Letter

Loop Closure Detection With Reweighting NetVLAD and Local Motion and Structure Consensus

Kaining Zhang, Jiayi Ma, *Senior Member, IEEE*, and Junjun Jiang, *Senior Member, IEEE*

Dear Editor,

Loop closure detection (LCD) is an important module in simultaneous localization and mapping (SLAM). In this letter, we address the LCD task from the semantic aspect to the geometric one. To this end, a network termed as AttentionNetVLAD which can simultaneously extract global and local features is proposed. It leverages attentive selection for local features, coupling with reweighting the soft assignment in NetVLAD via the attention map for global features. Given a query image, candidate frames are first identified coarsely by retrieving similar global features in the database via hierarchical navigable small world (HNSW). As global features mainly summarize the semantic information of images and lead to compact representation, information about spatial arrangement of visual elements is lost. To provide fine results, we further propose a feature matching method termed as local motion and structure consensus (LMSC) to conduct geometric verification between candidate pairs. It constructs local neighborhood structures of local features through motion consistency and manifold representation, and formulates the matching problem into an optimization model, enabling linearithmic time complexity via a closed-form solution. Experiments on several public datasets demonstrate that LMSC performs well in feature matching, and the proposed LCD system can yield satisfying results.

Related work: LCD aims to lessen the cumulative error of the pose estimation in the SLAM system by identifying reobservations during the navigation [1]. It is achieved by 1) first searching a connection between the current and the historic observations and 2) then regarding the recovered SE(3)/Sim(3) pose as a constraint to optimize the pose graph. In this letter, we focus on the former step in the visual SLAM system. Namely, we mainly solve how to find reliable image pairs to constitute loop-closing pairs.

The first step of LCD is commonly studied as an image retrieval task. But differently, the reference database in LCD is incremental while the size of that in image retrieval is generally fixed for a short period of time. In this step, it is important to determine how to generate a global descriptor for image representation. Notable early global image descriptors are dominated by keypoint detection coupling with aggregation of associated local descriptors, such as BoVW [2], VLAD [3] or ASMK [4]. These approaches rely on a visual dictionary, which can be trained off-line [5], [6], or on-line [7], [8]. Compared with the off-line manner, on-line ones are more scene-agnostic and become increasingly popular during recent years. In these aggregation-based approaches, inverted index [6] or voting [8], [9] technique is commonly exploited to accelerate the searching process. Recently, deep approaches gain increasing popularity in the context of image representation. These CNN models are trained based on ranking triplet [10], [11] or classification [12] losses, which can acquire deep semantic information of images and perform well even with large viewpoint changes. It is demonstrated that LCD approaches built on deep global representation can yield good performance [13]–[16]. In these approaches, HNSW [17] is always selected as a technique for searching acceleration.

After candidate pairs being identified in the first step, a geometric verification process is followed to guarantee precision. It is roughly achieved by first building the putative set based on the similarity between local descriptors and then rejecting false matches (i.e., outliers). Later, the preserved true matches (i.e., inliers) are exploited to recover fundamental or essential matrix. Little LCD literature has discussed this step, and most approaches use RANSAC [18] to achieve it. It assumes the transformation between image pairs is rigid, and the parametric model can be acquired by alternating between sampling and verification. RANSAC is vulnerable to dominated outliers and non-rigid deformation. Numerous non-parametric approaches have been extensively studied to address this issue, ranging from graph matching [19], generalized geometric constraints [20] to locality consistency assumptions [21], [22].

AttentionNetVLAD: Our feature extraction network is a follow-up of DELG [12]. As Fig. 1 shows, it can simultaneously extract global and local features of an image. The local part is the same as DELG, where keypoints are indicated via the attention map and the corresponding descriptors are the output of an auto-encoder. Since the global representation of DELG is trained via a classification loss to indicate which landmark the image belongs to, its distinguish capability is not sufficient for LCD. After all, scenes during a traversal may be similar and it may be regarded as " the same landmark". To address this issue, we incorporate NetVLAD [10], which is tailored for place recognition, within DELG for global representation of an image. Specifically, the dense descriptors $\mathcal{L}$ of size $W \times H \times C_1$ is first reshaped to $N \times C_1$ with $N = W \cdot H$, followed by an L2-normalization operation along the channel dimension. Later, a soft assignment $\mathcal{S} \in \mathbb{R}^{N \times J}$ is generated via a convolutional layer. Here, J indicates the number of cluster centers, and $\mathcal{S}_{n,j}$ serves as the probability of the feature $\mathcal{L}_{n,:}$ being assigned to the cluster j. To capture context information and adjust the impact of a local feature $\mathcal{L}_{n,:}$ at location n to the global representation, we reweight the soft assignment $\mathcal{S}$ with the attention map. We reshape the attention map to an N-dimensional column vector $\boldsymbol{a}$ with $N = W \cdot H$. Formally, for $N \times C_1$-dimensional features $\mathcal{L}$, the VLAD core $f_{VLAD} : \mathbb{R}^{N \times C_1} \rightarrow \mathbb{R}^{J \times C_1}$ is given by

$$f_{\mathrm{VLAD}}(\mathcal{L})(j,:) = \sum_{i=1}^{N} \boldsymbol{a}_i \mathcal{S}_{i,j} \left(\mathcal{L}_{i,:} - \mathbf{c}_j^T \right) \tag{1}$$

where the column vector $\mathbf{c}_j$ with dimension C_1 denotes the j-th cluster center, $\boldsymbol{a}_i$ is the i-th element of $\boldsymbol{a}$. After the aggregation operation, the resulting $J \times C_1$ matrix is first performed with intra (column)-wise normalization, unfolding to a $J \cdot C_1 \times 1$ vector, and then being L2-normalization along with whitening and L2-normalization.

The weights of common parts between our AttentionNetVLAD and DELG are initialized by the official pre-trained model (R50-DELG) and the whole AttentionNetVLAD is trained via knowledge distillation [14]. Specifically, the teacher network is VGG16+ NetVLAD and has been released in [10]. We train the student by minimizing the mean square error loss between its predictions and target global descriptors exported by the teacher on GLDv2 [23]. The number of cluster centers J is set to 64, and an FC layer is introduced additionally to make the dimension of the output of the student network equal to that of the teacher network (4096). Training details involving the input image size, batch size, learning rate, etc. are all referred to [12]. In this way, the new model can inherit the property of the teacher, i.e., being able to capture time-invariant visual clues at a high structural level, meanwhile with the training time being reduced drastically.

Corresponding author: Jiayi Ma.

Citation: K. N. Zhang, J. Y. Ma, and J. J. Jiang, "Loop closure detection with reweighting NetVLAD and local motion and structure consensus," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 6, pp. 1087–1090, Jun. 2022.

K. N. Zhang and J. Y. Ma are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: zkn19961212@whu.edu.cn; jyma2010@gmail.com).

J. J. Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: jiangjunjun@hit.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105635

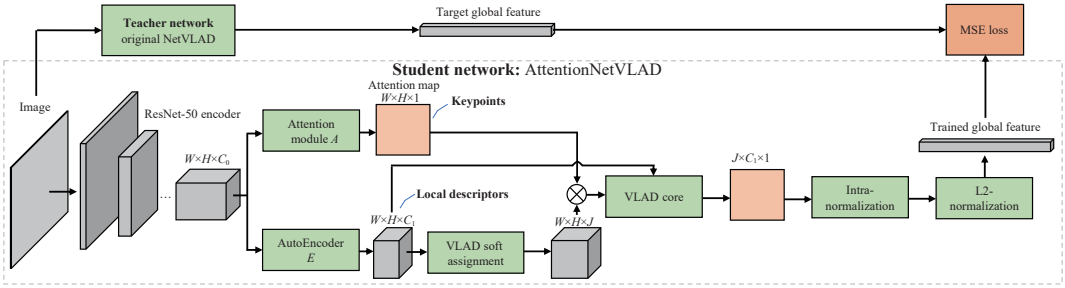
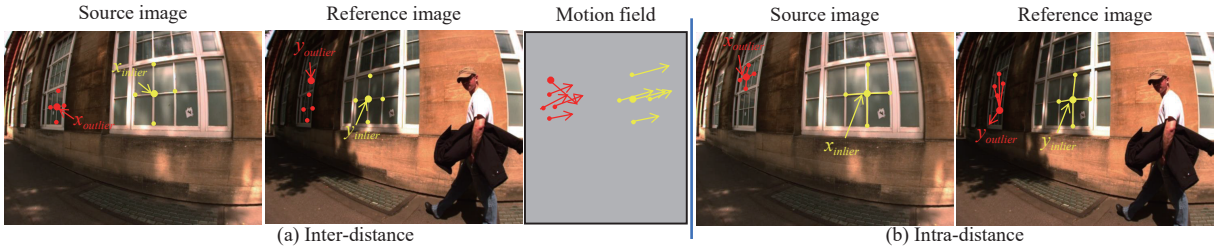


Fig. 1. AttentionNetVLAD and its training process.



(a) Inter-distance (b) Intra-distance

Fig. 2. Illustration of local inter-distance and intra-distance in LMSC. The example is shown in an ideal situation, i.e., no outliers existing in the local neighborhood ($K=4$). In the source image, $\mathbf{x}_{\text{inlier}}$, $\mathbf{x}_{\text{outlier}}$, $\mathcal{N}_{\mathbf{x}_{\text{inlier}}}$ and $\mathcal{N}_{\mathbf{x}_{\text{outlier}}}$ are presented. In the reference image, $\mathbf{y}_{\text{inlier}}$, $\mathbf{y}_{\text{outlier}}$, $C_{\mathbf{x}_{\text{inlier}}}$ and $C_{\mathbf{x}_{\text{outlier}}}$ are presented.

Local motion and structure consensus: The literature argues that local neighborhood structures of keypoints are not vulnerable to non-rigid deformation [21]. Thus we choose to conduct feature matching based on local information.

• Inter-distance based on motion consistency: Suppose a set of putative matches $\mathcal{S}=\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^N$ is established by the similarity between local descriptors, with $\mathcal{X}=\{\mathbf{x}_i\}_{i=1}^N$ and $\mathcal{Y}=\{\mathbf{y}_i\}_{i=1}^N$ being the coordinates of corresponding keypoints from the source and reference image. Meanwhile, we denote $\mathbf{m}_i=\mathbf{y}_i-\mathbf{x}_i$ as the motion vector, depicting the motion trajectory of a pairwise match across two images. If we denote the K-nearest neighbors of $\mathbf{x}_i$ in $\mathcal{X}$ as $\mathcal{N}_{\mathbf{x}_i}=\{\mathbf{x}_i^k\}_{k=1}^K$, with their corresponding motion vectors as $\{\mathbf{m}_i^k\}_{k=1}^K$, it can be concluded in Fig. 2(a) that $\mathbf{m}_i$ tends to be more consistent with $\{\mathbf{m}_i^k\}_{k=1}^K$ in both length and orientation for an inlier, while an outlier does not obey this rule. So, we can measure the motion consistency of $(\mathbf{x}_i,\mathbf{y}_i)$ by

$$d_{\text{inter}}^i=\frac{\max(|\mathbf{m}_i|,|\overline{\mathbf{m}}_i|)}{\min(|\mathbf{m}_i|,|\overline{\mathbf{m}}_i|)}\cdot\arccos\frac{\mathbf{m}_i\cdot\overline{\mathbf{m}}_i}{|\mathbf{m}_i|\cdot|\overline{\mathbf{m}}_i|} \tag{2}$$

where $\overline{\mathbf{m}}_i$ is the mean of $\{\mathbf{m}_i^k\}_{k=1}^K$. It is a weak local constraint and a smaller d_{inter}^i means higher motion consistency.

• Intra-distance based on manifold representation consistency: In our case, the local geometric structure of a sample $\mathbf{x}_i$ ($\mathbf{y}_i$) can be represented by $\mathbf{x}_i$ with $\mathcal{N}_{\mathbf{x}_i}=\{\mathbf{x}_i^k\}_{k=1}^K$ ($\mathbf{y}_i$ with $\mathcal{N}_{\mathbf{y}_i}=\{\mathbf{y}_i^k\}_{k=1}^K$, where $\mathcal{N}_{\mathbf{y}_i}$ is defined as the K-nearest neighbors of $\mathbf{y}_i$ in $\mathcal{Y}$). As Fig. 2(b) shows, for an inlier $(\mathbf{x}_i,\mathbf{y}_i)$, the local geometric structures of $\mathbf{x}_i$ and $\mathbf{y}_i$ tend to be similar, while those of an outlier appear dissimilar. Based on locally linear embedding [24], we reconstruct each data point $\mathbf{x}_i$ by N linear coefficients $\mathbf{W}_{i,:}^{\mathcal{N}_{\mathbf{x}}}$, enforcing $\sum_{j=1}^N\mathbf{W}_{i,j}^{\mathcal{N}_{\mathbf{x}}}=1$ and $\mathbf{W}_{i,j}^{\mathcal{N}_{\mathbf{x}}}=0$ if $\mathbf{x}_j$ does not belong to $\mathcal{N}_{\mathbf{x}_i}$. For simplicity, only non-zero values are preserved thus $\mathbf{x}_i$ is actually reconstructed by $\{\mathbf{W}_{i,k}^{\mathcal{N}_{\mathbf{x}}}\}_{k=1}^K$. Therefore, the local geometry structures of $\mathcal{X}=\{\mathbf{x}_i\}_{i=1}^N$ can be characterized by searching for the least square solution of the $N\times K$ matrix $\mathbf{W}^{\mathcal{N}_{\mathbf{x}}}$

$$\min_{\mathbf{W}^{\mathcal{N}_{\mathbf{x}}}}\sum_{i=1}^N\left\|\mathbf{x}_i-\sum_{k=1}^K\mathbf{W}_{i,k}^{\mathcal{N}_{\mathbf{x}}}\mathbf{x}_i^k\right\|_2^2. \tag{3}$$

According to the one-to-one correspondence M established in $\mathcal{S}=\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^N$, we can map each point $\mathbf{x}_i^k$ in $\mathcal{N}_{\mathbf{x}_i}$ to its corresponding one in $\mathcal{Y}=\{\mathbf{y}_i\}_{i=1}^N$, and denote the set as $C_{\mathbf{x}_i}=\{M(\mathbf{x}_i^k)\}_{k=1}^K$. Similar to (3), $\mathbf{y}_i$ can be reconstructed by $\mathbf{W}^{C_{\mathbf{x}}}$ defined as

$$\min_{\mathbf{W}^{C_{\mathbf{x}}}}\sum_{i=1}^N\left\|\mathbf{y}_i-\sum_{k=1}^K\mathbf{W}_{i,k}^{C_{\mathbf{x}}}M(\mathbf{x}_i^k)\right\|_2^2. \tag{4}$$

The similar definitions for $\mathbf{W}^{\mathcal{N}_{\mathbf{y}}}$ and $\mathbf{W}^{C_{\mathbf{y}}}$ lead to the intra-distance

$$d_{\text{intra}}^i=\left\|\mathbf{W}_{i,:}^{\mathcal{N}_{\mathbf{x}}}-\mathbf{W}_{i,:}^{C_{\mathbf{x}}}\right\|_q^q+\left\|\mathbf{W}_{i,:}^{\mathcal{N}_{\mathbf{y}}}-\mathbf{W}_{i,:}^{C_{\mathbf{y}}}\right\|_q^q. \tag{5}$$

We set $q=1/2$ for better robustness against outliers. This distance represents the similarity of intrinsic geometry between two points. So, it is a stronger local constraint compared with motion consistency.

• Problem formulation. The two above assumptions about the local consistency of inliers are built in an ideal case, i.e., there are no outliers in the K-nearest neighborhood of $\mathbf{x}_i$ and $\mathbf{y}_i$. Thus to construct a relatively clean correspondence set for consistency calculation, an iterative filtering strategy is used [25]. Based on the reliable set $\mathcal{U}$, the cost function is defined as

$$C(\mathcal{I};\mathcal{S},\mathcal{U},\lambda)=\sum_{i\in\mathcal{S}}\left(U_i\cdot d_{\text{intra}}^i+1-U_i\right)+\lambda(N-|\mathcal{I}|) \tag{6}$$

where $\mathcal{I}$ is the unknown inlier set and $\lambda>0$ aims to achieve a trade-off between the first item, i.e., penalizing any match with dissimilar local structure, and the second item, i.e., discouraging outliers. Regarding $U_i\in\{0,1\}$, it equals to 1 only when $d_{\text{inter}}^i<\alpha$ is satisfied, where α is set to the $\lfloor 1.5\cdot|\mathcal{U}|\rfloor$-th value after sorting $\{d_{\text{inter}}^i\}_{i=1}^N$ in the ascending order. In this case, the from-weak-to-strong local constraints are imposed, where samples in $\mathcal{S}$ that do not meet motion consistency would be penalized first.

To provide a closed-form solution for $\mathcal{I}$, we associate each sample $(\mathbf{x}_i,\mathbf{y}_i)$ with a binary value $p_i\in\{0,1\}$ to indicate whether a correspondence is an inlier ($p_i=1$) or an outlier ($p_i=0$). Therefore, if we denote the first item in (6) as $Dist(\mathbf{x}_i,\mathbf{y}_i)$, C in (6) can be rewritten as

$$C(\mathcal{I};\mathcal{S},\mathcal{U},\lambda)=\sum_{i=1}^N p_i(Dist(\mathbf{x}_i,\mathbf{y}_i)-\lambda)+\lambda N. \tag{7}$$

$Dist(\mathbf{x}_i,\mathbf{y}_i)$ can be calculated as long as $\mathcal{S}$ is established, thus the only unknow variable in (7) is p_i. Besides, $Dist(\mathbf{x}_i,\mathbf{y}_i)>\lambda$ would lead to a positive term, resulting in an increase of the cost, and vice versa. In this letter, we set $\lambda=0.17$. To minimize the cost, the optimal solution of $\{p_i\}_{i=1}^N$ can be decided by

$$p_i = \begin{cases} 0, & Dist(\mathbf{x}_i, \mathbf{y}_i) > \lambda, \\ 1, & Dist(\mathbf{x}_i, \mathbf{y}_i) \le \lambda, \end{cases} \quad i = 1, \ldots, N. \tag{8}$$

As such, the inlier set $\mathcal{I}$ is composed of putative matches with $p_i = 1$. The time and space complexity of LMSC is about $O(N \log N)$ and $O(N)$. Namely, LMSC has linearithmic time complexity and linear space complexity in regard to the scale of the given putative set.

Loop closure detection pipeline: When a query image I_i comes to the pipeline, we first extract its global and $\nu_{\max} = 500$ local features simultaneously via AttentionNetVLAD. Then, the global feature is used to retrieve the nearest neighbor image I_j in the reference database under the cosine distance. Specifically, the incremental database built by HNSW would ignore $\xi = f \cdot T$ neighboring images obtained before the query. Here, $T = 10$ is the consuming time in non-redundance path, while f is the frame rate shown in Table 1. The parameters in HNSW are referred to [13]. Only when the distance between two images $d(I_i, I_j)$ is greater than $\sigma_1 = 0.2$ would them be transferred into the second phase, i.e., geometric verification via LMSC. Later, if the cardinality of the inlier set $\mathcal{I}^*$ preserved by LMSC is sufficient enough, (i.e., $|\mathcal{I}^*| > \sigma_2$), the image pair is most likely to represent the same scene. Finally, the image pair is checked by a temporal constraint [26] to filter false positives that can not persist over time. Since global features involve semantic information of scenes while local features coupling with feature matching indicate geometric relationship between scenes, our LCD pipeline can be deemed semantic-to-geometric.

Table 1. Dataset Information. "# Images" Means the Number of Images

Dataset		# Images	Image resolution	Frame rate
KITTI	Sequence 00 (K00)	4541	1241 × 376	10
	Sequence 02 (K02)	4661		
St. Lucia	100909 14:10 (St1410)	20894	640 × 480	15
	180809 15:45 (St1545)	21434		
New College (NC)		52480	512 × 384	20

Experimental setup: We implement AttentionNetVLAD with TensorFlow, and run on an Intel(R) Core(TM) i9-9920X CPU @3.50 GHZ machine with three TITAN RTX GPUs. The information of six sequences selected for evaluation is presented in Table 1. The frame rates of St1410 and St1545 are downsampled to 3 Hz, meanwhile the right measurements of NC with the frame rate of 1 Hz are adopted. Ground truth (GT) which is present in the form of binary matrices and preserves the image-wise correspondence of datasets is provided by An *et al.* [13].

Results on feature matching: We select 33 loop-closing pairs according to GT from the datasets shown in Table 1 to evaluate LMSC. The putative set of each image pair is established based on SIFT [27], followed by GT generation through manual check of each putative match, which results in the average number of putative matches and the inlier ratio are 345 and 51.28%, respectively. As Fig. 3 shows, the runtime of LMSC scales up with the increase of K due to the linearithmic time complexity and F-score = (2·Precision·Recall)/(Precision + Recall) in the left plot is applied for comprehensive evaluation in feature matching. On the whole, $K = 12$ outperforms other cases with a relatively low time cost, thus we choose it for subsequent experiments.

We report the quantitative performance of LMSC in Fig. 4, involving six feature matching methods (i.e., RANSAC [18], GS [19], GMS [22], ICF [28], RFM-SCAN [29] and LPM [21]) for comparison. It can be seen that compared with Precision, LMSC is better at Recall, and outperforms other methods in F-score with relatively high efficiency.

Results on loop closure detection: We adopt maximum recall rate at 100% precision (MR) to evaluate the performance of LCD. Firstly, we perform it based on DELG, NetVLAD (teacher) and AttentionNetVLAD (student) respectively, along with LMSC for geometric check. Results are shown in Table 2. DELG is tailored for the instance recognition task, thus the performance of it degenerates with the similar structures occured in LCD scenes. NetVLAD is of low efficiency due to the requirement of additional local feature extraction. Its coupling with AL outperforms that with SURF may be caused by the fact that SURF cannot ignore redundant information in scenes [26], which reduces the accuracy of feature matching. Through training with knowledge distillation, our AttentionNetVLAD can at the same time learn prior knowledge of perceptual changes in NetVLAD, and be efficient enough for real-time requirements.

Secondly, on the basis of AttentionNetVLAD, different feature matching methods involving RANSAC, LPM and LMSC are embedded into our LCD pipeline, and the results are shown in Table 3. Records of runtime in the table exclude the process of putative set construction. Obviously, the pipeline with LMSC yields better performance on MR than RANSAC and LPM. This is because RANSAC performs poor when resampling in putative sets with low inlier ratio, while LPM cannot separate inliers from relatively low-precision noisy matches due to its weak local geometric constraint. When the process of geometric verification is considered, the average

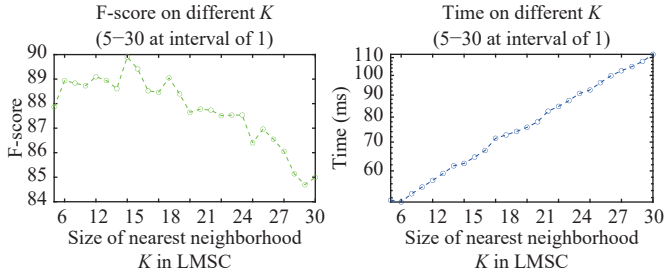


Fig. 3. The choice of the optimal K.

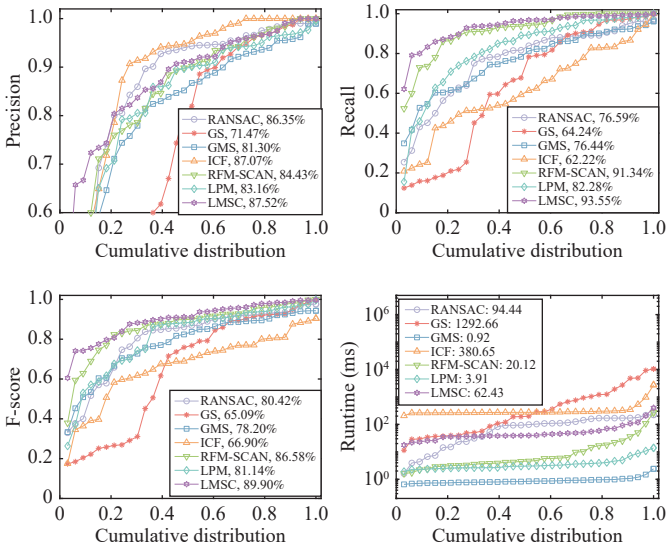


Fig. 4. Quantitative results of LMSC. From left to right: Precision, Recall, F-score and Runtime (ms) with respect to the cumulative distribution. The coordinate (x, y) on the curves means that there are $100 * x$ percent of image pairs which have precisions, recalls or runtime no more than y.

Table 2. LCD Results on Different Feature Extraction Approaches. "AL" Means the Local Part in AattentionNetVLAD

Dataset	DELG	NetVLAD+ SURF	NetVLAD+ AL	Attention-NetVLAD
	MR (%)/Runtime for feature extraction (ms)			
K00	92.51/38.7	92.77/93.0	94.16/109.5	94.29/45.1
K02	79.57/38.4	74.40/95.9	81.50/114.3	80.37/48.7
K05	86.99/38.5	91.33/83.3	88.43/101.8	91.57/39.5
St1410	72.00/34.6	74.43/75.4	79.71/89.2	80.39/35.0
St1545	85.31/34.7	79.55/75.4	86.78/89.8	88.47/34.7
NC	87.47/25.1	83.39/52.7	88.89/67.3	89.05/26.4

runtime of our system on the dataset with the maximum image resolution (K00) is about 116.14 ms/frame.

Table 3. LCD Results on Different Feature Matching Approaches

Approach	K00	K02	K05	St1410	St1545	NC
	MR (%)/Runtime for feature matching (ms)					
RANSAC	91.75/87.4	74.77/93.6	87.71/88.2	75.41/73.5	79.17/79.5	55.24/76.4
LPM	93.53/3.6	78.26/3.6	85.78/3.5	79.87/3.6	84.85/3.7	86.90/3.3
LMSC	94.29/58.2	80.37/61.0	91.57/63.4	80.39/58.8	88.47/59.0	89.05/67.0

Finally, we report the comparative results of our method with state-of-the-art LCD approaches in Table 4 . The results of ESA-VLAD [14] and Zhang *et al.* are cited from [26], while others are cited from [13]. Overall, our pipeline has satisfying performance on all datasets. Albeit our results are marginally worse than Zhang *et al.* [26], the runtime of our pipeline is about one time faster than them. Meanwhile, they need a pretrained visual dictionary to conduct candidate frame selection, while ours is completely operated online.

Table 4. Comparative Results

Approach	K00	K02	K05	St1410	St1545	NC
DLoopDetector [6]	72.43	68.22	51.97	14.87	31.36	47.56
Kazmi and Mertsching [30]	90.39	79.49	81.41	58.10	72.55	51.09
Tsintotas *et al*. [8]	93.18	76.01	94.20	9.77	15.07	52.44
FILD++ [13]	94.92	73.52	95.42	66.41	81.36	82.37
ESA-VLAD [14]	97.46	79.94	−	79.59	88.47	91.02
Zhang *et al*. [26]	94.29	81.00	91.81	82.88	88.59	−
Proposed	94.29	80.37	91.57	80.39	88.47	89.05

Conclusions: In this work, we conduct LCD in a semantic-to-geometric, coarse-to-fine manner. We first propose Attention-NetVLAD to achieve global and local feature extraction simultaneously. The global feature is used to perform candidate frame selection via HNSW, while the local one is exploited for geometric verification via LMSC. LMSC is the proposed feature matching method, which is able to identify reliable matches efficiently by imposing from-weak-to-strong local geometric constraints. Based on the above two components, the whole LCD system is at the same time high-performance and efficient compared with state-of-the-art approaches.

Acknowledgments: This work was supported by Key Research and Development Program of Hubei Province (2020BAB113), and the Natural Science Fund of Hubei Province (2019CFA037).

References

[1] W. Huang, G. Zhang, and X. Han, "Dense mapping from an accurate tracking slam," *IEEE/CAA J. Autom. Sinica* , vol. 7, no. 6, pp. 1565–1574, 2020.

[2] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, 2003, pp. 1470–1470.

[3] H. Jégou, M. Douze, C. Schmid, and P. Pérez, " Aggregating local descriptors into a compact image representation," in *Proc. CVPR*, 2010, pp. 3304–3311.

[4] G. Tolias, Y. Avrithis, and H. Jégou, " Image search with selective match kernels: Aggregation across single and multiple images," *Int. J. Comput. Vis.*, vol. 116, no. 3, pp. 247–261, 2016.

[5] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *Int. J. Rob. Res.* , vol. 27, no. 6, pp. 647–665, 2008.

[6] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.* , vol. 28, no. 5, pp. 1188–1197, 2012.

[7] E. Garcia-Fidalgo and A. Ortiz, " iBoW-LCD: An appearance-based loop-closure detection approach using incremental bags of binary words," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3051–3057, 2018.

[8] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "Assigning visual words to places for loop closure detection," in *Proc. ICRA*, 2018, pp. 1–7.

[9] D. Liu, Y. Cui, X. Guo, W. Ding, B. Yang, and Y. Chen, "Visual localization for autonomous driving: Mapping the accurate location in the city maze," in *Proc. ICPR*, 2021, pp. 3170–3177.

[10] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. CVPR*, 2016, pp. 5297–5307.

[11] D. Liu, Y. Cui, L. Yan, C. Mousas, B. Yang, and Y. Chen, "DenserNet: Weakly supervised visual localization using multi-scale feature aggregation," in *Proc. AAAI*, 2021, pp. 6101–6109.

[12] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Proc. ECCV*, 2020, pp. 726–743.

[13] S. An, H. Zhu, D. Wei, K. A. Tsintotas, and A. Gasteratos, " Fast and incremental loop closure detection with deep features and proximity graphs," *J. Field Robot.*, 2022. DOI: DOI: 10.1002/rob.22060

[14] Y. Xu, J. Huang, J. Wang, Y. Wang, H. Qin, and K. Nan, "ESA-VLAD: A lightweight network based on second-order attention and NetVLAD for loop closure detection," *IEEE Robot. Autom. Lett.* , vol. 6, no. 4, pp. 6545–6552, 2021.

[15] H. Wang, W. Wang, X. Zhu, J. Dai, and L. Wang, "Collaborative visual navigation," arXiv preprint arXiv: 2107.01151, 2021.

[16] H. Wang, W. Wang, T. Shu, W. Liang, and J. Shen, " Active visual information gathering for vision-language navigation," in *Proc. ECCV*, 2020, pp. 307–322.

[17] Y. A. Malkov and D. A. Yashunin, " Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, 2018.

[18] M. A. Fischler and R. C. Bolles, " Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM* , vol. 24, no. 6, pp. 381–395, 1981.

[19] H. Liu and S. Yan, " Common visual pattern discovery via spatially coherent correspondences," in *Proc. CVPR*, 2010, pp. 1609–1616.

[20] C. Leng, H. Zhang, G. Cai, Z. Chen, and A. Basu, " Total variation constrained non-negative matrix factorization for medical image registration," *IEEE/CAA J. Autom. Sinica* , vol. 8, no. 5, pp. 1025–1037, 2021.

[21] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, " Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, no. 5, pp. 512–531, 2019.

[22] J.-W. Bian, W.-Y. Lin, Y. Liu, L. Zhang, S.-K. Yeung, M.-M. Cheng, and I. Reid, " GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1580–1593, 2020.

[23] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2–a large-scale benchmark for instance-level recognition and retrieval," in *Proc. CVPR*, 2020, pp. 2575–2584.

[24] S. T. Roweis and L. K. Saul, " Nonlinear dimensionality reduction by locally linear embedding," *Science* , vol. 290, no. 5500, pp. 2323–2326, 2000.

[25] K. Zhang, Z. Li, and J. Ma, " Appearance-based loop closure detection via bidirectional manifold representation consensus," in *Proc. ICRA*, 2021, pp. 6811–6817.

[26] K. Zhang, X. Jiang, and J. Ma, " Appearance-based loop closure detection via locality-driven accurate motion field learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2350–2365, 2021.

[27] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[28] X. Li and Z. Hu, " Rejecting mismatches by correspondence function," *Int. J. Comput. Vis.*, vol. 89, no. 1, pp. 1–17, 2010.

[29] X. Jiang, J. Ma, J. Jiang, and X. Guo, " Robust feature matching using spatial clustering with heavy outliers," *IEEE Trans. Image Process.*, vol. 29, pp. 736–746, 2020.

[30] S. A. M. Kazmi and B. Mertsching, " Detecting the expectancy of a place using nearby context for appearance-based mapping," *IEEE Trans. Robot.*, vol. 35, no. 6, pp. 1352–1366, 2019.