# LSDNet: A Lightweight Self-Attentional Distillation Network for Visual Place Recognition

Guohao Peng*, Yifeng Huang*, Heshan Li, Zhenyu Wu, and Danwei Wang, *Fellow, IEEE*

***Abstract*— Visual Place Recognition (VPR) has become an indispensable capacity for autonomous vehicles and mobile robots to operate in large-scale environments. With the rapid development of the field, most recent methods focus on exploring high-performance encoding strategies, while few attempts are devoted to lightweight models with lower computational and memory costs. In this work, we propose a Lightweight Self-attentional Distillation Network (LSDNet) aiming to obtain dual advantages of both performance and efficiency. (1) From a performance perspective, an attentional encoding strategy is proposed to integrate crucial information in the scene. It extends the NetVLAD architecture with a self-attention module to facilitate the non-local information interaction between local features. Through further visual word vector rescaling, the final image representation can benefit from both non-local spatial integration and cluster-wise weighting. (2) From an efficiency perspective, LSDNet is built upon a lightweight backbone. To maintain comparable performance to large backbone models, a dual distillation strategy is proposed. It prompts LSDNet to learn both encoding patterns in the hidden space and feature distributions in the encoding space from the teacher model. Through distillation-augmented metric learning, LSDNet is able to rival the teacher model and outperform SOTA global representations with the same lightweight backbone. (3) Extensive experiments are conducted to verify the effectiveness of the proposed self-attentional encoding and dual distillation strategy. It demonstrates that LSDNet can greatly reduce resource consumption while maintaining high performance.**

## I. INTRODUCTION

The past decade has seen the growing popularity of unmanned vehicles and mobile robots in everyday life. This makes Visual Position Recognition (VPR) a research hotspot of widespread concern, as it can be used in robotic systems to identify historical locations to aid in geo-localization and pose estimation. Conventionally, VPR can be handled as either image retrieval [**?**], [1]–[4], place categorization [5], [6], or image-map matching [7] task. Considering the feasibility in large-scale environments, in this work, VPR is tackled as an image retrieval task. That is, given a query image, the best matching reference images are retrieved by traversing the database. To address the retrieve-based VPR, the core issue lies in how to formulate a compact image descriptor to represent an image. Taken into account practicality, resource consumption of descriptor generation and matching efficiency are also important indicators.

G. Peng, Y. Huang, H. Li, Z. Wu, and D. Wang are with School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore (email: peng0086@ntu.edu.sg)

* Co-first authorship and corresponding author

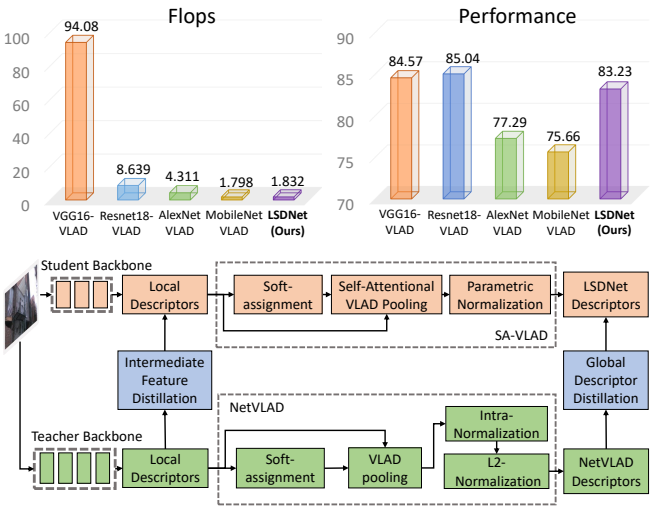


Fig. 1. Benefiting from the attentional encoding strategy and dual distillation augmented training, our LSDNet can achieve performance competitive with the teacher model at very low computational cost.

Among all the attempts in the field, aggregation-based methods have demonstrated their superiority. Traditional classic models include Bag-of-Words (BoW) [8], Vector of Locally Aggregated Descriptor (VLAD) [9], and Fisher Vector (FV) [10]. With the rise of deep learning, these models can be further re-formulated through learning-based architectures [3], [11], [12] with additional performance gains. Typically, inspired by VLAD, NetVLAD [3] has emerged as the most representative global descriptor for VPR. Taking the output feature maps of backbone model as deep local features, it characterizes the cluster-wise statistics of local features and enables further fine-tuning on task-specific datasets. In pursuit of higher performance, researchers manage to improve NetVLAD with additional functional modules, including contextual feature reweighting [4], semantic reinforced weighting [13] and spatial information integration [14], [15]. Although these variants achieve state-of-the-art performance, they rely on a large convolutional backbone such as VGG16 [16] or AlexNet [17] for local feature extraction. Despite the advantages, these large backbones with deeper architectures require more memory and computational cost, which places higher demands on hardware devices.

Due to the limited computing resources, the efficiency of VPR method must also be taken into account when adapting a mobile robot. Therefore, lightweight models with good

performance have attracted increasing attention. In general, lightweight VPR model design mainly focuses on model compression [18]–[22], typically with compact backbone replacement [20]–[22]. It can greatly reduce resource consumption, but at the cost of performance degradation. One mitigation is to learn better inference patterns of large models through Knowledge Distillation (KD), while little literature in the context of VPR explores from this perspective. The only two attempts [20], [21] use a descriptor approximation strategy for blunt representation imitation, which cannot sufficiently learn from the teacher model.

With the above motivations, we delve into model compression and distillation for VPR in this work. Specifically, a lightweight self-attentional distillation network, named LSDNet, is proposed. For local feature extraction with higher efficiency and lower memory cost, the more lightweight MobileNetV2 [23] is chosen as the backbone alternative to VGG16 [16]. To better integrate task-relevant information, a Self-Attentional VLAD encoding strategy, named SAVLAD, is proposed for global descriptor generation. It encompasses a self-attentional layer to facilitate non-local information interaction between cluster-wise local residuals. Besides leveraging spatial attention, visual word vector rescaling is further employed to highlight the different cluster significance in the final image representation. Overall, the backbone network and the SAVLAD pooling layer constitute the lightweight encoding architecture of LSDNet. To get the best performance of the LSDNet, we introduce a dual distillation strategy to enhance model training. On the one hand, for the lightweight backbone to learn better encoding patterns of the large teacher backbone, we employ intermediate feature distillation to increase the similarity between their output feature maps. On the other hand, for better descriptor distribution, we urge the topological relationship of the global descriptor triples of the LSDNet to approximate that inferred by the teacher model. Fine-tuned through triplet metric learning and dual knowledge distillation, LSDNet can achieve comparable performance to the teacher model, but with significantly lower cost (Fig.I) . Overall, our contributions can be summarized as follows:

- A Lightweight Self-attentional Distillation Network (LSDNet) is proposed for VPR from the perspective of model compression and knowledge distillation.
- An attentional VLAD pooling layer named SAVLAD is proposed, which incorporates a self-attentional residual refinement to utilize non-local spatial information.
- A dual distillation strategy is introduced to enhance model training. Intermediate feature distillation helps LSDNet to learn better encoding patterns in hidden space, while global descriptor distillation improves its topological distribution of triplets in encoding space.
- Ablation studies validate the effectiveness of our proposed SAVLAD and dual distillation strategy. Comparative experiments demonstrate that, through distillation-augmented training, LSDNet can achieve competitive performance with significantly lower cost.

## II. Related work

### A. Visual Place Recognition

The core of the retrieval-based VPR task lies in how to effectively describe the scene. Early encoding strategies are typified by Bag of Words (BOW) models [8], [24], which depict an image as a histogram of visual words. Later on, Vector of Locally Aggregated Descriptor (VLAD) [25] shows significant advantages in characterizing image details by aggregating cluster-wise local residuals.

With the rapid development of deep learning, traditional models, such as BoW and VLAD, are shown to achieve better performance when reconstructed with learning-based architectures. The typical case is NetVLAD [3], which extends the traditional VLAD encoding strategy as a trainable pooling layer. Following this seminal work, CRN [4] combines contextual weighting network to predict the importance of receptive regions. SPENetVLAD [26] builds a pyramid structure to encode the spatial VLAD feature of regional patches. SRALNet [13] introduces intra-cluster weighting with semantic constrained initialization. APPSVR [15] incorporates attentional pyramid pooling of salient local residuals and proposes cluster feature scaling. Although the aforementioned methods can achieve better performance, they rely on the large convolutional backbone VGG16 [16] for local feature extraction. Besides, the additional attention modules may also lead to a higher computational cost.

Considering efficiency in model designing, CAMAL [27] and Region-VLAD [28] employ middle convolutional layers of AlexNet [17] for feature extraction, and captures multi-layer attentions to enhance the descriptor robustness. HF-Net [20], [21] combines MobileNet [23] and NetLVAD into a lightweight encoding architecture, and introduces knowledge distillation to approximate its image representation to the teacher model. Also using MobileNet as the lightweight backbone, our LSDNet proposes a self-attentional VLAD pooling layer to exploit non-local spatial information. Rather than descriptor imitation in HF-Net, a dual distillation strategy is introduced, including intermediate feature distillation and topological approximation of global descriptor triples.

### B. Knowledge Distillation

Knowledge Distillation (KD) is first proposed for image classification by Hinton *et al.* [29]. The predicted logits of the teacher model are used as soft labels for supervising the student model training. Later on, researchers put forward the knowledge transfer between output features [30], [31] or attention maps [32] of teacher and student models. A typical case is Attention-based Feature Distillation (AFD) [31], which exploits the relative similarities between features to control distillation intensities of the feature pairs.

Although knowledge distillation has achieved great success on classification tasks, there has been little exploratory work on knowledge distillation for VPR tasks. In the context of image retrieval and person Re-ID, DarkRank [33] transfers cross-image similarities through rank matching between teacher and student models. Relational Knowledge

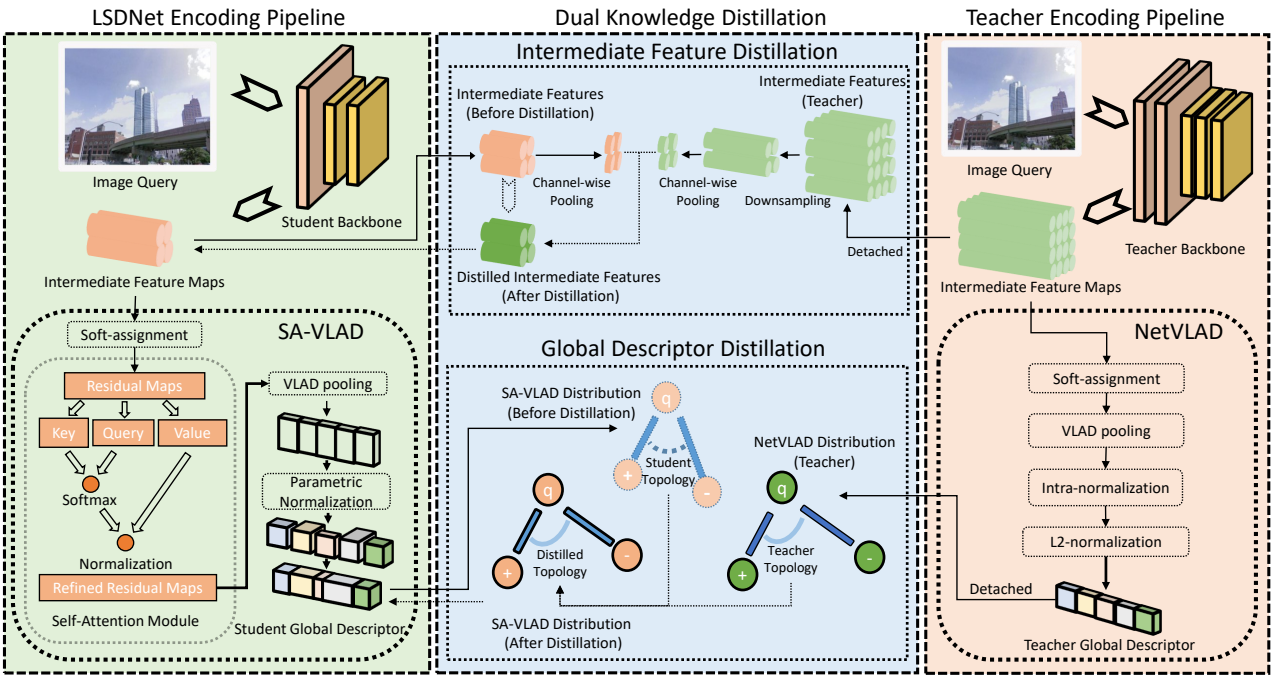


Fig. 2. Overall training and inference framework for LSDNet. On the left, the LSDNet encoding architecture consists of a lightweight backbone for local feature extraction and a SAVLAD pooling layer for attentional global descriptor generation. In the middle, a dual distillation strategy is introduced for better knowledge transfer between teacher NetVLAD and student LSDNet. including intermediate feature distillation and global descriptor distillation.

Distillation (RKD) [34] transfers the interrelationships of data samples, including mutual distances and angles. The seminal work that introduces KD into VPR tasks is HF-Net [20], [21]. It enforces a global descriptor approximation between MobileNetVLAD and pre-trained VGG16-NetVLAD by minimizing the mean squared error between their generated descriptors. However, blunt descriptor imitation cannot adequately learn the intrinsic inference patterns of the teacher model. Therefore, inspired by AFD [31] and RKD [34], LSDNet introduces a dual distillation strategy to metric learning. Intermediate feature distillation helps LSDNet to learn better encoding patterns in the hidden space, while global descriptor distillation improves its feature topological distribution in the encoding space.

## III. Proposed Method

In this section, the details of our proposed LSDNet will be elaborated. As illustrated in Fig.2, the whole distillation framework can be decomposed into three parts: (a) The student network LSDNet consists of a lightweight backbone for local feature extraction and a SAVLAD pooling layer for global descriptor generation. (b) The pre-trained NetVLAD with VGG16 backbone is employed as a teacher model to impart better inference patterns to the student model. (c) A dual distillation strategy is introduced for better knowledge transfer, including intermediate feature distillation and topological approximation of global descriptor triples.

### A. LSDNet Encoding Architecture

Designed for efficient global descriptor generation, LSDNet uses a lightweight backbone MobileNet [23] to extract deep local features, followed by a Self-Attentional VLAD pooling layer for attentional feature aggregation.

*1) Backbone Network:* In the original NetVLAD structure, a cropped VGG-16 [16] is exploited to extract deep local features. Although VGG16 with very deep architecture has excellent high-level representation, it is too computationally expensive to be deployed on devices with insufficient computing power. By contrast, MobileNetV2 [23] with more lightweight architecture utilizes depthwise separable convolution and inverted residual block to reduce parameters and improve efficiency. Therefore, MobileNetV2 [23] is employed as the backbone of LSDNet for deep local feature extraction. Cropped at the last convolutional layer, the spatial activations from the normalized feature maps $\mathbf{X} \in \mathbb{R}^{D\times H\times W}$ are regarded as deep local features $\mathbf{x} \in \mathbb{R}^{D\times 1\times 1}$.

*2) Self-Attentional VLAD Pooling Layer:* As an extension of NetVLAD, SAVLAD combines self-attentional residual refinement for non-local spatial information integration, and parametric normalization for visual cluster weighting.

Specifically, the deep local features extracted by the backbone network are first divided into $K$ clusters through soft-assignment [3] as in Eq.(1). The soft-assignment weight $\alpha_k(\mathbf{x_i})$ of a local feature $\mathbf{x}_i$ being allocated to the $k^{th}$ cluster is related to its proximity to the cluster centroids $\{\mathbf{c}_k\}_{k=1}^K$.

$$\alpha_k(\mathbf{x}_i) = \frac{e^{-a\|\mathbf{x}_i-\mathbf{c}_k\|^2}}{\sum_{j=1}^K e^{-a\left\|\mathbf{x}_i-\mathbf{c}_j\right\|^2}} \tag{1}$$

Since the centroids $\{\mathbf{c}_k\}_{k=1}^K$ represent the common characteristics of each feature cluster, the residual $\mathbf{r}_k = \mathbf{x}_i - \mathbf{c}_k$ between a local feature $\mathbf{x}_i$ and the centroid $\mathbf{c}_k$ can characterize its distinctiveness with respect to this cluster. Therefore, the

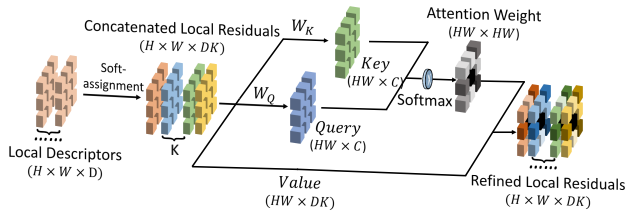


Fig. 3. Illustration of local residual refinement. Through self-attentional enhancement, a refined local residual can integrate information from all other residuals that are highly correlated with itself.

local residuals $\mathbf{R}_k \in \mathbb{R}^{D\times H\times W}$ in Eq.(2) can be regarded as the basic feature to describe the cluster-wise local details.

$$\mathbf{R}_k = \alpha_k(\mathbf{X})(\mathbf{X}-\mathbf{c}_k) \tag{2}$$

**Self-attentional Residual Refinement:** Unlike NetVLAD which directly aggregates local residuals into visual word vectors, we further introduce a self-attention module to refine local residuals through non-local information interaction.

As illustrated in Fig.3, the concatenated local residuals $\mathbf{R} \in \mathbb{R}^{DK\times HW}$ is first projected to the query and key vectors ($\mathbf{Q} \in \mathbb{R}^{C\times HW}$, $\mathbf{K} \in \mathbb{R}^{C\times HW}$) by corresponding projection matrix $\mathbf{W}_Q \in \mathbb{R}^{C\times DK}$, $\mathbf{W}_K \in \mathbb{R}^{C\times DK}$. Then the attention weight that reflects the correlation between local residuals is calculated by multiplying $W_Q$ and $W_K$, followed by a Softmax function. The refined residuals $\mathbf{R}'$ is finally obtained by weighted average of the original residuals $\mathbf{R}$ as in Eq.(4).

$$\mathbf{Q} = \mathbf{R}\mathbf{W}_Q^T, \quad \mathbf{K} = \mathbf{R}\mathbf{W}_K^T \tag{3}$$

$$\mathbf{R}' = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}})\mathbf{R} \tag{4}$$

The self-attentional enhancement facilitates the integration of spatial information, highlighting the crucial features with higher correlation to the entire image. After the residual refinement, visual word vectors $\mathbf{V}_k$ are generated by spatially aggregating the refined local residuals of each cluster.

**Visual Cluster Weighting:** Considering the different task-relevance of visual word clusters, we employ the parametric normalization proposed in our previous work [15] for visual cluster weighting in global descriptor generation.

Specifically, the trainable parameter $\gamma_k$ is introduced to quantify the significance of the $k^{th}$ visual cluster to the task. The cluster saliency $\tilde{\gamma} = [\tilde{\gamma}_1, \tilde{\gamma}_2, ..., \tilde{\gamma}_K]$ is first obtained by $L_2$-normalizing the trainable weights $\gamma = [\gamma_1, \gamma_2, ..., \gamma_K]$. Then through Eq.(5), the unit global image descriptor is generated by concatenating the normalized visual word vectors $\tilde{V}_K$ rescaled by their corresponding cluster saliency $\tilde{\gamma}_K$.

$$\mathbf{V} = [\tilde{\gamma}_1 \cdot \tilde{\mathbf{V}}_1, \tilde{\gamma}_2 \cdot \tilde{\mathbf{V}}_2, ..., \tilde{\gamma}_K \cdot \tilde{\mathbf{V}}_K] \tag{5}$$

It can be inferred that, in dot product matching of two descriptors, the variable $\tilde{\gamma}_k^2$ will distinguish the contribution of the $k^{th}$ visual word clusters to the final similarity score.

### *B. Dual Distillation Augmented Metric Learning*

To obtain the best performance of LSDNet, three losses are incorporated into the learning metrics.

*1) Triplet Ranking Loss:* Following traditional image representation learning for retrieval-based VPR [3], the triplet ranking loss [35]–[37] is adopted for tuple metric learning. Given a query image $I_q$, a set of tuple $(I_q, I_r^{p*}, \{I_r^n\})$ is mined in the same way as in [3]. $I_r^{p*}$ is the positive reference image, which is selected as the one with the smallest descriptor distance to the query. $\{I_r^n\}$ are $N$ hard negative samples. The triplet ranking loss $\mathscr{L}_{Tri}$ is formulated as Eq.(6), where the goal is to push the positive reference $I_r^p$ closer to the query $I_q$ than all negative references $\{I_r^n\}$ in encoding space. In Eq.(6), $[x]_+ = max(x, 0)$ and $m$ is a constant margin.

$$\mathscr{L}_{Tri}(I_q, I_r^{p*}, \{I_r^n\}) = \frac{1}{N}\sum_{j=1}^{N}[d^2(I_q, I_r^{p*}) - d^2(I_q, I_r^{nj}) + m]_+ \tag{6}$$

However, due to the shallow backbone, lightweight models usually perform poorly under unsupervised training with only triplet ranking loss. The solution to this problem lies in two aspects: improving the representation ability of the lightweight backbone and improving the encoding ability of the pooling layer. With the two motivations, two distillation losses are introduced for LSDNet supervised training.

*2) Intermediate Feature Distillation (IFD):* To endow the lightweight backbone with better representation ability, we facilitate knowledge transfer between intermediate features $\mathbf{X}$ output by the teacher and student backbones networks. The schematic diagram is illustrated in Fig.2.

Let $\mathbf{F}_{cat} = [\mathbf{X}_q, \mathbf{X}_p, .., \mathbf{X}_n]$ denote the tuple intermediate features concatenated across the channel dimension. To handle the asymmetric scale of teacher and student features, sampling and pooling operations as in AFD [31] are followed. The teacher feature $\mathbf{F}_{cat}^t$ is first down sampled to have the same height and width as the student feature $\mathbf{F}_{cat}^s$. Then channel-wise average pooling is performed on $\mathbf{F}_{cat}^t$ and $\mathbf{F}_{cat}^s$ to unify them into the same shape while integrating channel information. Finally, the intermediate feature distillation loss forms as Eq.(7), where $\phi^D$ and $\phi^C$ denotes the operation of down sampling and channel-wise average pooling.

$$\mathscr{L}_{IFD} = \left\|\phi^C(\mathbf{F}_{cat}^s) - \phi^C(\phi^D(\mathbf{F}_{cat}^t))\right\|_2 \tag{7}$$

*3) Global Descriptor Topological Distillation (GDTD):* Profiting from RKD [34], we introduce topological relationship transfer between global descriptor triples $(\mathbf{V}_q, \mathbf{V}_p, \mathbf{V}_n)$ of the teacher model and that of LSDNet. It aims to prompt the SAVLAD pooling layer to learn better encoding patterns.

Since the topological relationship of triples includes distance and angle, our topological distillation is also divided into two parts. The distance distillation term forms as Eq.(8), which compares the corresponding query-reference distances in the teacher and student triples.

$$\begin{aligned}\mathscr{L}_{GDTD-D} =& \mathscr{L}_{L_1}\left(\frac{1}{\mu_t}\left\|\mathbf{V}_q^t - \mathbf{V}_p^t\right\|_2, \frac{1}{\mu_s}\left\|\mathbf{V}_q^s - \mathbf{V}_p^s\right\|_2\right) + \\ & \mathscr{L}_{L_1}\left(\frac{1}{\mu_t}\left\|\mathbf{V}_q^t - \mathbf{V}_n^t\right\|_2, \frac{1}{\mu_s}\left\|\mathbf{V}_q^s - \mathbf{V}_n^s\right\|_2\right)\end{aligned} \tag{8}$$

$\mathscr{L}_{L_1}$ denotes smooth-$L_1$-loss in Eq.(9). $\mu$ is the average

| NetVLAD | PCA-W | Pitts30k-test | | | Pitts250k-test | | | Tokyo247 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| MobileNetVLAD | w/o | 75.66 | 89.33 | 92.97 | 73.86 | 87.07 | 91.10 | 36.51 | 53.33 | 60.22 |
| | 4096D | 82.19 | 91.97 | 94.76 | 80.58 | 90.77 | 93.61 | 43.49 | 59.68 | 64.71 |
| LSDNet-SA | w/o | 77.38 | 90.51 | 94.10 | 75.11 | 88.24 | 92.00 | 37.14 | 53.02 | 59.68 |
| | 4096D | 82.94 | 92.90 | 94.91 | 81.50 | 91.33 | 93.78 | 46.35 | 58.78 | 64.81 |
| LSDNet-SA-PN | w/o | 77.51 | 89.72 | 93.16 | 75.58 | 87.49 | 91.00 | 40.30 | 54.56 | 61.56 |
| | 4096D | 83.03 | 92.24 | 94.53 | 81.90 | 90.94 | 93.22 | 47.49 | 59.37 | 65.40 |
| LSDNet-SA-PN-Dis1 | w/o | 77.82 | 89.83 | 94.03 | 75.85 | 88.07 | 91.90 | 38.56 | 54.29 | 63.81 |
| | 4096D | 83.33 | 92.63 | 95.01 | 82.48 | 91.74 | 93.95 | 49.84 | 61.27 | 66.98 |
| LSDNet-SA-PN-Dis1&2 | w/o | 83.23 | 92.22 | 94.75 | 82.80 | 91.88 | 94.40 | 44.44 | 60.95 | 66.67 |
| | 4096D | **85.99** | **93.31** | **95.26** | **86.16** | **92.98** | **94.87** | **53.02** | **64.13** | **71.43** |

TABLE I

ABLATION STUDY ON THE PROPOSED COMPONENTS. THE COMPARISONS ARE MADE ON REPRESENTATIONS WITH ORIGIN-D AND 4096-D (BOLD).

query-reference distances within the tuple $(\mathbf{V}_q, \mathbf{V}_p, \{\mathbf{V}_n\})$.

$$\mathscr{L}_{\mathrm{L_1}}(x,y) = \begin{cases} 0.5\,(x-y)^2 & \text{if } |x-y| < 1 \\ |x-y| - 0.5 & \text{otherwise} \end{cases} \tag{9}$$

The angle distillation term is defined as Eq.(10). It penalizes angular differences between teacher and student triplets. As in Eq.(11), $\cos\varphi$ is cosine similarity of normalized query-positive and query-negative residual vectors in a triplet.

$$\mathscr{L}_{GDTD-A} = \mathscr{L}_{\mathrm{L_1}}(\cos\varphi_t, \cos\varphi_s), \tag{10}$$

$$\cos\varphi = \left\langle \frac{\mathbf{V}_q - \mathbf{V}_p}{\|\mathbf{V}_q - \mathbf{V}_p\|_2}, \frac{\mathbf{V}_q - \mathbf{V}_n}{\|\mathbf{V}_q - \mathbf{V}_n\|_2} \right\rangle \tag{11}$$

Combining triplet ranking and dual distillation terms, the overall loss function for LSDNet training forms as Eq.(12).

$$\mathscr{L} = \mathscr{L}_{Tri} + \lambda_1 \cdot \mathscr{L}_{IFD} + \lambda_2 \cdot (\mathscr{L}_{\mathrm{GDTD-D}} + \mathscr{L}_{\mathrm{GDTD-A}}) \tag{12}$$

## IV. EXPERIMENTS

### A. Datasets and Evaluation Metric

Three benchmark datasets, Pitts30k [3], Pitts250k [38], and Tokyo24/7 [39], are employed for comparative experiments. Pitts250k [38] contains 250k database images and 24k queries from different years and streets. Tokyo24/7 [39] has about 76k images and 315 queries, containing challenging images captured at night and sunset. Following SOTAs [15], [26], Pitts30k is used for training all models, while evaluation is performed on Pitts250k-test, Pitts30k-test, and Tokyo24/7.

Following the standard evaluation protocol for the employed datasets, the performance of models is evaluated by Recall@N. Given *N* positive candidates, a retrieval inference is correct once any retrieved place is within 25 meters from the query location. Besides, the number of parameters, FLOPS (floating-point operations per second), and latency time are used to evaluate the efficiency of models.

### B. Implementation Details

In this work, all experiments are conducted in PyTorch framework on an NVIDIA RTX 2080TI GPU. VGG-16 [16], AlexNet [17], ResNet-18 [40] and MobileNetV2 [23] are selected as the optional backbone networks in our experiments. Unlike the other three cropped at the last convolutional layer, ResNet-18 is cropped at the penultimate convolutional layer since we found this performs better. For all these backbone networks, only parameters from the last convolutional layer after cropping are set as trainable, while the others are fixed to their pretrained status. The number of visual clusters in all evaluated VLAD variants is set to 64. We use the SGD optimizer to minimize the loss function Eq.(12) for training LSDNet. Other non-distilled benchmark models, including VGG16-NetVLAD that is chosen as the teacher model for distillation, are trained following the same pipeline in [3]. For more compact representations (*e.g.*, 4096-D), PCA whitening (PCA-W) is optionally performed.

### C. Ablation Study

Compared with the baseline MobileNetVLAD, LSDNet encompasses a self-attention module (**SA**) for spatially residual refinement and parametric normalization (**PN**) for visual cluster weighting. In terms of training, a dual distillation strategy is combined into metric learning, including intermediate feature distillation (**Dis1**) and global descriptor topological distillation (**Dis2**). To evaluate the benefits of each component, we use abbreviations to denote their application to the plain LSDNet. Note that with all four components disabled, the plain LSDNet is structurally the same as MobineNetVLAD.

As shown in Table.I, applying each component incrementally results in steady performance improvements. This validates all the individual components, and shows that they can bring cumulative advantages to LSDNet. Specifically, the Recall@1 performance of LSDNet-SA with original-D representation surpasses MobileNetVLAD by 1.7%, 1.3%, and 1.6% respectively on the three datasets. It verifies the effectiveness of the introduced self-attention module for spatially residual refinement. LSDNet-SA-PN consistently outperforms MobileNetVLAD, showing the superiority of our proposed SAVLAD over NetVLAD. Further incorporating **Dis1** into training can see additional enhancement of LSDNet-SA-PN, which validates the intermediate feature distillation. Noticeably, applying **Dis2** to LSDNet-SC-PN-Dis1 brings remarkable performance gains by about 6% and 3% for the original and 4096-dimensional representations on all three datasets. It embodies the necessity of incorporating global topology distillation into model training.

| Teacher | Student | Dataset | Teacher | Stu_w/o_D | Stu_w/_D |
|---|---|---|---|---|---|
| VGG16-NetVLAD | LSDNet | Pitts30k | 84.6 | 77.5 | 83.2 |
| | | Pitts250k | 85.2 | 75.6 | 82.8 |
| | | Tokyo247 | 62.2 | 40.3 | 44.4 |
| Resnet18-NetVLAD | LSDNet | Pitts30k | 86.2 | 77.5 | 84.3 |
| | | Pitts250k | 86.8 | 75.6 | 83.1 |
| | | Tokyo247 | 52.1 | 40.3 | 44.8 |
| Resnet18-NetVLAD | Resnet18-NetVLAD | Pitts30k | 86.2 | 86.2 | 86.6 |
| | | Pitts250k | 86.8 | 86.8 | 86.8 |
| | | Tokyo247 | 52.1 | 52.1 | 61.0 |
| MobileNet VLAD | MobileNet VLAD | Pitts30k | 75.7 | 75.7 | 77.1 |
| | | Pitts250k | 73.9 | 73.9 | 74.4 |
| | | Tokyo247 | 36.5 | 36.5 | 42.2 |

TABLE II

DUAL DISTILLATION USING ALTERNATIVE TEACHER OR STUDENT MODELS. 'STU_W/O_D' DENOTES STUDENT MODEL TRAINED WITHOUT DUAL DISTILLATION STRATEGY. 'STU_W/_D' DENOTES THE OPPOSITE.

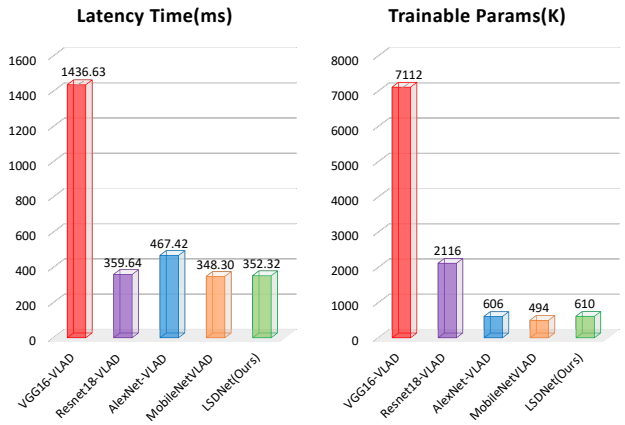


Fig. 4. Latency time and number of trainable parameters of different models. Combined with Fig.I, the advantages of LSDNet is pronounced.

| NetVLAD | Time(ms) | Number of Params | | FLOPS(G) |
|---|---|---|---|---|
| | | Trainable | Total | |
| VGG16-VLAD | 1436.63 | 7,112,192 | 14,747,456 | 94.08 |
| Resnet18-VLAD | 359.64 | 2,116,096 | 2,799,168 | 8.64 |
| AlexNetVLAD | 467.42 | 606,464 | 1,879,616 | 4.31 |
| MobileNetVLAD | 348.30 | 494,400 | 1,832,192 | 1.80 |
| LSDNet | 352.32 | 609,960 | 1,947,752 | 1.83 |

TABLE III

THE LATENCY TIME (*ms*), NUMBER OF PARAMETERS, AND FLOPS OF LSDNET AND OTHER BENCHMARK MODELS.

Overall, compared with the baseline MobileNetVLAD, our optimal model with all components enabled (LSDNet-SA-PN-Dis1&2) achieves a significant improvement of 4%, 5% and 9% on Pitts30k, Pitts250K and Tokyo247 respectively. It demonstrates the superiority of our proposed LSDNet.

### *D. More Results and Discussion*

**Potential of Dual Distillation:** To demonstrate the potential of the dual distillation strategy, we evaluate alternative teacher or student networks in our LSDNet distillation framework. As can be seen from the first two rows of Table.II, whether using VGG16-NetVLAD or ResNet18-NetVLAD as the teacher model, dual distillation can substantially improve the performance of our lightweight LSDNet on all three datasets by about 10%. The distilled student LSDNet is able to rival to the teacher VGG16-NetVLAD (only 1% worse on Recall@1), but with significantly lower computational cost. A consistent result can be seen when changing the teacher model to ResNet18-NetVLAD. It reveals that our dual distillation strategy allows the student network to sufficiently learn the intrinsic encoding patterns of the teacher model.

The last two rows of Table.II evaluates the self-distillation of two networks through dual distillation strategy. The self-distilled ResNet18-NetVLAD and MobileNetVLAD achieve steady improvements on all three datasets, especially on Tokyo247. Particularly, MobileNetVLAD shows a larger performance boost than ResNet18-NetVLAD. This suggests that the lightweight model may have greater potential to be tapped, and iterative self-supervised training may also be an effective way to enhance the model.

**Model Efficiency:** As shown in Table.III, our proposed LSDNet surpasses the teacher model VGG16-NetVLAD and most lightweight models in terms of model parameters, latency time, and FLOPS. Specifically, VGG16-NetVLAD requires the most resources, making it challenging to deploy on resource-constrained devices. By contrast, one forward inference of LSDNet only takes 352.3*ms*, which is four times faster than VGG16-NetVLAD. Moreover, LSDNet only has 600k trainable parameters, which is only 8.6% of VGG16-NetVLAD and 21.8% of ResNet18-NetVLAD. Furthermore, the FLOPS of LSDNet (1.832G) is the second smallest among all compared model, slightly larger than MobileNetVLAD due to the incorporated attention modules. It indicates that our proposed self-attentional enhancement and visual cluster weighting bring little efficiency burden to the model. Considering the competitive performance shown in Fig.I, it can be said that LSDNet has achieved a satisfying balance in performance and efficiency.

**Comparison with SOTAs:** In Table.IV, we compare our LSDNet with other SOTA global representations with lightweight MobileNetV2 backbone. The comparative models include NetVLAD [3], GhostVLAD-SC [41], CRN [4], SRAL [13], SPENetVLAD [26], and the global branch of HF-Net [21]. Built upon MobileNet, all the evaluated models are lightweight. Only HF-Net and LSDNet are knowledge distilled. As can be seen, GhostVLAD-SC, CRN, SRAL, SPENetVLAD, and LSDNet all surpass NetVLAD, which demonstrates the necessity of integrating attention into feature embedding. Only LSDNet and SPENetVLAD [26] achieve an over 80% performance on Recall@1, which can be attributed to their spatial information integration. HF-Net underperforms MobileNetVLAD and LSDNet, indicating that the distillation based on descriptor imitation cannot sufficiently learn from the teacher model. By contrast, our LSDNet convincingly outperforms all compared models. Significant improvements of 7.6% and 9,4% can be seen on both datasets compared to MobileNetVLAD, demonstrating the comprehensive advantages of our proposed attentional encoding and dual distillation strategy.

| Method | Pitts30k-test | | | | Pitts250k-test | | | |
|---|---|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@20 | Recall@1 | Recall@5 | Recall@10 | Recall@20 |
| HF-Net (global) [21] | 70.28 | 86.46 | 91.11 | 94.62 | 65.34 | 81.56 | 86.40 | 90.19 |
| NetVLAD [3] | 75.66 | 89.33 | 92.97 | 95.60 | 73.43 | 87.40 | 91.01 | 93.80 |
| GhostVLAD-SC [41] | 76.56 | 89.47 | 93.56 | 96.20 | 74.70 | 87.77 | 91.50 | 94.18 |
| CRN [4] | 76.14 | 89.14 | 92.80 | 95.32 | 73.48 | 87.23 | 90.81 | 93.68 |
| SRAL [13] | 76.88 | 89.64 | 93.78 | 96.25 | 74.43 | 87.64 | 91.52 | 94.23 |
| SPENetVLAD [26] | 80.99 | 91.64 | 94.41 | 96.39 | 80.44 | 90.05 | 94.18 | 95.98 |
| **LSDNet** | **83.23** | **92.22** | **94.75** | **96.38** | **82.80** | **91.88** | **94.40** | **96.24** |

TABLE IV

COMPARISONS WITH OTHER GENERALIZED VLAD POOLING LAYERS. THE BACKBONE OF ALL EVALUATED METHODS IS MOBILENETV2.

## V. CONCLUSIONS

Aiming to design a lightweight model with both performance and efficiency advantages, we propose a Lightweight Self-attentional Distillation Network (LSDNet) for VPR from the perspective of model compression and knowledge distillation. It contains a lightweight backbone network for local feature extraction, and a self-attentional VLAD (SAVLAD) pooling layer for global descriptor generation. To highlight crucial features in descriptor generation, SAVLAD incorporates a self-attention module for non-local spatial information integration, and visual word vector rescaling for visual cluster weighting. To obtain the best performance of LSDNet, a dual distillation strategy is introduced, including intermediate feature distillation and topological approximation of global descriptor triples. It motivates LSDNet to learn better encoding patterns in the hidden space and descriptor topological relationships in the encoding space from the teacher model. Through distillation-augmented metric learning, LSDNet is able to achieve comparable performance to the teacher model with significantly lower cost.

## REFERENCES


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

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

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

[4] 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, year=2017, pages=3251-3260*.

[5] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera relocalization by computing pairwise relative poses using convolutional neural network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 929–938.

[6] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited." in *BMVC*, vol. 1, no. 2, 2012, p. 4.

[7] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele, "Real-time image-based 6-dof localization in large-scale environments," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1043–1050.

[8] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, IEEE International Conference on*, vol. 3. IEEE Computer Society, 2003, pp. 1470–1470.

[9] 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.

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

[11] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *CoRR*, vol. abs/1706.06905, 2017. [Online]. Available: http://arxiv.org/abs/1706.06905

[12] E. Ong, S. S. Husain, M. Bober-Irizar, and M. Bober, "Deep architectures and ensembles for semantic video classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3568–3582, 2019. [Online]. Available: https://doi.org/10.1109/TCSVT.2018.2881842

[13] G. Peng, Y. Yue, J. Zhang, Z. Wu, X. Tang, and D. Wang, "Semantic reinforced attention learning for visual place recognition," in *IEEE International Conference on Robotics and Automation, ICRA China, May 30 - June 5, 2021*. IEEE, 2021, pp. 13 415–13 422.

[14] 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.

[15] G. Peng, J. Zhang, H. Li, and D. Wang, "Attentional pyramid pooling of salient visual residuals for place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 885–894.

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

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, pp. 84–90, 2012.

[18] A. Khaliq, S. Ehsan, M. Milford, and K. D. Mcdonald-Maier, "Camal: Context-aware multi-scale attention framework for lightweight visual place recognition," *ArXiv*, vol. abs/1909.08153, 2019.

[19] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes," *IEEE Transactions on Robotics*, vol. 36, no. 2, pp. 561–569, 2020.

[20] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, "Leveraging deep visual descriptors for hierarchical efficient localization," in *Conference on Robot Learning*. PMLR, 2018, pp. 456–465.

[21] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 708–12 717.

[22] Q. Gong, Y. Liu, L. Zhang, and R. Liu, "Ghost-dil-netvlad: A lightweight neural network for visual place recognition," 2021.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.

[25] 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.

[26] 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*, vol. 31, no. 2, pp. 661–674, 2019.

[27] A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "Camal: Context-aware multi-scale attention framework for lightweight visual place recognition," 2019.

[28] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns

for significant viewpoint and appearance changes," *IEEE transactions on robotics*, vol. 36, no. 2, pp. 561–569, 2019.

[29] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[30] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[31] M. Ji, B. Heo, and S. Park, "Show, attend and distill: Knowledge distillation via attention-based feature matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7945–7952.

[32] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017*, 2017.

[33] Y. Chen, N. Wang, and Z. Zhang, "Darkrank: Accelerating deep metric learning via cross sample similarities transfer," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 2852–2859. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17147

[34] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[35] 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.

[36] 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.

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

[38] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 883–890.

[39] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[41] Y. Zhong, R. Arandjelović, and A. Zisserman, "Ghostvlad for set-based face recognition," in *Asian conference on computer vision*. Springer, 2018, pp. 35–50.