2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Globalizing Local Features: Image Retrieval Using Shared Local Features with Pose Estimation for Faster Visual Localization

Wenzheng Song[1,2] Ran Yan[1] Boshu Lei[1,4] Takayuki Okatani[2,3]
[1]Megvii [2]GSIS, Tohoku University [3]RIKEN Center for AIP [4]Xi'an Jiaotong University
{song, okatani}@vision.is.tohoku.ac.jp yanran@megvii.com sobremesa121@gmail.com

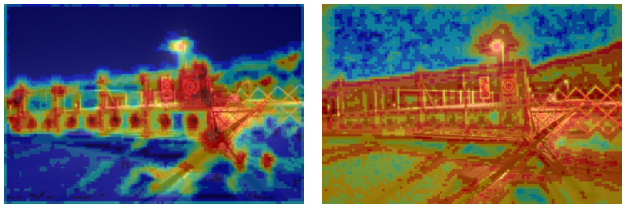
***Abstract*— Visual localization is an important sub-task in SfM and visual SLAM that involves estimating a 6-DoF camera pose for an input query image relative to a given 3D model of the environment. The most accurate approach is a hierarchical one that splits the task into two stages: image retrieval and camera pose estimation. Each stage requires different image features, with global features compactly encoding holistic image information for the first stage and local features encoding the appearance around salient image points for the second stage. While existing methods use independent networks to extract these features, one for global and one for local, this strategy is suboptimal in terms of computational efficiency. In this paper, we propose a novel approach that achieves state-of-the-art inference accuracy with significantly improved efficiency. Our approach's core component is SuperGF, a network that aggregates local features optimized for camera pose estimation to create a global feature that enables precise image retrieval. Through extensive experiments on the standard benchmark tests, we demonstrate that the method offers a better trade-off between accuracy and computational cost.**

I. INTRODUCTION

Visual localization entails estimating the 6-DoF camera pose based on an image of a scene, relative to a reference scene representation. This process is a critical component in computer vision tasks including structure-from-motion (SfM) and SLAM. It is foundational to a variety of applications including autonomous driving, mobile robotics, and augmented reality. As the range of applications expands, there is an increasing need for visual localization to operate reliably across both large-scale challenging indoor and outdoor environments, unaffected by variations in weather, lighting, or seasonal changes.

To achieve maximum accuracy, recent studies on visual localization adopt a hierarchical approach [1], [2], [3]. The process involves two steps: i) retrieving candidate images from an image database, and ii) estimating the query's 6-DoF camera pose by matching the query and candidate images. Specifically, in the image retrieval step, the closest multiple database images to the query are selected using holistic image similarity. In the camera pose estimation step, the query is matched to the found candidate images to establish 2D-3D point correspondences from the query to the given 3D scene model, thereby estimating the query's camera pose.

In this hierarchical approach, distinct image features are utilized at each step: global image features for image retrieval and local image features for camera pose estimation. Consequently, a majority of visual localization studies adopting this approach utilize separate methods for each step. For instance, NetVLAD [4] is often employed to extract global features from an input image, while image-matching tasks utilize well-established local features such as SIFT [6] or SuperPoint [5]. However, computational efficiency—encompassing factors like inference speed and memory footprint—is another critical factor, in addition to the accuracy of pose estimation. This underscores the motivation to enhance efficiency by integrating both processes, potentially through the sharing of certain computations.



(a) (b)

Fig. 1. Visualization of the final layer's activation maps for an identical image: (a) NetVLAD [4] and (b) SuperPoint [5]. Their differences imply the gap between (a) global features for image retrieval and (b) local features for image matching.

However, achieving this integration presents a challenge due to the differing requirements of the two types of features, as illustrated in Fig. 1. Global features must offer a comprehensive representation of input images by encoding the entirety of a scene. These features can be more effective when they focus on local regions with distinctive appearances while excluding irrelevant areas to enhance discriminability. Conversely, local features are tasked with encoding the localized appearances surrounding key points identified across entire images, facilitating precise camera pose estimation.

This paper addresses the challenge of integrating the processes of local feature extraction and global feature extraction. Specifically, we aim to transform a set of local features, which are beneficial for precise 6-DoF pose estimation, into global features that enhance the accuracy of image retrieval. To date, only a handful of studies have pursued a similar goal. HF-Net [1] seeks to unify global (e.g., NetVLAD [4]) and local (e.g., SuperPoint [5]) features for visual localization through multitask distillation. However, the employment of the teacher model for distillation can potentially decrease inference accuracy, particularly in challenging environments such as nighttime scenes. Similarly, Cao et al. [7] attempt

979-8-3503-8457-4/24/$31.00 ©2024 IEEE

to develop a unified model to extract both local and global features, but encounter a similar difficulty. Tolias et al. [8] propose aggregating learned local features to global ones for image retrieval employing ASMK [9]. However, the learned local features are unsuitable for an advanced visual localization pipeline.

Retrieving candidate images in the first step of hierarchical visual localization is equivalent to another task called visual place recognition (VPR). To enhance accuracy, several existing VPR methods ([7], [10], [11]) utilize a two-step approach akin to visual localization. Initially, these methods retrieve candidate images leveraging global image features, followed by a re-ranking based on geometric verification between the query and candidate images. However, while their structure is similar, these VPR methods cannot be applied to visual localization, as they rely on the assumption of a planar homography for geometric verification, which only roughly approximates 6-DoF camera pose changes. Consequently, these approaches often recycle features extracted from the intermediate layers of the network to extract the global features. Notably, these features do not correlate with keypoints, which are critical for accurate 6-DoF pose estimation.

In this paper, we present a novel method called SuperGF, which obtains a global feature capable of accurate image retrieval from an image's local features capable of accurate 6 DoF pose estimation. SuperGF achieves this by aggregating the local features, similar to the techniques employed in BoW [12], [13], ASMK [9] and Fisher Vector [14], but with low computational costs aimed at reducing the overall computational expenses. Inspired by recent successful applications in SOLAR [15] and TransVPR [10], we employ Transformer as the core component of SuperGF. By applying the Transformer's self-attention to the local features as tokens, we facilitate mutual interactions between them, enabling it to learn to extract good global representation capable of accurate retrieval from the local image features. We show the effectiveness of the approach through experimental results on several public benchmarks.

II. Related Works

A. Approaches for Visual Localization

Structure-based Localization. Previous visual localization approaches mainly rely on estimating correspondences between 2D keypoints in the query and 3D points in a sparse model using local descriptors. The map is usually composed of a 3D point cloud constructed via Structure-from-Motion (SfM), where each 3D point is associated with one or more local feature descriptors. The query pose is obtained by feature matching and solving a Perspective-n-Point problem (PnP) [16]. However, direct matching methods tend to be resource-intensive or fragile and challenging to apply in large-scale localization.

Image-based Localization. Visual localization in large-scale urban environments is often approached as an image retrieval problem. Specifically, the location of a given query image is predicted by transferring the geotag of the most similar image retrieved from a geotagged database [4], [17], [18], [19], [20], [21], [22]. This approach scales to entire cities thanks to compact image descriptors and efficient indexing techniques and can be further improved by spatial re-ranking [7], informative feature selection [23], [24] or feature weighting [25], [26], [20], [22]. Image-based localization approaches have recently shown promising results in robustness and efficiency but are not competitive in terms of accuracy [4], [27], which output only an approximate location of the query, not an exact 6-DoF pose.

Hierarchical Localization. Hierarchical localization takes an approach, dividing the problem into a global, coarse search followed by a fine pose estimation. It shows advantages in terms of efficiency and accuracy compared to the above two approaches, which can be applied to large-scale. The intermediate retrieval step of hierarchical localization limits the downstream feature matching to a reasonable range, which reduces the computational cost significantly while improving the localization performance by reducing the influence of feature repetition. [28] proposed to search at the map level using image retrieval and localize by matching hand-crafted local features against retrieved 3D points. However, its robustness and efficiency are limited by the underlying local descriptors and heterogeneous structure. Taira et al. applied learning-based features to camera pose estimation but in a dense, expensive manner [3]. Recently, HF-Net [1] integrated learning-based models of image retrieval and image-matching by model distillation that simultaneously predicts keypoints as well as global and local descriptors for accurate 6-DoF localization.

B. Global and Local Image Features

Before the emergence of deep learning, hand-crafted local features, such as SIFT [6], ORB [29], and SURF [30], are widely applied in computer vision fields such as image matching. Moreover, traditional aggregation methods [13], [14] are developed for generating global image features for image retrieval using these hand-crafted local features. However, hand-crafted local features are limited in invariance due to only involving low-level information.

Recent features emerging from convolutional neural networks (CNN) exhibit unrivaled robustness at a low computing cost. However, it tends to be task-specific. Specifically, task-specific local or global image features are generated using different models in an end-to-end manner. Even though they achieve superior performances in their respective domain, such as image retrieval [15], [10], [11], [7] or image matching [5], [31], [32], [33], [34], there are still problems in unifying for multi-task.

More recently, Transformer has been adopted for feature extraction in computer vision fields and achieved state-of-art performances [15], [10]. It benefits from the desirable property of the self-attention mechanism, which can naturally aggregate task-relevant features. Recent studies have applied Transformer to each component method for visual localization, i.e., image retrieval [15], [10], [35] and image matching [31], [36].

III. METHOD

A. Overview

In this section, we introduce SuperGF, a novel method that aggregates local features derived from an input image to obtain a global feature that encapsulates a comprehensive image representation. Figure 2 provides an overview of SuperGF.

Initially, sparse local features are extracted from the input image. SuperGF is designed to accommodate any type of local features; in our experiments, we opted for SuperPoint [5] owing to its exemplary performance and popularity. These features, consisting of a variable number of local descriptors identified at sparse keypoints in the input image, are fed into SuperGF.

Subsequently, the local descriptors are transformed into tokens, enriched with data such as image position and confidence score. This step aims to facilitate their processing within a Transformer network at a later stage. To enhance computational efficiency, we utilize the slot attention mechanism, reducing the number of resultant cluster centers to a smaller set, denoted as N, akin to "visual words." The slot attention was initially proposed for object detection [37], and later adopted for visual place recognition [35].

Following this, the N tokens are channeled through a Transformer encoder, encouraging interaction among them to derive an improved holistic image representation. In the final step, the output tokens undergo GeM pooling, culminating in a unified vector representing the global feature of the input image.

B. Augmenting and Compressing Local Features

Given an input image, we first extract a set of local features, which consists of n pairs of keypoints and descriptors. Let i be the index of keypoints ($i = 1, \ldots, n$). We denote the image position, descriptor, and confidence score of the i-th keypoint by $p_i \in \mathbb{R}^2$, $d_i \in \mathbb{R}^d$, and $r_i \in \mathbb{R}^1$, respectively. Here d is the size of the descriptors; e.g., $d = 256$ for SuperPoint.

We first augment the descriptor d_i with the side information, obtaining $t_i \in \mathbb{R}^d$ of i-th local feature. Specifically, we embed the position p_i and the score r_i into a d-dimensional vector t_i using a multi-layer perceptron (MLP) as

$$t_i = d_i + \mathrm{MLP}_{\mathrm{enc}}(p_i, r_i). \tag{1}$$

We use a two-layer fully-connected network with 256 hidden units in our experiments.

To improve the efficiency of subsequent computations while maintaining the token's representational power, we extract $N(< n)$ representatives from the augmented tokens $t_1, \ldots, t_n$. For this purpose, we adopt the slot attention mechanism; it was originally developed for object detection [37] and then utilized for VPR [35].

The slot attention acts as a clustering algorithm, finding cluster centers from a set of vectors. It is applied to our problem as follows. The slot attention mechanism receives $T = [t_1, \ldots, t_n]^\top (\in \mathbb{R}^{n \times d})$ and a set of 'templates' denoted by $Y^0 \in \mathbb{R}^{N \times D}$ as inputs, and then outputs updated templates $Y^L \in \mathbb{R}^{N \times D}$. Specifically, it iteratively updates Y^{l-1} to Y^l using cross-attention from Y^{l-1} to T for $l = 1, \ldots, L$; we set $L = 6$ in our experiments. In the cross attention, the query is obtained from Y^{l-1} with a linear mapping. The key and value are obtained from T with two different linear mappings. Along these, there is an MLP in the slot attention mechanism. Thus, the learnable parameters are i) the three linear mappings to obtain the query, key, and value, ii) the MLP, and iii) the initial templates Y^0. The final output is the updated template Y^L and a set of attention scores W^i ($i = 1, \ldots, N$). Y^L contains the N cluster centers, or equivalently N most representative vectors of the input set T of the augmented local features.

C. From Local to Global Features

The output Y^L of the slot attention is further processed with a self-attention mechanism to make them interact with each other to yield a better global representation. We employ the standard Transformer architecture for the self-attention, which consists of an encoder with three Transformer blocks with layer normalization, a multi-head self-attention (MSA) mechanism, an MLP, and another layer normalization, in that order from input to output ([38], [39]). In our experiments, we set the dimension D of tokens to 512 and the number of heads in MSA to 4; we use a two-layer network with 1,024 hidden units for the MLP. The set Y^L of tokens is fed into the Transformer encoder, yielding a set of updated tokens, denoted by $X \in \mathbb{R}^{N \times D}$. Finally, X is aggregated to a global image feature by generalized mean (GeM) pooling [40], which has a learnable parameter p initialized with $p = 3$. The aggregated global feature $O \in \mathbb{R}^D$ is given by:

$$O = \mathcal{N}_{L_2}(\mathrm{GeM}(X, p)), \tag{2}$$

where $\mathcal{N}_{L_2}$ is L_2 normalization and $\mathrm{GeM}(X; p) = (\sum_{i=1}^{N} (X^i)^p / N)^{1/p}$, where X^i is the i-th row vector of $X (\in \mathbb{R}^{N \times D})$.

D. Training

While SuperGF aims to enhance the computational efficiency of precise visual localization, its training objective is identical to that of the standard VPR. The distinction is that SuperGF employs local image features as inputs. The shared goal is to develop a model capable of generating a global image feature that encapsulates the unique characteristics of a scene depicted in an image. More precisely, when ranking images in a database based on their similarity to a query image, as measured by their global features, we want the correct image(s) to attain the highest possible rank in the sorted result.

In traditional VPR methods, it is common to presume the existence of a single positive sample for each query and to conduct metric learning by minimizing contrastive or triplet loss ([41], [42], [43]). This process aims to bring the query and the positive sample closer while distancing it from negative samples in the global feature space. However, as suggested in [44], a more nuanced evaluation of the

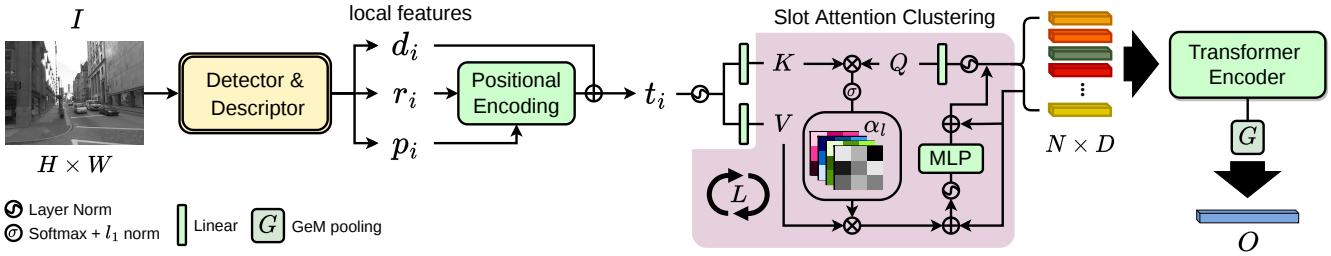


Fig. 2. The framework of SuperGF. SuperGF works on sparse local features used for image matching, i.e., keypoints (p_i), descriptors (d_i), and confidence scores (r_i). The modules indicated by green baskets contain learnable parameters.

ranking results' quality is desirable; see also [45], [33]. Furthermore, as noted in [27], the binary categorization of images into positive and negative samples for each query does not accurately reflect the spatial continuity observed in real-world scenes. In response to these findings, we employ the average precision (AP) loss of [44], replacing traditional metric learning losses, and adopt soft image similarity scores based on field-of-view (FoV) overlap utilized in [27], instead of employing binary scores.

The details are as follows. We first categorize the database images into three classes per query image, depending on the FoV overlap between images [27]. Specifically, images possessing a FoV overlap within the range of $[0.5, 1]$ are classified as positive samples, those with a FoV overlap falling within $(0, 0.5)$ are identified as soft negatives, and those with a FoV overlap of 0 are designated as (hard) negatives. We then choose for each query image a single positive sample, α soft negative samples, and β hard negative samples, with values $\alpha = 2$ and $\beta = 100$ utilized in our experiments. All these images are input to SuperGF to compute their global features. We use cosine similarity for the similarity measure. Let the similarity between the query image and the single positive and $\alpha + \beta$ negative samples be denoted by $S = [s_1, \ldots, s_{\alpha+\beta+1}]$. Denoting their ground truth similarity score by $\bar{S} = [\bar{s}_1, \ldots, \bar{s}_{\alpha+\beta+1}]$, we compute the AP loss [44] between S and $\bar{S}$ as $L_{\text{AP}} = 1 - \text{AP}(S, \bar{S})$.

We utilize the slot attention mechanism for computational efficiency, as explained in Sec. III-B. To facilitate its training, we use an attention decorrelation loss introduced in [35]. It aims to reduce the spatial correlation between attention maps, i.e., the output W^i $(i = 1, \ldots, N)$ from the slot attention mechanism. In other words, the loss encourages different maps to attend to different image regions. Specifically, for the N attention maps $[W^1, ..., W^N]$, the attention decorrelation loss is given by

$$L_{\text{attn}} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{W^i \cdot W^j}{\|W^i\|_2 \, \|W^j\|_2}, \tag{3}$$

where $i, j \in \{1, ..., N\}$.

IV. Experimental Results

We conducted experiments to compare SuperGF with existing state-of-the-art methods. Although our primary focus is on visual localization, we also tested the methods on VPR to demonstrate the independent effectiveness of SuperGF as a global feature.

A. Training Procedure

Following previous studies, we utilize the training set from the Mapillary Street-level Sequences (MSLS) [46], as mentioned in Sec.III-D. The training is conducted over 350 epochs; in each epoch, we randomly select 10,000 query samples from all sub-cities. The input image dimensions are set to 640×480 pixels. For SuperGF, we incorporate SuperPoint as the local feature and employ the AdamW optimizer with a weight decay parameter of 10^{-4}. The learning rate initiates at 10^{-4} and gradually decays to 10^{-6} over the training period. For all the other methods, we retain the settings as specified in the official codebases.

B. Results on Visual Localization

Datasets. Following recent studies, we evaluate methods on three datasets, the Aachen Day-Night [47], [48], Extended CMU [49], [50], and RobotCar Seasons datasets [51], [47]. The Aachen Day-Night dataset contains images captured using handheld cameras, while the RobotCar and Extended CMU datasets contain images captured using car-mounted cameras. These images were captured in different seasons, weather conditions, and locations of urban and rural areas.

Compared methods. Besides SuperGF, we consider hierarchical localization methods that combine NetVLAD with two different local descriptors, namely SIFT and SuperPoint, referred to as NV+SIFT and NV+SP, respectively. Specifically, NV+SIFT and NV+SP utilize global features extracted by NetVLAD and local features extracted by SIFT and SuperPoint, respectively. As in VPR, SuperGlue can be employed with SuperPoint. Therefore, we also investigate the method that uses SuperGlue for the pose estimation step, denoted by NV+SP+SG. Additionally, we consider two recent structure-based localization methods, Active Search (AS)[52], and City Scale Localization (CSL)[53]; see also Sec. II-A.

Experimental setting. Following previous studies, we adopt the hierarchical localization procedure of HLoc [1], provided by an open-source toolbox[1]. Specifically, we evaluate the performance of each method as follows. We first apply the method to obtain the global features of the query and the

[1] https://github.com/cvg/Hierarchical-Localization

TABLE I

LOCALIZATION RESULTS ON THREE BENCHMARKS. WE REPORT THE MEDIAN TRANSLATION (M) AND ROTATION (°) ERRORS AND THE AVERAGE RECALL [%] AT THREE THRESHOLDS, I.E., (0.25M, 2°), (0.5M, 5°), AND (5M, 10°). THE SYMBOL † DENOTES THAT THE NUMBER IS COPIED FROM THE ORIGINAL PAPER.

Method	Aachen Day-Nigh		RobotCar Seasons		Extended CMU Seasons		
	Day	Night	Day	Night	Urban	Suburban	Park
AS	85.3† / 92.2† / 97.9†	39.8† / 49.0† / 64.3†	50.9† / 80.2† / 96.6†	6.9† / 15.6† / 31.7†	81.0† / 87.3† / 92.4†	62.6† / 70.9† / 81.0†	45.5† / 51.6† / 62.0†
CSL	52.3† / 80.0† / 94.3†	29.6† / 40.8† / 56.1†	45.3† / 73.5† / 90.1†	0.6† / 2.6† / 7.2†	71.2† / 74.6† / 78.7†	57.8† / 61.7† / 67.5†	34.5† / 37.0† / 42.2†
NV+SIFT	82.8 / 88.1 / 93.1	30.6 / 43.9 / 58.2	54.5 / 79.2 / 95.4	7.1 / 13.6 / 21.4	75.4 / 81.4 / 88.0	59.0 / 66.6 / 77.2	37.2 / 43.2 / 53.2
NV+SP	86.0 / 93.7 / 96.8	68.4 / 83.7 / 94.9	55.9 / 80.7 / 96.6	9.3 / 16.2 / 24.9	89.5 / 94.2 / 97.9	76.5 / 82.7 / 92.7	57.4 / 64.4 / 80.4
NV+SP+SG	88.2 / 95.4 / **98.7**	86.7 / **92.9** / **100.0**	56.3 / **81.1** / **97.5**	19.5 / 37.0 / 50.3	**95.5** / **98.6** / **99.2**	**90.9** / 94.2 / 97.1	85.4 / 88.9 / 91.4
SuperGF-NN	87.4 / 94.4 / 97.6	70.3 / 85.7 / 96.9	55.6 / 81.0 / 95.5	13.3 / 21.2 / 39.6	89.9 / 94.3 / 97.7	75.4 / 82.5 / 93.5	66.2 / 74.7 / 85.2
SuperGF-SG	**89.4** / **95.4** / 98.6	**88.7** / 92.8 / **100.0**	**56.7** / **81.1** / 96.6	**23.7** / **47.6** / **69.4**	94.9 / 98.1 / 99.1	90.7 / **95.6** / **97.3**	**87.4** / **89.8** / **92.3**

TABLE II

COMPARISONS IN THE MEMORY REQUIREMENT AND THE AVERAGE LATENCY OF DIFFERENT METHODS.

Method	Memory (MB)	Extraction latency (ms)		
		Global	Local	Total
NetVLAD / SFRS + SP	150.30	14.4	6.5	20.9
ResNet50-GeM + SP	24.81	11.1	6.5	17.6
SOLAR + SP	57.45	19.5	6.5	26.0
HF-Net	32.66	–	–	**8.1**
SuperGF	**7.72**	5.2	6.5	11.7

database images. We perform nearest neighbor search in the space of the global feature as in VPR, obtaining 10, 20, and 50 candidates for the Extended CMU, RobotCar, and Aachen Day-Night datasets, respectively, following the original setting of HLoc. We then establish putative point correspondences from the query to the selected candidates using either nearest neighbor search in the local feature space or SuperGlue, denoted by '-NN' or '-SG', respectively, in the results shown below. We then perform PnP to estimate the camera pose of the query image using the matched points in the given 3D model.

Metrics. We adopt the standard evaluation metrics used in previous studies ([1], [3], [36], [54]), i.e., median translation error (in meters) and median rotation error (in degrees). We set three different thresholds for the errors and calculate the average recalls using these pairs of thresholds, i.e., (0.25m, 2°), (0.5m, 5°), and (5m, 10°).

Results. Table I presents the results, which allow us to make several observations. First, SuperGF performs comparably to the state-of-the-art NetVLAD variants across all datasets. Notably, it exhibits superior performance in challenging scenarios, particularly in night scenes of the RobotCar dataset, underscoring the effectiveness of SuperGF. Additionally, while two structure-based methods, AS and CSL, show competitive performance in easy cases, e.g., matching day and day images, where localization accuracy tends to saturate, they tend to be outperformed by NetVLAD and our SuperGF in more challenging cases.

Table II reports the running time and memory requirements of the compared methods. To assess the running time of each method, we randomly selected 1,000 query images from the MSLS [46] validation set and calculated the average time necessary to extract both global and local features using each method. It should be noted that the times reported do not include the time taken for matching local/global features. For a fair comparison, we standardized the input image size to 640×480 pixels.

As illustrated in Table II, SuperGF holds a distinct advantage over other methods in terms of both model size and latency. SuperGF efficiently generates the global feature of an input image solely using its local features, thereby reducing computational expenses. In contrast, other methods extract global features directly from input images, necessitating the additional step of extracting local features from the images, thereby increasing computational demands. While HF-Net [1] offers quicker inference speeds, it compromises inference accuracy, as indicated in [1]. This approach, which utilizes multitask distillation for training, yields less precise outcomes than its teacher model, namely, NV+SP. Considering the data in Table I, it becomes clear that SuperGF provides a superior balance between inference accuracy and computational efficiency, affirming the efficacy of our approach.

C. Results on Visual Place Recognition

We also evaluate SuperGF on the visual place recognition (VPR) task. Note that since VPR does not necessitate local features that are capable of precise 6-DoF pose estimation, SuperGF will be at a disadvantage when compared to native VPR methods.

Datasets. We employ several public benchmark datasets, namely MSLS [46], Pitts30k [22], Nordland [56], and Tokyo247 [21] to test the models trained as above. These are large-scale datasets containing diverse appearance variations, including day-night, weather, and seasonal changes, which pose significant challenges.

Compared methods. We experimentally compare several state-of-the-art and a few baseline methods, i.e., NetVLAD [4], several state-of-the-art CNN-based methods, including SFRS [55], ResNet50-GeM-GCL [27] and also two state-of-the-art Transformer-based methods, SOLAR [15] and TransVPR [10]. To evaluate each model, we follow previous studies ([7], [10], [11]). Specifically, we use L_2 distance in the global image feature space to retrieve candidate images from the database.

Metrics. Following previous studies of VPR, we use Recall@K metric, which computes the percentage of query

TABLE III

RESULTS FOR THE SINGLE-PASS RETRIEVAL OF VPR. WE REPORT RECALL@K WITH $K = 1$, 5, AND 10. THE SYMBOL † DENOTES THAT THE NUMBER IS COPIED FROM THE ORIGINAL PAPER.

Method	MSLS val			MSLS challenge			Pitts30k test			Nordland test			Tokyo247 test		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD [4]	58.2	71.4	75.2	33.5	45.1	49.4	85.1	92.2	94.5	21.0	33.5	39.2	67.0	77.8	80.3
SFRS [55]	69.2	80.3	83.1	41.5	52.0	56.3	**89.4**	**94.7**	**95.9**	18.8	32.8	39.8	**81.0**	**86.4**	**89.8**
ResNet50-GeM-GCL [27]	66.2†	78.9†	81.9†	43.3†	59.1†	65.0†	72.3†	87.2†	91.3†	27.2	41.1	49.2	44.1†	61.0†	66.7†
SOLAR [15]	78.3	87.2	89.6	44.6	58.6	63.2	85.4	92.6	94.8	41.8	58.4	66.1	76.2	84.4	88.3
TranVPR [10]	70.8†	85.1†	89.6†	48.0†	67.1†	73.6†	73.8†	88.1†	91.9†	15.9†	38.6†	49.4†	–	–	–
SuperGF	**78.5**	**88.6**	**91.3**	**56.8**	**69.9**	**76.0**	78.7	89.5	92.6	**57.5**	**77.4**	**87.5**	68.5	78.2	84.1

images that are correctly localized. The retrieval for a query image is considered successful if at least one of the top K ranked reference images is within a pre-defined distance threshold from the ground truth location of the query image. The distance thresholds are provided by the datasets ([11], [10], [4].) Following previous studies, we report Recall@K with $K = 1$, 5, and 10.

Results. Table III. Several key observations emerge from these findings. Firstly, the effectiveness of different models fluctuates markedly across various datasets, with some only achieving low accuracy. This underlines the inherent challenges of image retrieval and the limitations encountered without implementing geometric verification. Secondly, and of primary significance, our SuperGF approach outperforms competing methods on three datasets and matches them on another two, thus establishing its robust capabilities as a global feature extractor. This is notable, especially considering that it builds upon given local features instead of originating from an unprocessed image. We must note that the local feature leveraged here is SuperPoint, whose parameters were held constant during the SuperGF training phase.

D. Understanding How SuperGF Works

SuperGF aggregates local features to develop a global feature, thereby facilitating precise image retrieval. However, as previously highlighted, these local features are not inherently designed for holistic image feature extraction. The pertinent question is: how can SuperGF harness useful information from these local features for effective image retrieval?

A partial answer is found in SuperGF's method of generating and utilizing attention weights on the input local features, a process detailed in Sec. III. Figure 3 shows a visualization of the attention weights applied to the input local features (or keypoints) on several sample images. It is apparent that SuperGF prioritizes certain local features that potentially offer significant cues for image retrieval while disregarding less relevant ones, shedding light on the functioning of the SuperGF mechanism.

V. SUMMARY AND CONCLUSION

This paper introduced a novel method for the visual localization task, offering comparable inference accuracy to existing state-of-the-art methods while enhancing computational efficiency. The visual localization process unfolds in two steps: initial candidate image retrieval followed by camera pose estimation, referencing the scene's 3D model. The first step requires global image features, while the second leverages local image features.

Fig. 3. Visualization of the attention weights computed by SuperGF on the input local features. Red dots indicate local features with high weights and cyan dots indicate those with low weights.

Current methodologies achieve high accuracy through optimized feature extraction processes; however, they are computationally demanding, processing input images twice for feature extraction. To address this, we proposed SuperGF, a system that capitalizes on the local features used in the second step, aggregating them to formulate a global image feature. This global feature, which portrays the holistic image appearance, enables precise image retrieval. The ingenuity of SuperGF lies in the selection and combination of local features to create a global feature, facilitated through the use of Transformer and slot attention mechanisms, coupled with a well-constructed loss function for training.

Experiments on standard benchmark datasets demonstrate the effectiveness of the proposed approach. It is noteworthy, in the visual place recognition task, where the absence of local feature reliance seemingly puts SuperGF at a disadvantage, it still manages to deliver performance comparable to specialized VPR methods.

VI. ACKNOWLEDGMENTS

This work was partly supported by JSPS KAKENHI Grant Number 23H00482 and 20H05952.

REFERENCES

[1] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[2] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 817–833.

[3] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209.

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

[5] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. CVPRW*, 2018, https://github.com/magicleap/SuperPointPretrainedNetwork.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[7] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *European Conference on Computer Vision*. Springer, 2020, pp. 726–743.

[8] G. Tolias, T. Jenicek, and O. Chum, "Learning and aggregating deep local descriptors for instance-level recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 460–477.

[9] G. Tolias, Y. Avrithis, and H. Jégou, "To aggregate or not to aggregate: Selective match kernels for image search," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1401–1408.

[10] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 648–13 657.

[11] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.

[12] F. Radenović, G. Tolias, and O. Chum, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *European conference on computer vision*. Springer, 2016, pp. 3–20.

[13] ——, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *Proc. ECCV*, 2016.

[14] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[15] T. Ng, V. Balntas, Y. Tian, and K. Mikolajczyk, "Solar: second-order loss and attention for image retrieval," in *European conference on computer vision*. Springer, 2020, pp. 253–270.

[16] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, pp. 155–166, 2009.

[17] R. Arandjelović and A. Zisserman, "Dislocation: Scalable descriptor distinctiveness for location recognition," in *Asian conference on computer vision*. Springer, 2014, pp. 188–204.

[18] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, H. Chen, R. Vedantham, R. Grzeszczuk, and B. Girod, "Residual enhanced visual vectors for on-device image matching," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. IEEE, 2011, pp. 850–854.

[19] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2136–2145.

[20] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1582–1590.

[21] 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, pp. 1808–1817.

[22] 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.

[23] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[24] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *CVPR 2011*. IEEE, 2011, pp. 889–896.

[25] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla, "Learning and calibrating per-location classifiers for visual place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 907–914.

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

[27] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Data-efficient large scale place recognition with graded similarity supervision," *CVPR*, 2023.

[28] 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.

[29] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Proc. ICCV*, 2011.

[30] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. ECCV*, 2006.

[31] P. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proc. CVPR*, 2020, https://github.com/magicleap/SuperGluePretrainedNetwork.

[32] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint detection and description of local features," in *Proc. CVPR*, 2019.

[33] R. Jerome, W. Philippe, R. S. César, and H. Martin, "R2D2: repeatable and reliable detector and descriptor," in *Proc. NeurIPS*, 2019, https://github.com/naver/r2d2.

[34] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "Lf-net: learning local features from images," in *Proc. NeurIPS*, 2018.

[35] P. Weinzaepfel, T. Lucas, D. Larlus, and Y. Kalantidis, "Learning super-features for image retrieval," *arXiv preprint arXiv:2201.13182*, 2022.

[36] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[37] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 525–11 538, 2020.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

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

[42] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 459–474.

[43] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2495–2504.

[44] J. Revaud, J. Almazán, R. S. Rezende, and C. R. d. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5107–5116.

[45] K. Chen, W. Lin, J. Li, J. See, J. Wang, and J. Zou, "Ap-loss for accurate one-stage object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3782–3798, 2020.

[46] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2626–2635.

[47] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8601–8610.

[48] 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.

[49] H. Badino, D. Huber, and T. Kanade, "Visual topometric localization," in *2011 IEEE Intelligent vehicles symposium (IV)*. IEEE, 2011, pp. 794–799.

[50] C. Toft, W. Maddern, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, T. Pajdla, *et al.*, "Long-term visual localization revisited," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2074–2088, 2020.

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

[52] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1744–1756, 2016.

[53] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1455–1461, 2016.

[54] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[55] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li, "Self-supervising fine-grained region similarities for large-scale image localization," in *European conference on computer vision*. Springer, 2020, pp. 369–386.

[56] D. Olid, J. M. Fácil, and J. Civera, "Single-view place recognition under seasonal changes," *arXiv preprint arXiv:1808.06516*, 2018.