2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Semantic Reinforced Attention Learning for Visual Place Recognition

Guohao Peng[1], Yufeng Yue[2], Jun Zhang[1], Zhenyu Wu[1], Xiaoyu Tang[1] and Danwei Wang[1], *Fellow, IEEE*

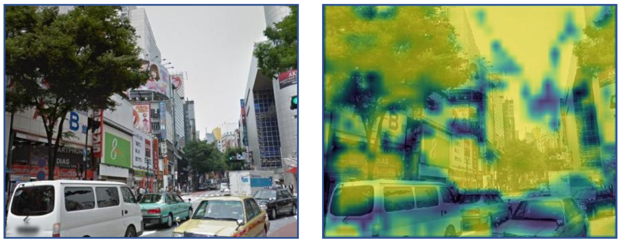
***Abstract*— Large-scale visual place recognition (VPR) is inherently challenging because not all visual cues in the image are beneficial to the task. In order to highlight the task-relevant visual cues in the feature embedding, the existing attention mechanisms are either based on artificial rules or trained in a thorough data-driven manner. To fill the gap between the two types, we propose a novel Semantic Reinforced Attention Learning Network (SRALNet), in which the inferred attention can benefit from both semantic priors and data-driven fine-tuning. The contribution lies in two-folds. (1) To suppress misleading local features, an interpretable local weighting scheme is proposed based on hierarchical feature distribution. (2) By exploiting the interpretability of the local weighting scheme, a semantic constrained initialization is proposed so that the local attention can be reinforced by semantic priors. Experiments demonstrate that our method outperforms state-of-the-art techniques on city-scale VPR benchmark datasets.**

I. Introduction

Visual place recognition (VPR) has been a crucial research field in computer vision [1]–[6] and robotics [7]–[11] communities, since it is the cornerstone of many popular applications including autonomous driving [12]–[14], geo-localization [15]–[18] and 3D reconstruction [19].

Typically, VPR is solved as an image retrieval task [1], [2], [6], [16], [20], [21], where the most similar reference images are retrieved when given a query image. City-scale VPR has always been challenging, because even the same scene may undergo great appearance changes due to different weather, illumination, and viewpoints. Partial occlusion and dynamic objects will also increase the task difficulty. Therefore, how to form a robust image representation has become the focus of research in the field.

Among all attempts to construct compact and powerful image representations in the past decades, aggregation-based methods have proven to be fruitful. Typical representatives range from Fisher Vector (FV) [22], Vector of Locally Aggregated Descriptors (VLAD) [23] to recent Convnet architectures that introduce multiple pooling strategies [24]–[28]. However, not all visual cues in the image are related to the task. Early methods quantify all local features indiscriminately, which may result in misleading information being encoded into the image representation.



Original image

Inferred attention

Fig. 1. The superimposed heat map illustrates which visual cues our model learns to suppress in feature embedding. As can be seen, the inferred attention is consistent with the human cognition that values long-term static structures and ignores unreliable objects. Going beyond the rough prior knowledge of "preserving building semantics", our model learns to adaptively retain billboards and suppress repeated structures on buildings.

To address this problem, attention mechanism is introduced to emphasize the task-relevant local features [2], [5], [29]. Recent attention-aware methods can be categorized as either data-driven or rule-based. The data-driven methods [20], [30], [31] integrate attention modules into an end-to-end encoding network for unified learning. However, they simply employ attention mechanism as the black box weighting of local features, which lacks the interpretability to reflect priors. The rule-based methods typically use semantic information to filter specific visual cues [32]–[34], while their performance is limited by prior knowledge and the generalization ability of the semantic segmentation algorithm. To narrow the gap between the two types, it's necessary to set up an interpretable attention module that can elegantly combine prior knowledge and data-driven learning, so that the attention can benefit from their complementary.

With this motivation, we propose an end-to-end architecture that integrates attention learned from both semantic priors and data-driven training. Specifically, the model incorporates an interpretable local weighting scheme which is constructed based on hierarchical feature distribution. Its intrinsic relationship with encoding space partition enables initial attention to be provided by prior knowledge. On this basis, a semantic constrained initialization is proposed, which equivalently provides better initial attention for the local weighting scheme. Through further fine-tuning, the ultimate local attention can benefit from the mutual promotion between semantic priors and data-driven learning. In accordance with the above statement, our contributions can be elaborated as follows:

- A novel attentional encoding architecture SRALNet is proposed for city-scale VPR, which incorporates comprehensive attention into feature embedding.
- An interpretable local weighting scheme(LW) is proposed to refine local features. In the field of VPR, to our best knowledge, this is the first attention mechanism that

[1]G. Peng, J. Zhang, Z. Wu, X. Tang and D. Wang are with School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore (email: peng0086@ntu.edu.sg)

[2]Y. Yue is with the School of Automation, Beijing Institute of Technology, Beijing 100081, China

978-1-7281-9077-8/21/$31.00 ©2021 IEEE

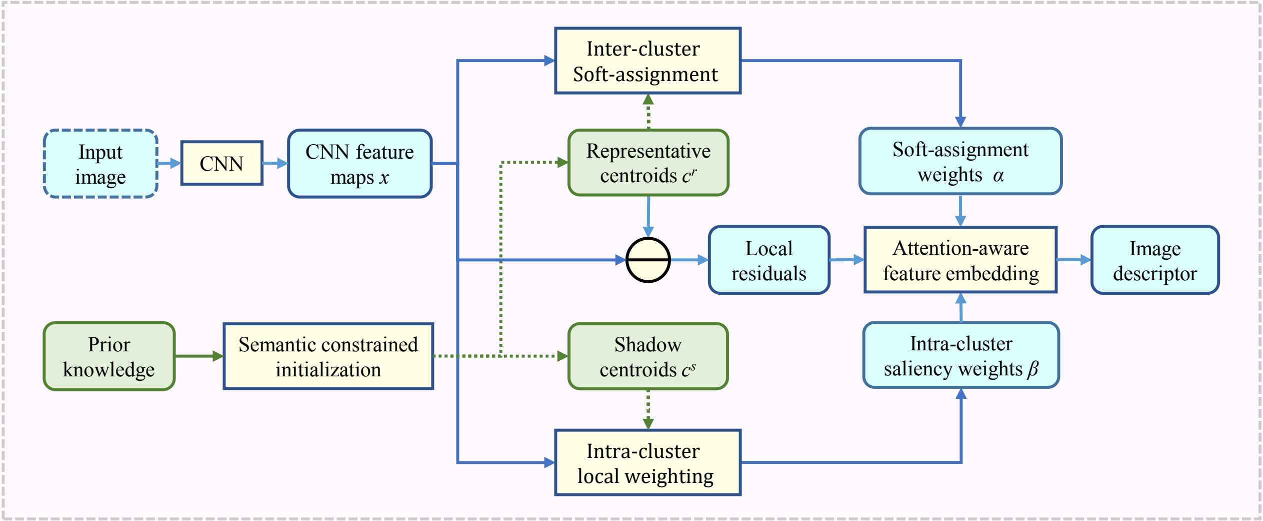

Input image
CNN
CNN feature maps x
Inter-cluster Soft-assignment
Representative centroids c^r
Soft-assignment weights α
Local residuals
Attention-aware feature embedding
Image descriptor
Prior knowledge
Semantic constrained initialization
Shadow centroids c^s
Intra-cluster saliency weights β
Intra-cluster local weighting

Fig. 2. The overall flowchart of the SRALNet encoding pipeline. The deep local features are clustered, refined, and encoded into the final image descriptor (the blue solid arrow). Semantic priors are introduced to enhance the hierarchical weighting module (the green dotted arrow).

can integratge semantic priors and data-driven learning.

- A semantic constrained initialization(SC) is proposed to reinforce the local weighting scheme. It paves the way to reflect semantic priors with initial weights.
- Experiments are conducted to verify the effectiveness of the proposed components, and our method outperforms SOTA techniques on all benchmark datasets.

II. Related Work

Early methodologies for VPR count on hand-crafted local features [35], [36] coupled with variant BoW [4]–[6] models. Later on, VLAD [23] and FV [22] have emerged as powerful alternatives. With the popularity of deep learning, Max pooling [37], sum pooling [23], VLAD [38] and FV [39] all show convincing advantages when encoding deep local features into compact image representations. Studies have also demonstrated that incorporating spatial information of pyramid patches [40] or selective regions [21], [41] can boost performance. A typical case is R-MAC [27] that aggregates the max pooled activations of multi-scale grids.

The recent researches [1], [42], [43] have shown that network fine-tuning on a task-specific dataset can significantly improve the performance. Arandjelovic *et al.* [1] propose a generalized VLAD pooling layer named NetVLAD, which is differentiable for end-to-end training. Concurrently, Gordo *et al.* [42] fine-tune the R-MAC [27] to fit for a large-scale landmark dataset. Yu *et al.* [40] propose SPENetVLAD, which enhances NetVLAD by encoding the spatial information in the stacked regional features. These methods recruit all the local features in feature embedding, which may result in misleading visual cues degrading the image representation.

To selectively embed task-relevant local features, Kim *et al.* [20] extend the NetVLAD [1] with a contextual weighting network(CRN), which utilizes the information from features' context. Zhu *et al.* [30] propose APAnet that integrates a cascaded attention scheme before pyramid aggregation of local features. More recently, high-level prior knowledge has proven conducible for scene classification and place recognition. In [32]–[34], semantic information is employed as supervision to harvest local features from man-made discriminative visual objects for the subsequent embedding. Noticeably, these attention modules are either artificial rule-based or harnessed with black-box weighting. In this paper, we manage to combine prior knowledge and data-driven learning in an interpretable local weighting scheme.

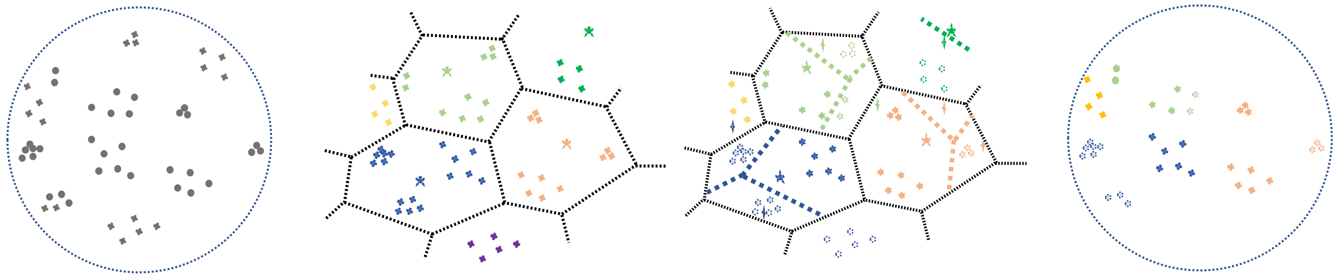
Outside the field of VPR, the most related work is GhostVLAD [44] proposed for face recognition. With the same architecture as NetVLAD, it specifies the clusters for characterizing vague faces and excludes them from feature embedding. Adapting GhostVLAD to the VPR tasks requires specific initialization. It is also imperfect to exclude or retain an entire cluster from the embedding, since local features from the same cluster cannot be all misleading or informative. SRALNet decouples clustering and filtering through a hierarchical weighting scheme, where more flexible refinement is performed within each cluster.

III. Semantic Reinforced Attention Learning

In order to highlight the divergent importance of visual cues to the task, we propose an attention-aware encoding architecture named SRALNet. Fig.2 presents the overall flowchart of the encoding pipeline. The following subsections will describe how we introduce local attention to suppress the misleading visual cues in the image representation.

A. Deep Local Feature Clustering

Following the common local feature representation [1], [20], [30], we exploit a deep convolutional neuron network VGG-16 [45] as the backbone for local feature extraction. Spatial activations $m \in R^{D\times1\times1}$ decomposed from the feature maps of the last convolutional layer $M \in R^{D\times H\times W}$ are treated



(a) Deep local features

(b) Inter-cluster soft-assignment

(c) Intra-cluster local weighting

(d) Cluster-wise refined local features

Fig. 3. The illustration of the hierarchical weighting within SRALNet. (a)∼(c) show how deep local features are clustered and refined through the inter-cluster soft-assignment and intra-cluster local weighting. (d) visualizes the local features that are reserved for the subsequent feature embedding after the hierarchical weighting.

as deep local features. Illustrated as Fig.3.a, the normalized local features $x \in R^D$ are scattered on the unit hypersphere.

After local feature extraction, we then adopt the soft-assignment [1] to divide the local features into K visual word clusters (Fig.3.b). Defined as in Eq.(1), the soft-assignment weight $\alpha_k(\mathbf{x_i})$ indicates the probability of a local feature $\mathbf{x}_i$ belonging to the k^{th} cluster. $\mathbf{c}_k^r$ denotes the representative centroid ($\star$ in Fig.3.b) of the k^{th} cluster. The constant a controls the decay of the response with the magnitude of the distance, which is initialized by a large positive number.

$$\alpha_k(x_i) = \frac{e^{-a\|x_i - c_k^r\|^2}}{\sum_{j=1}^{K} e^{-a\|x_i - c_j^r\|^2}} \tag{1}$$

B. Intra-cluster Local Weighting

To further suppress task-irrelevant features in each cluster, we propose a local weighting scheme based on the intra-cluster feature distribution. As visual cues that describe similar semantics and appearance usually have consistent task relevance and are mapped to adjacent locations in the feature space, we hypothesize the Voronoi cell of a cluster can be separated into an informative area and multiple ambiguous areas (Fig.3.c). Each ambiguous area is represented by a shadow centroid $\mathbf{c}_{kl}^s$ ($+$ in Fig.3.c). The intra-cluster saliency weight $\beta_k(\mathbf{x}_i)$ is defined as the probability of a local feature $\mathbf{x}_i$ from the k^{th} cluster being located in the informative area I. Assuming that sub-clusters are uniformly distributed and each one conforms to a Gaussian with equal covariance matrix, $\beta_k(\mathbf{x}_i)$ can be derived through the Bayesian theorem as in Eq.(2):

$$\begin{aligned}\beta_k(\mathbf{x}_i) &= P(I|x_i, c_k) \\ &= \frac{P(x_i|I,c_k)P(I|c_k)}{\sum_{l=1}^{S} P(x_i|s_l,c_k)P(s_l|c_k) + P(x_i|I,c_k)P(I|c_k)} \\ &= \frac{e^{-a\|\mathbf{x}_i - \mathbf{c}_k^r\|^2}}{\sum_{l=1}^{S} e^{-a\|\mathbf{x}_i - \mathbf{c}_{kl}^s\|^2} + e^{-a\|\mathbf{x}_i - \mathbf{c}_k^r\|^2}}\end{aligned} \tag{2}$$

According to Eq.(2), local features located in ambiguous areas will be assigned with a low saliency weight. Although a total of S shadow centroids are initialized for each cluster, some of them could be cast away from the cluster boundary after optimization and become ineffective. This provides a flexible internal partition for each cluster. Essentially, the local weighting acts as the re-allocation of local features into sub-clusters where informative area is highlighted and ambiguous ones are suppressed. This interpretability makes it possible to introduce prior attention by initializing the encoding space partition, which is intrinsically determined by the centroids $\mathbf{c}_k^r$ and $\mathbf{c}_k^s$ according to Eq.(1) and Eq.(2). This characteristic will be further leveraged in Section.III-E.

C. Unified Implementation of The Hierarchical Weighting

It can be noticed that by expanding the square terms, both Eq.(1) and Eq.(2) can be simplified by canceling out $e^{-a\|\mathbf{x}_i\|^2}$ between the numerator and the denominator. Additionally, by using the abbreviated symbol $\mathbf{c}_{kn}$ to represent $\mathbf{c}_k^r$ and $\{\mathbf{c}_{kl}^s\}$ in the k^{th} cluster, where n=0 denotes the representative centroid and the others denote the shadow ones, $\alpha_k(\mathbf{x}_i)$ and $\beta_k(\mathbf{x}_i)$ can be further derived as in Eq.(3) and Eq.(4). Since both transforms take local features as inputs, they can be implemented through a unified convolutional layer, followed by the Softmax function across the specified channels.

$$\alpha_k(\mathbf{x_i}) = \frac{e^{2a\mathbf{c}_{k0}^{\mathrm{T}}\mathbf{x_i} - a\|\mathbf{c}_{k0}\|^2}}{\sum_{j=1}^{K} e^{2a\mathbf{c}_{j0}^{\mathrm{T}}\mathbf{x_i} - a\|\mathbf{c}_{j0}\|^2}} = \frac{e^{\mathbf{w}_{k0}^{\mathrm{T}}\mathbf{x}_i + b_{k0}}}{\sum_{j=1}^{K} e^{\mathbf{w}_{j0}^{\mathrm{T}}\mathbf{x}_i + b_{j0}}} \tag{3}$$

$$\beta_k(\mathbf{x}_i) = \frac{e^{2a\mathbf{c}_{k0}^{\mathrm{T}}\mathbf{x}_i - a\|\mathbf{c}_{k0}\|^2}}{\sum_{l=0}^{S} e^{2a\mathbf{c}_{kl}^{\mathrm{T}}\mathbf{x}_i - a\|\mathbf{c}_{kl}\|^2}} = \frac{e^{\mathbf{w}_{k0}^{\mathrm{T}}\mathbf{x}_i + b_{k0}}}{\sum_{l=0}^{S} e^{\mathbf{w}_{kl}^{\mathrm{T}}\mathbf{x}_i + b_{kl}}} \tag{4}$$

D. Attention-aware Image Representation

After the division and refinement of local features, the k^{th} visual word vector V_k can be calculated as a spatial aggregation of the double-weighted local residuals.

$$\mathbf{V_k} = \sum_{i=1}^{HW} \alpha_k(\mathbf{x}_i)\beta_k(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{c}_k^r) \tag{5}$$

Intuitively, $\beta(\mathbf{x}_i)$ scales down the residual norm of misleading local features in the ambiguous areas, thereby suppressing them in the feature embedding.

As illustrated in Fig.4, the final image descriptor is the concatenation of K visual word vectors followed by intra-normalization [46] and L_2-normalization. To achieve the most discriminative representation, the optimization of trainable parameters $\{\mathbf{W}\}$, $\{b\}$ and $\{\mathbf{c}_k^r\}$ in Eq.(3)∼(5) is intrinsically equivalent to explore the optimal hierarchical encoding space partition and feature allocation.

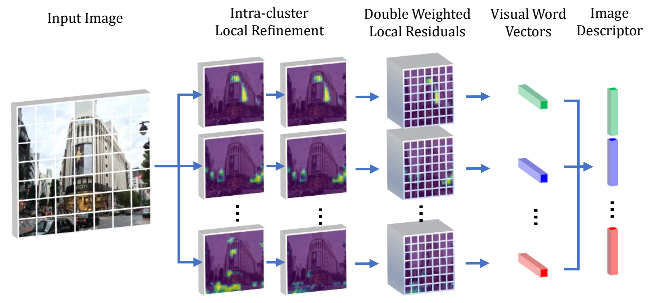

Input Image
Intra-cluster Local Refinement
Double Weighted Local Residuals
Visual Word Vectors
Image Descriptor

Fig. 4. The diagram of the attention-aware descriptor embedding.

E. Semantic Constrained Initialization

According to Eq.(3)∼(5), the initial parametric model is determined by the representative centroids and shadow centroids. Thus a normal initialization could be achieved by mimicking traditional VLAD as [1] does: the representative centroids $\{\mathbf{c}_k^r\}$ are initialized by the clustering centers of the sampled deep local features, while shadow centroids $\{\mathbf{c}_{kl}^s\}$ are randomly initialized to be distant from their corresponding $\mathbf{c}_k^r$. However, not all sampled local features are task-relevant. Consequently, some of the representative centroids may be initialized by misleading visual cues. Therefore, we introduce semantics as the prior constraints to select specific local features for initializing $\{\mathbf{c}_k^r\}$ and $\{\mathbf{c}_{kl}^s\}$ respectively.

Specifically, we adopt the DeepLabV3 [47] pre-trained on Cityscapes dataset [48] to provide common semantic classes under urban driving scenes. The activations before Softmax prediction are first scaled to the same size of our feature map through max pooling. Then Softmax is implemented to predict the labels of local features. Features labeled as static objects, including 'building','road','traffic signs' and 'vegetation', are filtered and sampled for generating K representative centroids. While those dynamic or task-irrelevant semantics, such as 'sky','person' and 'vehicle', are used for generating N shadow candidates. For each cluster, the S shadow centroids are initialized by the top S candidates that have the closest Euclidean distances with the representative centroid.

The semantic constrained initialization essentially partitions the encoding space based on semantic priors, which equivalently provides better initial attention for the local weighting scheme. On this basis, we allow the network to fit the optimal attention through end-to-end training. Thereby, the ultimate local attention can benefit from the mutual promotion between semantic priors and data-driven learning. As illustrated in Fig.1 and Fig.5, although no precise pixel-level semantic annotations are required for supervision, the learned attention still turns out to be largely consistent with human cognition that inhibits task-irrelevant semantics.

IV. Training Pipeline

Let f_θ be the image representation and θ be the parameters to be trained. To make a positive reference I_r^p closer to the query I_q than any negative candidate I_r^n in feature space, we adopt the triplet ranking loss [43], [49]–[52] as in Eq.(6) for metric learning.

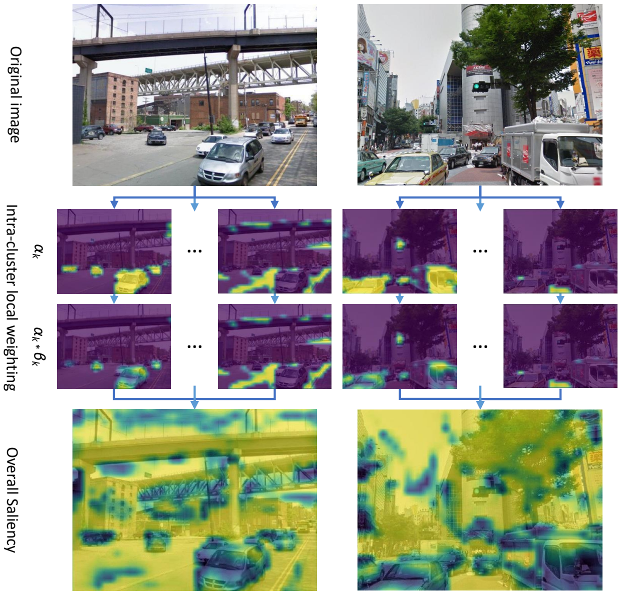


Fig. 5. Attention inference of two examples from Pittsburgh and TokyoTM. The second row shows the cluster-wise saliency of visual cues before and after intra-cluster weighting, from which it can be observed that the vehicles are suppressed and the road markings are preserved.

$$l_\theta(I_q, I_r^p, I_r^n) = [d^2(f_\theta(I_q), f_\theta(I_r^p)) \\ - d^2(f_\theta(I_q), f_\theta(I_r^n)) + m]_+ \tag{6}$$

$[x]_+ = max(x, 0)$ and m denotes the empirical margin. We follow the same positive and hard negative mining in [1] to prepare a set of tuples $(I_q, I_r^{p*}, \{I_r^n\})$ for training. A tuple consists of 1 query, 1 positive and N negatives, which can be further divided into N triplets $(I_q, I_r^{p*}, I_r^{nj})$. The loss for each tuple can be expressed as:

$$L_\theta(I_q, I_r^{p*}, \{I_r^n\}) = \frac{1}{N}\sum_{j=1}^{N} l_\theta(I_q, I_r^{p*}, I_r^{nj}) \tag{7}$$

By minimizing Eq.(7), the parametric model is trained under an weakly supervised manner, with only GPS tags used for threshold-based tuple mining for query images.

V. Experiments

This section describe how we conduct experiments to validate our proposed method.

A. Datasets and Evaluation Metric

In this work, three benchmark datasets of retrieval-based VPR [1], [20], [30], [53] are employed for the evaluation. **Pitts250k** [54] contains 254k images captured in Pittsburgh area, which are geographically divided into three subsets for training, validation and testing. **Pitts30k** [1] is a refined subset of Pitts250k, with train/val/test sets all containing a 10k database and around 7k queries. **Tokyo 24/7** [1], [6] is a more challenging dataset, which contains 76k database images and 315 query images captured at daytime, sunset and

TABLE I

Evaluation of the proposed components (**LW** and **SC**). The comparison is made with other generalized VLAD pooling layers. All representations are based on VGG-16 architecture without PCA whitening.

Method	LW	SC	Pitts30k-test			Pitts250k-test			Tokyo24/7		
			r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10
NetVLAD [1]	×	×	83.6	92.2	94.0	84.1	92.5	94.5	60.0	76.2	79.7
GhostVLAD [44]	√	×	83.7	92.5	94.7	84.1	92.7	95.1	61.6	76.5	80.6
CRN [20]	√	×	84.0	92.6	94.9	84.7	92.9	95.3	61.9	75.6	79.7
Ours: SRALNet	√	×	84.4	92.5	94.8	84.8	93.1	95.4	63.5	77.8	81.0
Ours: SRALNet-SC	√	√	**85.1**	**93.3**	**95.2**	**85.8**	**94.1**	**95.9**	**68.6**	**80.0**	**83.8**

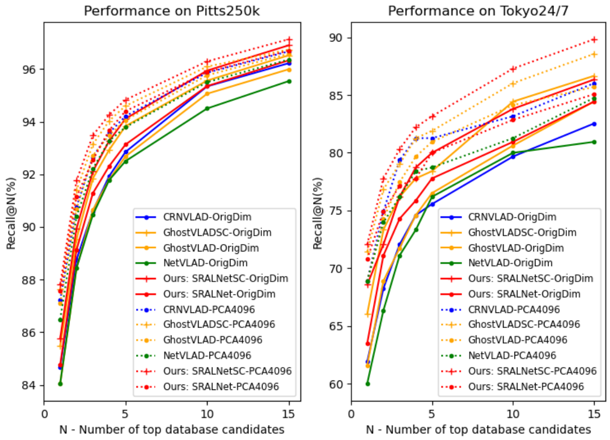


Fig. 6. Representations applying **SC**(-+-) steadily outperform those without **SC**(-·-). Performing **PCA-W** to reduce original dimensionality(–) to 4096-D(···) improves the performance of all models. Our SRALNet(red) surpasses all corresponding counterparts(yellow, blue, green) in all cases.

night. Same as the latest SOTA [40], we employ Pitts30k-train as the only training set, and test the trained models on Pitts250k-test, Pitts30k-test, Tokyo24/7 respectively.

Following the standard evaluation protocol [1], [20], [30], the performance of a retrieval inference is measured by the recall given N potential positive candidates ($Recall@N$). The indexing of a query is deemed successful as long as one of the candidates falls within a distance of d_r=25m from the geographic location of the query image.

B. Benchmark Methods

Five state-of-the-art models proposed for VPR are selected for comparison: **NetVLAD** [1] is the seminal generalized VLAD pooling layer. On top of it, **CRN** [20] introduces an attention layer to estimate local saliency according to semi-global context. **SPENetVLAD** [40] stacks the regional VLAD features to retain the spatial information. **R-MAC** [27] aggregates the max-pooled activations of multi-scale rigid grids. **APAnet** [30] aggregates spatial pyramid features weighted by cascaded attention blocks. Besides, we adapt **GhostVLAD** [44], a similar architecture proposed for face recognition, to the VPR tasks.

C. Implementation Details

Since different deep learning frameworks were used in the previous SOTAs [1], [20], [30], [40], we re-implement all comparative models in Pytorch for fair comparison. The pre-trained VGG-16 [45] cropped at the last convolutional layer is employed as the base network for local feature extraction. Benchmark models and ours are implemented as a subsequent pooling layer appended to the base network. The number of visual words K in evaluated VLAD variants is uniformly set to 64. All models are trained and evaluated using the same pipeline. We use Stochastic Gradient Descent (SGD) optimizer (with initial learning rate 0.01, momentum 0.9 and weight decay 0.001) to minimize the loss function Eq.(7), in which the margin m is chosen as 0.1. All models are trained for 30 epochs, with the learning rate reduced by a factor of 2 every 5 epochs. An early stop is triggered once the best validation $Recall@1$ stagnates for 10 epochs. For more compact representations, we also perform PCA whitening (**PCA-W**) and L_2-normalization on the baselines and our method.

TABLE II

Evaluate the separate contribution of semantic constrained initialization (**SC**) and data-driven fine-tuning (**FT**).

Method	SC	FT	Pitts250k-test			Tokyo24/7		
			r@1	r@5	r@10	r@1	r@5	r@10
NetVLAD [1]	-	×	68.2	81.7	85.6	42.9	62.9	69.8
	-	√	**84.1**	**92.5**	**94.5**	**60.0**	**76.2**	**79.7**
GhostVLAD [44]	×	×	68.2	81.7	85.5	44.1	62.9	70.8
	√	×	71.7	84.6	88.3	54.0	69.2	74.6
	×	√	84.1	92.7	95.1	61.6	76.5	80.6
	√	√	**85.5**	**93.8**	**95.6**	**66.0**	**78.4**	**84.4**
Ours: SRALNet	×	×	68.7	82.2	86.0	43.2	64.1	70.8
	√	×	72.6	85.5	89.2	56.5	70.5	76.8
	×	√	84.8	93.1	95.4	63.5	77.8	81.0
	√	√	**85.8**	**94.1**	**95.9**	**68.6**	**80.0**	**83.8**

D. Ablation Study

As elaborated in Section III, comprehensive attention has been integrated into the encoding strategy through the proposed components. To validate the local weighting scheme (**LW**) and semantic constrained initialization (**SC**) respectively, the plain SRALNet (with **LW** only) is set as the base model while **SC** is a applicable option for ablation study.

We compare the performance of our method with other generalized VLAD pooling layers. GhostVLAD has the same architecture as NetVLAD, but excludes the specified clusters from the descriptor embedding. SRALNet and CRN differ from NetVLAD with an additional weighting for local features: CRN acts as a black-box estimator for predicting the local saliency, while the local weighting scheme within

TABLE III

Performance comparison with other generalized VLAD pooling layers in original-D and 4096-D representations.

Method	PCA-W	Pitts250k-test			Tokyo24/7		
		r@1	r@5	r@10	r@1	r@5	r@10
NetVLAD [1]	w/o	84.1	92.5	94.5	60.0	76.2	79.7
	4096D	**86.5**	**93.8**	**95.5**	**68.9**	**78.7**	**81.3**
GhostVLAD [44]	w/o	84.1	92.7	95.1	61.6	76.5	80.6
	4096D	**87.1**	**94.1**	**95.8**	**68.9**	**81.0**	**84.1**
CRN [20]	w/o	84.7	92.9	95.3	61.9	75.6	79.7
	4096D	**87.2**	**94.2**	**95.9**	**68.9**	**81.3**	**83.2**
Ours: SRALNet	w/o	84.8	93.1	95.4	63.5	77.8	81.0
	4096D	**87.6**	**94.4**	**95.9**	**70.8**	**80.0**	**85.1**
Ours: SRALNet-SC	w/o	85.8	94.1	95.9	68.6	80.0	83.8
	4096D	**87.8**	**94.8**	**96.3**	**72.1**	**83.2**	**87.3**

TABLE IV

Performance comparison with 512-D representations. All comparative models are reimplemented in Pytorch, trained and evaluated using the same protocol as our model.

Method	PCA-W	Pitts250k-test			Tokyo24/7		
		r@1	r@5	r@10	r@1	r@5	r@10
Max pooling [37]	512D	39.3	59.0	67.2	11.8	23.5	33.3
R-Mac [27]	512D	54.7	72.8	78.9	27.9	49.2	56.8
Sum pooling [23]	512D	70.7	84.1	88.5	28.6	43.8	53.0
APAnet [30]	512D	76.7	88.8	91.7	51.1	66.7	71.1
NetVLAD [1]	512D	83.3	92.3	94.5	55.2	68.9	74.9
GhostVLAD [44]	512D	83.9	92.6	95.1	56.5	71.8	76.5
SPENetVLAD [40]	512D	84.4	93.1	94.8	57.1	72.4	79.7
CRN [20]	512D	84.5	92.9	95.0	59.1	73.7	76.8
Ours: SRALNet-SC	512D	**84.8**	**93.5**	**95.6**	**60.6**	**76.5**	**80.0**

SRALNet is interpretable which is derived from the feature distribution. The top retrieved results of different models are given in Table I. As can be seen, the models with local weighting mechanism (SRALNet, CRN and GhostVLAD) all surpass NetVLAD on both benchmarks. One can easily judge that introducing attention in local feature refinement conduces to the discriminability of generated descriptors. Besides, both SRALNet and CRN outperforms plain GhostVLAD, which shows the advantages of decoupling clustering and filtering. SRALNet surmounts CRN in the retrieval performance, which indicates the superiority of our interpretable local weighting scheme. Furthermore, applying semantic constrained initialization can bring stable performance improvements. As shown in the Table I and Fig.6, SRALNet-SC convincingly outperforms all other baselines on both benchmarks, especially for Tokyo24/7 where a great improvement of 9% has been achieved in Recall@1 index compared with NetVLAD. It demonstrates that endowing the local weighting scheme with semantic attention priors brings more robustness to the embedded descriptor, and its advantages are more pronounced in more challenging scenarios.

The outstanding performance of SRALNet-SC can be attributed to the mutual promotion between semantic priors and data-driven fine-tuning. To evaluate their contribution separately, we first compare the off-the-shelf models with and without semantic prior enhancement. Then we compare their performance with and without further fine-tuning. As can be seen in Table II, the off-the-shelf NetVLAD, GhostVLAD and SRALNet perform similarly, it is because they are all initialized to mimic traditional VLAD. Through semantic constrained initialization, the local features predicted to have misleading semantics are basically located in shadow areas/ghost clusters and suppressed. Thereby, the off-the-shelf SRALNet-SC/GhostVLAD-SC can be regarded as 'rule-based filter + VLAD'. The results show that introducing semantic priors for local feature filtering steadily raises the performance for both models, while further fine-tuning can bring another significant improvement. Besides, SRALNet surpasses GhostVLAD in all cases, which demonstrates the greater optimization potential of our hierarchical architecture.

E. Comparison with SOTA methods

Table III shows the the comparison of our model with other generalized VLAD variants in two dimensional representations. One can infer that the reduction into 4096 dimensions using PCA whitening consistently improves the retrieval performance of all evaluated architectures. It can be attributed to the fact that PCA whitening could penalize the co-occurance over-counting [55] while preserve the energy distribution. Table IV represents the comparison of performance on a more compact representation with 512 dimensions, where our method outperforms all the baseline counterparts. Combining the results in Table I$\sim$IV, it can be seen that our proposed SRALNet has shown compelling advantages in different dimensional representations.

VI. Conclusions

In this paper, we propose an attentional encoding architecture named SRALNet for VPR. To suppress misleading visual cues in the representation, we propose an interpretable local weighting scheme that can elegantly integrate semantic priors and data-driven learning. Experiments show that the comprehensive attention incorporated in the feature embedding can greatly enhance the image representation. On the benchmark datasets for city-scale VPR, our SRALNet is proven to be superior to the state-of-the-art methods in different dimensional representations.

References

[1] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016.

[2] R. Arandjelovic and A. Zisserman, "Dislocation: Scalable descriptor distinctiveness for location recognition," in *ACCV*, 2014.

[3] S. Cao and N. Snavely, "Graph-based discriminative learning for location recognition," in *CVPR*, 2013.

[4] G. Schindler, M. A. Brown, and R. Szeliski, "City-scale location recognition," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2007.

[5] T. Sattler, M. Havlena, F. Radenović, K. Schindler, and M. Pollefeys, "Hyperpoints and fine vocabularies for large-scale location recognition," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2102–2110, 2015.

[6] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *CVPR*, 2015.

[7] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," *CVPR 2011*, pp. 737–744, 2011.

[8] M. J. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *I. J. Robotics Res.*, vol. 27, pp. 647–665, 2008.

[9] Z. Wu, M. Wen, G. Peng, X. Tang, and D. Wang, "Magnetic-assisted initialization for infrastructure-free mobile robot localization," in *2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*. IEEE, 2019, pp. 518–523.

[10] M. Cummins and P. M. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *I. J. Robotics Res.*, vol. 30, no. 9, pp. 1100–1123, 2011. [Online]. Available: https://doi.org/10.1177/0278364910385483

[11] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," *ArXiv*, vol. abs/1805.07703, 2018.

[12] E. Chalmers, E. B. Contreras, B. Robertson, A. Luczak, and A. J. Gruber, "Learning to predict consequences as a method of knowledge transfer in reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 2259–2270, 2018.

[13] C. McManus, W. Churchill, W. P. Maddern, A. D. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 901–906, 2014.

[14] R. Mur-Artal, J. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, pp. 1147–1163, 2015.

[15] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera relocalization by computing pairwise relative poses using convolutional neural network," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 920–929, 2017.

[16] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited," in *BMVC*, 2012.

[17] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele, "Real-time image-based 6-dof localization in large-scale environments," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1043–1050, 2012.

[18] Z. Wu, Y. Yue, M. Wen, J. Zhang, G. Peng, and D. Wang, "MSTSL: Multi-sensor based two-step localization in geometrically symmetric environments," in *2021 International Conference on Robotics and Automation (ICRA)*. IEEE, to be published, 2021.

[19] D. J. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher, "Discrete-continuous optimization for large-scale structure from motion," in *CVPR*, 2011.

[20] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3251–3260, 2017.

[21] A. Khaliq, S. Ehsan, M. Milford, and K. D. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for severe viewpoint and appearance changes," *ArXiv*, vol. abs/1811.03032, 2018.

[22] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3384–3391, 2010.

[23] A. Babenko and V. S. Lempitsky, "Aggregating local deep features for image retrieval," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1269–1277, 2015.

[24] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marqués, and X. Giró, "Bags of local convolutional features for scalable instance search," in *ICMR*, 2016.

[25] E.-J. Ong, S. Husain, and M. Bober, "Siamese network of deep fisher-vector descriptors for image retrieval," *ArXiv*, vol. abs/1702.00338, 2017.

[26] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Visual instance retrieval with deep convolutional networks," *CoRR*, vol. abs/1412.6574, 2014.

[27] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *CoRR*, vol. abs/1511.05879, 2015.

[28] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *ECCV Workshops*, 2015.

[29] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *ECCV*, 2010.

[30] Y. Zhu, J. Wang, L. Xie, and L. Zheng, "Attention-based pyramid aggregation network for visual place recognition," in *ACM Multimedia*, 2018.

[31] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3476–3485, 2016.

[32] N. Piasco, D. Sidibé, V. Gouet-Brunet, and C. Demonceaux, "Learning scene geometry for visual localization in challenging conditions," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9094–9100, 2019.

[33] A. Mousavian, J. Kosecka, and J.-M. Lien, "Semantically guided location recognition for outdoors scenes," *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4882–4889, 2015.

[34] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2614–2620, 2017.

[35] M. C. Dorst, "Distinctive image features from scale-invariant keypoints," 2011.

[36] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, 2008.

[37] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "A baseline for visual instance retrieval with deep convolutional networks," in *ICLR 2015*, 2014.

[38] J. Y.-H. Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 53–61, 2015.

[39] T. Uricchio, M. Bertini, L. Seidenari, and A. D. Bimbo, "Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 1020–1026, 2015.

[40] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2019.

[41] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Robotics: Science and Systems*, 2015.

[42] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *ECCV*, 2016.

[43] F. Radenović, G. Tolias, and O. Chum, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *ECCV*, 2016.

[44] Y. Zhong, R. Arandjelovic, and A. Zisserman, "Ghostvlad for set-based face recognition," in *ACCV*, 2018.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[46] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3304–3311, 2010.

[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[48] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[49] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision*, vol. 124, pp. 237–254, 2016.

[50] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[51] N. Gupta, S. Mujumdar, S. Samanta, and S. Mehta, “Learning an order preserving image similarity through deep ranking,” in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug 2018, pp. 1115–1120.

[52] Y. Gu, K. Vyas, M. Shen, J. Yang, and G. Yang, “Deep graph-based multimodal feature embedding for endomicroscopy image retrieval,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2020.

[53] L. Liu, H. Li, and D. Yu-chao, “Stochastic attraction-repulsion embedding for large scale image localization,” 2018.

[54] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, “Visual place recognition with repetitive structures,” *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 883–890, 2013.

[55] H. Jégou and O. Chum, “Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening,” in *ECCV*, 2012.