2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Voronoi-based Second-order Descriptor with Whitened Metric in LiDAR Place Recognition

Jaein Kim[1], Hee Bin Yoo[2*], Dong-Sig Han[3*], and Byoung-Tak Zhang[1,4]

***Abstract*— The pooling layer plays a vital role in aggregating local descriptors into the metrizable global descriptor in the LiDAR Place Recognition (LPR). In particular, the second-order pooling is capable of capturing higher-order interactions among local descriptors. However, its existing methods in the LPR adhere to conventional implementations and post-normalization, and incur the descriptor unsuitable for Euclidean distancing. Based on the recent interpretation that associates NetVLAD with the second-order statistics, we propose to integrate second-order pooling with the inductive bias from Voronoi cells. Our novel pooling method aggregates local descriptors to form the second-order matrix and whitens the global descriptor to implicitly measure the Mahalanobis distance while conserving the cluster property from Voronoi cells, addressing its numerical instability during learning with diverse techniques. We demonstrate its performance gains through the experiments conducted on the Oxford Robotcar and Wild-Places benchmarks and analyze the numerical effect of the proposed whitening algorithm.**

I. Introduction

Place recognition is a problem for searching the nearest place to the query place from the memory of the autonomous system [1], [2]. It has gained its renown as one of the significant topics in the robotics field due to its correlation to loop closure detection and global localization in the SLAM or navigation systems [3], [4]. The research on place recognition is distinguishable concerning the sensory format of the observed place information. Especially, LiDAR Place Recognition (LPR) utilizes three-dimensional LiDAR to capture geometric information [4] and is known to be more robust to illuminative noise [4], [5].

Deep learning recently has become a dominant approach in the LPR; it utilizes the expressiveness of neural networks and has surpassed traditional approaches [6]. It generally reframes the task as the retrieval problem and learns the compressed representation of places with the application of contrastive learning techniques [3], [7], [8]. This representation is referred to as a global descriptor, which is obtainable by aggregating a set of local features while preserving local information to ensure the separability [9]–[11].

This work was partly supported by grants from the IITP (RS-2021-II211343-GSAI/10%, RS-2022-II220951-LBA/15%, RS-2022-II220953-PICA/15%), the NRF (RS-2024-00353991-SPARC/15%, RS-2023-00274280-HEI/15%), the KEIT (RS-2025-25453780/15%), and the KIAT (RS-2025-25460896/15%), funded by the Korean government.

[1]Interdisciplinary Program in Neuroscience, Seoul National University
[2]Département d'Informatique, École Normale Supérieure (ENS)
[3]Department of Computing, Imperial College London
[4]Dept. of Computer Science and Engineering, Seoul National University
*The majority of the work for this publication was done while these authors were in Seoul National University.

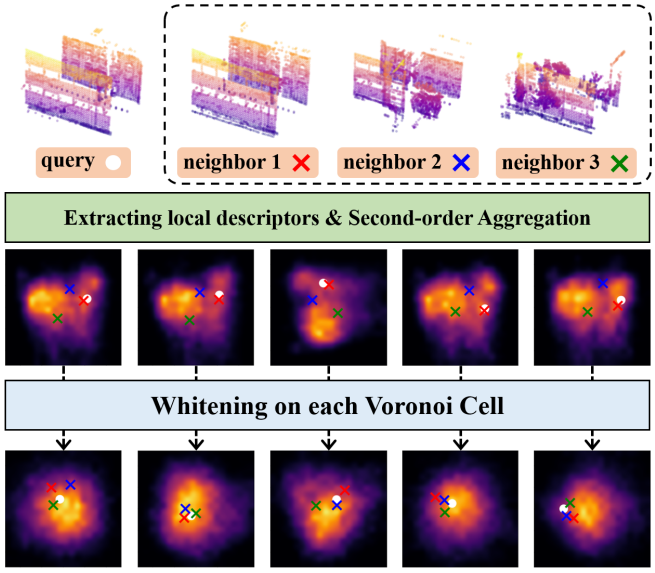

Fig. 1. The illustration of feature space of each Voronoi cell learned by our method. We select the query place and its top-3 closest neighbors metrized by our method from the Oxford [22] and visualize each cell of their descriptors reduced by ICA. White dots and colored cross marks denote query and neighbors. Our whitening transforms each Voronoi cell more homogeneous and suitable for Euclidean distancing.

The most representative pooling methods in the LPR are NetVLAD [12] and GeM [13], adopted from the Visual Place Recognition (VPR) studies. These pooling methods are recognized as the first-order pooling in the field [14], [15], which utilizes the first-order statistics, e.g., average or maximum, to compress the set of local descriptors. Although many studies have demonstrated the effectiveness of first-order pooling methods in the LPR [5], [10], [11], [16], the first-order pooling has a limitation that it cannot capture the higher-order interactions between local features [9], [17].

The second-order pooling utilizes second-order statistics between local features and embeds the pairwise correlation that cannot be modeled in first-order pooling [15], [17]–[19]. The limitation of previous methods is that they only exploit naive second-order statistics, e.g. the max or average operation over the outer products of local descriptors, and adhere to conventional normalization techniques irrelevant to the descriptor property. However, [20] have recently suggested to interpret NetVLAD as the second-order pooling in contrast to the traditional perspective rooted in VLAD [21], showing the equivalence between NetVLAD and the second-order operation with a condition on the soft-assignment.

The above interpretation expands the second-order pooling

beyond a naive outer-product on local descriptors and provides an inductive bias to design more plausible normalization technique. Our principal inductive bias from NetVLAD is that the clusters of global descriptor form independent subspaces with their own unique metrics, in accordance with the fact that the clusters of VLAD are considered as embedded Voronoi cells [12], [23]–[25]. Despite such a property, existing methods using VLAD or second-order pooling naively apply L^2 normalization or linear maps to the global descriptor. These might project the global descriptor onto a compact hypersphere or reduce the dimensionality, but they violate the topology of Voronoi cells and distort the metric of the descriptor. We claim that a more relevant normalization exists for the descriptor space embedding Voronoi cells by relaxing the compactness condition.

This paper proposes a novel second-order pooling method applicable to a learning-based place recognition model. It adopts the interpretation by [20] to justify the relation between NetVLAD and second-order pooling and designs simple neural networks to model the pooling modules. From the understanding of the global descriptor as a set of disjoint clusters, we propose to metrize the descriptor by the Mahalanobis distance considering the metric of each Voronoi cell, which is achievable through whitening the feature of each Voronoi cell homogeneous using its covariance as Figure 1.

Our core contribution is to realize the ZCA whitening [26] with additional techniques to stabilize numerical errors, enabling the exact whitening on the learnable global descriptor in an end-to-end manner. Our method demonstrates the state-of-the-art performance in Oxford RobotCar [22] and Wild-Places [27] with the integration of second-order aggregation and suitable whitening to metrize the descriptor in Voronoi cells. Numerical analyses also verify the positive effect of our whitening on the structure of each Voronoi cell.

II. Related Works

a) Learning-based LiDAR Place Recognition: PointNetVLAD [5] and MinkLoc3D [10] are some of the representative works in the LPR. PointNetVLAD popularizes a general pipeline of learning-based retrieval models in the LPR with the utilization of NetVLAD [12], and MinkLoc3D introduces a sparse tensor convolution based on the MinkowskiEngine library [28] and GeM pooling [13]. Most of the progress in the LPR has focused on proposing a novel backbone architecture [11], [16], [29]–[31] while adhering to the established first-order pooling methods. As aforementioned, capturing more complex interactions is available through second-order statistics, and such approaches have shown its validity in the recent VPR studies [8], [20].

Second-order pooling, however, has not been studied as thoroughly as first-order pooling methods in the LPR. Some representative works are Locus [17] and LoGG3D-Net [19] that aggregate local descriptors by O2P [15] and apply the square root over the eigenvalues of the pooled descriptor. Our approach extends prior methods by introducing a novel framework that establishes an association between second-order statistics and the cluster assumption.

b) VLAD and Descriptor Normalization: Vector of Locally Aggregated Descriptors (VLAD) [21], inspired by *bag-of-features* and Fisher kernel [32], accumulates the residuals of local descriptors to their centroids obtained by K-Means. Some post-normalization techniques, e.g., signed square rooting [33], [34] and intra-normalization [23], have been proposed to compensate for VLAD by suppressing visual burstiness of overestimated features [35]. However, their element-wise operations or L^2 normalization break the property of the VLAD descriptor as a set of clusters.

Meanwhile, a post-normalization by multivariate whitening has been studied previously; they whiten the global descriptor by PCA with a perspective coinciding with ours, viewing VLAD as the set of independent descriptors [24], [34]. However, they still normalize the global descriptor by L^2 at the last and are not suitable for learning neural networks in an end-to-end manner. Our method relaxes the compactness condition and proposes to learn the metric with neural networks from scratch.

III. Background: VLAD to Second-order Pooling

This section introduces the derivation by [20] that associates NetVLAD [12] with the second-order pooling, based on its equivalence to the bilinear pooling [36]. Given the local descriptors projected on the Voronoi diagram $\mathbf{F} = \begin{bmatrix}\mathbf{f}_1, \cdots, \mathbf{f}_L\end{bmatrix} \in \mathbb{R}^{C\times L}$, the centroids of clusters $\mathbf{C} = \begin{bmatrix}\mathbf{c}_1, \cdots, \mathbf{c}_M\end{bmatrix} \in \mathbb{R}^{C\times M}$, and the soft assignment of each local descriptor to clusters $\mathbf{P} = \begin{bmatrix}\mathbf{p}_1, \cdots, \mathbf{p}_L\end{bmatrix} \in \mathbb{R}^{M\times L}$, the global descriptor $\widetilde{\mathbf{X}}$ by NetVLAD is formulated as follows:

$$\widetilde{\mathbf{X}} = \sum_{i=1}^{L} \begin{bmatrix}\mathbf{f}_i - \mathbf{c}_1, \cdots, \mathbf{f}_i - \mathbf{c}_M\end{bmatrix} \odot \underbrace{\begin{bmatrix}\mathbf{p}_i, \cdots, \mathbf{p}_i\end{bmatrix}}_{C}^{\top}, \quad (1)$$

where $\odot$ is an element-wise product. NetVLAD computes $\mathbf{X}$ and $\mathbf{P}$ from the input local descriptors and parametrizes $\mathbf{C}$ as learnable weights. Equation (1) is reducible further when the additional constraint on $\mathbf{P}$ is given, saying $\sum_{i=1}^{L}\mathbf{p}_i = \mathbf{v}$ *is a constant vector*. Then,

$$\begin{aligned}(1) &= \mathbf{F}\,\mathbf{P}^{\top} - \mathbf{C} \odot \sum_{i=1}^{L}\begin{bmatrix}\mathbf{p}_i, \cdots, \mathbf{p}_i\end{bmatrix}^{\top} \\ &= \mathbf{F}\,\mathbf{P}^{\top} - \mathbf{C} \odot \begin{bmatrix}\mathbf{v}, \cdots, \mathbf{v}\end{bmatrix}^{\top},\end{aligned} \quad (2)$$

where the cluster centroids $\mathbf{C}$ become independent to the input local descriptors. We can declare that the global descriptor is pooled by second-order statistics and interpretable as clusters in Voronoi cells simultaneously, thanks to the disregardable centroids: $\widetilde{\mathbf{X}} \triangleq \mathbf{F}\,\mathbf{P}^{\top}$. However, [20] remain to normalize the descriptor with heuristic approach based on an approximated matrix power normalization and intra-normalization. We propose a suitable normalization technique that reflects the metrics of each Voronoi cell.

IV. Methods

Our pooling method consists of the feature projection $\mathcal{F}_{\text{proj}}(\cdot)$ and soft-assignment score $\mathcal{F}_{\text{score}}(\cdot)$ networks that correspond to $\mathbf{F}$ and $\mathbf{P}$ respectively, and the whitening

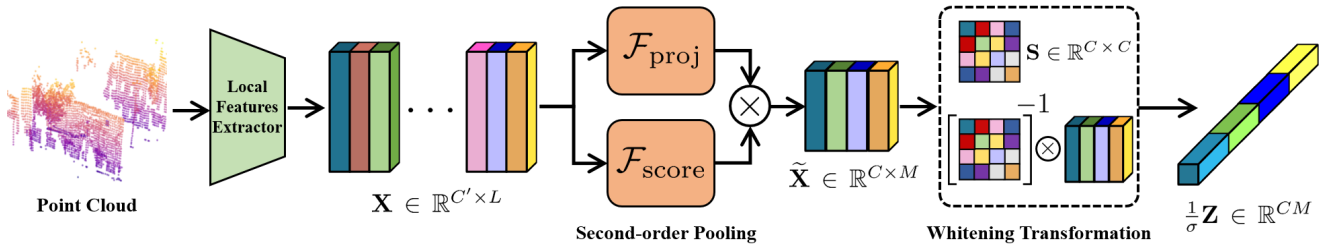

Fig. 2. The overall architecture. First, the input local descriptors $\mathbf{X}$ are mapped by networks $\mathcal{F}_{\text{proj}}$ and $\mathcal{F}_{\text{score}}$, multiplied into the global descriptor $\widetilde{\mathbf{X}}$. Our whitening module implemented with ZCA whitening [26] computes the standard deviation matrix $\mathbf{S}$ from $\widetilde{\mathbf{X}}$ and normalizes each cluster's feature by multiplying $\mathbf{S}^{-1}$ over $\widetilde{\mathbf{X}}$. Finally, the global descriptor $\mathbf{Z}$ is vectorized and scaled down to $\frac{1}{\sigma}\mathbf{Z}$.

module. As shown in Figure 2, the local descriptors $\mathbf{X}$ are mapped into the compressed features and the assignment scores to each Voronoi cell. Then the global descriptor $\widetilde{\mathbf{X}}$ is acquired from the summation of the projected features weighted by the score. Finally, each cluster feature of the global descriptor is whitened, and the whole descriptor vector is scaled by a positive coefficient σ.

A. The Second-order Descriptor in Voronoi Diagram and its Mahalanobis Distancing

Let us consider that $\mathcal{F}_{\text{proj}}(\cdot)$ and $\mathcal{F}_{\text{score}}(\cdot)$ are neural networks with two linear layers whose hidden layers are followed by batch normalization [37] and GELU [38] activation. Given $\mathbf{X} = \left[\mathbf{x}_1, \cdots, \mathbf{x}_L\right]$, which is the sequence of local descriptors of length L, these networks map each local descriptor in $\mathbf{X}$ with shared weights and output $\mathcal{F}_{\text{proj}}(\mathbf{X})$ and $\mathcal{F}_{\text{score}}(\mathbf{X})$. Then the weighted summation of projected features in $\mathcal{F}_{\text{proj}}(\mathbf{X})$ by the scores toward each cluster $\mathcal{F}_{\text{score}}(\mathbf{X})$ is computed as Equation (3):

$$\begin{aligned} \widetilde{\mathbf{X}} &= \mathcal{F}_{\text{proj}}(\mathbf{X})\ \text{softmax}(\mathcal{F}_{\text{score}}(\mathbf{X}))^\top, \\ \text{s.t. } \mathcal{F}_{\text{proj}}(\mathbf{X}) &\in \mathbb{R}^{C \times L},\ \mathcal{F}_{\text{score}}(\mathbf{X}) \in \mathbb{R}^{M \times L}, \end{aligned} \tag{3}$$

where $\widetilde{\mathbf{X}}$ is the aggregated global descriptor, and $\text{softmax}(\cdot)$ is applied along the sequence dimension of $\mathcal{F}_{\text{score}}(\mathbf{X})$ to meet the constraint mentioned in Section III, following [20].

The global descriptor $\widetilde{\mathbf{X}}$ is interpretable as the set of features in the Voronoi diagram. Since the Voronoi cells are independent of each other, let us assume that the distribution of each cell follows Gaussian with its own unique covariance $\mathbf{\Sigma}_i \in \mathbb{R}^{C \times C}$, where i is an index of the cell. The distribution of the entire descriptor space is explainable as the joint distribution of independent cells. Hence, we suggest to distance global descriptors $\widetilde{\mathbf{X}}_1$ and $\widetilde{\mathbf{X}}_2$ by Mahalanobis distance, with the abuse of notation that omits the vectorization:

$$\begin{aligned} d^2_{\text{Mahal}}(\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2; \mathbf{\Sigma}) &= (\widetilde{\mathbf{X}}_1 - \widetilde{\mathbf{X}}_2)^\top \mathbf{\Sigma}^{-1} (\widetilde{\mathbf{X}}_1 - \widetilde{\mathbf{X}}_2) \\ &= \sum_{i=1}^{M} (\widetilde{\mathbf{x}}_{1i} - \widetilde{\mathbf{x}}_{2i})^\top \mathbf{\Sigma}_i^{-1} (\widetilde{\mathbf{x}}_{1i} - \widetilde{\mathbf{x}}_{2i}), \\ \text{s.t. } \widetilde{\mathbf{X}}_j = \left[\widetilde{\mathbf{x}}_{j1}, \cdots, \widetilde{\mathbf{x}}_{jM}\right],\ \mathbf{\Sigma} &= \begin{bmatrix} \mathbf{\Sigma}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Sigma}_M \end{bmatrix}, \end{aligned} \tag{4}$$

where the covariance of joint distribution $\mathbf{\Sigma}$ is a block diagonal matrix. Due to the independence between the Voronoi cells, the entire operation is decomposable into the summation of Mahalanobis distance inside each cell with the reduced cost compared to the operation by the full $\mathbf{\Sigma}$.

B. Whitening Transformation for Mahalanobis Distancing

We implement Equation (4) with the whitening transformation instead of explicitly computing the inverse of covariances. It efficiently converts the Mahalanobis distance into the Euclidean distance:

$$\begin{aligned} d^2_{\text{Mahal}}(\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2; \mathbf{\Sigma}) &= \sum_{i=1}^{M} \|\mathbf{z}_{1i} - \mathbf{z}_{2i}\|_2^2\,, \\ \text{s.t. } \mathbf{S}_i\mathbf{S}_i^\top &= \mathbf{\Sigma}_i\,,\ \ \mathbf{z}_{ji} = \mathbf{S}_i^{-1}\mathbf{x}_{ji}. \end{aligned} \tag{5}$$

At a glance, one can acquire covariance $\mathbf{\Sigma}_i$ with the mini-batch samples for each cell; however, the statistics acquired from the batch-wise samples during training may not reflect the domain shift during the evaluation [39], [40]. As we measure the descriptor metric by Mahalanobis distance, the batch-wise covariance at the final layer can distort the metric during the evaluation. Therefore, we propose to modify the whitening with the assumption that *every cluster cell has the identical distribution parameters* as follows:

$$\begin{aligned} d^2_{\text{Mahal}}(\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2) &= \sum_{i=1}^{M} \|\mathbf{z}_{1i} - \mathbf{z}_{2i}\|_2^2\,, \\ \text{s.t. } \mathbf{z}_{ji} &= \mathbf{S}_j^{-1}(\mathbf{x}_{ji} - \hat{\mu}_j)\,,\ \ \hat{\mu}_j = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x}_{ji} \\ \mathbf{S}_j\mathbf{S}_j^\top &= \frac{1}{M}\sum_{i=1}^{M}(\mathbf{x}_{ji} - \hat{\mu}_j)(\mathbf{x}_{ji} - \hat{\mu}_j)^\top. \end{aligned} \tag{6}$$

Samples for the estimates are taken *within the instance* thanks to the assumption, free from tracking the learnable distribution parameters during the training.

Equation (6) is realizable with the estimation of the distribution parameters among Voronoi cells and the application of ZCA whitening over each cell's feature. However, a computation of the standard deviation matrix from a naive sample covariance results in the rank deficiency since the number of clusters is smaller than the dimension of cluster features, providing only the limited amount of information.

We resolve this issue by implementing ZCA whitening with the Rao-Blackwell Ledoit-Wolf (RBLW) [41] algorithm

TABLE I

THE AVERAGE RECALL AT TOP-1 AND TOP-1% RETRIEVALS IN THE OXFORD [22]. †: THE RESULTS QUOTED FROM THE ORIGINAL PAPER.

Methods	Oxford		U.S.		R.A.		B.D		Average	
	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%
PointNetVLAD [5]	74.88	87.89	68.09	81.37	64.05	74.66	63.42	70.63	67.61	78.64
TransLoc3D [16]	93.99	98.13	83.50	93.23	80.11	90.39	80.01	87.05	84.40	92.20
MinkLoc3D [10]	93.76	97.91	86.01	95.04	81.11	91.19	82.66	88.48	85.89	93.16
MinkLoc3Dv2 [11]	96.26	98.87	90.85	96.65	86.49	93.75	86.26	91.15	89.97	95.11
CASSPR [29]	95.84	98.75	92.91	98.00	89.44	94.69	87.25	92.36	91.36	95.95
SelFLoc [30]†	96.0	98.8	93.2	98.3	88.8	94.8	88.4	92.4	91.6	96.1
Ours (C=16, M=16)	96.63	98.87	91.04	97.36	86.95	93.89	87.55	91.76	90.54	95.47
Ours (C=128, M=64)	**97.97**	**99.40**	**94.97**	**98.58**	**90.99**	**95.90**	**90.98**	**94.33**	**93.73**	**97.05**

Algorithm 1: The proposed ZCA whitening.

Input: $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \cdots, \widetilde{\mathbf{x}}_M] \in \mathbb{R}^{C \times M}$, a small noise $\varepsilon = 1\text{e}{-5}$

Output: $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_M] \in \mathbb{R}^{C \times M}$

1. $\overline{\mathbf{X}} = [\widetilde{\mathbf{x}}_1 - \hat{\mu}, \cdots, \widetilde{\mathbf{x}}_M - \hat{\mu}]$ from $\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} \widetilde{\mathbf{x}}_i$
2. Initialize $\hat{\mathbf{\Sigma}} = \frac{1}{M}\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top$ and $\hat{\mathbf{F}} = \frac{\mathrm{Tr}(\hat{\mathbf{\Sigma}})}{C}\mathbf{I}$
3. $\rho_{\text{RBLW}} = \min\left(\frac{(\frac{M-2}{M})\mathrm{Tr}(\hat{\mathbf{\Sigma}}^2) + \mathrm{Tr}^2(\hat{\mathbf{\Sigma}})}{(M+2)\left(\mathrm{Tr}(\hat{\mathbf{\Sigma}}^2) - \frac{1}{C}\mathrm{Tr}^2(\hat{\mathbf{\Sigma}})\right)}, 1\right)$
4. $\hat{\mathbf{\Sigma}}_{\text{RBLW}} = \rho_{\text{RBLW}}\hat{\mathbf{F}} + (1 - \rho_{\text{RBLW}})\hat{\mathbf{\Sigma}}$
5. $\hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^\top = \texttt{SVDPI}(\hat{\mathbf{\Sigma}}_{\text{RBLW}} + \varepsilon\mathbf{I})$
6. $\mathbf{Z} = \hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}^{-\frac{1}{2}}\hat{\mathbf{Q}}^\top\overline{\mathbf{X}}$ `/* ZCA Whitening */`
7. **return** $\mathbf{Z}$

for the covariance shrinkage as Algorithm 1. The RBLW algorithm estimates the coefficient ρ_{RBLW} that minimizes the difference between the true covariance, which is supposed to be full rank, and the combination of $\hat{\mathbf{F}}$ and $\hat{\mathbf{\Sigma}}$.

However, our pooling method, which applies ZCA whitening over the descriptor learned with neural networks in an end-to-end manner, is vulnerable to the gradient instability. It is known that a naive eigendecomposition causes a gradient explosion when multiple eigenvalues have close values [42]–[44]. Thus, we replace it with an algorithm stable regarding the backpropagation [44] (`SVDPI` in Algorithm 1), which decomposes the covariance using SVD without auto differentiation and backpropagates the gradient using the iterative method. Please refer to its original paper [44] for the details.

C. The Configuration of Place Recognition Model

We apply our pooling method to the backbone network from MinkLoc3Dv2 [11] as a default model. The number of clusters M and the dimension of cluster features C are configured differently with regard to the size of global descriptor and datasets. We also divide the whitened descriptor $\mathbf{Z}$ by a scalar hyperparameter σ to normalize the norm of the entire descriptor. Since the divided descriptor follows the Gaussian distribution with diagonal covariance, $\frac{1}{\sigma}\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\sigma^2}\mathbf{I})$, its Euclidean distancing still implies the the Mahalanobis. The final metric is optimized by the Truncated Smooth-AP loss with the identical setups to [11], but we substitute a learning rate scheduler with the cosine annealing scheduler [45].

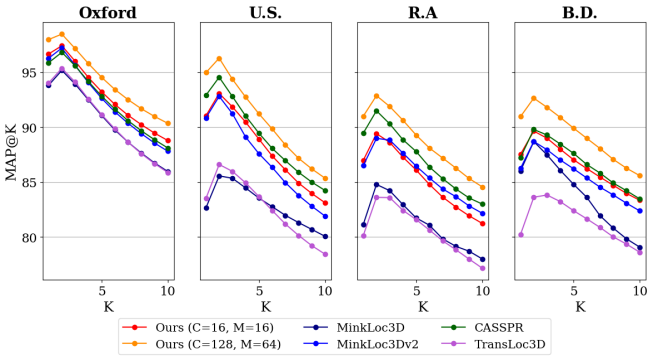

Fig. 3. Mean Average Precision at K (MAP@K) plotted up to 10 predictions in the Oxford and In-house datasets.

V. EXPERIMENTAL RESULTS

We conducted the experiments with two place recognition benchmarks: Oxford [22] and Wild-Places [27] concerning the urban and natural environments. Following the training and evaluation protocol from [5], the evaluations were also conducted with the In-house datasets which are out-of-distribution from the Oxford. In the Wild-Places, we followed the protocol from [27] and quantitatively assessed models with respect to the loop-closure detection and the global localization. Ablation studies and analyses were also conducted to verify the backbone-agnostic applicability and the validity of numerical techniques in the ZCA whitening.

A. Experiments on Oxford RobotCar

We compared our models to previous methods using NetVLAD (PointNetVLAD [5], TransLoc3D [16]) or GeM (MinkLoc3D [10], MinkLoc3Dv2 [11], CASSPR [29], SelFLoc [30]) as an aggregator in the Oxford experiments. They utilized the pretrained weights published by their authors or were trained from the scratch. Meanwhile, our methods had two variations; we configured the one have 8192 dimensions of the global descriptor ($C = 128,\ M = 64$) to emulate the conventional setup of VLAD and another have 256 dimensions (C=16, M=16) for fair comparison to baselines. Otherwise, they have identical training configurations with $\sigma = \sqrt{M}$. The performances of baselines were measured with Recalls at top-1 (R@1) and top-1% (R@1%) retrievals.

TABLE II

THE INTRA-SEQUENCE AND INTER-SEQUENCE RESULTS IN THE WILD-PLACES [27]. †: THE RESULTS QUOTED FROM THE ORIGINAL PAPER.

Methods	Descriptor Dimensions	Intra-sequence								Inter-sequence			
		V-03		V-04		K-03		K-04		Venman		Karawatha	
		F1	R@1	F1	R@1	F1	R@1	F1	R@1	R@1	MRR	R@1	MRR
ScanContext [46]	1200	05.54	12.79	39.28	43.33	31.33	43.55	58.09	61.72	43.15	45.23	52.79	56.40
TransLoc3D [16]	256	22.92	35.50	75.00	64.03	43.50	38.95	75.80	75.53	62.85	74.92	54.32	69.08
MinkLoc3Dv2 [11]	256	49.85	49.94	82.19	71.61	51.39	50.97	80.00	71.18	75.77	84.86	67.81	79.20
LoGG3D-Net [19]	1024	54.03	62.40	80.37	72.47	64.26	64.05	84.54	80.26	79.84	86.55	74.67	82.66
ForestLPR† [31]	1024	**64.15**	**76.53**	78.62	82.33	65.01	74.89	81.97	76.73	77.15	-	**79.02**	-
Ours (C=16, M=16)	256	49.67	51.38	90.13	84.95	65.27	67.73	91.41	89.53	81.02	88.08	73.83	82.86
Ours (C=32, M=32)	1024	55.33	52.48	**92.29**	**87.62**	67.56	70.56	93.33	92.31	83.94	90.05	76.12	84.67
Ours (C=128, M=64)	8192	59.33	54.02	90.49	84.95	**70.12**	**77.91**	**96.38**	**97.03**	**86.03**	**91.20**	78.83	**86.24**

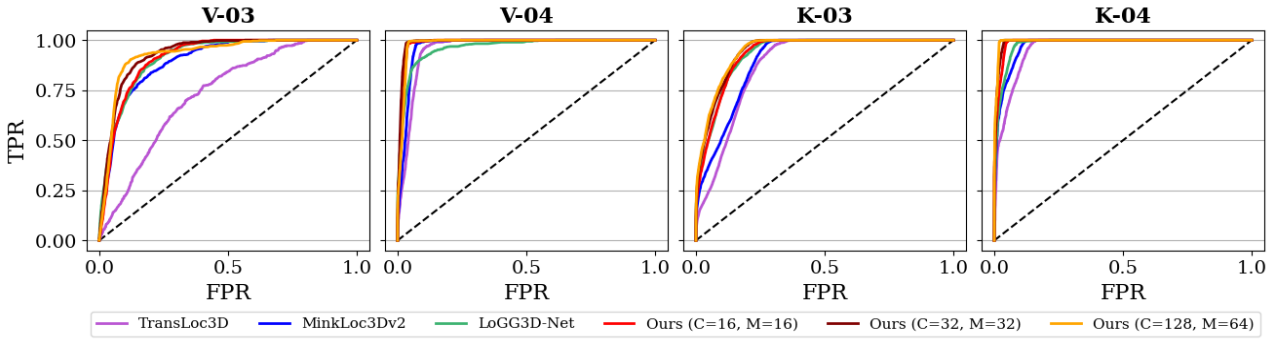

Fig. 4. The ROC curves by place recognition models in the Wild-Places intra-sequence evaluations. False positive rate (FPR) and true positive rate (TPR) were measured with the decision threshold in range $[0, 1]$ metrized in the descriptor space.

Table I demonstrates the benefit of our method against previous first-order poolings. Our method with 8192 dimensions descriptor surpassed recent baselines with advanced backbone architectures, i.e., CASSPR [29] and SelFLoc [30], in every measurement. Meanwhile, our method with 256 dimensions also achieved decent performance gains compared to MinkLoc3Dv2, implying that the benefit of our method does not rely on the large descriptor size and is manifestable even with the limited capacity of a single cluster.

Moreover, we plotted the Mean Average Precision up to top-10 predictions (MAP@K) in Figure 3, since Recall@K only indicates whether a correct retrieval exists among top-K predictions but does not reflect how the descriptor space is metrized overall [47]. Our methods showed similar performance tendency to Recall@K; the one with 8192 dimensions achieved the best, and the one with 256 dimensions also outperformed MinkLoc3Dv2 in the Oxford, U.S., and B.D. It supplements that our pooling method induces well-separated descriptor space than the existing first-order methods.

B. Experiments on Wild-Places

The Wild-Places experiments were comprised of two evaluation protocols: the intra-sequence and the inter-sequence. The intra-sequence evaluates models on the loop-closure detection within a single sequence, using F1-max and recall@1 measured with the distance in the descriptor space and real-world each. The inter-sequence assesses models on the global localization with recall@1 and mean rank reciprocal (MRR). A retrieved scan is considered true when its geographic distance to the query is smaller than 3 m. While most of the baselines were from [27], we also considered ForestLPR [31], one of the state-of-the-art methods that utilizes the preprocessing and bird's-eye-view projection with the prior on the natural environment. Since these baselines have different dimensions of descriptor, e.g., 256 or 1024, our methods with $\sigma = M$ have three variations with respect to the dimension as well: 256, 1024, and 8192.

Table II shows that our method with the 8192 dimensions achieved the best results among the baselines in the most of inter- and intra-sequence tasks. Our method with the 1024 dimensions also surpassed LoGG3D-Net and ForestLPR in the V-04 and K-04 intra-sequences and the Venman inter-sequence, while achieving comparable performances in other evaluations. Even with the smaller 256 dimensions, our method outperformed MinkLoc3Dv2 and demonstrated comparable performance to the 1024 dimensions. These results demonstrate that our method is more effective to model complex features from the unstructured environment than relying on the descriptor size or the environment prior.

Furthermore, the ROC curves in Figure 4, which were measured during the intra-sequence evaluations while altering the decision threshold in the descriptor space, show that our methods formed better curves than other baselines regardless of the descriptor size. They validate that the better metrizability of the overall descriptor space is inducible with our approach in the natural environment as well.

TABLE III
ABLATION STUDIES IN THE OXFORD AND IN-HOUSE DATASETS.
(TOP) ABLATIONS ON THE ZCA WHITENING.
(BOTTOM) ABLATIONS ON THE BACKBONE ARCHITECTURES.

Methods	Oxford		U.S.		R.A.		B.D	
	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%
RBLW ✗, SVDPI ✗	93.25 (-3.38)	97.60 (-1.27)	82.86 (-8.18)	93.23 (-4.13)	70.83 (-16.13)	83.33 (-10.56)	74.72 (-12.83)	82.50 (-9.26)
RBLW ✓, SVDPI ✗	Diverged							
RBLW ✗, SVDPI ✓	95.82 (-0.81)	98.57 (-0.30)	89.18 (-1.86)	95.75 (-1.61)	84.95 (-2.01)	91.47 (-2.42)	85.12 (-2.42)	89.94 (-1.82)
w/ PointNetVLAD [5]	77.82 (+2.94)	88.59 (+0.70)	70.35 (+2.26)	82.67 (+1.30)	66.07 (+2.02)	78.91 (+4.25)	68.19 (+4.77)	74.87 (+4.24)
w/ MinkLoc3D [10]	94.88 (+1.12)	98.26 (+0.35)	88.27 (+2.26)	95.75 (+0.71)	84.47 (+3.36)	91.46 (+0.27)	83.89 (+1.23)	88.78 (+0.30)

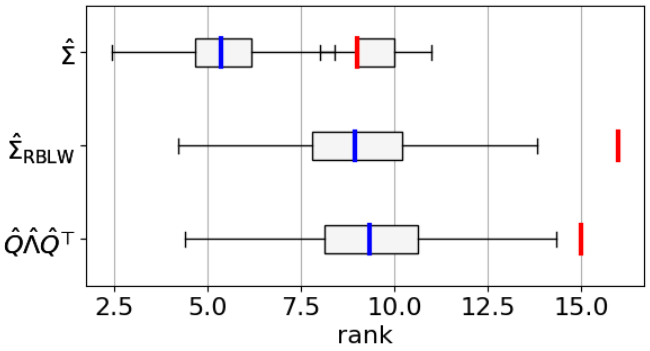

Fig. 5. We measured the matrix rank and effective rank [48] of covariances before the whitening ($\hat{\mathbf{\Sigma}}$), after RBLW shrinkage ($\hat{\mathbf{\Sigma}}_{\text{RBLW}}$), and after the whitening ($\hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^{\top}$) using our default ($C = 16,\ M = 16$) method. Red and blue lines denote the medians of matrix ranks and effective ranks each.

C. Ablation Studies and Analyses of Our ZCA Whitening

We conducted quantitative ablation studies in two aspects: (1) the validity of numerical techniques used in our ZCA whitening, and (2) the backbone-agnostic benefit of our approach. First, we prepared three variants which RBLW is omitted or SVDPI is substituted to a naive eigendecomposition with clipping small eigenvalues. These variants were trained and evaluated using the identical setups to ours (C=16, M=16) in the Oxford. Second, we also replaced the aggregation layers of PointNetVLAD [5] and MinkLoc3D [10] to ours (C=16, M=16). We modified some hyperparameters during their training, e.g., epochs, learning rate, scheduler, and triplet margin, but their backbone architectures and the loss functions remained unchanged.

The results of ablation studies are in Table III with the performance variance compared to our default method (top) and their original previous methods (bottom) from Table I. In the ablation of whitening, every alternative method performed worse than our default whitening, and method with RBLW but without SVDPI even diverged during the training. Furthermore, our method enhanced the performance of backbones from PointNetVLAD and MinkLoc3D by substituting their first-order aggregators, supporting its backbone-agnostic benefit to any place recognition models.

We further analyzed the effect of the numerical techniques in our whitening and the cause of the training failure in Figure 5. It shows the change of matirx rank and effective rank [48] of Voronoi cell covariances during the whitening ($\hat{\mathbf{\Sigma}}$, $\hat{\mathbf{\Sigma}}_{\text{RBLW}}$, $\hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^{\top}$ from Algorithm 1), where the matrix rank was the number of eigenvalues larger than a small threshold ($1\text{e}{-5}$) and the effective rank was calculated from the clipped eigenvalues using the identical threshold.

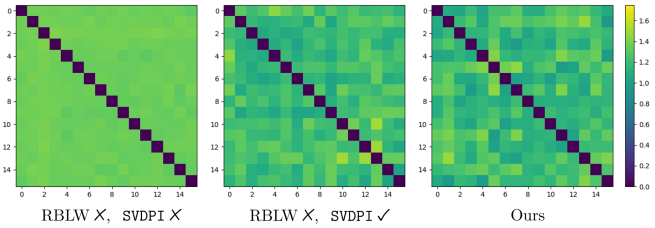

Fig. 6. We measured the empirical Wasserstein-2 distances between the Voronoi cells of descriptors acquired by ablation models in Table III. The brighter color denotes the farther distributional distance between cells.

It reveals that RBLW made $\hat{\mathbf{\Sigma}}_{\text{RBLW}}$ full rank and increased its effective rank as well, implying that $\hat{\mathbf{\Sigma}}_{\text{RBLW}}$ embeds more information than $\hat{\mathbf{\Sigma}}$. However, it also necessitates SVDPI for decomposing the shrunk covariance. As the discrepancy between the matrix and effective ranks was larger in $\hat{\mathbf{\Sigma}}_{\text{RBLW}}$, its repetitive eigenvalues were prone to destabilize the gradient by a naive decomposition than $\hat{\mathbf{\Sigma}}$, whose gradient by low eigenvalues would be discarded during the low-rank clipping.

Moreover, we verified that SVDPI induces more homogeneous Voronoi cell during the training. Figure 6 plots the empirical Wasserstein-2 distances between the Voronoi cells on the descriptors by models from ablation studies, demonstrating that Voronoi cells whitened without SVDPI have larger distributional differences to others. In summary, both RBLW and SVDPI contribute to the enhanced performance of our ZCA whitening, and SVDPI especially plays an essential role in stabilizing the descriptor learning.

VI. CONCLUSION

This paper suggests a novel Voronoi-based second-order pooling method, following the interpretation of connecting NetVLAD and second-order statistics. Our approach computes outer-products of the local descriptors mapped by separate neural networks, followed by normalization of the global descriptor using ZCA whitening. We implement informative and numerically stable ZCA whitening by applying RBLW shrinkage and SVDPI, which enables Mahalanobis metrization across the Voronoi topology. Extensive quantitative evaluations in the Oxford and Wild-Places demonstrated that the proposed method is performant and forms a well-metrizable descriptor space. We also verified its backbone agnostic applicability and the validity of RBLW and SVDPI through numerical analyses. We further expect that our approach will serve as a foundation for future researches in the LPR and VPR as well with its effective metric space regularization and backbone agnostic benefits.

REFERENCES

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *ieee transactions on robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[2] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" *arXiv preprint arXiv:2103.06443*, 2021.

[3] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, vol. 9, pp. 19 516–19 547, 2021.

[4] Y. Zhang, P. Shi, and J. Li, "Lidar-based place recognition for autonomous driving: A survey," *ACM Computing Surveys*, vol. 57, no. 4, pp. 1–36, 2024.

[5] M. A. Uy and G. H. Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4470–4479.

[6] C. Chen, B. Wang, C. X. Lu, N. Trigoni, and A. Markham, "Deep learning for visual localization and mapping: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[7] J. Ruiz-del Solar, P. Loncomilla, and N. Soto, "A survey on deep learning methods for robot vision," *arXiv preprint arXiv:1803.10862*, 2018.

[8] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2024, pp. 17 658–17 668.

[9] S. Rahman and P. Moghadam, "Learning compact channel correlation representation for lidar place recognition," *arXiv preprint arXiv:2409.15919*, 2024.

[10] J. Komorowski, "Minkloc3d: Point cloud based large-scale place recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1790–1799.

[11] J. Komorowski, "Improving point cloud based place recognition with ranking-based loss and large batch training," in *2022 26th international conference on pattern recognition (ICPR)*. IEEE, 2022, pp. 3699–3705.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[13] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[14] Y. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in vision algorithms," in *Proc. International Conference on Machine learning (ICML'10)*, vol. 28, 2010, p. 3.

[15] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *European conference on computer vision*. Springer, 2012, pp. 430–443.

[16] T.-X. Xu, Y.-C. Guo, Z. Li, G. Yu, Y.-K. Lai, and S.-H. Zhang, "Transloc3d: Point cloud based large-scale place recognition using adaptive receptive fields," *arXiv preprint arXiv:2105.11605*, 2021.

[17] K. Vidanapathirana, P. Moghadam, B. Harwood, M. Zhao, S. Sridharan, and C. Fookes, "Locus: Lidar-based place recognition using spatiotemporal higher-order pooling," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5075–5081.

[18] T.-Y. Lin and S. Maji, "Improved bilinear pooling with cnns," *arXiv preprint arXiv:1707.06772*, 2017.

[19] K. Vidanapathirana, M. Ramezani, P. Moghadam, S. Sridharan, and C. Fookes, "Logg3d-net: Locally guided global descriptor learning for 3d place recognition," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2215–2221.

[20] Q. Qiu, S. Zhang, H. Gao, H. Yang, H. Ying, W. Wang, and X. He, "Emvp: Embracing visual foundation model for visual place recognition with centroid-free probing," *Advances in Neural Information Processing Systems*, vol. 37, pp. 120 928–120 950, 2024.

[21] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3304–3311.

[22] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[23] R. Arandjelovic and A. Zisserman, "All about vlad," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585.

[24] A. Chadha and Y. Andreopoulos, "Voronoi-based compact image descriptors: Efficient region-of-interest retrieval with vlad and deep-learning-based descriptors," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1596–1608, 2017.

[25] F. Lu, X. Zhang, C. Ye, S. Dong, L. Zhang, X. Lan, and C. Yuan, "Supervlad: Compact and robust image descriptors for visual place recognition," *Advances in Neural Information Processing Systems*, vol. 37, pp. 5789–5816, 2024.

[26] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *The American Statistician*, vol. 72, no. 4, pp. 309–314, 2018.

[27] J. Knights, K. Vidanapathirana, M. Ramezani, S. Sridharan, C. Fookes, and P. Moghadam, "Wild-places: A large-scale dataset for lidar place recognition in unstructured natural environments," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 11 322–11 328.

[28] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3075–3084.

[29] Y. Xia, M. Gladkova, R. Wang, Q. Li, U. Stilla, J. F. Henriques, and D. Cremers, "Casspr: Cross attention single scan place recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 8461–8472.

[30] Q. Qiu, W. Wang, H. Ying, D. Liang, H. Gao, and X. He, "Selfloc: Selective feature fusion for large-scale point cloud-based place recognition," *Knowledge-Based Systems*, vol. 295, p. 111794, 2024.

[31] Y. Shen, T. Tuna, M. Hutter, C. Cadena, and N. Zheng, "Forestlpr: Lidar place recognition in forests attentioning multiple bev density images," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6659–6669.

[32] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," *Advances in neural information processing systems*, vol. 11, 1998.

[33] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 9, pp. 1704–1716, 2011.

[34] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening," in *European conference on computer vision*. Springer, 2012, pp. 774–787.

[35] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1169–1176.

[36] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[38] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[39] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," *arXiv preprint arXiv:1603.04779*, 2016.

[40] S. Choi, T. Kim, M. Jeong, H. Park, and C. Kim, "Meta batch-instance normalization for generalizable person re-identification," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 3425–3435.

[41] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for mmse covariance estimation," *IEEE transactions on signal processing*, vol. 58, no. 10, pp. 5016–5029, 2010.

[42] C. Ionescu, O. Vantzos, and C. Sminchisescu, "Matrix backpropagation for deep networks with structured layers," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2965–2973.

[43] C. Ionescu, O. Vantzos, and C. Sminchisescu, "Training deep networks with structured layers by matrix backpropagation," *arXiv preprint arXiv:1509.07838*, 2015.

[44] W. Wang, Z. Dang, Y. Hu, P. Fua, and M. Salzmann, "Backpropagation-friendly eigendecomposition," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[45] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[46] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4802–4809.
[47] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," in *European Conference on Computer Vision*. Springer, 2020, pp. 681–699.
[48] O. Roy and M. Vetterli, "The effective rank: A measure of effective dimensionality," in *2007 15th European signal processing conference*. IEEE, 2007, pp. 606–610.