target pairs (C_i, l_i) ($i = 1 \rightarrow N$), which can be computed by:

$$\mathcal{J}_2(C_i, l_i) = -\sum_{i=1}^{N}\sum_{r=1}^{c}\tau(l_i, r) \times \log\frac{e^{\mathbb{P}_{E+1}^r V_{E+1}(\mathcal{T}_i)}}{\sum_o e^{\mathbb{P}_{E+1}^o V_{E+1}(\mathcal{T}_i)}}, \quad (24)$$

where $\mathbb{P}_{E+1}^r$ indicates the r-th row of the projection matrix $\mathbb{P}_{E+1}$ and $V_{E+1}(\mathcal{T}_i)$ is the vectorized modality of the i-th output $\mathcal{H}_{E+1}(\mathcal{T}_i)$ of the $(E+1)$-th SMAE.

Due to space limitation, please kindly refer to Section I of our supplementary material for the details of the Riemannian matrix backpropagation algorithm used to train the proposed network.

E. Image Set Classification

In the test phase, a given test image set X_{te} is firstly encoded as an SPD matrix C_{te} using Eq. (5). Then, the well-trained SPD network is exploited to transform C_{te} into the low-dimensional and compact representation $\mathcal{H}_e(\mathcal{T}_{te})$ at the LogMap layer. Afterward, the classification probability of X_{te} belonging to class r can be computed by:

$$P_e^r(X_{te}) = \frac{e^{\mathbb{P}_e^r V_e(\mathcal{T}_{te})}}{\sum_o e^{\mathbb{P}_e^o V_e(\mathcal{T}_{te})}}, \quad (25)$$

where $r = 1 \rightarrow c$, $e = 1 \rightarrow (E+1)$, and $V_e(\mathcal{T}_{te})$ denotes the vectorized form of $\mathcal{H}_e(\mathcal{T}_{te})$. Once the classification probabilities $P_e^r(X_{te})$ become available from all the $(E+1)$ classifiers, we use the voting mechanism (analogous to [19]) to determine the output label of X_{te}. Specifically, the vote $v_e(X_{te})$ of the e-th classifier is cast for the class with the highest probability, *i.e.*:

$$v_e(X_{te}) = \arg\max_r P_e^r(X_{te}). \quad (26)$$

The votes cast by all the classifiers of X_{te} are tallied and the class receiving the maximum number of votes is declared as the label of X_{te}. This can be formulated as:

$$l_{te} = \arg\max_r \sum_e \omega_r(v_e(X_{te})) \text{ with,} \quad (27)$$

$$\omega_r(v_e(X_{te})) = \begin{cases}1, & v_e(X_{te}) = r, \\ 0, & \text{otherwise.}\end{cases} \quad (28)$$

IV. EXPERIMENTS

In this section, we study the effectiveness of the proposed approach[1] on three typical visual classification tasks using three different benchmarking datasets, *i.e.,* video-based facial emotion recognition on the AFEW dataset [43], dynamic scene classification on the MDSD dataset [45], and skeleton-based hand action recognition on the FPHA dataset [49], respectively.

A. Implementation Details

To build the backbone of our model, we use seven layers: $C_i \rightarrow f_b^{(1)} \rightarrow f_{rc}^{(2)} \rightarrow f_{re}^{(3)} \rightarrow f_b^{(4)} \rightarrow f_{rc}^{(5)} \rightarrow f_{re}^{(6)} \rightarrow f_b^{(7)}$, where f_b, f_{rc}, and f_{re} denote the BiMap, ReCov, and ReEig layers, respectively. The stacked SPD manifold autoencoder (SSMAE) constructed on the tail of the backbone is comprised of E SMAEs (the later experiments show that setting E to 2 is reasonable), and each SMAE contains two branch networks. The first branch is designed for feature encoding and reconstruction, which is made up of five layers: f_b (input) $\rightarrow$ $f_{rc} \rightarrow f_{re} \rightarrow f_b$ (hidden) $\rightarrow f_b$ (reconstruction). The second branch, consisting of three layers: $f_l \rightarrow f_{\mathrm{F}} \rightarrow f_s$, is connected to the hidden layer of each SMAE to produce Euclidean representations for classification. Here, f_l, f_{F}, and f_s represent the layers of LogMap, FC, and Softmax loss, respectively. The recommended values of the network parameters, such as the learning rate ξ, the thresholds ζ, ρ_1, ρ_2, and the tradeoff coefficients $\lambda_1, \lambda_2, \lambda_3, \beta$ on the three used datasets are listed in Table II. In the experiments, the values of ρ_3, ρ_4, λ_4, λ_5, and η are configured to be the same as those of ρ_1, ρ_2, λ_1, λ_3, and β. To train the suggested network, an i7-9700 (3.4GHz) PC with 16GB RAM is utilized. Besides, on the AFEW, MDSD, and FPHA datasets, the batch size $\mathbb{B}$ is set to 30, 20, and 30, respectively. This configuration respectively takes about 4.25 minutes, 0.32 minutes, and 0.78 minutes per training epoch for the AFEW, MDSD, and FPHA datasets.

B. Comparative Methods and Settings

For the evaluation of our model, the following representative image set classification methods are selected for comparison, which can be grouped into three categories:

(1) SPD matrix learning-based methods: Covariance Discriminative Learning (CDL) [5], Riemannian Sparse Representation (RSR) [36], Log-Euclidean Metric Learning (LEML) [11], SPD Manifold Learning (SPDML) Based on AIM and Stein divergence [20], Deep Second-Order Pooling Network (DeepO2P) [22], SPD Manifold Neural Network (SPDNet) [25], SPDNet using Riemannian Batch Normalization (SPDNetBN) [26], and Lightweight SPD Manifold Neural Network (SymNet) [27].

(2) Linear subspace learning-based methods: Grassmann Discriminant Analysis (GDA) [1], Grassmannian Graph-Embedding Discriminant Analysis (GEDA) [33], Projection Metric Learning (PML) [30], Graph Embedding Projection Metric Learning (GEPML) [32], and Graph Embedding Multi-Kernel Metric Learning (GEMKML) [10].

[1]The source code will be released on: https://github.com/GitWR/SMTNet

Fig. 4. Facial emotion images of the AFEW dataset

(3) Multiple Riemannian matrix learning-based methods: Localized Multi-Kernel Metric Learning (LMKML) [37], Hybrid Euclidean-and-Riemannian Metric Learning (HERML) [39], and Multiple Riemannian Manifolds Metric Learning (MRMML) [9].

The experimental results of all the comparative methods on the three datasets are obtained by running the source codes provided by the original authors, except for DeepO2P. The recognition score of DeepO2P on the AFEW dataset is provided by [25]. For a fair comparison, we empirically tuned the parameters of the baseline systems according to the recommendations of the original papers. For CDL, the perturbation parameter was set to $10^{-3} \times trace(C)$. In PML, the trade-off coefficient α was set in line with [30]. In LEML, we searched the values of η and ζ in the scopes of $[0.1, 1, 10]$ and $[0.1 : 0.1 : 1]$, respectively. For SPDNet and SPDNetBN, the learning rate, batch size, and sizes of the transformation matrices were determined by cross-validation on the MDSD and the FPHA datasets, and the settings on the AFEW dataset were consistent with [25]. For SymNet, the sizes of the connection weights and the values of thresholds ϵ, η were configured as recommended in [27]. For RSR, the value of λ were searched in the range of $[0.0001, 0.001, 0.01, 0.1]$. For SPDML, GEPML, and GEMKML, the number of intra- and inter-class nearest neighbors of a given anchor point were determined by cross-validation on all the datasets. In HERML, the proper values of γ and ζ were selected from the sets $[0.001, 0.01, 0.1, 1, 10, 100, 100]$ and $[0.1 : 0.1 : 1]$, respectively. In LMKML, the learning rate α was set to 10^{-6}. For MRMML, we applied cross-validation to choose the appropriate value for d_w.

C. Datasets Description and Settings

AFEW dataset. This dataset involves 1,345 video sequences of natural facial expressions collected from movies. Fig. 4 presents some examples of this dataset. For the evaluation, we follow the standard protocols of [25] to first split these training videos into 1,746 small clips for data augmentation. Then, each video frame is shaped into a 20×20 gray-scale image, such that a 400×400 SPD matrix can be computed for video characterization. Since the groundtruth of the test set has not been publicly available, we finally report the classification accuracy on the validation set. The sizes of the transformation matrices of the proposed network on this dataset are set to 400×200, 200×100, 100×50, 50×100, 100×50, 50×100, and 100×50, respectively.

MDSD dataset. This dataset makes up of 13 different categories of dynamic scenes, each of which containing 10 video sequences collected in unconstrained scenarios. Fig. 5 illustrates some dynamic scene images of this dataset. To be

Fig. 5. Dynamic scene images of the MDSD dataset

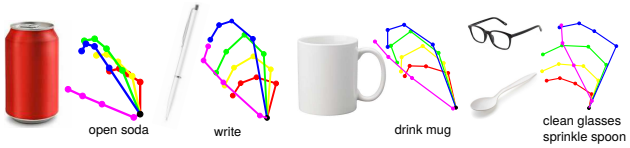

Fig. 6. Some 3D hand pose instances of the FPHA dataset

TABLE II
RECOMMENDED VALUES OF THE KEY PARAMETERS.

Datasets	ξ	ζ	ρ_1	ρ_2	λ_1	λ_2	λ_3	β
AFEW	0.010	1E-4	-1.0	1.0	0.1	1E-2	1.0	0.2
MDSD	0.013	1E-5	-0.1	0.1	1.0	1E-4	1.0	0.2
FPHA	0.010	1E-4	-1.0	1.0	1.0	1E-1	1.0	0.2

compatible with the previous works [9], [10], [27], we first resize all the video frames into 20×20 intensity images. Then, the SPD matrix of size 400×400 is computed to represent each video clip. Finally, the seventy-thirty-ratio (STR) protocol, which typically selects 7 videos for training and the remaining 3 for a query set per category, is applied to build the gallery and probes. On this dataset, the network filters are configured to be of the same sizes as those of the AFEW dataset.

FPHA dataset. This dataset is comprised of 1,175 hand action videos representing 45 different categories, performed by 6 actors in the first-person view. Some hand action instances of the FPHA dataset are displayed in Fig. 6. To make a fair comparison, we follow the standard protocol of [10] to first transfer each hand action frame into a 63-dimensional vector using the 3D coordinates of 21 hand joints provided, such that a 63×63 SPD matrix can be computed for the modeling of an action sequence. Then, the 1:1 setting is applied for evaluation, *i.e.*, 600 action clips are designated for training and the remaining 575 are used for testing. On this dataset, the sizes of the network weights are set to 63×53, 53×43, 43×33, 33×43, 43×33, 33×43, and 43×33, respectively.

D. Results and Discussions

The experimental results achieved by different methods on the AFEW and MDSD datasets are presented in Table III. From this table, the following interesting observations can be made. Firstly, the classification scores of LMKML, HERML, and MRMML are higher than most of the competitors on these two datasets, demonstrating that the complementarity of multiple statistics in image set characterization enables the network to learn more powerful geometric features for effective decision making. Besides, the classification performance of LMKML is inferior to that of HERML and MRMML. The fundamental reason is that LMKML applies an Euclidean kernel function to the high-order statistics (they typically lie in the non-Euclidean spaces) to perform kernel spaces embedding, which will distort the geometry of the original data manifold.

TABLE III
ACCURACY COMPARISON (%) ON THE AFEW AND MDSD DATASETS.

Methods	Year	AFEW	MDSD
GDA [1]	2008	29.11	30.51
GEDA [33]	2011	29.45	30.37
CDL [5]	2012	31.81	31.28
RSR [36]	2012	27.49	31.62
LMKML [37]	2013	-	32.37
HERML [39]	2015	32.14	33.59
PML [30]	2015	28.98	29.67
LEML [11]	2015	25.13	29.30
DeepO2P [22]	2015	28.54	-
SPDNet [25]	2017	34.23	32.05
SPDML-AIM [20]	2018	26.72	30.04
SPDML-Stein [20]	2018	24.55	27.69
SPDNetBN [26]	2019	36.12	35.26
GEPML [32]	2021	33.78	35.33
GEMKML [10]	2021	35.71	35.89
MRMML [9]	2022	35.71	36.67
SymNet [27]	2022	32.70	35.58
Proposed		**37.74**	**42.05**

Secondly, it is evident that the classification ability of LEML and SPDML-AIM/Stein are convincingly surpassed by SPDNet, SPDNetBN, and SymNet on the AFEW and MDSD datasets. The primary reason is that both LEML and SPDML-AIM/Stein exploit a shallow linear architecture to perform feature transformation, which is incapable of faithfully respecting the Riemannian geometry of the original data manifold in the resulting space. In contrast, SPDNet, SPDNetBN, and SymNet generalize the shallow linear scheme for SPD matrix learning to the nonlinear function through RiemNets, being qualified to capture fine-grained geometric features. This is also the fundamental reason why the classification performance of PML and GEPML is inferior to that of GEMKML. Thirdly, the experimental comparison between SPDNet, SPDNetBN, and SymNet shown in Table III demonstrates that the shallow optimization algorithm-based lightweight SPD network, *i.e.*, SymNet, is more suitable for the classification tasks with limited data, as pivotal data variations can be captured for visual scene parsing. As can be clearly seen from Table III, the proposed method obtains the highest classification performance on the AFEW and MDSD datasets, confirming its effectiveness in enhancing the within-class compactness and boosting the between-class separability of the learned features.

Next, we compare the suggested network with some representative hand action recognition models on the FPHA dataset, such as the convolutional two-stream network (Two streams) [52], Novel View [56], Lie Group [63], hierarchical recurrent neural network (HRNN) [59], LSTM [49], jointly learning heterogeneous features (JOULE) [62], Gram Matrix [53], transition forests (TF) [64], temporal convolutional network (TCN) [60], spatial-temporal graph convolutional network (ST-GCN) [61], unified hand and object model (H+O) [66], and temporal transformer network (TTN) [65]. The recognition scores of different methods on the FPHA dataset are reported in Table IV. It is evident that the methods that performing hand action recognition in the context of Riemannian mani-

TABLE IV
ACCURACY COMPARISON (%) ON THE FPHA DATASET.

Methods	Year	Color	Depth	Pose	Acc.
Lie Group [63]	2014	✗	✗	✓	82.69
HRNN [59]	2015	✗	✗	✓	77.40
JOULE-pose [62]	2015	✗	✗	✓	74.60
JOULE-all [62]	2015	✓	✓	✓	78.78
Two streams [52]	2016	✓	✗	✗	75.30
Novel View [56]	2016	✗	✓	✗	69.21
Gram Matrix [53]	2016	✗	✗	✓	85.39
SPDNet [25]	2017	✗	✗	✓	86.26
TF [64]	2017	✗	✗	✓	80.69
TCN [60]	2017	✗	✗	✓	78.57
LSTM [49]	2018	✗	✗	✓	80.14
ST-GCN [61]	2018	✗	✗	✓	81.30
SPDML-AIM [30]	2018	✗	✗	✓	76.52
H+O [66]	2019	✓	✗	✗	82.43
TTN [65]	2019	✗	✗	✓	83.10
SPDNetBN [26]	2019	✗	✗	✓	86.83
GEMKML [10]	2021	✗	✗	✓	81.75
MRMML [9]	2022	✗	✗	✓	83.33
SymNet [27]	2022	✗	✗	✓	82.96
Proposed		✗	✗	✓	**89.39**

folds (*e.g.*, Lie Group, Gram Matrix, SPDML-AIM, SPDNet, SPDNetBN, SymNet, MRMML, and GEMKML) show competitive recognition performance on this dataset. This provides further demonstration of the merit of Riemannian geometry in modeling the global spatiotemporal fluctuations of a sequence of data. From Table IV, we can also note that the classification scores of SPDNet and SPDNetBN are higher than those of LEML, SPDML-AIM, and SymNet, further certifying that the mechanism of end-to-end deep-embedding learning is more effective than shallow learning scheme in SPD matrix analysis. Table IV shows that the suggested SPD manifold deep metric learning framework is still the best performer on the FPHA dataset, again certifying its availability in mining useful geometric information for visual scene description.

E. Ablation Study of the Number E of the Stacked SMAEs

As the designed SSMAE is a crucial subsystem of the proposed model, in this subsection, we carry out cross-validation experiments on the AFEW, MDSD, and FPHA datasets to explore its suitable architecture by measuring the impact of the number E of the stacked SMAEs on the overall performance of our approach. The experimental results are depicted in Fig. 7, where E takes values from the set $\{0, 1, 2, 3\}$. Here, $E = 0$ indicates that the proposed network does not contain the SSMAE module and its associated two metric learning stages, and is just trained by the classification error term $\mathcal{J}_2$ (its structure is equivalent to that of SPDNet embedded with the ReCov layers). From Fig. 7, it is evident that when E changes from 0 to 2, the classification scores of our method are increasing on all three datasets. The underlying reasons come from two aspects: 1) the capability of the SSMAE architecture in generating deep representations with richly structured semantic information; 2) the efficacy of the metric learning terms in narrowing the intra-subject diversity and magnifying the inter-subject dissimilarity by modeling and learning the data distribution information. From Fig. 7, we

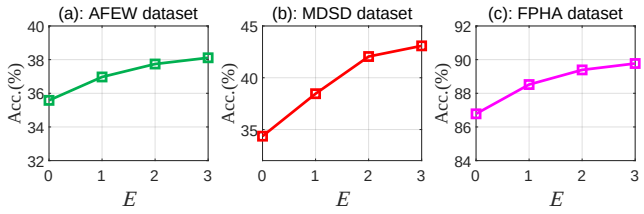

Fig. 7. Comparison (%) under different number E of the stacked SMAEs.

can also observe that the classification performance of the suggested approach is promoted slightly when increasing E from 2 to 3. The essential reason is that with increase of the number E of the stacked SMAEs, the reconstruction error terms of SSMAE will gradually make the residual information between two adjacent SMAEs tend to 0, so that the gradient does not change significantly when transmitting between upper layers. In addition, it is easy to know that the value of E is proportional to the computation time (this will be studied later). To sum up, we set E to 2 in this paper as a compromise. In what follows, some specific discussions on the effectiveness of each component of the designed model are carried out.

F. Ablation Study of the SSMAE Module

In this subsection, we make experiments to evaluate the efficacy of the designed SSMAE module, choosing the AFEW dataset as an example. The baseline architecture is the same as the aforementioned SPDNet. From Table V and Fig. 8, we have three major findings. Firstly, the situation is reversed with the operation of SSMAE, *i.e.*, P-$E2$ (19-layer) performs better than P-$E1$ (11-layer). Here, 'P-Ex' *represents that the number E of the stacked SMAEs included in the proposed framework is x*. More importantly, the classification score of P-$E2$ is somewhat higher than that of P-$E1$. This demonstrates that the degradation problem can be conquered in this design, supporting us to successfully obtain accuracy gains from increased depth. The consistent observations can be illustrated between P-$E2$ and P-$E3$ (27-layer).

Secondly, in this experiment, we also explore two deep models, *i.e.*, P-$E4$ and P-$E8$, of over 30 and 65 layers respectively. It is evident that our approach has no optimization difficulty, and the 35/67-layer SPD networks are able to achieve fairly good recognition scores (36.12% and 35.85%) on the AFEW dataset. Thirdly, compared Fig. 8 with Fig. 1, we can also observe that the convergence speed of our 11/19/27/35/67-layer SPD networks are faster than that of the 8/18-layer SPDNets. This demonstrates that the suggested embedding function of SSMAE helps to train the networks that are deeper than those used previously. The underlying reason is that the stacked SMAEs and the associated reconstruction error terms enable the richer gradient information of the upper layers to be easily transmitted to the lower layers. The experimental evidences mentioned above also manifest that setting E to 2 on the AFEW dataset is rational, considering the training time, the number of parameters, and the convergence behavior.

However, the test result of P-$E8$ is 0.54% and 0.27% lower than that of P-$E3$ and P-$E4$, respectively. We argue that the overfitting problem is one of the reasons for the inferiority of

TABLE V
COMPARISON UNDER DIFFERENT VALIDATION METRICS.

Methods	SPDNet	P-$E1$	P-$E2$	P-$E3$	P-$E4$	P-$E8$
Acc. (%)	34.23	35.03	35.31	36.39	36.12	35.85
s/epoch	19.82	20.97	25.63	31.17	38.53	61.20
#params	0.12M	0.13M	0.16M	0.18M	0.21M	0.32M

TABLE VI
ACCURACY COMPARISON (%) UNDER DIFFERENT METRIC LEARNING STAGES.

Datasets	AFEW	MDSD	FPHA
backbone-SSMAE	35.31	36.41	86.09
backbone-SSMAE-1stMLS	36.66	38.46	88.17
backbone-SSMAE-2ndMLS	35.95	37.44	86.96
backbone-SSMAE-1st&2ndMLS	**37.19**	**39.49**	**89.04**

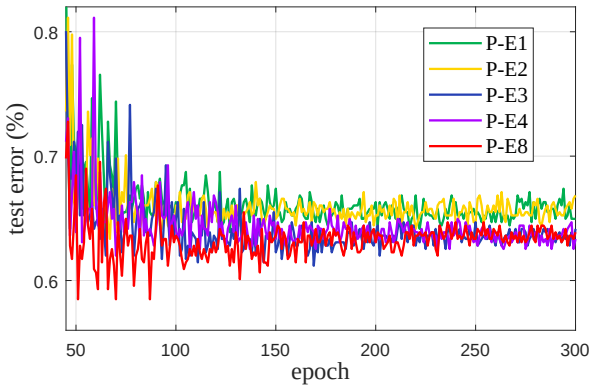

Fig. 8. The test error of different subarchitectures of the proposed model versus the number of training epochs on the AFEW dataset.

P-E8, as the 67-layer network may be a bit large (0.32M) for this dataset. For the basic reason of the inferiority of P-$E4$ and P-$E8$ compared to P-$E3$, we believe that it is due to the loss of some pivotal structural information embedded in the input SPD matrices during multi-stage SMAE transformation. Since the reconstruction loss is introduced as the learning objective, the solution for the early stages of SSMAE network will tend to diagonalise the respective SPD matrices. This follows from the well known fact that the optimal solution to the problem of minimising the signal (image set) approximation error using a reduced number of basis functions are the eigenvectors of the signal covariance matrix associated with the largest eigenvalues. The SPD matrix reconstruction problem is a proxy to the signal approximation problem. However, once an off-diagonal element of an SPD matrix becomes zero, it will never contribute to the generation of the lower-dimensional SPD matrices. This will considerably reduce the number of variables involved in the generation of these matrices. Due to the nonlinear activation function (ReCov+ReEig) used in this article plays the role of SPD matrix regularization, it allows the off-diagonal elements to contribute to the learning process. Nevertheless, with increasing the number E of the stacked SMAEs, the ReCov and ReEig operations will gradually exacerbate the amount of adjustment to the input SPD matrices. In this scenario, the more transformation stages of SSMAE, the more distortion of the original statistical information will be. As a consequence, the classification performance of P-$E4$ and P-$E8$ is not as good as that of P-$E3$.

The above analysis indicates that with the network depth increases, the potential degradation of structural information hinders the embedding function of the proposed network from approaching an identity mapping. Inspired by the philosophy of ResNet [50], in the future, we plan to study how to carry out residual learning in the context of SPD manifolds. Since the skip connections can make the current learning stage access the informative feature maps of the previous stages easily, the Riemannian residual learning may open up new possibilities for solving the degradation problem.

G. Ablation Study of the Two-Stage Metric Learning

As introduced in Section III-C and Section III-D, the purpose of designing the two progressive metric learning stages is to enforce the samples from the same class to be closer to each other and those from different classes to be separated by a large manifold margin after the network embedding. Accordingly, in this section, we carry out experiments on the AFEW, MDSD, and FPHA datasets to evaluate the impact of each metric learning stage (MLS) on the classification performance of backbone-SSMAE studied in Section IV-E. The experimental results are reported in Table VI, where 'backbone-SSMAE-1stMLS' and 'backbone-SSMAE-2ndMLS' signify that the network only includes the first (1st) and the second (2nd) metric learning stages, respectively. From Table VI, one can note that both of these two MLSs can improve the classification performance of backbone-SSMAE on all the datasets, showing the effectiveness of the suggested deep metric learning strategy in exploiting the geometry of the learned feature manifolds for finding a more discriminative decision space. Besides, we can also observe that the classification scores of backbone-SSMAE-1stMLS are respectively 0.71%, 1.02%, and 1.21% higher than those of backbone-SSMAE-2ndMLS. The fundamental reason is that the data variations conveyed by the network inputs is inconspicuously embodied in the top-level feature maps, which restricts the 2ndMLS from sufficiently encoding and analyzing the geometric distributions of the learned deep representations. In contrast, since the 1stMLS consists of E metric learning instances that parse the data distributions in different hidden layers of SSMAE, thus being qualified to produce geometric features with reinforced discrimination. More importantly, the combination of the 1stMLS and the 2ndMLS results in a further enhancement of the classification ability of the studied network, confirming their complementary role in SPD matrix discriminative learning.

To intuitively measure the discriminatory power of the features learned by the proposed approach under the 1stMLS and the 2ndMLS respectively, we provide the 2-D visualization results, choosing the MDSD dataset as an example. It can be observed from Fig. 9 that the holistic between-class separability reflected in Fig. 9(b) is better than that shown in Fig. 9(a). All in all, these experimental evidences indicate that injecting metric learning into the proposed network is beneficial to magnify the inter-class diversity and shrink the discrepancy of the intra-class distributions.

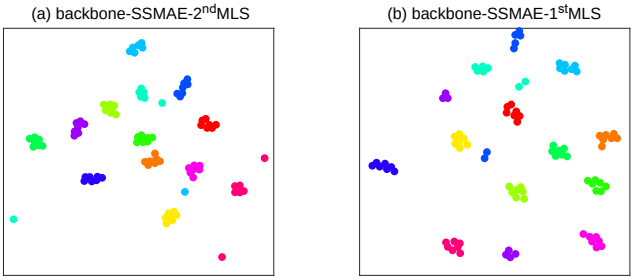

Fig. 9. 2-D visualization of the learned representations on the MDSD dataset, where colors and points denote categories and image set samples, respectively.

H. Ablation Study of the ReCov Layer

On the basis of the above-mentioned experiments, in this subsection, we experiment on the three used datasets to verify the effectiveness of the introduced ReCov regularization by measuring the effect of the threshold ϵ on the classification performance of the proposed method. For the selection of the value of $-\epsilon$ in this paper, we first check the approximate value range of the elements in the input SPD matrices. In practice, only a few elements have relatively large negative correlation values. According to Theorem 1, a sufficiently large negative correlation value is then selected as an anchor point. Afterward, the cross-validation is exploited to search a proper value for $-\epsilon$ around the picked anchor point. The experimental results tabulated in Table VII show that the ReCov operation can play a part in ameliorating the discrimination of the learned deep representations. From Table VII, we can also find that the suggested method is less sensitive to ϵ when it takes values from the set $\{0, 1\text{E}-6, 1\text{E}-5, 1\text{E}-4\}$. However, when the value of ϵ is configured as $1\text{E}-3$, the obtained classification scores are lower than those of other cases on the three used datasets. Based on Theorem 1, we conjecture that the current setting (1E−3) does not meet the condition of being sufficiently small. As a result, some of the main structural information of the learned features will be lost. For a new dataset, we still recommend to exploit the procedures mentioned above for parameter selection because of its rationality and simplicity.

Furthermore, we choose the MDSD dataset as an example to perform 2-D visualization experiments, which enables us to gain an intuitive feeling about the efficacy of the ReCov operation. The final visual results, obtained via the t-SNE technique [47], are illustrated in Fig. 10. Compared with Fig. 10(a), both the intra-subject compactness and the inter-subject separability of the samples reflected in Fig. 10(b) are further enhanced after integrating the ReCov layers into the proposed architecture. In addition, the comparison between Fig. 10(a) and Fig. 9 intuitively certifies that the integration of the two metric learning stages helps probe a more discriminative feature space for classification.

In this part, we design another activation function for the ReCov layer to better verify the rationality and validity of the Eq. (8)-based ReCov operation. This function is similar to Eq. (8), with the only difference being that it adjusts the negative

TABLE VII
COMPARISON (%) UNDER DIFFERENT VALUES OF ϵ.

Datasets	0	1E-6	1E-5	1E-4	1E-3
AFEW	37.19	36.66	37.47	**37.74**	36.39
MDSD	39.49	39.49	**42.05**	40.26	38.46
FPHA	89.04	88.52	**89.39**	88.00	87.13

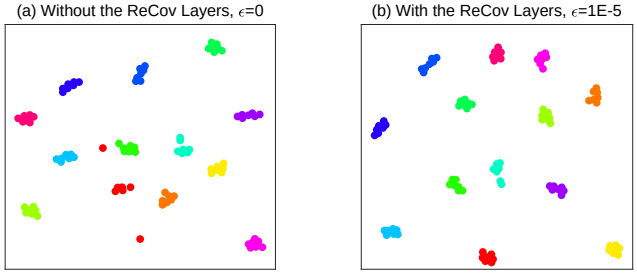

Fig. 10. 2-D visualization of the learned representations on the MDSD dataset, where colors and points denote categories and image set samples, respectively.

TABLE VIII
COMPARISON (%) UNDER DIFFERENT ACTIVATION FUNCTIONS.

Datasets	Proposed-Eq. (8)	Proposed-Eq. (7)	Proposed-Eq. (29)
AFEW	**37.74**	32.88	36.39
MDSD	**42.05**	34.36	35.90
FPHA	**89.39**	74.44	88.70

elements in the interval of $(-\epsilon, 0]$ to 0, *i.e.*,

$$\hat{C}_k(i,j) = \begin{cases} 0, & \text{if } i \neq j \text{ and } \hat{C}_{k-1}(i,j) \in (-\epsilon, 0], \\ \hat{C}_{k-1}(i,j), & \text{otherwise.} \end{cases} \tag{29}$$

The experimental results of the proposed approach achieved on the AFEW, MDSD, and FPHA datasets under different activation functions are given in Table VIII. It is worth noting that the classification performance of Proposed-Eq. (7) is significantly inferior to that of Proposed-Eq. (8) and Proposed-Eq. (29). This further justifies that the negative correlation values of SPD matrix play a crucial role in characterizing the intra-subject data distributions. Imposing sparseness, *i.e.*, Eq. (7), on the elements of SPD matrices will lead to the loss of plenty of structural information (inferred from Theorem 1). Table VIII also reports that the classification results of Proposed-Eq. (29) are lower than those of Proposed-Eq. (8) on the three datasets. This experimentally confirms that performing local statistical information regularization in the scope of $(-\epsilon, 0]$ using Eq. (8) can intensify the negative statistical relevance of some local feature regions in the original visual scene, thus enabling the metric learning terms to be more effective in capturing and learning the inconspicuous yet useful data variability information that hide in such an interval. Hence, the discrepancy of the intra-class data distributions will be narrowed to a certain extent. All in all, these experimental evidences certify the effectiveness of the ReCov regularization in promoting the discriminatory power of the structured representations learned by the proposed SPD network.

I. Parameter Selection

To measure the impact of the trade-off parameters (*i.e.*, λ_1, λ_2, and λ_3) in Eq. (11) on the classification performance of the

TABLE IX
STUDY THE INFLUENCE OF THE TRADE-OFF PARAMETERS λ_1 AND λ_2 ON THE CLASSIFICATION PERFORMANCE OF THE PROPOSED METHOD ON THE AFEW DATASET

	λ_1=10			
	λ_2=1E-4	λ_2=1E-3	λ_2=1E-2	λ_2=1E-1
Acc. (%)	33.69	35.58	35.85	35.31
	λ_1=1.0			
	λ_2=1E-4	λ_2=1E-3	λ_2=1E-2	λ_2=1E-1
Acc. (%)	35.58	35.85	36.93	36.39
	λ_1=0.1			
	λ_2=1E-4	λ_2=1E-3	λ_2=1E-2	λ_2=1E-1
Acc. (%)	36.93	37.20	37.74	37.47
	λ_1=0.01			
	λ_2=1E-4	λ_2=1E-3	λ_2=1E-2	λ_2=1E-1
Acc. (%)	36.66	37.74	37.20	37.47

proposed approach, we make cross-validation experiments on the AFEW dataset as an example. The purpose of introducing these three trade-off parameters to Eq. (11) is to balance the magnitude of the classification term, metric learning term, and reconstruction error term, so as to train effective classifiers for improved classification. Accordingly, we fix the value of λ_3 to 1 and assign a relatively small value to λ_1 and λ_2 to fine-tune the classification performance in this paper. To study the efficacy of λ_1 and λ_2 in regulating the model capacity, the candidate sets of them are respectively set to $\{10, 1.0, 0.1, 0.01\}$ and $\{1\mathrm{E}-1, 1\mathrm{E}-2, 1\mathrm{E}-3, 1\mathrm{E}-4\}$ in the experiments. From Table IX, we have some interesting observations. Firstly, when the value of λ_1 is fixed, our model generally shows a first increasing trend and then a decreasing trend as the value of λ_2 changes. The fundamental reason is that when λ_2 is assigned to a smaller value, the degradation problem will impact the classification ability of our method. In contrast, a larger value of λ_2 makes the learning system focus on deep reconstruction learning, which is also disadvantageous for training discriminative classifiers. Another important finding in Table IX is that when λ_1 takes values from the set $\{0.1, 0.01\}$, the classification accuracy of our method is higher than that of λ_1 whose values are configured as 10 and 1.0, respectively. This is mainly attributed to that assigning a comparatively larger value to λ_1 results in a higher order of magnitude of the metric learning term, compared to the classification term. In this case, the softmax classifier may not be able to fit the probability distribution of different categories learned by the suggested SPD network well, thus causing misclassification. Finally, Table IX delivers that the proposed method is less sensitive to these two trade-off parameters, supporting our assertion that the reconstruction and metric learning terms help to fine-tune the classification performance.

All in all, these experimental observations confirm the complementarity of these two terms in guiding our model to learn more informative features for better decision making. As studied above, our guideline for choosing their values on the four used datasets is to ensure that the metric learning regularizer and the reconstruction error term are at least an order of magnitude lower than the classification term. With this criterion, the softmax classifier can better integrate the gradient information of the metric learning and reconstruction terms to learn a hypersphere with a more reasonable probability distribution for different categories. On the AFEW dataset, the eligible values of λ_1 and λ_2 are set to 0.1 and 0.01, respectively. Table II shows their values on the remaining two datasets. For a new dataset, based on the values of the cross-entropy loss, the aforementioned principle can help the readers quickly determine the initial value ranges of λ_1 and λ_2.

For other ablation studies, please kindly refer to Section II, Section III, Section IV, and Section V of our supplementary material.

J. Discussion

Since the SPD matrix encodes the statistical information between different feature dimensions (attributes) in the original image set data, the multi-stage data compressed sensing will inevitably lead to the loss of some main structural information of the input feature matrices, thus preventing the existing SPD neural networks from going deeper. As stated above, the column full-rank transformation matrices W_k are imposed on the semi-orthogonality, such that optimizing W_k over a compact Stiefel manifold could render optimal solutions. Since $W_k^T W_k = I$, inspired by the paradigm of Euclidean autoencoder, if one can design an autoencoder network in the domain of SPD manifolds, the function composition of successive SPD matrix upsampling and downsampling layers would be able to asymptotically approach an identity mapping (IM) theoretically. Therefore, we build a stacked SPD manifold autoencoder (SSMAE) on the tail of the backbone network with a series of reconstruction error terms (RTs) to train. For simplicity, we denote $\mathcal{M}_2 = W_2^T \pi(W_1 \mathcal{M}_1 W_1^T) W_2$ as the resulting SPD matrix after one upsampling and downsampling operation. As the ReCov and ReEig operations bring about minor perturbations to the eigenvalue space, the nonlinear activation function π (ReEig+ReCov) could preserve the main structural information of the input data. In addition, under the supervision of RTs, W_1 and W_2 may close to each other, resulting in that $||\mathcal{M}_2||_{\mathrm{F}} \rightarrow ||\mathcal{M}_1||_{\mathrm{F}}$. These factors make it possible to create an IM on the SPD manifolds. The experimental results listed in Fig. 8 and Table V demonstrate that our design can speed up the training of deep networks and solve the degradation problem to a certain extent. Besides, the following Table X reports the average F-norm of the feature matrices in the hidden layer of each SMAE on the AFEW dataset. From this table, we can compute that the variance between the three values is 1.15E-09, which is negligible, further supporting our speculation experimentally.

In the Euclidean deep networks, the information degradation during multi-level data transformation is also inevitable. Besides, the ReLU operation that sets all the negative values to 0 also leads to irreversible information loss. As studied in ResNet [50], to make a deeper model produce no higher training error than its shallower counterpart, an ideal solution is to enable the added layers to be constructed as IMs. However, the following two challenges make it difficult to approximate an IM for the underlying mapping $H(\mathcal{E})$: 1) information degradation mentioned above; 2) different from

TABLE X
THE AVERAGE F-NORM OF THE FEATURE MATRICES IN THE HIDDEN LAYER OF EACH SMAE ON THE AFEW DATASET.

Layers	6^{th} layer	14^{th} layer	22^{th} layer
F-norm	1.66E−4	1.21E−4	9.94E−5

SPD networks, it is impossible for the plain network studied in [50] to drive the weights of the multiple nonlinear layers toward (semi-) orthogonal. Hence, instead of approximating a desired $H(\mathcal{E})$, the authors equivalently let these stacked layers learn a residual function: $\mathcal{F}(\mathcal{E}) := \mathcal{H}(\mathcal{E}) - \mathcal{E}$. In this case, if the optimal function is closer to an IM than to a zero mapping, it would be easier for the solver to seek the perturbations with reference to IM, than to learn the function as a new one. The experimental results illustrated in Fig. 7 of [50] support the authors' hypothesis that the residual functions might be generally close to zero.

In short, the designed SPD network tends to approach the IM more easily, while for the residual learning framework, it seems to be easier for the residual function to fit an approximate zero mapping. Other differences between the proposed SPD network and ResNet include: 1) both the inputs and the outputs of our model are structured SPD matrices, rather than the image features of ResNet. In other words, the suggested model is strictly defined on the Riemannian manifolds, other than the Euclidean vector space; 2) the parameter optimization of the designed network is realized by exploiting the stochastic gradient descent (SGD) setting on the Stiefel manifolds with the Riemannian matrix backpropagation for characterizing and preserving the Riemannian geometry of SPD data points, while the conventional SGD-based Euclidean backpropagation is used in ResNet.

V. APPLICATION TO SKELETON-BASED HUMAN ACTION RECOGNITION WITH UNMANNED AERIAL VEHICLES

Due to the rapid motion and the constantly changing attitudes and altitudes of the UAVs during flight, the video sequences captured by UAVs exhibit large variations in viewpoint, resolution, illumination, background information, and geometrical morphology of the object. These factors make the UAV-based computer vision tasks, *e.g.*, person re-identification and pose estimation, rather challenging and gaining increasing attention. In this section, we experiment on the large-scale UAV-Human dataset [68] to further examine the effectiveness of the designed SPD network for the task of skeleton-based human pose recognition.

This dataset consists of 67,428 annotated video sequences for human behavior understanding. Among them, 22,476 videos belonging to 155 action categories are specified for the task of human pose estimation. For the evaluation, we follow the practice introduced in [67] to first convert each action frame into a 51-dimensional feature vector, as each person is marked by 17 body joints with 3D coordinates (illustrated in Fig. 11). Furthermore, a number of 305 action frames in each video clip were selected, resulting in that each video sequence can be represented by an image set matrix of size 51×305. Since some of the action categories in the UAV-Human dataset are performed interactively by two persons, for simplicity, the PCA technique is utilized to transform the 102-dimensional ($17 \times 3 \times 2$) feature vectors into 51-dimensional ones with preserving 99% energy of the data. In this scenario, a total of 22,476 SPD matrices of size 51×51 can be computed for image set modeling. Then, the training and test sets were constructed from the randomly singled out 16,723 SPD matrices using the seventy-thirty-ratio (STR) protocol. On this dataset, the sizes of the network filters are respectively configured as 51×43, 43×37, 37×31, 31×37, 37×31, 31×37, and 37×31. In addition, the learning rate ξ, batch size $\mathbb{B}$, rectification thresholds (ϵ and ζ), margin thresholds (ρ_1 and ρ_2), and trade-off parameters (β, λ_1, λ_2, and λ_3) are configured to be 0.01 (attenuate by a factor of 0.8 every 50 epochs), 30, (1E-5 and 1E-5), (-1.0 and 1.0), and (0.2, 1.0, 0.1, and 1.0), respectively.

Table XI reports the recognition scores of the different methods on the UAV-Human dataset. Note that we run the open-source codes of these reference methods on this dataset, with parameter tuning to obtain optimal classification results currently. According to Table XI, the following observations can be drawn. Firstly, the classification performance of HRGEML and HERML are superior to most of the single geometric model-based classification methods on this dataset. This again certifies that performing image set encoding and learning from a multi-geometric perspective is beneficial to generate a more discriminative subspace for classification. Secondly, the classification scores of PML, LEML, and SPDML-AIM are lower than those of SymNet and GEMKML, further demonstrating that compared with shallow linear learning scheme, deep nonlinear metric learning is more effective in alleviating intra-class diversity and inter-class ambiguity of input data. Another meaningful finding from Table XI is that the learning capacity of GEMKML and SymNet is inferior to that of GrNet and SPDNet on this dataset, respectively. This further illustrates that compared with the lightweight Riemannian networks that do not require optimization by matrix backpropagation, the end-to-end Riemannian deep learning mechanism is equipped to mine fine-grained geometric features with better discriminability for the original image set data, especially on the large-scale datasets.

Finally, the proposed method achieves an accuracy of 45.54% on the UAV-Human dataset, which is 3.23% higher than that of SPDNet. Besides, when the two metric learning stages are removed from the proposed SSMAE module, the classification score obtained by the simplified network (called Proposed-nMLS) on this dataset is 44.73%, still outperforming SPDNet and other competitors. These experimental evidences not only confirm that the designed SPD matrix learning method can also learn useful geometric information for improved classification on large-scale dataset, but also further verify that the complementary role of the designed SSMAE and metric learning modules is an important factor to enhance the learning capacity of the baseline network.

VI. CONCLUSION

This paper carries out a basic exploration of the combination of deeper neural networks and Riemannian metric

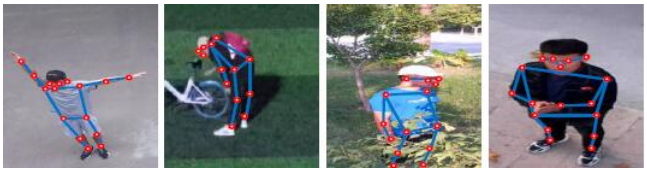

Fig. 11. Some instances of the UAV-Human dataset.

TABLE XI
ACCURACY COMPARISON (%) ON THE UAV-HUMAN DATASET.

Methods	CDL [5]	PML [30]	LEML [11]	HERML [39]
Acc.	31.11	10.66	21.83	34.18
Methods	SPDNet [25]	SPDML-AIM [20]	GrNet [24]	SPDNetBN [26]
Acc.	42.31	22.69	35.23	43.28
Methods	GEMKML [10]	HRGEML [67]	SymNet [27]	**Proposed**
Acc.	34.67	36.10	35.89	**45.54**

learning on the SPD manifolds to improve the image set classification performance by addressing the issues of intra-class diversity and inter-class similarity. Specifically, we first build a stacked SPD manifold autoencoder (SSMAE) on the tail of the SPD backbone to explore a feasible way to increase the depth of representations without causing model degradation. Whereafter, the SSMAE module is equipped with two successive metric learning stages to learn the intra- and inter-class variations of deep representations generated by the proposed network. This not only facilitates supervising a powerful manifold-to-manifold transforming network, but also helps to train effective classifiers. In this article, we also introduce the ReCov layer for the designed architecture to perform local statistics nonlinear regularization of SPD data. Extensive experiments show that compared with the SOTA methods, the propsoed approach is an effective candidate in improving the image set classification performance, even with limited data. Besides, a series of ablation studies justify the significance of each component of our model in promoting the learning capacity of the baseline network.

Since the designed metric learning terms need to traverse all the neighboring points within a batch of samples and compute the intra- and inter-class scatter matrices, the computational burden of the propsoed model is inevitably higher than that of the baseline network. Therefore, in the future, more efforts will be made to introduce new metric learning frameworks with comparatively higher computational efficiency. Considering that the temporal information is an important factor in describing video data, another possible future work is to integrate a spatial-temporal SPD aggregation module into the designed network to perform coarse-to-fine temporal modeling. In addition, we plan to generalize the proposed method to other computer vision tasks, such as visual object tracking and person re-identification, to support wide applications.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (62020106012, U1836218, 61672265, 62106089, 62006097), the 111 Project of Ministry of Education of China (B12018), the Natural Science Foundation of Jiangsu Province (BK20200593), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21-2006), the UK EPSRC Grant MVSE (EP/V002856/1), EPSRC/dstl/MURI project EP/R018456/1, and the National Key Research and Development Program of China (2017YFC1601800).

REFERENCES

[1] Hamm. J. H and Lee. D. D. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *ICML*, 2008, pp. 376-383.
[2] Tuzel. O, Porikli. F and Meer. P. Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008, pp. 1713-1727.
[3] Tosato. D, Farenzena. M, Spera. M, Murino. V and Cristani, M. Multi-class classification on riemannian manifolds for video surveillance. In *ECCV*, 2010, pp. 378-391.
[4] Sanin. A, Sanderson. C, Harandi. M and Lovell. B. C. Spatio-temporal covariance descriptors for action and gesture recognition. In *WACV Workshop*, 2013, pp. 103-110.
[5] Wang. R, Guo. H, Davis. L. S and Dai. Q. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, 2010, pp. 2496-2503.
[6] Harandi. M and Salzmann. M. Riemannian coding and dictionary learning: Kernels to the rescue. In *CVPR*, 2015, pp. 3926-3935.
[7] Jayasumana. S, Hartley. R, Salzmann. M, Li. H and Harandi. M. Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In *CVPR*, 2013, pp. 73-80.
[8] Gao. Z, Wu. Y, Harandi. M and Jia. Y. A robust distance measure for similarity-based classification on the SPD manifold. *IEEE Trans. Neural Netw. Learn. Syst.,* 2020, pp. 3230-3244.
[9] Wang. R, Wu. X.-J, Chen K.-X and Kittler. J. Multiple Riemannian Manifold-valued Descriptors based Image Set Classification with Multi-Kernel Metric Learning. *IEEE Trans. Big Data*, 2022, pp. 753-769.
[10] Wang. R, Wu. X.-J and Kittler. J. Graph embedding multi-kernel metric learning for Image set classification with Grassmannian manifold-valued features. *IEEE Trans. Multimedia*, 2021, pp. 228-242.
[11] Huang. Z, Wang. R, Shan. S, Li. X and Chen. X. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *ICML*, 2015, pp. 720-729.
[12] Zhou. L, Wang. L, Zhang. J, Shi. Y and Gao. Y. Revisiting metric learning for spd matrix based visual representation. In *CVPR*, 2017, pp. 3241-3249.
[13] Lu. J, Liong. V. E, Zhou. J. Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, pp. 1979-1993.
[14] Zhang. G, Sun. H, Zheng. Y, Xia. G, Feng. L and Sun. Q. Optimal discriminative projection for sparse representation-based classification via bilevel optimization. *IEEE Trans. Circuits Syst. Video Technol.*, 2020, pp. 1065-1077.
[15] Zhang. G, Ge. Y, Dong. Z, Wang. H, Zheng. Y and Chen. S. Deep high-resolution representation learning for cross-resolution person re-Identification. *IEEE Trans. Image Process.*, 2021, pp. 8913-8925.
[16] Chakraborty. R, Bouza. J, Manton. J and Vemuri. B. C. Manifoldnet: A deep neural network for manifold-valued data with applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022, pp. 799-810.
[17] Barachant. A, Bonnet. S, Congedo. M and Jutten. C. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, 2013, pp. 172-178.
[18] Wright. J, Yang. A. Y, Ganesh. A, Sastry. S. S and Ma. Y. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008, pp. 210-227.
[19] Cheng. G, Zhou. P and Han. J. W. Duplex metric learning for image set classification. *IEEE Trans. Image Process.*, 2017, pp. 281-292
[20] Harandi. M, Salzmann. M and Hartley. R. Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, pp. 48-62.
[21] Huang. Z, Wang. R, Shan. S, Van Gool. L and Chen. X. Cross Euclidean-to-Riemannian metric learning with application to face recognition from video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, pp. 2827-2840.
[22] Ionescu. C, Vantzos. O and Sminchisescu. C. Training deep networks with structured layers by matrix backpropagation. *arXiv preprint arXiv:1509.07838*, 2015
[23] Wang. W, Wang. R, Huang. Z, Shan. S and Chen. X. Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets. *IEEE Trans. Image Process.*, 2018, pp. 151–163.

[24] Huang. Z, Wu. J and Van. G. L. Building deep networks on Grassmann manifolds. In *AAAI*, 2018, pp. 3279-3286.
[25] Huang. Z and Van. G. L. A riemannian network for spd matrix learning. In *AAAI*, 2017, pp. 2036-2042.
[26] Brooks. D, Schwander. O, Barbaresco. F, Schneider. J. Y and Cord. M. Riemannian batch normalization for SPD neural networks. In *NeurIPS*, 2019.
[27] Wang. R, Wu. X.-J and Kittler. J. SymNet: A simple symmetric positive definite manifold deep learning method for image set classification. *IEEE Trans. Neural Netw. Learn. Syst.*, 2022, pp. 2208-2222.
[28] Lu. J, Wang. G and Moulin. P. Localized multifeature metric learning for image-set-based face recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 2016, pp. 529-540.
[29] Gao. S, Zeng. Z, Jia. K, Chan. T. H and Tang. J. Patch-set-based representation for alignment-free image set classification. *IEEE Trans. Circuits Syst. Video Technol.*, 2016, pp. 1646-1658.
[30] Huang. Z, Wang. R, Shan. S and Chen. X. Projection metric learning on Grassmann manifold with application to video based face recognition. In *CVPR*, 2015, pp. 140-149.
[31] Huang. Z, Wang. R, Li. X, Liu. W, Shan. S, Van. G. L and Chen. X. Geometry-aware similarity learning on SPD manifolds for visual recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 2018, pp. 2513-2523.
[32] Wang. R, Wu. X.-J, Liu. Z and Kittler. J. Geometry-aware graph embedding projection metric learning for image set classification. *IEEE Trans. on Cogn. Develop. Syst.*, 2021, doi: 10.1109/TCDS.2021.3086814.
[33] Harandi. M, Sanderson. C, Shirazi. S and Lovell. B. C. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. In *CVPR*, 2011, pp. 2705-2712.
[34] Pennec. X, Fillard. P and Ayache. N. A riemannian framework for tensor computing. *Int. J. Comput. Vis.*, 2006, pp. 41-66
[35] Sra. S. Positive definite matrices and the S-divergence. *Proc. Amer. Math. Soc.*, 2016, pp. 2787-2797.
[36] Harandi. M, Sanderson. C, Hartley. R and Lovell. B. C. Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In *ECCV*, 2012, pp. 216-229.
[37] Lu. J, Wang. G and Moulin. P. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, 2013, pp. 329-336.
[38] Duan. Y, Lu. J, Feng. J and Zhou. J. Deep localized metric learning. *IEEE Trans. Circuits Syst. Video Technol.*, 2018, pp. 2644-2656.
[39] Huang. Z. W, Wang. R. P, Shan. S. G and Chen. X. L. Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning. *Pattern Recognit.*, 2015, pp. 3113-3124.
[40] Absil. P. A, Mahony. R and Sepulchre. R. Optimization algorithms on matrix manifolds. *Princeton University Press*, 2009.
[41] Edelman. A, Arias. T. A and Smith. S. T. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 1998, pp. 303-353.
[42] Nguyen. X. S, Brun. L, Lézoray. O and Bougleux. S. A neural network based on SPD manifold learning for skeleton-based hand gesture recognition. In *CVPR*, 2019, pp. 12036-12045.
[43] Dhall. A, Goecke. R, Joshi. J, Sikka. K and Gedeon. T. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *ICMI*, 2014, pp. 461-466.
[44] Sun. H, Zhen. X, Zheng. Y, Yang. G, Yin. Y and Li. S. Learning deep match kernels for image-set classification. In *CVPR*, 2017, pp. 3307-3316.
[45] Shroff. N, Turaga. P and Chellappa. R. Moving vistas: Exploiting motion for describing scenes. In *CVPR*, 2010, pp. 1911-1918.
[46] Arsigny. V, Fillard. P, Pennec. X and Ayache. N. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.*, 2007, pp. 328-347.
[47] Maaten. L. V. D and G. Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 2008, pp. 2579-2605.
[48] Fisher. W. D, Camp. T. K and Krzhizhanovskaya. V. V. Anomaly detection in earthdam and levee passive seismic data using support vector machines andautomatic feature selection. *J. Comput. Sci.*, 2017, pp. 143-153.
[49] Garcia-Hernando. G, Yuan. S, Baek. S and Kim. T. K. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *CVPR*, 2018, pp. 409-419.
[50] He. K, Zhang. X, Ren. S and Sun. J. Deep residual learning for image recognition. In *CVPR*, 2016, pp. 770-778.
[51] Shen. Z, Wu. X.-J and Xu. T. FEXNet: foreground extraction network for human action recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 2022, pp. 3141-3151.
[52] Feichtenhofer. C, Pinz. A and Zisserman. A. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016, pp. 1933-1941.
[53] Zhang. X, Wang. Y, Gou. M, Sznaier. M and Camps. O. Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In *CVPR*, 2016, pp. 4498-4507.
[54] Ohn-Bar. E and Trivedi. M. M. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Trans. Intell. Transport. Syst.*, 2014, pp. 2368-2377.
[55] Zanfir. M, Leordeanu. M and Sminchisescu. C. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013, pp. 2752-2759.
[56] Rahmani. H and Mian. A. 3D action recognition from novel viewpoints. In *CVPR*, 2016, pp. 1506-1515.
[57] Oreifej. O and Liu. Z. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*, 2013, pp. 716-723.
[58] Nair. V and Hinton. G. E. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010, pp. 807-814.
[59] Du. Y, Wang. W and Wang. L. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015, pp. 1110-1118.
[60] Kim. T. S and Reiter. A. Interpretable 3d human action analysis with temporal convolutional networks. In *CVPRW*, 2017, pp. 1623-1631.
[61] Yan. S, Xiong. Y and Lin. D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018, pp. 7444-7452.
[62] Hu. J.-F, Zheng. W.-S, Lai. J and Zhang. J. Jointly learning heterogeneous features for RGB-D activity recognition. In *CVPR*, 2015, pp. 5344-5352.
[63] Vemulapall. R, Arrate. F and Chellappa. R. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 2014, pp. 588-595.
[64] Garcia-Hernando. G and Kim. T. K. Transition forests: Learning discriminative temporal transitions for action recognition and detection. In *CVPR*, 2017, pp. 432-440.
[65] Lohit. S, Wang. Q and Turaga. P. Temporal transformer networks: Joint learning of invariant and discriminative time warping. In *CVPR*, 2019, pp. 12426-12435.
[66] Tekin. B, Bogo. F and Pollefeys. M. H+O: Unified egocentric recognition of 3d hand object poses and interactions. In *CVPR*, 2019, pp. 4511-4520.
[67] Chen. Z, Xu. T, Wu. X.-J, Wang. R and Kittler. J. Hybrid riemannian graph-embedding metric learning for image set classification. *IEEE Trans. Big Data*, 2021, doi: 10.1109/TBDATA.2021.3113084.
[68] Li. T, Liu. J, Zhang. W, Ni. Y, Wang. W and Li. Z. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *CVPR*, 2021, pp. 16266-16275.
[69] Yang. X, Wang. Y, Chen. K, Xu. Y and Tian. Y. Fine-grained object classification via self-supervised pose alignment. In *CVPR*, 2022, pp. 7399-7408.