Hierarchical Gaussian Descriptors with Application to Person Re-Identification

Tetsu Matsukawa[†], Takahiro Okabe[†], Einoshin Suzuki, and Yoichi Sato[‡] [†] *Member,* [‡] *Senior Member, IEEE*

Abstract—Describing the color and textural information of a person image is one of the most crucial aspects of person re-identification (re-id). Although a covariance descriptor has been successfully applied to person re-id, it loses the local structure of a region and mean information of pixel features, both of which tend to be the major discriminative information for person re-id. In this paper, we present novel meta-descriptors based on a hierarchical Gaussian distribution of pixel features, in which both mean and covariance information are included in patch and region level descriptions. More specifically, the region is modeled as a set of multiple Gaussian distributions, each of which represents the appearance of a local patch. The characteristics of the set of Gaussian distributions are again described by another Gaussian distribution. Because the space of Gaussian distribution is not a linear space, we embed the parameters of the distribution into a point of Symmetric Positive Definite (SPD) matrix manifold in both steps. We show, for the first time, that normalizing the scale of the SPD matrix enhances the hierarchical feature representation on this manifold. Additionally, we develop feature norm normalization methods with the ability to alleviate the biased trends that exist on the SPD matrix descriptors. The experimental results conducted on five public datasets indicate the effectiveness of the proposed descriptors and the two types of normalizations.

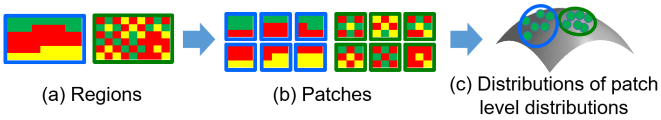
Index Terms—Person re-identification, image feature descriptor, Gaussian distribution, Riemannian geometry, symmetric positive definite matrices, log-Euclidean Riemannian metric

1 Introduction

Appearance matching of person images observed in disjoint camera views, referred to as person re-identification (re-id), is receiving increasing attention, mainly because of its broad range of applications [1], [2], [3]. In this task, the person images are captured from various viewpoints and under different illuminations, resolutions, human poses, and against various background environments. These large intra-personal variations in person images cause considerable difficulties during attempts to match the person. In addition, similar clothes among different persons add further challenges.

Person images are low in resolution and are further characterized by large pose variations; consequently, only coarse information of person appearance would be robustly described. It has been proven that the most important clue for person re-id is color information such as color histograms and color name descriptors [4]. Because they cannot sufficiently differentiate different persons of similar colors, textural descriptors such as a Local Binary Pattern (LBP) and the responses of filter banks are often combined with color descriptors [5], [6], [7]. To enhance the robustness against complex combination effects of the variations, supervised learning methods are often applied to the descriptors [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15].

A covariance descriptor [16] describes a region of interest as a covariance matrix of pixel features. The covariance matrix describes a statistical dependency between elements within pixel features and provides a natural way to fuse different modalities, *e.g.,* the color and texture, of pixel features into a single meta-descriptor. Because the covariance matrix is obtained by averaging the features inside the region, it remedies the effects of noise and spatial misalignments, *e.g.,* variations caused by pose changes. In addition, the descriptor only requires coarse information around pixels, which makes it suitable for processing low-quality images captured by surveillance cameras [17]. Consequently, the covariance descriptor has been successfully applied to person re-id [18], [19].



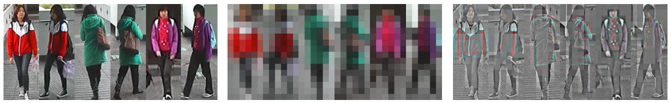
(a) Regions (b) Patches (c) Distributions of patch level distributions

Fig. 1. Importance of hierarchal distribution: (a) Regions that have the same distribution (mean/covariance) of pixel features (each color indicates the same feature vector). (b) Local patches with a different pixel feature distribution inside the regions. (c) Regions can be distinguished via distributions of patch level distributions.

In this paper, we propose novel meta-descriptors based on the hierarchical Gaussian distribution of pixel features. More specifically, our descriptors densely extract local patches inside a region and regard the region as a set of local patches. The region is firstly modeled as a set of multiple Gaussian distributions, each of which represents the appearance of one local patch. We refer to such a Gaussian distribution representing each local patch as a *patch Gaussian*. The characteristics of the set of patch Gaussians are again described by another Gaussian distribution. We refer to this Gaussian distribution as a *region Gaussian*. In both steps, we embed the parameters of one Gaussian distribution into a point on the manifold of Symmetric Positive Definite (SPD) matrices where several Riemannian metrics on the manifold are defined [20], [21].

Our motivation for the use of a hierarchical distribution is to develop discriminative meta-descriptors by focusing on the

- *T. Matsukawa and E. Suzuki are with the Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan. E-mail: {matsukawa, suzuki}@inf.kyushu-u.ac.jp*
- *T. Okabe is with the Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology, Fukuoka, Japan. E-mail: okabe@ai.kyutech.ac.jp*
- *Y. Sato is with the Institute of Industrial Science, The University of Tokyo, Tokyo, Japan. E-mail: ysato@iis.u-tokyo.ac.jp*
  Corresponding author: T.Matsukawa.



(a) Original (b) Local mean (c) Mean removed

Fig. 2. Importance of mean: (a) Original images. (b) Mean RGB values of local patches. (c) Mean removed images. Determining the same persons from (b) is easy, whereas difficult from (c).

structural appearance of person images. A person's clothes consist of local parts, each of which has local color/texture structures. The spatial arrangement of these parts determines the global structural appearance. However, most of the existing meta-descriptors [16], [22], [23], [24], [25], [26] are based on the global distribution of pixel features inside a region, and thus the local structure of the person's image is lost. In contrast, we describe the global distribution using the local distribution of the pixel features. As illustrated in Fig. 1, this enables us to distinguish colors with the same global distribution but different local structures.

We use the Gaussian distribution as a base component of the hierarchy. The motivation for the use of this distribution originates from the importance of the mean color of local parts. Although hierarchical covariance representations have been proposed [27], [28], each hierarchy lacks the mean information. The absence of the mean information is crucial when applied to person re-id. This is because the clothes a person wears tend to consist of a small number of colors in each local part, and therefore the mean color in the local parts tends to be the major information that enables to distinguish persons. Because the mean color is calculated by averaging the local part pixel values, it is robust to noise and represents the global description of the local part. From Fig. 2, we can also see that the mean color represents the most distinguished color feature of the person image, but for the mean removed images, all of these distinguished color features have gone.

We name the proposed meta-descriptors **H***ierarchical* **G***aussian* **D***escriptors* (**HGD**s). Although its concept is simple, the HGDs introduce a new challenge in handling the SPD matrix manifold hierarchically. This paper also provides an analysis of this topic through our attempt to improve the supervised person re-id. The main contributions of this paper[1] are summarized below: (Ⅰ) We present effective handcrafted descriptors for person re-id. The HGDs provide a conceptually simple and consistent way to generate discriminative features that describe the color and textural information simultaneously. (Ⅱ) We propose the hierarchical use of Gaussian embedding of pixel features (GOG). We experimentally validate the importance of both the hierarchical distribution and the mean information of pixel features. (Ⅲ) We define a scale normalization of an SPD matrix and validate its importance for HGDs. Based on this normalization, we develop a new zero-mean Gaussian embedding. The HGDs based on this embedding can achieve performance close to the HGDs using the original embedding of Lovrić *et al.* [29] with smaller dimensionality and computational cost (ZOZ). (Ⅳ) With the aim of normalizing the norm of HGDs, we point out the biased trend of the SPD matrices in the Log-Euclidean (LE) tangent space. To alleviate this effect, we propose norm normalization methods accompanied with bias removal of the SPD matrix manifold. In the previous version [30], we validated the effectiveness of this normalization with the extrinsic statistics of the Riemannian manifold. In this paper, we extend this normalization with the intrinsic statistics.

2 Related Work

Feature representation in person re-id. Several feature representations have been sought to obtain invariances against challenging variations in person appearances. The localization of a person's body parts [31] and local feature accumulation exploiting the symmetry of a person [32] were proposed to enhance the invariance to pose changes. A color invariant signature [33] and a color name descriptor [4] were proposed to enhance the invariance of color under different illumination conditions. Self-similarity among local covariance matrices was proposed to enhance the invariances to both illumination and background variations [34]. Rare appearances, which usually remain in different camera views, *e.g.,* rare-colored coats, are matched by saliency learning [35]. Attribute-based descriptors obtain a lingual description of person images, which are also invariant under the imaging conditions [36]. These invariant features discard much information on person images, which can vary under different camera views, yet are helpful to distinguish different persons.

Meanwhile, feature representations for metric learning are typically rather naïve such that they retain discriminative information of person appearances, *e.g.,* high-dimensional features composed of densely sampled color histograms, LBPs, and SIFTs [5], [6], [7], [10], [11], [12]. Although color histograms are sensitive to variations in illumination, metric learning learns the variations under different camera views from training data, and typically outperforms color normalization [8], [11]. Because the person images are coarsely aligned in a vertical direction, creating histograms with horizontal strips is a common successful strategy to enhance the view invariance [3], [4], [13]. Local Maximal Occurrence (LOMO) [11] extended the local discrimination of this approach, by utilizing a two-stage representation in which local histograms of pixel features within local patches are first constructed, and then maximal histogram bins are taken along the horizontal strips. Unfortunately, the max-pooling step discards all of the non-maximal local histogram values along a strip, thereby losing much of the information of multiple local patches. This limits the ability to discriminate among different persons who wear clothes consisting of parts that are visually different.

Lately, the use of Convolutional Neural Networks (CNNs) [37] has gradually been improving the accuracy of person re-id [38], [39], [40], [41], [42], [43], [44]. The high performance of CNN relies on its deep hierarchal architecture with millions of parameters. It typically requires a considerable number of labeled training samples and the model is exposed to over-fitting risk when sufficient training samples are unavailable [13]. Several recent studies circumvented this issue by transferring knowledge from large pre-training datasets, *e.g.,* extracting the features of lower layers from a pre-trained model [13] or conducting fine-tuning on the target dataset [3], [45]. Nevertheless, deep CNN still requires an expensive GPU with large memory capacity to store the model parameters.

In this work, we enhance the discriminative ability of LOMO-like two-stage representation for supervised person re-id. Enhancing the representation only with first-order statistics in Euclidean space is inherently difficult, *e.g.,* simple average pooling on local histograms coincides with a global histogram. Therefore, we focus on the hierarchical distribution on the Riemannian manifold. In contrast to CNN, HGDs require neither training samples nor model parameters (except for optional norm normalization).

Meta-descriptors of local features. A covariance descriptor sum-

[1] This paper is a substantially extended version of our paper in the CVPR2016 proceedings [30]. The HGDs refer to the generic name including the GOG descriptor proposed in [30] and its simplification ZOZ descriptor.

marizes the local features within a region via second-order statistics [16]. Its advantages are robustness against noise and changes in pose/illumination. Several studies extended the linear relation between the feature elements of a covariance matrix to nonlinear relations [46], [47], [48], [49], *e.g.,* Brownian covariance, which was used to measure the degree of all kinds of possible relations between feature elements [49]. Nevertheless, a distinct drawback, *i.e.,* the absence of the mean information, remains unsolved in these extensions.

A natural way to endow the covariance descriptor with the mean information would be an extension to the Gaussian distribution. The earliest studies [23], [25] measured the distance between Gaussian descriptors by the metric of an affine transformation matrix [23] and the α-divergence [50]. Both of these distances entangle with the parameters of the two Gaussian distributions, thereby complicating their processing unlike in the Euclidean space. Recent Gaussian descriptors solved this limitation by using the metrics on the SPD matrix manifold [22], [26], [51]. Notably, because of its convenience in the joint presence of the mean vector and the covariance matrix in one SPD matrix, the embedding proposed by Lovrić *et al.* [29] is rising in popularity [52], [53], [54]. Differently from the above embeddings, Li *et al.* introduced the structure of Lie group into the space of Gaussian distributions, *i.e.,* the geometric structure of the Riemannian manifold and the algebraic structure of the smooth group [55]. Based on the isomorphisms of Lie group, they developed two embedding methods. One method directly embeds the Gaussian distribution into Euclidean space from a subgroup of upper triangular matrices via the matrix logarithm. The other method first embeds Lie subgroup into the space of SPD matrices and then into Euclidean space. Interestingly, the latter embedding shares a similar SPD matrix to that of Lovrić *et al.* [29]. In an application for unsupervised person re-id, Ma *et al.* [24] accounted for a slight concern regarding the sensitivity of the mean color to illumination changes by applying gray world color normalization as preprocessing of the Gaussian descriptor.

As hierarchical meta-descriptors, several summarized representations on local covariance matrices were proposed. In L^{2}ECM [27], a vector map was presented, which was obtained by mapping the local covariance matrices into the LE tangent space. The covariance matrix on the vectorized covariance matrices was used for image representation. In another approach [56], [57], [58], [59], image representation was obtained by employing coding-based summaries *e.g.,* Bag-of-Words [60], Fisher Vector (FV) [61], and VLAD [62]. With the help of feature distributions on the training data, these summaries flexibly describe the distributions of local features. Unfortunately, the accuracies of the coding-based summaries are highly dependent on the training data used for codebook learning. This drawback is not preferable for person re-id because persons who need to be matched are not necessarily included in the training dataset. Although L^{2}ECM was extended into a vector map of local Gaussian distributions [55] concurrently to our work [30], the coding-based summary on local Gaussian distributions was still used for image representation.

In contrast to the existing meta-descriptors, HGDs include both mean and covariance information in each hierarchy and do not require data-dependent codebook learning.

Metrics for SPD matrix manifold. The Affine Invariant Riemannian Metric (AIRM) [20] and Log-Euclidean Riemannian Metric (LERM) [21] are well-known metrics for the SPD matrix manifold. AIRM entangles two input matrices for distance calculation to achieve the invariance to affine transformation. In general, the tangent space locally approximates the geodesic on the manifold in the Euclidean space. Tuzel *et al.* confirmed that the tangent space on the mean point of the given SPD matrices minimizes the approximation error of the geodesic distances of AIRM [63]. Tosato *et al.* proved that the SPD matrix manifold is a homogeneous space which means that any tangent pole preserves the neighborhood relation between the points on the manifold [17]. From a computational point of view, they suggested that the best choice of the tangent pole is the identity matrix. In fact, this tangent space is equivalent to that of LERM [21]. Tosato *et al.* also showed that detecting the role of curvature by a Campbell-Baker-Haussdorff expansion improves the Euclidean distance in the LE tangent space [17]. Except for this work, previous SPD matrix-based meta-descriptors treated the LE tangent space only as in the Euclidean space without any concern [22], [26], [27], [51].

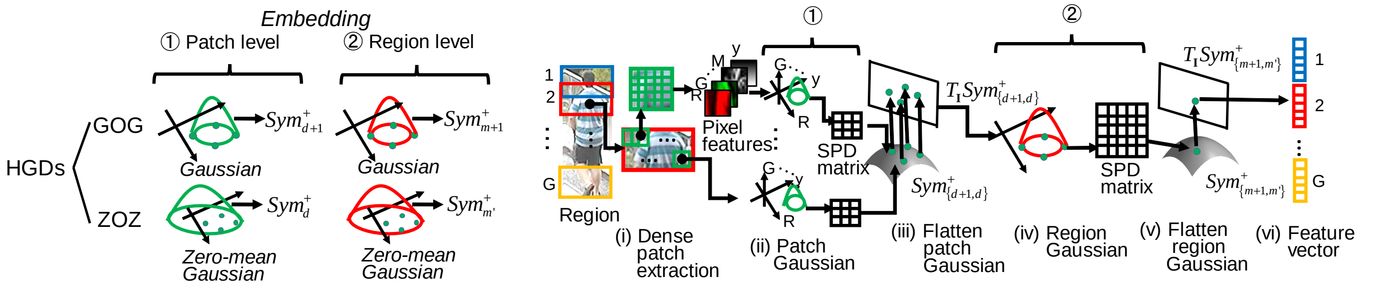
Apart from these two Riemannian metrics, recent studies of embedding CNN features have shown the superiority of Matrix Power Normalization (MPN) against LERM [64], [65], [66]. Covariance matrices that were applied MPN enable a robust covariance estimation [65]; at the same time, MPN corresponds to Power Euclidean Metric (PoEM) on the Riemannian manifold [67]. Li *et al.* proved that PoEM approximates LERM [65] and the distance of PoEM is also decoupled. Notably, PoEM allows non-negative eigenvalues for the embedding matrix, whereas LERM requires strictly positive values. From a computational perspective, the matrix power operation is more suitable for backpropagation than the matrix logarithm operation of LERM [65], [66]. We note that, in this work, strictly positive definite matrices are processed without involving backpropagation. In such a case, LERM is known to be a superior Riemannian metric than PoEM [68].

Unfortunately, our analysis reveals LERM can produce largely biased values in the tangent space, and this property degrades the hierarchal representation on the Riemannian manifold. To overcome this problem, we define a matrix scale normalization and propose bias removal before feature norm normalization.

3 Hierarchical Gaussian Descriptors

In this section, we propose two **H***ierarchical* **G***aussian* **D***escriptors* (**HGD**s). Both HGDs are based on a common pipeline that follows two motivations: ($\mathbb{I}$) The hierarchical distribution is discriminative because it can distinguish colors with a similar global distribution but a different local distribution of pixel features (Fig. 1). ($\mathbb{II}$) When constructing the hierarchal descriptor, the mean information of pixel features in local parts tends to be major discriminative information (Fig. 2). We extract two HGDs named **G***aussian* **O***f* **G***aussians* (**GOG**) and **Z***ero-mean Gaussian* **O***f* **Z***ero-mean Gaussians* (**ZOZ**) descriptors by changing the base Gaussian embedding of the common pipeline (Fig. 3(a)). Each of the HGDs provides the features of regions in vector form to enable conventional metric learning methods to be easily applied.

We outline the common pipeline of HGDs in Fig. 3 (b). The feature representation of a person image is obtained by adopting a part-based model, which divides a person image into G regions. The division of a person image is arbitrary, *e.g.,* the estimation of human body parts could be used. In this paper, we assume the G regions are given in advance. As regions, we use fixed horizontal stripes to enhance the view invariance [11]. For each region, we characterize each pixel by low-level features such as its color and gradient. We summarize them in a two-level (patch/region)



(a) Descriptors (b) Common pipeline

Fig. 3. **Hierarchical Gaussian Descriptors (HGDs):** (a) The two descriptors we extract by changing the Gaussian embeddings of patch/region levels. (b) Common pipeline for each descriptor: (i) Densely extract local patches located inside each region. (ii) Describe each of these local patches via a Gaussian distribution of pixel features which we refer to as a patch Gaussian. (iii) Flatten and vectorize each of the patch Gaussians by considering their underlying Riemannian geometry. (iv) Summarize the patch Gaussians inside a region into a region Gaussian. (v) Flatten the region Gaussian and create a feature vector. (vi) Concatenate the feature vectors extracted from all regions into one vector.

hierarchical distribution. In each hierarchy, we summarize the feature distribution by one of the following two embeddings: the Lovrić's Gaussian embedding [29] and the Zero-mean Gaussian (ZmG) embedding, which we propose in §3.3.

3.1 Pixel Features

Let us focus on one of the G regions of a person image. We describe the local structure of the region by densely extracting squared (k × k pixels) patches with p pixel intervals (Fig. 3 (b-i)). In order to characterize each pixel in the patch, we extract a d-dimensional feature vector $\boldsymbol{f}_i$ for every pixel i. The feature vector can consist of any type of features, such as the color, intensity, gradient orientation, and filter response.

Because the number of pixels in each patch is small, the dimension d is preferable to be low to ensure that the estimation of the covariance matrices of the patch Gaussians in the next step is robust. In this work, we extract eight-dimensional pixel features defined as:

$$\boldsymbol{f}_i = [y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, R, G, B]^T, \qquad (1)$$

where y is the pixel location in the vertical direction, $M_{\theta \in \{0^\circ, \dots, 270^\circ\}}$ are the magnitudes of the pixel intensity gradient along four orientations, and R, G, B are the color channel values. Each dimension of $\boldsymbol{f}_i$ is linearly stretched to the range [0, 1] to equalize the scales of the different feature values.

The pixel location is introduced to leverage the spatial information within each region. Our use of the vertical image location only originates from the analysis in [24]: the person images tend to be well aligned in the vertical direction, whereas changes in the pose/viewpoint cause a large misalignment in the horizontal direction. Note that it would be preferable to set y_i from the top (or center) of the current region as in [22]. However, each pixel belongs to multiple regions, and such a setting would increase the computational complexity. Because person images are coarsely aligned, we directly set the axis of y_i from the top of the image.

The gradient information is introduced to provide the textural information of clothes. The gradient orientation $O = \arctan(I_y/I_x)$ is calculated from the x- and y-derivatives I_x, I_y of the intensity I. We quantize the orientation into four bins: $O_{\theta \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}}$. To compensate the loss of information by the quantization, we use soft voting into two nearby orientation bins. The voting weights are linearly determined based on the distances from the quantized orientations. We focus on high gradient edges by multiplying the gradient magnitude $M = \sqrt{I_y^2 + I_y^2}$ by the quantized orientation O_θ to obtain the oriented gradient magnitude: $M_\theta = M O_\theta$.

Color information is the most important clue for person re-id. We use the color channel values of the most basic color space: RGB. We extend our pixel features to other color spaces, *e.g.,* Lab, HSV, and nRGB in §3.5.

3.2 Patch Level Summarization

After extracting the pixel features inside a patch, we summarize them via the most classical parametric distribution, which has the mean and covariance as parameters: Gaussian distribution (Fig. 3 (b-ii)). For every patch s, we model the feature vectors as the patch Gaussian $\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ defined as,

$$\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{f} - \boldsymbol{\mu}_s)^T \boldsymbol{\Sigma}_s^{-1} (\boldsymbol{f} - \boldsymbol{\mu}_s)\right)}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_s|}, \qquad (2)$$

where $\boldsymbol{\mu}_s$ is the mean vector, $\boldsymbol{\Sigma}_s$ is the covariance matrix of the sampled patch s, and $|\cdot|$ is the determinant of the matrix. The mean vector and the covariance matrix are respectively estimated by: $\boldsymbol{\mu}_s = \frac{1}{n_s} \sum_{i \in \mathcal{L}_s} \boldsymbol{f}_i$ and $\boldsymbol{\Sigma}_s = \frac{1}{n_s - 1} \sum_{i \in \mathcal{L}_s} (\boldsymbol{f}_i - \boldsymbol{\mu}_s)(\boldsymbol{f}_i - \boldsymbol{\mu}_s)^T$, where $\mathcal{L}_s$ is the area of the sampled patch s and n_s denotes the number of pixels in $\mathcal{L}_s$.

Note that the densely sampled mean vectors and covariance matrices can be efficiently calculated by using integral images [63]. Because regions can overlap each other, we construct the integral images of the pixel features for the entire person image rather than creating them for each region.

A Gaussian Mixture Model (GMM) might be used for a more precise description. Because a local patch is expected to consist of a small number of colors/textures, we assume the unimodal Gaussian to be sufficient for describing its feature distribution.

Gaussian Embedding (Lovrić)

As explained above, our descriptors are summarized representations of the patch Gaussians inside a region. This summarization requires mathematical operations to obtain the mean or covariance of the Gaussian. From the viewpoint of information geometry, the space of probability distributions is considered as a Riemannian manifold to which the Euclidean operation cannot be applied directly [50]. The space of $d \times d$ SPD matrices Sym_d^+ is a special type of Riemannian manifold, and LERM [21] provides a solid approach to map a point on the manifold to the Euclidean tangent space via a principal matrix logarithm.

To leverage LERM, we embed the patch Gaussians in the SPD matrix in a manner similar to a previous approach [52]. According to an analysis in the information geometry literature [29], the space of d-dimensional multivariate Gaussians can be embedded into the space of $d+1$-dimensional SPD matrices denoted by Sym^+_{d+1}. We represent the d-dimensional patch Gaussian $\mathcal{N}(\boldsymbol{f};\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s)$ into Sym^+_{d+1} as $\boldsymbol{P}_s$:

$$\mathcal{N}(\boldsymbol{f};\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s) \sim \boldsymbol{P}_s = |\boldsymbol{\Sigma}_s|^{-\frac{1}{d+1}} \begin{bmatrix} \boldsymbol{\Sigma}_s + \boldsymbol{\mu}_s\boldsymbol{\mu}_s^T & \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_s^T & 1 \end{bmatrix}. \tag{3}$$

The covariance matrix of the local patch often becomes singular due to the lack of a sufficient number of pixels within the patch. We overcome this problem by adding a d-dimensional identity matrix $\boldsymbol{I}_d$ to $\boldsymbol{\Sigma}_s$ with a small positive constant value ϵ_s: $\boldsymbol{\Sigma}_s \leftarrow \boldsymbol{\Sigma}_s + \epsilon_s\boldsymbol{I}_d$.

In order to describe the region distribution in a Euclidean operation, we map each of the patch Gaussians $\boldsymbol{P}_s$ into a tangent space via a principal matrix logarithm (Fig. 3 (b-iii)). Note that different Gaussian embeddings [55] could be used as alternatives of the embedding above.

We then store the upper triangular (or equivalent lower triangular) part of the mapped matrix as a vector because the matrix is symmetric. By considering the off-diagonal entries as being counted twice during the norm computation [63], the matrix of the patch Gaussian $\boldsymbol{P}_s$ becomes an $m = \frac{1}{2}(d+1)(d+2)$ dimensional vector $\boldsymbol{g}_s$, defined as,

$$\boldsymbol{g}_s = \mathrm{vec}(\mathrm{log}\boldsymbol{P}_s) = [\mathrm{diag}(\mathrm{log}\boldsymbol{P}_s)^T \quad \sqrt{2}\mathrm{offdiag}(\mathrm{log}\boldsymbol{P}_s)^T]^T, \tag{4}$$

where $\mathrm{diag}(\cdot)$ and $\mathrm{offdiag}(\cdot)$ respectively represent the operator to reshape the diagonal elements and the upper-triangular (half) off-diagonal elements of a symmetric matrix into a vector form.

3.3 Alternative Summarization

Because the dimensionality of the patch Gaussian vector grows quadratically *w.r.t.* the size of the row or column of the SPD matrix, a hierarchical use of this embedding drastically increases the dimensionality. It is desirable to retain the size of the SPD matrix as small as possible, even as small as one dimension. Thus, we develop an alternative embedding method. We assume a Gaussian distribution of which mean vector is fixed to the zero vector *i.e.,* $\boldsymbol{\mu}_s = \boldsymbol{0} = (0,\ldots,0)^T$. The **Z***ero*-**m***ean* **G***aussian* (**ZmG**) distribution $\mathcal{N}(\boldsymbol{f};\boldsymbol{0},\boldsymbol{\Sigma}_s)$ is given by,

$$\mathcal{N}(\boldsymbol{f};\boldsymbol{0},\Xi_s) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{f}^T\Xi_s^{-1}\boldsymbol{f}\right)}{(2\pi)^{d/2}|\Xi_s|}, \tag{5}$$

where the covariance matrix is estimated by $\Xi_s = \frac{1}{n_s-1}\sum_{i\in\mathcal{L}_s}\boldsymbol{f}_i\boldsymbol{f}_i^T$. Note that the covariance matrix Ξ_s coincides with the raw (non-central) moment [51] and is often referred to as the autocorrelation matrix [69]. Summarizing local features with Ξ_s is empirically known as an effective embedding method [22], [51]. Here, we view it from a connection to a Gaussian distribution.

The autocorrelation matrix naturally holds the mean information of the pixel features and almost coincides with the upper-left block of the Gaussian matrix in Eq.(3)[2]. Namely,

$$\Xi_s = \frac{1}{n_s-1}\sum_{i\in\mathcal{L}_s}\boldsymbol{f}_i\boldsymbol{f}_i^T = \boldsymbol{\Sigma}_s + \frac{n_s}{n_s-1}\boldsymbol{\mu}_s\boldsymbol{\mu}_s^T. \tag{6}$$

Although the ZmG distribution also includes the mean information of features in its covariance matrix, it restricts the center of distribution to the origin of the feature space. Thus, ZmG models a broader area of the feature space. Consequently, even though an embedding based on ZmG can be expected to be less discriminative than the Gaussian embedding, it could be considered to be more robust to changes in the feature vectors.

Zero-mean Gaussian Embedding (ZmG)

Using the same Gaussian embedding, we can represent $\mathcal{N}(\boldsymbol{f};\boldsymbol{0},\Xi_s)$ into Sym^+_{d+1} as $\boldsymbol{D'}_s$:

$$\mathcal{N}(\boldsymbol{f};\boldsymbol{0},\Xi_s) \sim \boldsymbol{D'}_s = |\Xi_s|^{-\frac{1}{d+1}} \begin{bmatrix} \Xi_s & \boldsymbol{0} \\ \boldsymbol{0}^T & 1 \end{bmatrix}. \tag{7}$$

The autocorrelation matrix can be regularized as $\Xi_s \leftarrow \Xi_s + \epsilon_s\boldsymbol{I}_d$ to ensure that the matrix is an SPD matrix.

The eigendecomposition of the diagonal block matrix and the definition of the principal matrix logarithm are used to derive the matrix values on the LE tangent space as follows:

$$\mathrm{log}\boldsymbol{D'}_s = \begin{bmatrix} \mathrm{log}\Xi_s - \frac{\mathrm{Tr}(\mathrm{log}\Xi_s)}{d+1}\boldsymbol{I}_d & \boldsymbol{0} \\ \boldsymbol{0}^T & -\frac{\mathrm{Tr}(\mathrm{log}\Xi_s)}{d+1} \end{bmatrix}. \tag{8}$$

A $\frac{1}{2}d(d+1)+1$-dimensional patch Gaussian vector may be obtained by taking only the independent elements.

Because we assumed that the mean vectors of the Gaussian are zero, as is commonly the case, another natural choice to embed a ZmG distribution into an SPD matrix is to use the autocorrelation matrix Ξ_s directly as in the 2AvgP [22]. However, as we verify in §4.1, the scaling of the SPD matrix is important for HGDs. Based on the analogy to Gaussian embedding, which adopts scale normalization[3], we propose to represent the d-dimensional patch Gaussian $\mathcal{N}(\boldsymbol{f};\boldsymbol{0},\Xi_s)$ into Sym^+_d as the following $\boldsymbol{D}_s$:

$$\mathcal{N}(\boldsymbol{f};\boldsymbol{0}_s,\Xi_s) \sim \boldsymbol{D}_s = |\Xi_s|^{-\frac{1}{d}}\Xi_s. \tag{9}$$

Similarly to the case of the Gaussian matrix, we apply the principal matrix logarithm and half-vectorization to the matrix $\boldsymbol{D}_s$ and obtain an $m' = \frac{1}{2}d(d+1)$-dimensional vector $\boldsymbol{g}'_s = \mathrm{vec}(\mathrm{log}\boldsymbol{D}_s)$.

The embeddings $\boldsymbol{D'}_s$ and $\boldsymbol{D}_s$, which are similar except for an additional dimension, were found to perform similarly. Because the size is more compact, we use $\boldsymbol{D}_s$ as the ZmG embedding.

3.4 Region Level Summarization

As a result of the pose variation in person images, the positions of local parts vary in different observations. This leads us to summarize the local patches into an orderless representation. More specifically, we summarize the flattened patch Gaussians in the previous subsections into a region distribution (Fig.3 (b-iv)). For this summarization, we also use a Gaussian distribution that not only has the ability to describe the covariance but also the mean.

The use of a Gaussian distribution for summarization entails considering the spatial property of patches as follows. A person image often contains background regions that significantly differ in places. We therefore suppress the effect of background regions by introducing a weight for each patch in a manner similar to that of weighted color histograms [32]. In most cases, the person is centered in each image; thus, a higher value is assigned to the

[2] In the LE tangent space, this block no longer coincides with $\mathrm{log}\Xi_s$ because $\boldsymbol{\mu}_s$ in the last row or column in Eq.(3) affects the result of the principal matrix logarithm.

[3] There is an equivalence of the determinant $|\boldsymbol{G}_s| = |\boldsymbol{\Sigma}_s|$ where $\boldsymbol{G}_s = \begin{bmatrix} \boldsymbol{\Sigma}_s + \boldsymbol{\mu}_s\boldsymbol{\mu}_s^T & \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_s^T & 1 \end{bmatrix}$ is a non-scaled Gaussian matrix excluding the scaling term of Eq. (3). The proof is given in Appendix A.

patches that are closer to the center of the y-axis of an image: $w_s = \exp(-(x_s - x_c)^2/2\sigma^2)$, where $x_c = W/2$ and $\sigma = W/4$. Here, x_s denotes the x-coordinate of the center pixel of patch s and W is the image width. Then we define the weighted mean vector and covariance matrix as

$$\boldsymbol{\mu}^{\mathcal{G}} = \frac{1}{\sum_{s\in\mathcal{G}} w_s} \sum_{s\in\mathcal{G}} w_s \boldsymbol{g}_s, \tag{10}$$

$$\boldsymbol{\Sigma}^{\mathcal{G}} = \frac{1}{\sum_{s\in\mathcal{G}} w_s} \sum_{s\in\mathcal{G}} w_s (\boldsymbol{g}_s - \boldsymbol{\mu}^{\mathcal{G}})(\boldsymbol{g}_s - \boldsymbol{\mu}^{\mathcal{G}})^T, \tag{11}$$

where $\mathcal{G}$ is the region in which the patch Gaussians are summarized. Similarly, the weighted autocorrelation matrix is defined by

$$\boldsymbol{\Xi}^{\mathcal{G}} = \frac{1}{\sum_{s\in\mathcal{G}} w_s} \sum_{s\in\mathcal{G}} w_s \boldsymbol{g}'_s \boldsymbol{g}'^T_s. \tag{12}$$

Here, we regularize the covariance and autocorrelation matrices $\boldsymbol{\Sigma}^{\mathcal{G}}$ and $\boldsymbol{\Xi}^{\mathcal{G}}$ with the parameter $\epsilon^{\mathcal{G}}$, *e.g.,* $\boldsymbol{\Sigma}^{\mathcal{G}} \leftarrow \boldsymbol{\Sigma}^{\mathcal{G}} + \epsilon^{\mathcal{G}} \boldsymbol{I}_m$. Using the mean vector and covariance matrix, we represent the region as the region Gaussian $\mathcal{N}(\boldsymbol{g}; \boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$ or ZmG $\mathcal{N}(\boldsymbol{g}'; \boldsymbol{0}, \boldsymbol{\Xi}^{\mathcal{G}})$.

In terms of matching among region descriptors, the region Gaussian is conveniently mapped into Euclidean space on which most of the matching methods such as metric learning are designed. For this purpose, we embed an $\{m, m'\}$-dimensional region Gaussian into SPD matrices in the same manner as in Eq.(3) or Eq.(9): $\mathcal{N}(\boldsymbol{g}; \boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}) \sim \boldsymbol{Q} \in Sym^+_{m+1}$ or $\mathcal{N}(\boldsymbol{g}'; \boldsymbol{0}, \boldsymbol{\Xi}^{\mathcal{G}}) \sim \boldsymbol{R} \in Sym^+_{m'}$, respectively. We then map $\{\boldsymbol{Q}, \boldsymbol{R}\}$ into LE tangent space and half-vectorize it to form an r-dimensional feature vector $\boldsymbol{z}$, where $r = \{\frac{1}{2}(m+1)(m+2), \frac{1}{2}m'(m'+1)\}$, respectively, for $\{\boldsymbol{Q}, \boldsymbol{R}\}$ (Fig.3 (b-v)).

By extracting the region Gaussian for each of the G regions, we obtain feature vectors $\{\boldsymbol{z}_g\}_{g=1}^G$. In order to maintain the spatial location of these vectors, we concatenate them. A person image is represented by a feature vector $\boldsymbol{z} = [\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_G^T]^T$ (Fig.3(b-vi)).

3.5 Parameter and Dimension

We specify the parameters we empirically tuned for person re-id. First, we resize each input image to 128×48 pixels to facilitate evaluation with the common parameters. We then set the height of each horizontal strip as one-quarter of the image height, *i.e.,* the size of a strip is 32×48 pixels. We slide the strip in a vertical direction such that half of each strip overlaps with another strip from the top of the image until any pixel in the strip exceeds the scope of the input image. In this way, we obtain seven overlapping horizontal strips (regions with $G = 7$). We extract local patches at two-pixel intervals ($p = 2$) in each region by considering the trade-off between the computational time and predictive accuracy. In addition, we set the local patch size to 7×7 pixels ($k = 7$). Finally, we set the regularization parameters of patch/region Gaussians as $(\epsilon, \epsilon^{\mathcal{G}}) = (10^{-4}, 10^{-2})$.

It has been proven that descriptors extracted from different color spaces are complementary to each other [4]. We extract GOG or ZOZ by replacing the RGB color space in the pixel feature in Eq.(1) with three alternative color spaces $\{$Lab, HSV, nRGB$\}$ and concatenate them. Here, nRGB is the normalized color space (*e.g.,* nR = R/(R+G+B)). Because this space includes redundancy, we only use $\{$nR, nG$\}$. Thus, the dimensions of the nRnG color space are $(d, m, r) = (7, 36, 703)$ for GOG and $(d, m', r') = (7, 28, 403)$ for ZOZ, whereas the dimensions of each $\{$RGB, Lab, HSV$\}$ color space are $(d, m, r) = (8, 45, 1081)$ for GOG and $(d, m', r') = (8, 36, 666)$ for ZOZ. The dimensions of the fusion descriptor of each $\{$GOG, ZOZ$\}$ is the sum of G (regions) × $\{r, r'\}$ (dim.) in the four color spaces. Because we use seven regions ($G = 7$), the dimensions of GOG and ZOZ become 27,626 and 16,828, respectively.

4 Normalization of HGDs

HGDs are the hierarchal representations in the LE tangent space, which corresponds to the tangent space of the identity matrix. In this section, we explain two types of normalizations to enhance the representation of HGDs based on the properties of this space: $(\mathbb{I})$ We explain that the scale normalization of the SPD matrix adopted in the Gaussian and ZmG embeddings alleviates the biased diagonal elements within each of a patch/region Gaussian matrix. $(\mathbb{II})$ We propose the bias removal before norm normalization to alleviate the largely biased elements among the region Gaussians of different images in the LE tangent space.

4.1 Matrix Scale Normalization

In the empirical observation, we found that the diagonal elements of a symmetric matrix tend to be biased in the LE tangent space (§5.3). Such a property introduces a bias in the flattened vectors on both the patch/region Gaussians. Notably, the bias in the patch Gaussians is undesirable because HGDs correspond to their summarized representation. In this section, we explain that the scale normalization of the SPD matrix adopted in the Gaussian and ZmG embeddings alleviates the biased diagonal elements within the patch/region Gaussian matrices.

Let $\boldsymbol{X} \in Sym^+_e$ be a general SPD matrix including the non-scaled patch/region Gaussian matrices $\{\boldsymbol{G}, \boldsymbol{\Xi}, \boldsymbol{G}^{\mathcal{G}}, \boldsymbol{\Xi}^{\mathcal{G}}\}$ and e be the column (row) size of $\boldsymbol{X}$. Let $\boldsymbol{X} = \boldsymbol{U}\mathrm{Diag}(\lambda_i)\boldsymbol{U}^T$ be its eigendecomposition where $\mathrm{Diag}(\lambda_i)$ is a diagonal matrix formed from the eigenvalues $\lambda_1, \ldots, \lambda_e$ and $\boldsymbol{U} \in \mathbb{R}^{e\times e}$ is the eigenvectors. The principal matrix logarithm of $\boldsymbol{X}$ is defined as:

$$\mathrm{log}\boldsymbol{X} = \boldsymbol{U}\mathrm{Diag}\left(\ln\lambda_i\right)\boldsymbol{U}^T. \tag{13}$$

In LERM, the principal matrix logarithm in Eq.(13) corresponds to the tangent space mapping. The scale normalization of HGDs is related to the following property of LERM.

Property (Logarithmic linearity). LERM inherits the logarithmic linearity of the scalar space as follows.

$$\log(a\boldsymbol{X}) = \mathrm{log}\boldsymbol{X} + \ln(a)\boldsymbol{I}_e. \tag{14}$$

Here $a \in \mathbb{R}^1$ is an arbitrary scalar value that satisfies $a > 0$.

Proof. We can confirm that $\log(a\boldsymbol{X}) = \boldsymbol{U}\mathrm{Diag}(\ln(a\lambda_i))\boldsymbol{U}^T = \boldsymbol{U}\mathrm{Diag}(\ln(\lambda_i) + \ln(a))\boldsymbol{U}^T = \log(\boldsymbol{X}) + \ln(a)\boldsymbol{U}\boldsymbol{U}^T = \log(\boldsymbol{X}) + \ln(a)\boldsymbol{I}_e$. □

Based on this property, the scale normalization defined below (which is adopted in the Gaussian embedding in Eq.(3) and ZmG embedding in Eq.(9)) adjusts the diagonal elements in the LE tangent space.

Matrix Scale Normalization (MSN)
Given an SPD matrix $\boldsymbol{X} \in Sym^+_e$, the matrix scale normalization $\eta : Sym^+_e \to Sym^+_e$ is defined as follows:

$$\eta(\boldsymbol{X}) = |\boldsymbol{X}|^{-\frac{1}{e}}\boldsymbol{X}. \tag{15}$$

Because $|\boldsymbol{X}|^{-\frac{1}{e}} = \left(\prod_{j=1}^e \lambda_j\right)^{-\frac{1}{e}}$, we see

$$\begin{aligned}\log\eta(\boldsymbol{X}) &= \boldsymbol{U}\mathrm{Diag}\left(\ln\left(\left(\textstyle\prod_{j=1}^{e}\lambda_j\right)^{-\frac{1}{e}}\lambda_i\right)\right)\boldsymbol{U}^T\\ &= \boldsymbol{U}\mathrm{Diag}\left(\ln\lambda_i\right)\boldsymbol{U}^T + \ln\left(\left(\textstyle\prod_{j=1}^{e}\lambda_j\right)^{-\frac{1}{e}}\right)\boldsymbol{U}\boldsymbol{U}^T\\ &= \log\boldsymbol{X} - \left(\frac{1}{e}\textstyle\sum_{j=1}^{e}\ln\lambda_j\right)\boldsymbol{I}_e. \end{aligned} \quad (16)$$

In this way, $\eta(\cdot)$ adjusts the diagonal elements of $\log\boldsymbol{X}$.

Note that the MSN changes the eigenspectrum of $\boldsymbol{X}$ as $\lambda_i \leftarrow \lambda_i/(\prod_{j=1}^{e}\lambda_j)^{\frac{1}{e}}, i = 1, \ldots, e$. When the dimensionality e is high, many eigenvalues could be small *e.g.,* less than 1. In such a case, it can cause division by zero because the denominator becomes too small due to repeated multiplication by small eigenvalues. We avoid this problem by directly calculating the matrix values in the LE tangent space by the last equation of Eq.(16).

4.2 Norm and Power Normalizations

Norm normalization is important to equalize the ranges of the feature vectors and control the distance measure between them [70]. The research work on FV representations [61] pointed out that the norm normalization can help to improve the recognition accuracies for any of the high-dimensional features.

Because the HGDs are high dimensional, we normalize the descriptor by using the L2 norm normalization, which is the most widely adopted form of normalization.

The feature vectors in the LE tangent space can be largely biased because the origin of this space is the identity matrix. We assume that the SPD matrix has a common structure, *e.g.,* the last row or column in the Gaussian matrix is the mean vector in the Gaussian embedding. In such a case, the embedded SPD matrices will be similar apart from the identity matrix. As a result, the cosine distance, *i.e.,* the Euclidean distance on the L2 normalized features, would be dominated by the bias and therefore decreases the discriminative ability.

As a remedy to these biased values, we investigate two types of statistics on the SPD matrix manifold to remove the bias before performing L2 norm normalization. For the fusion descriptor, we normalize each of the HGDs extracted on the four color spaces before concatenating them.

L2 Norm Normalization with Extrinsic Statistics (E-L2/E*-L2)

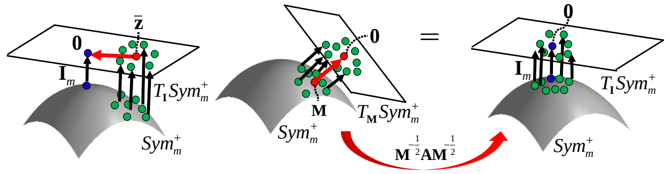
In the first normalization, we simply remove the biased component directly in the LE tangent space, of which calculation has a small computational cost (Fig. 4 (a)). The E-L2 normalization becomes the following:

$$\hat{\boldsymbol{z}} = (\boldsymbol{z} - \overline{\boldsymbol{z}})/\|\boldsymbol{z} - \overline{\boldsymbol{z}}\|_2, \quad (17)$$

where $\overline{\boldsymbol{z}}$ is the sample mean of the GOG or ZOZ acquired from the training samples.

A similar normalization was proposed for the Bag-of-Words representation to reflect co-missing words for cosine similarity [71]. In contrast, we employ the normalization to remedy the effect of the bias in the LE tangent space.

As a variant of the E-L2 normalization, we conduct further tests to adjust dimensions with broad and narrow ranges. Let σ_i be the standard deviation of the i-th dimension of $\boldsymbol{z}$ in training samples. We remove the sample mean and scale all dimensions to have the unit standard deviation as $z_i^* = \frac{z_i - \bar{z}_i}{\sigma_i}$. Subsequently, we normalize the L2 norm of normalization of $\boldsymbol{z}^*$. We refer to this normalization as E*-L2.



(a) Extrinsic bias removal (b) Intrinsic bias removal

Fig. 4. Two counter-bias methods applied before feature norm normalization. (a) Extrinsic and (b) Intrinsic bias removals. Note that the sample points here are the region Gaussian matrices of different training images.

L2 Norm Normalization with Intrinsic Statistics (I-L2)

In the second normalization, we consider the intrinsic statistics of the Riemannian manifold [20]. Here, we use the region Gaussian/ZmG matrices before vectorization. If all matrices are similar apart from the identity matrix, the vector on the tangent space will be biased. A natural choice to remove this bias is to use the tangent space of the mean point of training matrices.

Let $\boldsymbol{A}$ be one of the region Gaussian/ZmG matrices $\{\boldsymbol{Q}, \boldsymbol{R}\}$ and m be the column (row) size of $\boldsymbol{A}$. Given N_T training SPD matrices $\{\boldsymbol{A}_i \in Sym_m^+\}_{i=1}^{N_T}$, the Riemannian center of mass $\boldsymbol{M}$ is the point on Sym_m^+ that minimizes the sum of the squares of the Riemannian distance:

$$\hat{\boldsymbol{M}} = \underset{\boldsymbol{M}\in Sym_m^+}{\arg\min} \sum_{i=1}^{N_T} D_{\mathrm{geo}}^2\left(\boldsymbol{A}_i, \boldsymbol{M}\right), \quad (18)$$

where $D_{\mathrm{geo}}(\boldsymbol{A}, \boldsymbol{M})$ is the geodesic distance between $\boldsymbol{M}$ and $\boldsymbol{A}$. We use the AIRM distance. The optimization procedure of Eq.(18) is found in Ref. [20].

We map the SPD matrix $\boldsymbol{A}$ into the tangent space of the mean matrix $\boldsymbol{M}$. By taking the orthogonal coordinates on the tangent space, the half vectorized representation is given by [20]:

$$\boldsymbol{z}' = \mathrm{vec}\left(\log\left(\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{M}^{-\frac{1}{2}}\right)\right). \quad (19)$$

Note that the matrix $\boldsymbol{M}^{-\frac{1}{2}}$ is full rank because $\boldsymbol{M}$ is an SPD matrix and thus the transformed matrix $\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{M}^{-\frac{1}{2}}$ is also an SPD matrix. Therefore, we can interpret the space of $\boldsymbol{z}'$ as being the LE tangent space of transformed matrices (Fig.4(b)). Consequently, under the LERM, we can interpret the Euclidean distance on the tangent space of the pole as being the geodesic of the transformed matrices.

We independently estimate the Riemannian mean for each region of the horizontal strips. We map the region Gaussian matrices to the tangent space of each region and apply half-vectorization. Because the mean vector on the training data is zero in the tangent space, we directly normalize the L2 norm of $\boldsymbol{z}'$.

Power Normalization (PN)

Applying an element-wise Power Normalization (PN) before conducting L2 norm normalization often improves performance [61]. We thus further test the PN for the normalization of HGDs. PN enlarges the small magunitudes and reduces large magnitudes in the elements by taking the signed power of each element z in $\boldsymbol{z}$ as: $z \leftarrow \mathrm{sign}(z)|z|^{\rho}$ with $0 < \rho \leq 1$. Following the previous work [61], we set the power value as $\rho = 0.5$.

To combine with the bias removal methods, we apply PN for each of the bias removed vectors, *i.e.,* $\{\boldsymbol{z} - \bar{\boldsymbol{z}}, \boldsymbol{z}^*, \boldsymbol{z}'\}$ respectively for { E-L2, E*-L2, I-L2 }. Subsequently, we normalize their L2 norm.

5 Experiments

5.1 Datasets and evaluation protocol

We use five benchmark datasets to evaluate our method: VIPeR [72], GRID [73], CUHK01 [9], CUHK03 [38], and Market-1501 [74]. Example images of each of these datasets are shown in Fig. 5. All of these datasets are challenging because the images contain large variations regarding their viewpoints, pose, illumination, occlusion, and background clutter. We evaluate the performance by using the Cumulative Matching Characteristic (CMC) curves, which visualizes an expectation of finding the correct person in the top r matches [72]. As a measure to evaluate the entire CMC curves, we report the Proportion of Uncertainty Removed (PUR), which represents the uncertain reduction by a given algorithm from the random ranking [10]. For the Market-1501 dataset, we report the mean Average Precision (mAP), which considers both the precision and recall of the retrieval process [74] because the gallery contains multiple images of one person.

The ***VIPeR*** dataset contains 1,264 images of 632 persons captured in disjoint camera views. The ***GRID*** dataset contains 1,275 images with 250 annotated persons and an additional 775 gallery images of persons except those included as annotated persons. Both the VIPeR and GRID datasets contain one image of each person with one camera view. Thus, we evaluate the performance with *single-shot* matching. We report an average of 10 random training/test splits, in which each split image only contains one-half of the people. The ***CUHK01*** dataset contains 3,884 images of 971 persons. There are two images of each person in each camera view. Thus, we carry out the evaluation with *multi-shot* matching, in which we calculate the distances between two persons by averaging the corresponding cross-view image pairs. We report the average of 10 random 485/486 person splits for the training/test sets. The ***CUHK03*** dataset contains 13,164 images of 1,360 persons with an average of 4.8 images of each person in each view. We use the images that are *automatically detected* by the person detector and evaluate the performance with *multi-shot* matching. We report the average result of 20 random 1,260/100 person splits for the training/test sets. The ***Market-1501*** dataset contains 32,668 bounding boxes of 1,501 persons. Each person is captured by six cameras at most and two cameras at least. During testing, for each person, one query image in each camera is selected. We use a fixed 750/751 person split for the training/test set and report the results of the *single-query* evaluation on 3,386 query images.

We evaluate the proposed descriptors by learning ***three distance metrics***: the Keep It Simple and Straightforward MEtric (KISSME) [5], Cross-view Quadratic Discriminant Analysis (XQDA) [11], and Null Foley-Sammon Transform (NFST) [12]. KISSME learns a Mahalanobis-like distance by a likelihood ratio test of similar or dissimilar pairs. For KISSME, we first project feature vectors in the PCA subspace where 98% of the energy is maintained. XQDA learns a discriminative subspace and a distance metric simultaneously and can select the optimal dimensionality automatically. NFST seeks intersection space of the null space of within-class distances, and non-zero between-class distances, it is also free to select the subspace dimensionality. For NFST, we use the RBF kernel of which bandwidth is equivalent to the mean pairwise distances of the training samples.

5.2 Evaluation of the hierarchical Gaussian embedding

We evaluate the following aspects of HGDs: (1) Embedding methods; (2) Regularization parameters; (3) Patch/region sizes.

(a) VIPeR (b) GRID (c) CUHK01 (d) CUHK03 (e) Market-1501

Fig. 5. Example images from the person re-id datasets. For each dataset, images in the same column represent the same person.

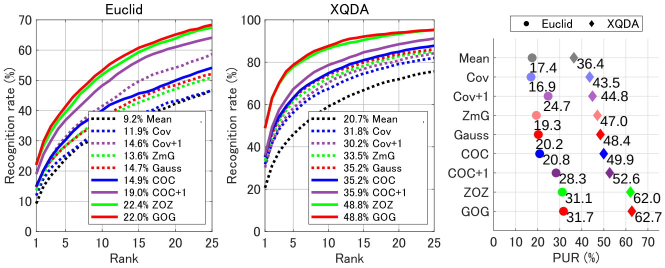


Fig. 6. Feature embedding analysis on the VIPeR dataset. All methods use the E-L2 normalization without MSN and PN. The figures on the CMC curves indicate the rank-1 rates.

Impact of embedding methods. We compare different embedding methods for feature summarization including both global and hierarchical methods. For the global distribution embeddings of pixel features inside each region, we carry out a comparison with the mean vector (Mean), covariance matrices (Cov and Cov$_{+1}$[4]), Zero-mean Gaussian (ZmG), and Gaussian (Gauss). We ensure a fair comparison by commonly adopting the weighted embedding for all descriptors, defined as follows:

- Mean: $\boldsymbol{\mu}' = \frac{1}{\sum_{i \in \mathcal{G}} w_i} \sum_{i \in \mathcal{G}} w_i \boldsymbol{f}_i$,
- Cov: $\boldsymbol{\Sigma}' = \frac{1}{\sum_{i \in \mathcal{G}} w_i} \sum_{i \in \mathcal{G}} w_i (\boldsymbol{f}_i - \boldsymbol{\mu}')(\boldsymbol{f}_i - \boldsymbol{\mu}')^T$,
- ZmG: $\boldsymbol{\Xi}' = \frac{1}{\sum_{i \in \mathcal{G}} w_i} \sum_{i \in \mathcal{G}} w_i \boldsymbol{f}_i \boldsymbol{f}_i^T$,
- Cov$_{+1}$: $\boldsymbol{\Sigma}'_{+1} = \begin{bmatrix} \boldsymbol{\Sigma}' & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}$,
- Gauss: $\boldsymbol{G}' = \begin{bmatrix} \boldsymbol{\Sigma}' + \boldsymbol{\mu}' \boldsymbol{\mu}'^T & \boldsymbol{\mu}' \\ \boldsymbol{\mu}'^T & 1 \end{bmatrix}$.

Here the pixel weight w_i is determined in the same manner as w_s.

For two-level embeddings, we compare the performance with two descriptors referred to as Covariance-Of-Covariances (COC and COC$_{+1}$) in which the Cov and Cov$_{+1}$ embeddings, respectively, are used in both patch and region-level embeddings. As in the HGDs, we use the patch weight for region-level summarization.

We apply LERM for all descriptors except Mean. In the same manner, as for the HGDs, we regularize the covariance/autocorrelation matrices and concatenate the feature vectors of seven regions. Fig. 6 shows the results on the VIPeR dataset. In this comparison, we apply the E-L2 normalization without MSN and PN. In addition to the XQDA distance metric, the figures show the performance with the Euclidean distance to determine the fundamental property of each embedding. For the sake of generality, we evaluate the performances using different normalization methods proposed in §4. Table 1 lists the results on the five datasets evaluated with XQDA and NFST metrics[5].

[4] Cov$_{+1}$ and COC$_{+1}$, respectively, correspond to the mean removed version of Gauss and GOG, *e.g.*, Cov$_{+1}$ is the case of $\boldsymbol{\mu}' = \mathbf{0}$ in Gauss.

[5] We omitted the results of KISSME because it showed similar trends to XQDA. Although NFST outperformed XQDA on LOMO features [12], our evaluation on HGDs showed different trends depending on the datasets.

TABLE 1
The impact of the hierarchical Gaussian embedding evaluated with MSN

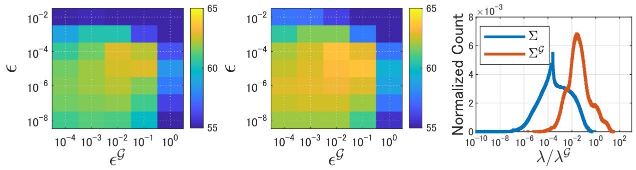
				VIPeR (PUR)				**GRID** (PUR)				**CUHK01** (PUR)				**CUHK03** (PUR)				**Market-1501** (mAP)			
				w/o PN		**w/ PN**		**w/o PN**		**w/ PN**		**w/o PN**		**w/ PN**		**w/o PN**		**w/ PN**		**w/o PN**		**w/ PN**	
Norm	**Methods**	**Hie.**	**Mean**	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST
E-L2	Mean		✓	36.4	28.7	31.9	26.4	24.6	23.5	21.2	24.0	37.1	9.2	36.1	8.0	33.2	1.9	33.3	1.4	9.7	4.9	8.6	3.4
	Cov			45.7	50.4	43.2	52.9	31.7	38.4	33.9	39.4	58.5	50.8	61.3	60.7	54.8	24.4	60.2	42.5	25.9	22.3	27.4	26.2
	COC	✓		54.7	54.3	53.5	52.5	37.1	33.5	36.3	33.6	65.6	70.2	70.5	72.9	80.3	79.2	82.6	81.3	35.3	37.4	36.0	36.1
	ZmG		✓	49.5	52.3	45.1	52.9	34.1	41.3	35.6	41.7	61.8	53.9	64.2	62.9	58.0	23.8	63.1	39.7	26.4	22.5	27.9	26.1
	ZOZ	✓	✓	**62.8**	**62.7**	**61.0**	**60.2**	**47.0**	**44.3**	**44.9**	**42.1**	**76.1**	**78.9**	**78.7**	**80.3**	**81.3**	**80.1**	**83.2**	**82.9**	**40.8**	**44.0**	**41.7**	**43.4**
	Cov_{+1}			44.8	51.1	43.7	54.4	30.9	36.5	32.8	37.8	58.3	49.8	61.8	61.8	56.5	23.7	62.9	43.9	26.0	21.7	27.9	26.9
	COC_{+1}	✓		50.1	51.1	50.0	49.7	38.4	35.4	36.4	34.1	68.6	72.9	71.6	73.0	76.8	78.1	80.3	81.3	32.6	35.2	32.8	33.5
	Gauss		✓	50.6	53.4	49.6	55.5	34.0	40.8	36.0	41.1	61.9	53.8	65.4	63.8	58.1	25.7	63.1	41.5	27.6	24.7	29.7	28.8
	GOG	✓	✓	**63.6**	**62.6**	**62.1**	**60.4**	**46.0**	**43.0**	**44.7**	**41.5**	**76.5**	**78.3**	**79.5**	**79.8**	**81.9**	**81.3**	**84.4**	**83.5**	**41.4**	**44.3**	**43.0**	**44.5**
E^*-L2	Mean		✓	33.3	22.0	30.6	20.8	26.6	24.9	22.2	23.5	37.8	8.8	36.7	7.2	32.4	1.3	33.3	1.0	9.2	2.8	8.1	2.0
	Cov			41.5	50.7	40.1	50.6	29.5	35.7	31.4	36.7	63.5	64.9	62.1	64.8	60.3	44.3	62.4	50.6	28.4	26.5	27.8	26.4
	COC	✓		48.8	48.7	48.7	48.5	33.3	31.3	33.6	32.0	70.4	72.4	70.7	72.8	81.7	78.4	82.8	79.8	34.7	33.4	34.2	32.4
	ZmG		✓	42.7	50.6	41.2	50.0	30.8	38.3	32.8	38.6	65.2	66.2	64.2	66.3	62.0	40.1	64.2	46.3	28.4	25.6	28.4	26.1
	ZOZ	✓	✓	**56.5**	**56.1**	**56.6**	**55.8**	**41.0**	**38.4**	**41.0**	**39.2**	**78.5**	**80.0**	**78.7**	**79.9**	**82.5**	**82.1**	**83.4**	**83.0**	**40.2**	**40.8**	**40.2**	**40.2**
	Cov_{+1}			41.6	52.7	40.3	52.2	30.1	36.6	30.5	37.1	64.3	67.2	63.2	66.9	62.6	49.4	64.7	54.6	29.6	28.2	28.6	27.5
	COC_{+1}	✓		44.0	44.6	44.3	44.9	32.9	31.6	33.4	31.6	70.4	71.6	70.2	71.3	80.4	77.1	81.0	79.0	31.1	30.1	30.5	29.1
	Gauss		✓	45.1	53.1	44.9	53.3	31.5	38.1	33.1	38.3	66.1	67.0	65.8	67.5	62.2	42.2	64.5	48.5	30.2	28.2	30.3	29.0
	GOG	✓	✓	**58.2**	**56.7**	**57.9**	**56.2**	**41.6**	**38.2**	**41.4**	**38.7**	**79.3**	**79.3**	**79.4**	**79.4**	**83.9**	**82.4**	**84.6**	**83.3**	**41.8**	**41.4**	**41.7**	**41.5**
I-L2	Mean		✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Cov			46.5	51.6	43.6	53.5	32.6	39.0	34.5	39.7	59.4	52.5	59.4	52.5	56.0	26.9	60.6	43.4	26.7	23.5	28.0	26.8
	COC	✓		53.1	53.2	53.0	52.0	37.2	33.9	36.9	33.7	67.7	71.7	67.7	71.7	81.2	80.2	83.1	81.8	36.0	38.2	36.2	36.5
	ZmG		✓	51.0	53.4	46.7	53.1	34.2	41.4	35.7	41.1	61.6	54.6	61.6	54.6	59.5	26.8	62.8	39.2	27.3	24.1	27.9	25.8
	ZOZ	✓	✓	**63.9**	**63.5**	**62.0**	**61.1**	**48.2**	**45.9**	**46.3**	**43.7**	**78.0**	**80.4**	**79.9**	**81.8**	**81.3**	**81.6**	**83.6**	**84.2**	**41.8**	**45.6**	**42.7**	**45.4**
	Cov_{+1}			45.4	52.8	43.8	55.1	31.5	37.5	33.8	38.2	59.4	53.0	59.4	53.0	58.1	28.2	63.7	47.0	27.0	23.4	28.4	27.7
	COC_{+1}	✓		52.8	53.1	50.4	50.6	38.5	35.5	37.1	34.6	69.7	73.6	69.7	73.6	77.5	78.8	80.6	81.7	33.0	35.8	33.1	34.0
	Gauss		✓	52.2	54.8	49.6	55.2	34.8	41.4	35.1	40.7	62.4	55.1	62.4	55.1	60.5	28.9	63.7	41.0	28.9	26.2	29.9	28.5
	GOG	✓	✓	**65.0**	**64.2**	**63.6**	**62.1**	**48.4**	**45.0**	**46.9**	**43.8**	**79.1**	**80.2**	**81.1**	**81.4**	**82.8**	**82.6**	**84.5**	**84.9**	**43.7**	**47.8**	**45.1**	**48.2**

The mark ✓ indicates methods that contain hierarchy/mean information. The **red**/**blue** *scores show the* **first**/**second** *best scores in each normalization.*

The results indicate that: (1) The performance trend is similar among the Euclidean distance and the two metrics: hierarchical and Gaussian embeddings (GOG and ZOZ) perform the best. These results indicate that these embedding methods are comparatively discriminative and robust. (2) The hierarchical embeddings perform more effectively than global embeddings in the same base embedding, *e.g.,* COC outperforms Cov. These results confirm one of our motivations, *i.e.,* hierarchical distributions are more discriminative than global distributions. (3) Gaussian embeddings perform more effectively than covariance embeddings. Among the global embeddings, ZmG and Gauss outperform Cov and Mean. In addition, among the hierarchical embeddings, ZOZ and GOG outperform COC and COC_{+1}. These results confirm another motivation, *i.e.,* the importance of using both the mean and covariance information. (4) The performance of ZmG and ZOZ is close to that of Gauss and GOG, respectively, especially on the E-L2 and E^*-L2 normalizations. These results confirm that ZmG approximates the Gaussians distribution well. In several cases on the GRID and CUHK01 datasets, ZOZ outperforms GOG. These results may be because severe changes of mean information are included in these datasets. Compared with ZmG, the Gaussian embedding has additional independent dimensions of the mean vector; thus, Gauss can be more discriminative combined with metric learning, whereas the influence of the mean information is more substantial than ZmG. (5) The I-L2 normalization tends to improve the performance of GOG more than ZOZ. This could probably be attributed to the fact that the bilinear transformation in the I-L2 normalization adjusts the different feature magnitudes among the mean and other components in the Gaussian embedding.

Effect of regularization parameters. We examine the sensitivity of performance *w.r.t.* the regularization parameters (ϵ, $\epsilon^{\mathcal{G}}$) of the covariance matrices. In this analysis, we use the GOG descriptor normalized by the E-L2 normalization without PN, and evaluate the performance with the XQDA distance metric. Fig. 7 (a) shows the PUR of GOG on the VIPeR dataset in cases without and with applying MSN in the parameter ranges $\epsilon \in \{10^{-8}, 10^{-7}, \dots, 10^{-2}\}$ and $\epsilon^{\mathcal{G}} \in \{10^{-4}, 10^{-3}, \dots, 10^{0}\}$. Fig. 7 (b) shows the eigenvalue distribution of covariance matrices.

Overly large ϵ covers the characteristics of eigenvalues of each patch/regions. Conversely, with overly small ϵ, the descriptors become sensitive to the difference of small eigenvalues, because the logarithmic function magnifies small eigenvalues less than 1. We see that the best parameters are located around $(\epsilon, \epsilon^{\mathcal{G}}) = (10^{-5} - 10^{-4}, 10^{-1} - 10^{-2})$ and these are near the center of eigenvalue distribution. Based on these analyses, we set the parameters $(\epsilon, \epsilon^{\mathcal{G}}) = (10^{-4}, 10^{-2})$.



(a) PUR (%) [left: w/o MSN, right: w/ MSN] (b) Eigenvalues

Fig. 7. Analysis of regularization parameters: (a) PUR scores of the GOG descriptor in cases without/with applying MSN. (b) Eigenvalue distribution of patch/region covariance matrices (w/o MSN). To show a typical example, we used patch Gaussian matrices of $\epsilon = 10^{-4}$ for $\mathbf{\Sigma}^{\mathcal{G}}$.

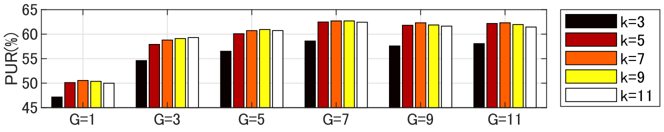
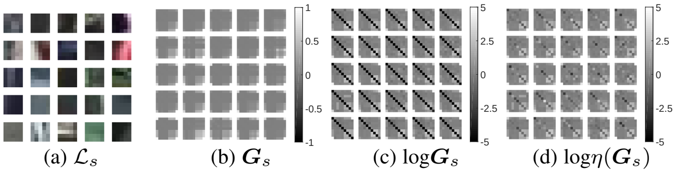


Fig. 8. Analysis of patch/region sizes on the VIPeR dataset.

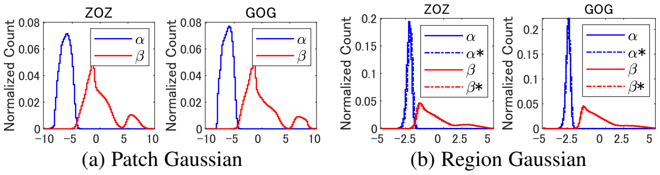
Effect of patch/region sizes. We examine the sensitivity of performance *w.r.t.* the number of regions G and patch size k. We use the GOG descriptor normalized by the E-L2 normalization without MSN and PN. We test the regions with the number $G \in \{1, 3, 5, 7, 9, 11\}$ in which the heights of horizontal strips are set as $\{\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}\}$, respectively of the image height, and half of each strip overlaps with another strip. The patch size varies in the range $k \in \{1, 3, 5, 7, 9, 11\}$. Fig. 8 shows the results on the VIPeR dataset evaluated with the XQDA metric.

The results indicate that: (1) The performance increases as the number of regions G increases. These results are probably due to the fact that the person is relatively well aligned in a horizontal direction; thus, more detailed information along the height can increase the discriminative ability of different persons. At the same time, a large number of G increases the dimensionality of HGDs and the performances saturate at approximately $G = 7$. (2) The performances with patch sizes $k > 5$ are higher than $k = 3$. At the same time, the performance slightly decreases in $k = 9, 11$ for several settings of G. By considering the above results, we set the parameters $(G, k) = (7, 7)$.



(a) $\mathcal{L}_s$ (b) $\boldsymbol{G}_s$ (c) $\log\boldsymbol{G}_s$ (d) $\log\eta(\boldsymbol{G}_s)$

Fig. 9. Visualization of patch Gaussian matrices extracted from randomly sampled patches on the VIPeR dataset: (a) Image patches (7×7 pixels). (b) Patch Gaussian matrices (9×9 dims). (c)/(d) After obtaining the principal matrix logarithm of (b) without/with applying MSN.



(a) Patch Gaussian (b) Region Gaussian

Fig. 10. Histogram of decomposed logarithmic eigenvalues (α and β) on the VIPeR dataset: (a)/(b) Non-scaled patch/region Gaussian matrices. In (b), $(\alpha^*, \beta^*)/(\alpha, \beta)$ represent the values in the case without/with applying MSN to the patch Gaussians. The relation $|\alpha| \gg |\beta|$ implies that a large diagonal bias occurs in the matrix.

5.3 Evaulation of Normalizations

We evaluate the proposed normalizations in terms of the following aspects: (1) MSN; (2) PN and norm normalizations.

Analysis of MSN. We show how MSN improves the representations of HGDs. Empirical observation showed that LE tangent space mapping causes bias in the diagonal elements of the symmetric matrix. Fig. 9 illustrates this effect on the patch Gaussian matrices. In these figures, we randomly sampled image patches on the VIPeR dataset (Fig. 9 (a)) and extracted their non-scaled patch Gaussian matrices $\boldsymbol{G}_s$ with the pixel features in Eq.(1) (Fig. 9 (b)). Fig. 9 (c) and (d), respectively, show the patch Gaussian matrices in the LE tangent space with and without applying MSN (Note that $\boldsymbol{P}_s = \eta(\boldsymbol{G}_s)$). We see that the diagonal elements of the matrices commonly have large negative values in the LE tangent space in Fig. 9 (c) and MSN alleviates them as shown in Fig. 9 (d).

We explain the cause of the biased diagonal elements. Let $\boldsymbol{X} \in Sym_e^+$ be a general SPD matrix including the patch/region Gaussian matrices and $(\ln\lambda_1, \ldots, \ln\lambda_e)$ be the logarithm of its eigenvalues. Let us decompose $\log\boldsymbol{X}$ using the mean eignenvalue $\alpha = \frac{1}{e}(\sum_{j=1}^{e} \ln\lambda_j)$ and the residual $(\beta_1, \ldots, \beta_e)$ as follows:

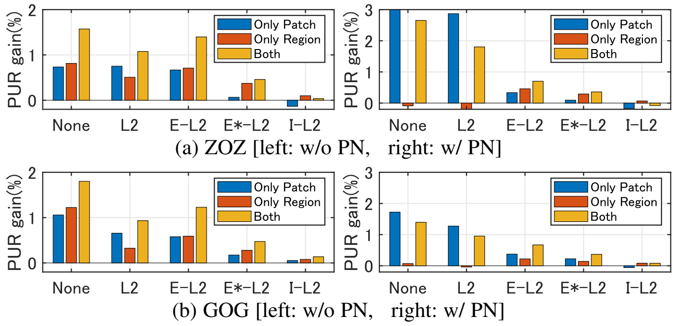
$$\begin{aligned}\log\boldsymbol{X} &= \boldsymbol{U}\mathrm{Diag}(\beta_i + \alpha)\boldsymbol{U}^T \\ &= \boldsymbol{U}\mathrm{Diag}(\beta_i)\boldsymbol{U}^T + \alpha\boldsymbol{U}\boldsymbol{U}^T = \log\hat{\boldsymbol{X}} + \alpha\boldsymbol{I}_e. \quad (20)\end{aligned}$$

This decomposition implies that if the magnitude of the mean value is larger than the residual, *i.e.,* $|\alpha| \gg |\beta_j|, j = 1, \ldots, e$, then the diagonal elements of $\log\boldsymbol{X}$ are expected to have larger magnitudes than the off-diagonal elements. This situation often occurs in HGDs because local patches/regions consist of fewer pixels/patches. Thus, only a few eigenvalues of an SPD matrix are expected to be large, whereas most of the remaining values are very small. Fig. 10 (a) and (b), respectively, show the histograms of the decomposed logarithmic eigenvalues of patch and region Gaussians on all images of the VIPeR dataset[6]. We see that α tends to be much smaller than β in both patch/region Gaussians. These trends are common in both GOG and ZOZ.

HGDs construct the Gaussian/autocorrelation matrix upon the patch Gaussians with the *outer product* of the flattened vectors. Recall the vectorization operation in Eq.(4) and our concern with the case in which α dominates the diagonal elements of $\log\boldsymbol{G}$, *i.e.,* $\mathrm{diag}(\log\boldsymbol{G}) \approx \alpha\mathbf{1}_{d+1}$, where $\mathbf{1}_{d+1} \in \mathbb{R}^{d+1} = (1, \ldots, 1)^T$. In this case, we have $\boldsymbol{g} = \mathrm{vec}(\log\boldsymbol{G})$ = $[\mathrm{diag}(\log\boldsymbol{G}) \quad \sqrt{2}\mathrm{offdiag}(\log\boldsymbol{G})^T]^T \approx [\alpha\mathbf{1}_{d+1}^T \ \boldsymbol{e}^T]^T$, where $\boldsymbol{e}$ is a vector of which elements have smaller magnitudes than α. In this way, α dominates the patch Gaussian vectors in the LE tangent space, and also the statistical values over the vectors. For example, the autocorrelation matrix (similar to the upper-left block of the Gaussian matrix) of $\boldsymbol{g}$ becomes,

$$\boldsymbol{\Xi}'^{\mathcal{G}} \approx \frac{1}{\sum_{s\in\mathcal{G}} w_s} \sum_{s\in\mathcal{G}} w_s \begin{bmatrix} \alpha_s^2\mathbf{1}_{d+1}\mathbf{1}_{d+1}^T & \alpha_s\mathbf{1}_{d+1}\boldsymbol{e}_s^T \\ \alpha_s\boldsymbol{e}_s\mathbf{1}_{d+1}^T & \boldsymbol{e}_s\boldsymbol{e}_s^T \end{bmatrix}. \quad (21)$$

Because it is expected to be $|\alpha_s^2| > |\alpha_s e_s| > |e_s e_s|$, where e_s is an arbitrary component of $\boldsymbol{e}_s$, the region autocorrelation matrix is expected to be largely dominated by α_s^2. This means that only the mean logarithmic eigenvalue of the patch Gaussians is mostly reflected in the region Gaussian. Therefore, the removal of the diagonal bias $\alpha_s\boldsymbol{I}_{d+1}$ from the patch Gaussian vector can enhance the representation of HGDs. In fact, MSN cancels this component (Eq.(16)). We note that α_s may also contain the characteristic properties of a patch because it contains the mean information of the logarithmic eigenvalues. Namely, if the advantage of removing the biased components exceeds the potential risk of losing the discriminative information contained in α_s, MSN can improve the performance. The same discussion holds true when we apply MSN to region Gaussians.



(a) ZOZ [left: w/o PN, right: w/ PN]

(b) GOG [left: w/o PN, right: w/ PN]

Fig. 11. Performance gain by applying MSN to either one or both of the patch/region Gaussians on the VIPeR dataset: (a) ZOZ and (b) GOG.

Fig. 11 compares the performance gain by applying MSN to either one or both of the patch/region Gaussians under different norm and power normalizations[7]. In this comparison, we evaluate the performance on the VIPeR dataset using three distance metrics and report their average results. The results indicate that: (1) When PN is not used, applying MSN to either one of the patch/region Gaussians tends to improve the performance. This performance improvement is ascribed to the effects of removing the dominating α_s values in Eq.(21) and diagonal bias elements in the region Gaussians. Because the application of MSN to either the patch or region Gaussians improves the performance, applying MSN to both Gaussians improves the performance further. (2) The effect of MSN on the region Gaussians disappears when PN is applied, especially when unnormalized features or the standard L2 normalization is used. This may be because the improvement in the region Gaussians resulted from adjusting the magnitude of diagonal and off-diagonal elements and PN also has this ability. (3) MSN tends to fail to improve the performance when the I-L2 normalization is used. The reason is probably that the bilinear

[6] We use the regularization parameters $\epsilon = 10^{-4}$ and $\epsilon^{\mathcal{G}} = 10^{-2}$; thus the lowest logarithmic values of patch and region Gaussians are $\ln 10^{-4} \approx -9.2$ and $\ln 10^{-2} \approx -4.6$, respectively.

[7] L2 represents the standard normalization without bias removal.

TABLE 2
Evaluating the impact of MSN, PN, and norm normalizations

			VIPeR (PUR)				**GRID** (PUR)				**CUHK01** (PUR)				**CUHK03** (PUR)				**Market-1501** (mAP)			
			ZOZ		**GOG**		**ZOZ**		**GOG**		**ZOZ**		**GOG**		**ZOZ**		**GOG**		**ZOZ**		**GOG**	
Norm	**PN**	**MSN**	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST	XQDA	NFST
None			59.8	61.6	60.8	61.5	44.4	43.6	43.5	42.0	74.9	77.7	74.3	77.1	77.0	77.3	76.9	78.8	41.1	44.3	41.6	45.2
		✓	**60.8**	**62.5**	**61.6**	**62.6**	**45.6**	**44.7**	**44.9**	**43.3**	**75.0**	**78.2**	**74.3**	**77.9**	76.8	77.3	76.5	**78.9**	41.0	44.1	41.4	43.8
	✓		54.1	55.6	57.2	57.3	37.6	36.1	38.5	36.6	70.4	77.0	72.1	77.2	68.9	80.3	59.7	80.5	30.2	38.2	26.1	40.4
		✓	**55.4**	**57.2**	**58.1**	**58.1**	**38.9**	**37.4**	**40.3**	**38.2**	**70.7**	76.8	**72.8**	**77.2**	**69.6**	**81.1**	**60.9**	**81.2**	**31.1**	**38.5**	**27.3**	**40.5**
L2			59.3	61.6	60.6	61.9	44.7	43.2	43.8	41.9	75.3	78.0	75.1	77.3	75.5	76.3	75.5	77.7	38.8	44.1	39.0	45.1
		✓	**60.5**	**62.4**	**61.5**	**62.6**	**45.9**	**44.3**	**45.1**	**43.1**	73.8	**78.2**	74.5	**77.8**	75.3	**76.7**	**76.3**	**78.7**	**40.7**	44.0	**41.7**	43.4
	✓		54.2	55.6	57.3	57.3	37.9	35.7	38.3	36.4	71.5	77.0	73.3	77.1	73.7	80.2	73.1	80.6	33.7	38.2	34.5	40.4
			55.5	**57.0**	**58.5**	**58.0**	**38.6**	**37.5**	**40.2**	**38.2**	71.5	76.9	**74.2**	**77.4**	72.4	**81.4**	71.2	**81.5**	32.5	**38.7**	33.2	40.1
E-L2			62.0	61.3	62.7	61.3	46.3	43.4	45.1	41.7	75.9	78.5	75.6	77.5	80.5	79.3	81.2	80.7	40.6	43.7	41.3	44.7
		✓	**62.8**	**62.7**	**63.6**	**62.6**	**47.0**	**44.3**	**46.0**	**43.0**	**76.1**	**78.9**	**76.5**	**78.3**	**81.3**	**80.1**	**81.9**	**81.3**	**40.8**	**44.0**	**41.4**	44.3
	✓		60.8	59.5	61.8	59.7	44.1	41.5	43.8	40.2	78.7	80.1	79.2	79.4	83.1	82.2	84.1	83.0	41.8	43.5	43.1	44.9
		✓	**61.0**	**60.2**	**62.1**	**60.4**	**44.9**	**42.1**	**44.7**	**41.5**	78.7	**80.3**	**79.5**	**79.8**	**83.2**	**82.9**	**84.4**	**83.5**	41.7	43.4	43.0	44.5
E*-L2			56.1	55.4	57.8	56.0	39.8	37.3	39.9	37.0	78.6	79.8	79.1	78.9	82.7	81.5	83.6	81.6	40.3	40.8	42.4	41.9
		✓	**56.5**	**56.1**	**58.2**	**56.7**	**41.0**	**38.4**	**41.6**	**38.2**	78.5	**80.0**	**79.3**	**79.3**	82.5	**82.1**	**83.9**	**82.4**	40.2	40.8	41.8	41.4
	✓		56.3	55.3	57.6	55.7	40.4	37.6	40.5	36.9	78.9	79.7	79.2	79.1	83.4	82.5	84.9	82.6	40.4	40.4	42.2	41.8
		✓	**56.6**	**55.8**	**57.9**	**56.2**	**41.0**	**38.6**	**41.4**	**38.3**	78.7	**79.9**	**79.4**	**79.4**	83.4	**83.0**	84.6	**83.3**	40.2	40.2	41.7	41.5
I-L2			64.1	63.4	64.8	63.8	49.0	46.1	48.9	45.0	78.6	80.6	78.7	79.7	80.8	81.2	82.4	82.2	42.3	46.3	43.5	47.6
		✓	63.9	**63.5**	**65.0**	**64.2**	48.2	45.9	48.4	45.0	78.0	80.4	**79.1**	**80.2**	**81.3**	**81.6**	**82.8**	**82.6**	41.8	45.6	**43.7**	**47.8**
	✓		62.7	61.3	63.6	61.8	47.2	44.3	47.3	43.7	80.6	81.7	81.2	81.4	83.5	84.4	84.5	85.0	43.2	46.3	45.3	48.4
		✓	62.0	61.1	63.6	**62.1**	46.3	43.7	46.9	**43.8**	79.9	**81.8**	81.1	81.4	**83.6**	84.2	84.5	84.9	42.7	45.4	45.1	48.2

The mark ✓ indicates the usage of MSN/PN. The improved scores by MSN in the case of {**w/o PN**, **w/ PN**} *are marked in* {**blue**, **red**}*, respectively.*

transformation in the I-L2 normalization also enables elimination of the large bias caused by α_s.

Table 2 lists the results on all five datasets when MSN was applied to both the patch/region Gaussians. We see that MSN tends to improve the performance except for the I-L2 normalization[8].

Analysis of PN and the norm normalizations. We conducted an in-depth analysis of PN and the norm normalizations in Table 2. The results indicate that: (1) Both the E-L2 and I-L2 normalizations improve the performance of the original features, whereas the L2 normalization shows either no or a slight improvement. These results confirm the effect of removing the large bias that exists on the descriptors. (2) The E*-L2 normalization decreases the performance of the original features in several datasets. These results are probably obtained because HGDs have essentially uncorrelated covariance components in pixel features, *e.g.,* M_{90} and M_{180}. Enlarging narrowly ranged dimensions can emphasize these irrelevant dimensions. (3) PN decreases the performance on several datasets. The reason is probably the same as for the E*-L2 normalization. Furthermore, in Table 2, PN shows improved performance on the CUHK01 and CUHK03 datasets. These results are probably attributed to the smaller effect of magnifying small feature elements than the E*-L2 normalization. Thus, PN performs well when the magnitudes of discriminative elements are very small. (4) Normalization with intrinsic bias removal (I-L2) improves the performance more than normalization with extrinsic bias removal (E-L2 and E*-L2), which suggests that a greater respect for Riemannian geometry enables higher performance.

Fig. 12 compares the distance distribution of the GOG descriptor on all possible image pairs of the VIPeR dataset. We confirm that by applying the bias removal, the range of distance distributions becomes broader than the standard L2 normalization, both in the case of the extrinsic and intrinsic bias removals (E-L2 and I-L2). Furthermore, the I-L2 normalization has a slightly broader distance distribution than the E-L2 normalization. Apart from this, the distance distribution of the E*-L2 normalization is narrower than that of the E-L2 normalization. A plausible reason is that similar elements occur in small ranges of values in HGDs. Enlarging these ranges would enable the cosine distance to be dominated by similar elements. For the same reason, PN narrows the distance distribution. Moreover, we confirm that the bias removal of the E-L2 and I-L2 normalizations broadens the distance distribution, also in the case of PN.

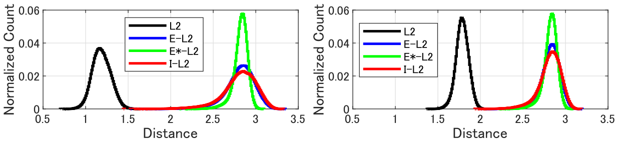


Fig. 12. Histogram of distances of GOG descriptor on all image pairs of the VIPeR dataset (Left w/o PN, Right w/ PN).

5.4 Performance Comparison

We compare the following aspects of HGDs with other state-of-the-art approaches: (1) Meta-descriptors; (2) Person re-id descriptors; (3) State-of-the-art methods on person re-id.

Comparison with existing meta-descriptors. We compare HGDs with existing meta-descriptors: Local Descriptors encoded by Fisher Vector (LDFV) [75], Riemannian(R)-VLAD [57], and FV-L^{2}EMG [55]. LDFV encodes pixel features using FV coding, which encodes the difference of pixel features from pre-trained GMM means. By following the recommended setting [75], we set the number of GMM components to 16. R-VLAD is the VLAD coding on the Riemannian manifold data. We used the Stein divergence [76] as metric to encode local covariance matrices. We set the number of codebooks to 32. FV-L^{2}EMG summarizes patch Gaussian matrices by FV coding. We confirmed that the influence of the number of GMM components is small and set it to 16.

We extract each of the meta-descriptors from the same horizontal strips as HGDs using the same pixel features of the four color spaces. For a fair comparison, we commonly use the three metric learnings and apply the mean removal and L2 norm normalization (equivalent to the E-L2 for SPD matrix descriptors). Because of the inapplicability to several descriptors, we evaluate the performance without the patch weight for all descriptors.

Table 3 (a) compares the results of the different methods. The hierarchical descriptors, R-VLAD and FV-L^{2}EMG clearly outperform the descriptor based on single-layered distribution, LDFV. The ZOZ and GOG outperform R-VLAD because they include the mean information, which is missing from the local covariance matrices used in R-VLAD. Additionally, ZOZ and GOG outperform FV-L^{2}EMG. These results may be understood by considering that the FV coding assumes diagonal covariance and thus the correlations within the local Gaussian vectors are absent in encoding.

Comparison with existing re-id descriptors. We compare the HGDs with several state-of-the-art descriptors used in supervised person re-id: gBiCov [34], Color Histogram(CH)+LBP [6], LOMO [11], Fine-Tuned(FT) CNN [45], and Histogram of In-

[8] MSN also tends to fail to improve the scores on the Market-1501 dataset probably because our approach leaves large misalignments unaddressed.

tensity Pattern and Ordinal Pattern (HIPHOP) [13]. We used the public codes with default parameters for LOMO and gBiCov. We used 75 regions to extract 28 bin CH and two uniform LBPs for CH+LBP, which was the best setting in the previous work [6]. FTCNN conducts a fine-tuning of the pre-trained AlexNet [37] using a dataset consisting of pedestrian attributes and extracts features from the fc6 layer. HIPHOP is a concatenation feature of two histogram patterns extracted from the lower layers of the pre-trained AlexNet. In addition, we improve the baseline by using CNN features of a Residual Network (ResNet) [77]. We trained a 50-layered ResNet using the training split of the Market-1501 dataset with a triplet loss function [78]. Following the previous work [78], we use the features of the last fully connected layer (fc2). Because the fully connected layer may overfit the trained dataset of CNN, we also use the last convolutional features (conv5_x). With average pooling, we accumulate the conv5_x features in the same seven horizontal strips as HGDs. The detailed settings are given in Appendix F. We apply the three metric learnings for all the descriptors. To show the superiority of the HGDs against the existing re-id descriptors, we use the final model, *i.e.,* MSN and the I-L2 normalization for HGDs. We do not use PN because it decreased the performance on several datasets.

Table 3 (b) and (c), respectively, present the results on hand-crafted and CNN features. Although LOMO and CH+LBP use a larger number of spatial regions and higher dimensional pixel features, both ZOZ and GOG outperform these descriptors by a large margin. The superiority of these descriptors originates from their hierarchical use of the mean and covariance information of pixel features, whereas LOMO uses only the mean information. In addition, ZOZ and GOG outperform CNN features such as HIPHOP and FTCNN. On the Market-1501 dataset, ResNet outperforms GOG and ZOZ in both fc2 and conv5_x features. On other datasets, the conv5_x features largely outperform fc2 features. Nevertheless, GOG and ZOZ show higher performances than the conv5_x features. Notably, the combination of ResNet and GOG/ZOZ shows significant improvements from both features. These results are probably obtained because the low-level features of HGDs are complementary to high-level features of CNN.

Comparison with state-of-the-art approaches. We compare the best combination of HGDs and distance metrics with state-of-the-art person re-id approaches. In this comparison, we refer HGD to as GOG+XQDA, ZOZ+XQDA, ZOZ+NFST, GOG+NFST, and GOG+NFST, respectively, for the VIPeR, GRID, CUHK01, CUHK03, and Market-1501 datasets. We also use the combination of HGD and ResNet, which we refer to as HGD+ResNet. The normalizations we used are the same as those in the previous comparison.

Table 4 (a) lists the results of the state-of-the-art metric learning approaches. We see that HGD alone outperforms the results of DNS [12] and MESP [79]. We note that NK3ML [15] uses the combination of GOG and LOMO features for the input of metric learning. On the GRID dataset, HGD outperforms NK3ML due to the I-L2 normalization we developed. With the E-L2 normalization, HGD performed 27.1% rank-1 rate, whereas it increased to 28.2% by the I-L2 normalization. We note that both CRAFT [13] and IRS [14] use the fusion of hand-crafted features and CNN features; thus, these methods show higher performances than HGD alone. Nevertheless, we see that the combination of ResNet and HGD outperforms these results.

Table 4 (b) lists the results of the state-of-the-art deep learning approaches. On the VIPeR and GRID datasets, HGD outperforms Gated-SCNN [39], Fused-CNN [40], Spindle-net [41], and Multi-level similarity-CNN [43]. These results are attributed to the lack of a sufficiently large number of training samples to train deep models. Typically, a deep model contains millions of parameters. Additionally, the nonlinear activation function in each layer of the CNN may discard appearance patterns that are not covered in the training samples. For example, the VIPeR dataset contains only one person who is in yellow jeans (Fig 5 (a)). Although such rare appearances are discriminative, deep learning methods may discard such discriminative appearance patterns. In contrast, HGD has no dependency on training samples; thus, it describes the uncovered appearance patterns in training data more appropriately.

On the CUHK03 and the Market-1501 datasets, deep learning approaches [42], [43], [44], [80] outperform HGDs. The highly adaptive abilities of the deep models enable complex variations to be learned that commonly exist in both the training and test sets. For example, although the Market-1501 dataset contains a significant difference in the bounding box of persons (Fig 5 (e)), these misalignments are common in the training and test sets. In contrast, HGDs use fixed horizontal strips without employing the mechanism of handling large variations of the bounding boxes. Although metric learning learns common variation in the training and test sets, this ability is limited compared with deep models because it basically learns a linear transformation of the extracted features. Thus, on these large-scale datasets, deep learning can improve the performance.

We emphasize that the focus of the study presented in this paper is the latest advance of handcrafted descriptors. As we have shown, HGDs would complement the recent advances of CNNs. This work is also foreseen to complement the advances of feature augmentation and metric learning methods such as CRAFT [13].

6 Conclusions

We have proposed novel hierarchical meta-descriptors for person re-id, which model both the mean and covariance information of the pixel features in both of the patch and region hierarchies. Extensive experiments confirmed the importance of both the mean information of pixel features and the hierarchal distribution for supervised person re-id. We also verified that the scale normalization of Gaussian matrices enhances the hierarchical descriptors. Based on the normalization, we proposed zero-mean Gaussian embedding, which achieves similar performance to the original Gaussian embedding with smaller dimensionality. Additionally, we investigated the feature norm normalization of SPD matrix-based descriptors. The normalization with the intrinsic statistics of the Riemannian manifold exhibited more accurate re-id results.

The normalization results have encouraged us to develop metric learnings of the SPD matrix descriptor for person re-id, *e.g.,* [81], [82]. Another future direction is to add more hierarchies to HGDs. In our preliminary experiments, this approach showed only slight improvements, whereas the dimensionality of the descriptors increased drastically. We expect that by developing supervised descriptor learning for HGDs, the difficulty in handling more hierarchies would be solved. Further, we consider enhancing the robustness of HGDs against large changes in the viewpoint and/or pose and occlusions. In this direction, obtaining pose-normalized regions by the recently advanced pose estimation technique seems promising [41], [42], [80]. We anticipate that the highly discriminative ability of HGDs would facilitate the description of features on the pose-normalized regions.

TABLE 3
Comparison of state-of-the-art descriptors: (a) Meta-descriptors; (b) Hand-crafted Re-id descriptors; (c) Re-id features using CNN

			VIPeR (PUR)			**GRID** (PUR)			**CUHK01** (PUR)			**CUHK03** (PUR)			**Market-1501** (mAP)		
	Methods	Dim.	KISSME	XQDA	NFST	KISSME	XQDA	NFST	KISSME	XQDA	NFST	KISSME	XQDA	NFST	KISSME	XQDA	NFST
(a)	LDFV [75]	6,944	43.9	44.5	48.4	35.3	36.4	37.2	53.9	58.1	61.0	54.8	66.5	56.4	16.7	21.5	24.1
	R-VLAD [57]	30,464	46.6	46.6	45.4	33.6	33.9	30.1	67.3	68.3	69.2	**77.2**	**82.0**	**82.7**	23.1	30.6	31.7
	FV-L^{2}EMG [55]	38,304	49.6	49.9	48.8	40.6	40.7	37.3	64.3	66.7	68.7	61.4	72.4	71.6	19.4	27.3	30.6
	ZOZ(w/o patch weight, E-L2)	16,828	**60.8**	**61.3**	**61.3**	**43.9**	**44.4**	**41.3**	**72.4**	**72.7**	**76.0**	**77.2**	81.0	80.7	**32.5**	**36.8**	**40.3**
	GOG(w/o patch wieght, E-L2)	27,622	**61.5**	**61.9**	**60.9**	**43.4**	**43.8**	**40.2**	**72.7**	**73.3**	**75.5**	**78.3**	**81.5**	**81.9**	**33.6**	**38.1**	**41.8**
(b)	gBiCov [34]	5,940	38.2	40.4	45.6	28.1	28.4	29.1	45.5	47.6	49.9	46.6	45.8	49.8	10.4	10.6	11.7
	CH+LBP [6]	32,250	43.0	45.0	46.1	36.3	36.7	36.8	54.1	57.2	59.8	49.5	56.6	56.4	15.3	16.0	22.0
	LOMO [11]	26,960	56.5	56.9	57.0	36.4	36.7	34.4	72.1	72.0	73.9	71.6	69.2	76.9	24.1	21.0	29.6
	ZOZ	16,828	**64.2**	**64.3**	**63.7**	**48.1**	**48.1**	**45.9**	**77.9**	**79.1**	**80.9**	**75.3**	**81.7**	**82.3**	**33.2**	**41.7**	**45.5**
	GOG	27,622	**64.7**	**64.9**	**64.3**	**48.0**	**48.0**	**44.7**	**77.9**	**78.9**	**79.9**	**76.4**	**82.6**	**83.4**	**32.3**	**42.8**	**46.5**
(c)	HIPHOP [13]	84,096	62.2	62.6	61.6	41.2	41.3	38.6	73.9	73.6	75.8	66.0	63.4	71.7	28.4	26.2	38.5
	FTCNN [45]	4,096	56.6	57.4	56.4	43.1	43.5	42.0	68.7	69.0	68.5	69.4	70.2	62.9	28.2	28.7	30.4
	ResNet (fc2) [78]	128	44.2	45.2	40.6	31.6	31.6	33.0	44.4	45.6	23.5	48.5	49.5	44.4	65.6	65.9	52.2
	ResNet (conv5_x) [78]	14,336	59.6	61.9	59.6	42.3	42.1	40.2	62.5	66.0	66.1	61.6	65.0	53.3	65.8	66.5	63.0
	ResNet (conv5_x)+**ZOZ**	31,164	**71.2**	**71.6**	**70.7**	**53.9**	**54.2**	**51.3**	**81.7**	**82.1**	**83.5**	**80.6**	**84.4**	**84.2**	**66.2**	**68.6**	**69.0**
	ResNet (conv5_x)+**GOG**	41,958	**72.1**	**72.4**	**70.9**	**54.0**	**54.0**	**50.1**	**81.8**	**82.4**	**83.2**	**81.7**	**85.3**	**84.8**	**66.9**	**69.1**	**69.5**

*The **red**/**blue** scores show the **first**/**second** best scores in each comparison.*

TABLE 4
State-of-the-art results (CMC@rank-r/mAP):(a) Feature extraction + metric learning approaches; (b) Deep learning approaches

			VIPeR				**GRID**			
	Methods	Ref.	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
(a)	DNS [12]	CVPR'16	51.2	82.1	90.5	95.9	-	-	-	-
	MESP [79]	IJCV'17	-	-	-	-	23.5	42.3	52.4	62.2
	CRAFT [13]	PAMI'18	54.2	**82.4**	**91.5**	**96.9**	26.0	**50.6**	**62.5**	**73.3**
	NK3ML [15]	ECCV'18	-	-	-	-	27.2	-	61.0	71.0
	IRS [14]	IJCV'18	**54.6**	-	90.3	95.7	-	-	-	-
	HGD	**Ours**	52.0	81.1	89.8	95.2	**28.2**	49.7	60.6	71.5
	HGD+ResNet	**Ours**	**63.4**	**86.7**	**93.7**	**97.3**	**33.7**	**60.4**	**70.8**	**79.0**
(b)	Gated-SCNN [39]	ECCV'16	37.8	66.9	77.4	-	-	-	-	-
	Fused-CNN [40]	NIPS'16	-	-	-	-	19.2	38.4	53.6	66.4
	Spindle [41]	CVPR'17	53.8	74.1	83.2	92.1	-	-	-	-
	ML-Sim. [43]	CVPR'18	50.1	73.1	84.4	-	-	-	-	-

			CUHK01			**CUHK03**			**Market-1501**	
	Methods	Ref.	r=1	r=10	r=20	r=1	r=10	r=20	r=1	mAP
(a)	DNS [12]	CVPR'16	-	-	-	54.7	94.8	-	61.0	35.7
	MESP [79]	IJCV'17	-	-	-	-	-	-	53.1	26.7
	CRAFT [13]	PAMI'18	78.8	95.3	97.8	-	-	-	71.8	45.5
	NK3ML [15]	ECCV'18	76.8	95.6	98.0	-	-	-	-	-
	IRS [14]	IJCV'18	**80.8**	**96.9**	**98.7**	**83.3**	**97.9**	**98.6**	**73.9**	**49.4**
	HGD	**Ours**	74.9	95.1	97.5	83.9	97.6	98.7	71.7	47.8
	HGD+ResNet	**Ours**	**80.3**	**97.0**	**98.7**	**88.5**	**98.4**	**99.2**	**87.0**	**70.9**
(b)	PDC [42]	ICCV'17	-	-	-	78.3	97.2	98.4	84.1	63.4
	ML-Sim. [43]	CVPR'18	-	-	-	86.5	99.1	-	-	-
	PN-GAN [80]	ECCV'18	67.7	91.8	-	79.8	98.6	-	89.4	72.6
	PA-Bilinear [44]	ECCV'18	-	-	-	88.0	98.6	99.0	90.2	76.0

*The **red**/**blue** scores shows the **first**/**second** best scores in (a).*

REFERENCES

[1] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-Identification*. Advances in Computer Vision and Pattern Recognition, Springer, 2014.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv:1610.02984*, 2016.

[3] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: feature, metrics, and datasets," *IEEE TPAMI*, vol. 41, no. 3, pp. 523–536, 2019.

[4] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *ECCV*, 2014.

[5] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*, 2014.

[6] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*, 2014.

[7] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE TPAMI*, vol. 35, no. 3, pp. 653–668, 2013.

[8] M. Dikmen, E. Akbas, T. Haung, and N. Ahuja, "Pedestrian recognition with learned metric," in *ACCV*, 2010.

[9] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012.

[10] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013.

[11] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[12] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016.

[13] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE TPAMI*, vol. 40, no. 2, pp. 392–408, 2018.

[14] H. Wang, X. Zhu, S. Gong, and T. Xiang, "Person re-identification in identity regression space," *Int. Journal of Computer Vision*, vol. 126, no. 12, pp. 1288–1310, 2018.

[15] T. Ali and S. Chaudhuri, "Maximum margin metric learning over discriminative nullspace for person re-identification," in *ECCV*, 2018.

[16] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *ECCV*, 2006.

[17] D. Tosato, M. Spera, M. Cristani, and V. Murino, "Characterizing humans on Riemannian manifolds," *IEEE TPAMI*, vol. 35, no. 8, pp. 1972–1984, 2013.

[18] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat, "Learning to match appearances by correlations in a covariance metric space," in *ECCV*, 2012.

[19] S. Bak, E. Corvée, F. Brémond, and M. Thonnat, "Boosted human re-identification using Riemannian manifolds," *Image and Vision Computing*, vol. 30, no. 6-7, pp. 443–452, 2012.

[20] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *Int. Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006.

[21] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM J. Matrix Analysis Applications*, vol. 29, no. 1, pp. 328–347, 2006.

[22] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Free-form region description with second-order pooling," *IEEE TPAMI*, vol. 37, no. 6, pp. 1177–1189, 2015.

[23] L. Gong, T. Wang, and F. Liu, "Shape of Gaussians as feature descriptors," in *CVPR*, 2009.

[24] B. Ma, Q. Li, and H. Chang, "Gaussian descriptor based on local features for person re-identification," in *ACCV Workshop*, 2014.

[25] H. Nakayama, T. Harada, and Y. Kuniyoshi, "Global Gaussian approach for scene categorization using information geometry," in *CVPR*, 2010.

[26] G. Serra, C. Grana, M. Manfredi, and R. Cucchiara, "GOLD: Gaussians of local descriptors for image representation," *Computer Vision and Image Understanding*, vol. 134, pp. 22–32, 2015.

[27] P. Li and Q. Wang, "Local log-Euclidean covariance matrix (L2ECM) for image representation and its applications," in *ECCV*, 2012.

[28] G. Serra, C. Grana, M. Manfredi, and R. Cucchiara, "Covariance of covariance features for image classification," in *ICMR*, 2014.

[29] M. Lovrić, M. Min-Oo, and E. A. Ruh, "Multivariate normal distributions parametrized as a Riemannian symmetric space," *Journal of Multivariate Analysis*, vol. 74, no. 1, pp. 36–48, 2000.

[30] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchal Gaussian descriptor for person re-identification," in *CVPR*, 2016.

[31] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *BMVC*, 2011.

[32] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130–144, 2013.

[33] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE TPAMI*, vol. 35, no. 7, pp. 1622–1634, 2013.

[34] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6, pp. 379–390, 2014.

[35] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by saliency learning," *IEEE TPAMI*, vol. 39, no. 2, pp. 356–370, 2017.

[36] R. Layne, T. M. Hospedales, and S. Gong, "Person re-identification by attributes," in *BMVC*, 2012.

[37] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[38] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[39] R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016.

[40] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *NIPS*, 2016.

[41] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017.

[42] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017.

[43] Y. Guo and N.-M. Cheung, "Efficient and deep person re-identification using multi-level similarity," in *CVPR*, 2018.

[44] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *ECCV*, 2018.

[45] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *ICPR*, 2016.

[46] M. S. Biagio, M. Crocco, M. Cristani, S. Martelli, and V. Murino, "Heterogeneous auto-similarities of characteristics (HASC): exploiting relational information for classification," in *ICCV*, 2013.

[47] M. Harandi, M. Salzmann, and F. Porikli, "Bregman divergences for infinite dimensional covariance matrices," in *CVPR*, 2014.

[48] H. Minh, M. Biagio, and V. Murino, "Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces," in *NIPS*, 2014.

[49] S. Bak, M. S. Biagio, R. Kumar, V. Murino, and F. Brémond, "Exploiting feature correlations by Brownian statistics for people detection and recognition," *IEEE TSMC*, vol. 47, no. 9, pp. 2538–2549, 2017.

[50] S. Amari and H. Nagaoka, *Methods of Information Geometry*. American Mathematical Society, 2001.

[51] X. Hong, G. Zhao, S. Zafeiriou, M. Pantic, and M. Pietikáinen, "Capturing correlations of local features for image representation," *Neurocomputing*, vol. 184, no. 5, pp. 99–106, 2016.

[52] P. Li, Q. Wang, and L. Zhang, "A novel earth mover's distance methodology for image matching with Gaussian mixture models," in *ICCV*, 2013.

[53] Q. Wang, P. Li, L. Zhang, and W. Zuo, "Towards effective codebookless model for image classification," *Pattern Recognition*, vol. 59, pp. 63–71, 2016.

[54] Q. Wang, P. Li, W. Zuo, and L. Zhang, "RAID-G: Robust estimation of approximate infinite dimensional gaussian with application to material recognition," in *CVPR*, 2016.

[55] P. Li, Q. Wang, H. Zeng, and L. Zhang, "Local log-Euclidean multivariate Gaussian descriptor and its application to image classification," *IEEE TPAMI*, vol. 39, no. 4, pp. 803–817, 2017.

[56] M. Faraki, M. T. Harandi, . A. Wiliem, and B. C. Lovell, "Fisher tensors for classifying human epithelial cells," *Pattern Recognition*, vol. 47, no. 7, pp. 2348–2359, 2014.

[57] M. Faraki, M. T. Harandi, and F. Porikli, "More about VLAD: A leap from Euclidean to Riemannian manifolds," in *CVPR*, 2015.

[58] J. Sánchez and J. Redolfi, "Exponential family Fisher vector for image classification," *Pattern Recognition Letters*, vol. 59, pp. 26–32, 2015.

[59] I. Ilea, L. Bombrun, C. Germain, R. Terebes, M. Borda, and Y. Berthoumieu, "Texture image classification with Riemannian Fisher vectors," in *ICIP*, 2016.

[60] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. d. Bray, "Visual categorization with bags of keypoints," in *ECCV Workshop*, 2004.

[61] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[62] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE TPAMI*, vol. 34, no. 9, pp. 1704–1716, 2012.

[63] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE TPAMI*, vol. 30, no. 10, pp. 1713–1727, 2008.

[64] Q. Wang, P. Li, and L. Zhang, "G^{2}DeNet: Global Gaussian distribution embedding network and its application to visual recognition," in *CVPR*, 2017.

[65] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition?" in *ICCV*, 2017.

[66] P. Li, J. Xie, Q. Wang, and Z. Gao, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *CVPR*, 2018.

[67] I. Dryden, A. Koloydenko, and D. Zhou, "Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging," *The Annals of Applied Statistics*, vol. 3, no. 3, pp. 1102–1123, 2009.

[68] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel methods on Riemannian manifolds with Gaussian RBF kernels," *IEEE TPAMI*, vol. 37, no. 12, pp. 2464–2477, 2015.

[69] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 1972.

[70] L. Xie, Q. Tian, and B. Zhang, "Simple techniques make sense: Feature pooling and normalization for image classification," *IEEE TCSVT*, vol. 26, pp. 1251–1264, 2016.

[71] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening," in *ECCV*, 2012.

[72] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008.

[73] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *Int. Journal of Computer Vision*, vol. 90, no. 1, pp. 106–129, 2010.

[74] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[75] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by Fisher vectors for person re-identification," in *ECCV Workshop*, 2012.

[76] S.Sra, "A new metric on the manifold of kernel matrices with application to matrix geometric means," in *NIPS*, 2012, pp. 144–152.

[77] K. He, X. Zhang, S. Ren, and S. J, "Deep residual learning for image recognition," in *CVPR*, 2016.

[78] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv:1703.07737*, 2017.

[79] D. Chen, Z. Yuan, J. Wang, B. Chen, G. Hua, and N. Zheng, "Exemplar-guided similarity learning on polynomial kernel feature map for person re-identification," *Int. Journal of Computer Vision*, vol. 123, no. 3, pp. 392–414, 2017.

[80] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *ECCV*, 2018.

[81] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen, "Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification," in *ICML*, 2015.

[82] M. Harandi, M. Salzmann, and R. Hartley, "Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods," *IEEE TPAMI*, vol. 40, no. 1, pp. 48–62, 2018.

Tetsu Matsukawa received his B.E., M.E., and D.E. degrees from University of Tsukuba in 2006, 2008 and 2011, respectively. He worked as a project research associate at Institute of Industrial Science, the University of Tokyo from 2011 to 2014. He is currently an assistant professor at Kyushu University. His research interests include computer vision and pattern recognition.

Takahiro Okabe received his B.S. and M.S. degrees in physics, and his Ph.D. degree in information science and technology from the University of Tokyo in 1997, 1999, and 2011 respectively. After working at Institute of Industrial Science, the University of Tokyo, he joined Kyushu Institute of Technology in 2013, where he is currently a professor. He was a visiting scholar at Tübingen University from 2011 to 2012. His research interests include computer vision, computational photography, and their related fields in particular their physical and mathematical aspects.

Einoshin Suzuki received his B.E., M.E., and D.E. degrees from the University of Tokyo in 1988, 1990 and 1993, respectively. His research fields include data mining and machine learning. He has been a professor at Kyushu University from 2006. He is currently on the editorial board of JIIS. He has served as a PC member or several kinds of Chairs at KDD, ICDM, SDM, ECML/PKDD, and CIKM.

Yoichi Sato is a professor at Institute of Industrial Science, the University of Tokyo. He received his B.S. degree from the University of Tokyo in 1990, and his M.S. and Ph.D. degrees in robotics from School of Computer Science, Carnegie Mellon University in 1993 and 1997. His research interests include physics-based vision, reflectance analysis, first-person vision, and gaze sensing and analysis. He served/is serving in several conference organization and journal editorial roles including IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, Computer Vision and Image Understanding, ECCV 2012 Program Co-Chair, ACCV 2016 Program Co-Chair, ACCV 2018 General Co-Chair, and MVA 2013 General Chair.