POLITECNICO DI TORINO
Repository ISTITUZIONALE

Adaptive-Attentive Geolocalization From Few Queries: A Hybrid Approach

Original
Adaptive-Attentive Geolocalization From Few Queries: A Hybrid Approach / Paolicelli, V; Berton, G; Montagna, F; Masone, C; Caputo, B. - In: FRONTIERS IN COMPUTER SCIENCE. - ISSN 2624-9898. - 4:(2022). [10.3389/fcomp.2022.841817]

Availability:
This version is available at: 11583/2971030 since: 2022-09-07T10:02:57Z

Publisher:
FRONTIERS MEDIA SA

Published
DOI:10.3389/fcomp.2022.841817

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

24 April 2024

frontiers | Frontiers in Computer Science

ORIGINAL RESEARCH
published: 13 June 2022
doi: 10.3389/fcomp.2022.841817

Adaptive-Attentive Geolocalization From Few Queries: A Hybrid Approach

Valerio Paolicelli[1†], *Gabriele Berton*[1*†], *Francesco Montagna*[1], *Carlo Masone*[1,2] *and Barbara Caputo*[1]

[1] *Visual and Multimodal Applied Learning Laboratory, Politecnico di Torino, Department of Control and Computer Engineering, Turin, Italy,* [2] *Consorzio Interuniversitario Nazionale per l'Informatica, Rome, Italy*

We tackle the task of cross-domain visual geo-localization, where the goal is to geo-localize a given query image against a database of geo-tagged images, in the case where the query and the database belong to different visual domains. In particular, at training time, we consider having access to only few unlabeled queries from the target domain. To adapt our deep neural network to the database distribution, we rely on a 2-fold domain adaptation technique, based on a hybrid generative-discriminative approach. To further enhance the architecture, and to ensure robustness across domains, we employ a novel attention layer that can easily be plugged into existing architectures. Through a large number of experiments, we show that this adaptive-attentive approach makes the model robust to large domain shifts, such as unseen cities or weather conditions. Finally, we propose a new large-scale dataset for cross-domain visual geo-localization, called SVOX.

Keywords: domain adaptation (DA), domain generalization, visual place recognition (VPR), few-shot domain adaptation, visual geolocalization

OPEN ACCESS

Edited by:
Sarah Adel Bargal,
Boston University, United States

Reviewed by:
Hongfu Liu,
Brandeis University, United States
Ben Usman,
Boston University, United States

****Correspondence:***
Gabriele Berton
gabriele.berton@polito.it

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Computer Vision, a section of the journal Frontiers in Computer Science

Received: *22 December 2021*
Accepted: *25 April 2022*
Published: *13 June 2022*

Citation:
Paolicelli V, Berton G, Montagna F, Masone C and Caputo B (2022) Adaptive-Attentive Geolocalization From Few Queries: A Hybrid Approach. Front. Comput. Sci. 4:841817. doi: 10.3389/fcomp.2022.841817

1. INTRODUCTION

Visual Geo-localization (VG) is defined as the task of coarsely localizing an image taken from a known environment on the basis of its visual content (Arandjelovic et al., 2018). Such an ability is relevant for numerous applications, either as a self-standing function or in conjunction with additional processing steps for more accurate localization methods (Pion et al., 2020). Examples of these applications include assistive devices (Cheng et al., 2020), autonomous robots and vehicles (Cummins and Newman, 2008; Milford and Wyeth, 2012; McManus et al., 2014; Tomita et al., 2021), the cataloging of archival imagery (Aubry et al., 2014), and augmented reality (Middelberg et al., 2014).

From a methodological perspective, VG is commonly approached as an image retrieval problem where the single image to be localized (*query*) is matched to a large collection of images (*database*) that are tagged with geographical information (usually GPS coordinates and, if available, compass measurements) (He et al., 2015; Kim et al., 2017; Noh et al., 2017; Arandjelovic et al., 2018; Lou et al., 2018; Nakka and Salzmann, 2018; Wang et al., 2019; Cao et al., 2020). The intuition of this approach is that, by retrieving images that depict the same scene as the query and from approximately the same viewpoint, their geo-tags can be taken as hypotheses of the query's location. A crucial part of this process is how to represent the images for the retrieval, i.e., how to extract descriptors that are informative of the location of the images. All the recent state-of-the-art VG methods perform the description step using deep learning techniques (Masone and Caputo, 2021; Zhang et al., 2021)

which, combined with the increasing availability of data for VG (Sünderhauf et al., 2013; Chen et al., 2017a; Maddern et al., 2017; Arandjelovic et al., 2018; Sattler et al., 2018; Torii et al., 2018; Warburg et al., 2020), have led to remarkable results in comparison to non-learned methods. However, these methods perform well when they are tested with queries that come from the same distribution as the training data, but they may not translate well to queries that come from unseen domains. The domain shift problem, which is ubiquitous across deep learning tasks, is particularly important in VG because the appearance of a place naturally and cyclically changes over time, even when viewed from the same perspective. In fact, a scene may appear drastically different due to changes in weather conditions (e.g., rainy/sunny), illumination (e.g., day/night cycle), and season (e.g., summer/winter), as shown in the example depicted in **Figure 1**. Indeed, several studies have shown that these appearance shifts can significantly harm the geo-localization results (Sattler et al., 2018; Zaffar et al., 2019).

So far, the domain shift problem in VG due to environmental changes has been addressed indirectly, either by robustifying the learned visual descriptors (Garg et al., 2018a; Oertel et al., 2020; Peng et al., 2021a,b) or with *ad-hoc* solutions for specific cases (Garg et al., 2018b). Yet, only a few studies have explicitly addressed the problem of visual geo-localization as a task across domains, minimizing the domain discrepancy with an adversarial training (Porav et al., 2018; Anoosheh et al., 2019; Hu et al., 2021) or a multi-kernel Maximum Mean Discrepancy loss (Wang et al., 2019). These previous studies require a non-negligible number of images from the target domain—at least a few hundred—in order to train the model. Considering that the process of collecting images is expensive, it would be advisable to reduce the number of required target images in order to make the VG models more scalable and easier to deploy. In light of these considerations, we propose the first few-shot domain adaptation architecture for VG. The two variants of our method, named AdAGeo and AdAGeo-Lite, combine three main ideas:

- A domain-driven data augmentation module that transfers the style from few unlabeled target domain images to the labeled training queries in the source domain. This effectively creates a set of pseudo-target labeled images that can be used for training.
- A domain adaptation module that aligns the features extracted by the model on the source and target images, making them invariant (to some extent) to the domain shift.
- An attention module based on class-specific activation maps that induce the model to focus on the elements in the scene that are most informative and stable across domains (e.g., buildings).

Given the lack of large-scale cross-domain visual geo-localization datasets, we also build and publicly release the first dataset of such kind, to foster future research in the field. This dataset, named Street View Oxford (SVOX), has been built by collecting images of the city of Oxford from Google Street View, as a source domain (database and training queries). Instead, the target domain queries are taken from Oxford RobotCar (Maddern et al., 2017), which contains images from multiple weather conditions (snow, rain, sun, night, and overcast).

Through extensive experiments, we demonstrate that our solution exceeds the current state-of-the-art visual geo-localization methods across domains and that the three core ideas of the architecture (domain-driven data augmentation, attention, domain adaptation) provide orthogonal increments to the performance.

This article extends the study presented in Berton et al. (2021b) with several contributions:

- It replaces the domain-driven data augmentation solution adopted in Berton et al. (2021b), which was inspired by Cohen and Wolf (2019) and used two parallel autoencoders to learn a bi-directional mapping between source and target domain. Instead, we resort to a non-learnable style transfer module based on a Fourier transform (Yang and Soatto, 2020). This new domain-driven data augmentation module, which constitutes the main novelty in AdAGeo-Lite with respect to AdAGeo, not only removes the need for the two-step training procedure used in Berton et al. (2021b), greatly simplifies the pipeline, but it also produces geo-localization results that are either comparable or better than those reported in Berton et al. (2021b).
- It includes an extended suite of experiments, aimed at assessing the performance of our solution when deployed to a target domain of which only few unlabeled queries were seen during training (domain adaptation) and when deployed to completely different cities and domains unseen during training (domain generalization).

2. RELATED WORKS

Visual Geo-Localization Across Domains

The task of visual geo-localization, also known by the name Visual Place Recognition (Lowry et al., 2016), is a long lasting topic of research, originally using handcrafted features to perform the retrieval (Murillo and Kosecka, 2009; Johns and Guang-Zhong, 2011; Sünderhauf and Protzel, 2011; Kim et al., 2015), and more recently combining deep convolutional neural networks as feature extractors with trainable aggregation modules (Gordo et al., 2017; Arandjelovic et al., 2018) or pooling layers (Radenovic et al., 2019) to produce global image descriptors. The success of learning based methods in visual geo-localization is tightly connected to the availability of increasingly large and diverse datasets for the task (Chen et al., 2017a; Maddern et al., 2017; Warburg et al., 2020). At the same time, these recent datasets that include diverse meteorologic/illumination conditions have exposed the lack of robustness of VG methods with respect to environmental changes. The problem of cross-domain VG has mostly been addressed indirectly and with a limited scope, with approaches that are based on heuristics (e.g., selecting features corresponding to man-made structures Naseer et al., 2017), on regions of interest (Chen et al., 2017b), on additional semantic information (Peng et al., 2021a,b), use attention to focus on robust structures (Kim et al., 2017; Noh et al., 2017; Lou et al., 2018; Nakka and

FIGURE 1 | The appearance of a place viewed from the same perspective can drastically change in different environmental conditions. **(A)** Image from the city of Oxford taken from Google StreetView. **(B–F)** Images taken from of the Oxford RobotCar dataset (Maddern et al., 2017), that show the same place as in A but in different conditions, i.e., snow, rain, sun, night and overcast, respectively.

FIGURE 2 | (A) Area of the city of Oxford covered by SVOX and RobotCar (Maddern et al., 2017) datasets, respectively. **(B–D)** pairs of images collected from Google StreeetView Time Machine, which view the same places but in different years, demonstrating how long-term temporal variations Iter the appearance of a place.

Salzmann, 2018; Zhu et al., 2018), or are tailored for a specific domain shift (e.g., day/night Garg et al., 2018b; Torii et al., 2018).

Only few previous studies have explicitly tackled the appearance shift in VG as a domain adaptation problem. In particular, Porav et al. (2018) and Anoosheh et al. (2019) both use GANs to replace the query with a synthetic image that depicts the same scene but with the appearance of the source domain. These methods not only require hundreds or more target images, but they may also require a small portion of target images to be aligned with source images (Anoosheh et al., 2019). The authors of Wang et al. (2019) instead use MK-MMD (Gretton et al., 2012) for domain adaptation and allow the localization of old grayscale photos against a gallery of present-day images. Both source and target datasets are not available at the time of this writing. Hu et al. (2021) instead use adversarial training to address a synthetic-to-real shift, which arises because their architecture

requires depth and pixel-wise semantic labels which are not readily available in VG datasets with real world images. While these prior studies use either a generative approach or a domain adaptation method, we combine both paradigms and show that they provide complementary improvements. Moreover, unlike previous methods, AdAGeo and AdAGeo-Lite are truly few-shot domain adaptation solutions that require as few as 5 unlabeled and not aligned images from the target domain to produce convincing results.

Domain Adaptation

Unsupervised domain adaptation attempts to reduce the shift between the source and target distribution of the data by relying only on labeled source data and unlabeled target data. There are two typical approaches that are used for unsupervised domain adaptation in computer vision. The first approach is based on learning a style-transfer transformation to map images from one domain to the other. The cross-domain mapping is usually learned through GANs, as in Hong et al. (2018), Huang et al. (2018), or autoencoders (Shang et al., 2017). Zhu et al. (2017) propose to use a cycle-consistency constraint to learn a meaningful translation, which has since been used in several tasks (Benaim and Wolf, 2017; Hoffman et al., 2018; Russo et al., 2018; Fu et al., 2019). Recently, it has also been shown that a simple non-learned alignment of the low-level statistics between the source and target distributions can improve performance in UDA (Yang and Soatto, 2020).

The second approach is based on learning domain-invariant features from the data, building on the idea that a good cross-domain representation contains no discriminative information about the origin (i.e., domain) of the input. This approach was introduced by Ganin and Lempitsky (2015), where a domain discriminator network and the gradient reversal layer (GRL) forces the feature extractor to produce domain-invariant representations. This method found successful applications in many tasks, such as object detection (Ganin and Lempitsky, 2015), semantic segmentation (Bolte et al., 2019), and video classification (Chen et al., 2019). As an alternative, Xu et al. (2019) shows that features with larger norms are more transferable across domains and propose to increasingly enlarge the norms of the embedding during training.

In this study, we integrate both kinds of approaches in a unique pipeline that only needs few samples from the target domain.

3. DATASET

In order to address the cross-domain VG problem, we need a dataset that supports different domains between database (source) and queries (target). In recent years, there have been few VG datasets that include multiple ambient conditions (weather, seasons, lighting) (Sünderhauf et al., 2013; Chen et al., 2017a; Maddern et al., 2017; Arandjelovic et al., 2018; Sattler et al., 2018; Warburg et al., 2020). However, we find that each of these datasets comes with some limitations: Tokyo 24/7 (Arandjelovic et al., 2018) has a limited number of domains, and few queries; Oxford RobotCar (Maddern et al., 2017) spans a limited geographical area; Mapillary SLS (Warburg et al., 2020) is a collection of sequences, and does not densely cover a given area; Nordland (Sünderhauf et al., 2013) is built from sequences collected by a train-mounted car, with very little urban scenery. To fill the void for a multi-domain dataset that densely covers a large urban environment, we propose the Street View Oxford (SVOX) dataset which encloses the whole city of Oxford (refer to **Figure 2A**).

TABLE 1 | Sizes of SVOX dataset and Oxford RobotCar (Maddern et al., 2017) from 5 different scenarios.

	SVOX		RobotCar				
	Gallery	**Queries**	**Snow**	**Rain**	**Sun**	**Night**	**Overcast**
Train	22,232	11,294	750	714	712	702	705
Val	17,226	14,698	-	-	-	-	-
Test	17,166	14,278	937	870	854	823	872

For the source domain, we relied on Google StreetView: we took images from 2012 for the database, and images from 2014 as training queries (refer to **Table 1**), making sure that for each query there is at least one positive sample in the gallery. Using gallery and queries taken from 2 years apart helps to train the networks in a robust way, and ensures that they focus on long-term elements, instead of short-term or changing elements such as vegetation or scaffolding. **Figure 2** illustrates some examples of these long-term temporal variations. Given that Street View imagery provides 360° equirectangular panoramas at various resolutions, we cropped two rectangles on opposite sides for each panorama, corresponding to the front and rear views of the car.

To investigate VG across domains we also need images of the city of Oxford taken in different environmental conditions. For this purpose, we use samples taken from the Oxford RobotCar dataset (Maddern et al., 2017), in which images are conveniently tagged according to their weather or lighting conditions. Specifically, RobotCar provides five domains: Snow, Rain, Sun, Night, and Overcast (refer to **Figure 1** for an example of the differences among these domains and the source). Note that, besides the environmental conditions, the RobotCar images also differ from the source domain for the viewpoint, which is clear by the fact that the hood of the car is visible in the foreground (cf., **Figure 1**). Similarly to Piasco et al. (2019), we take one image every 5 m in order to avoid redundant data, e.g., collected when the car was still at a traffic light. We further note that the RobotCar dataset was collected between 2014 and 2015, few years apart from the SVOX database, which adds more temporal variations between source and target domains. In the process of sampling the images from RobotCar, we ensure that for each target query there is at least one positive sample in the database built from StreetView, i.e., an image within 25 m of distance. Eventually, this procedure results in roughly 1,500 images per domain which are then divided in to two sets: few images, without labels, are used for domain adaptation, whereas the rest are used to test the models on different target domains.

Finally, we provide train, validation, and test splits of SVOX, as reported in **Table 1**. The images sampled from the RobotCar dataset (Maddern et al., 2017) are included

only in the train set (without labels, to be used for domain adaptation) and in the test set (with labels, to assess the inference across domains). All images, both from source and target domains, were isotropically resized to 512x384 pixels. The dataset can be found at this link https://drive.google.com/file/d/16iuk8voW65GaywNUQlWAbDt6HZzAJ_t9/view.

4. METHOD

In this section, we present our solution for domain adaptive and attentive visual geo-localization of outdoor images. Given the nature of the task, test-time images are likely to come from different domains that the source, where the most common domain shifts are caused by illumination (day/night) and weather changes. These domain shifts can lead to drops in accuracy when the model is deployed in the real world, which represents an open and challenging problem for previous visual geo-localization methods (Arandjelovic et al., 2018), which typically train their network on a single-domain dataset with little variability (e.g., StreetView imagery). Given that unlabeled target-domain images are cheap to obtain and can be massively collected, we propose to tackle the issue by combining multiple strategies: i) a domain-driven data augmentation module, to generate images close to the target domain, ii) an attention layer that provides robustness to domain shifts, iii) a domain adaptation layer, to maximize the similarity between features of the different domains.

These strategies are combined in a modular architecture, that is composed of two parts (**Figure 3**):

- A domain-driven data augmentation module (Section 4.1) that transfers the style of the target domain to the source images. In particular, we present two variants of this module: one that uses two autoencoders and the second that uses a non-learned style transfer method. We call AdAGeo the overall architecture using the learned DDDA and AdAGeo-Lite the one using the non-learned DDDA.
- A network that produces the image descriptors. This network is composed of a CNN encoder, to extract features, an attention layer to give focus on the salient parts of the scene (Section 4.2), and an aggregation layer (Section 4.3) that builds the final embedding. This network is trained with a domain adaptation strategy (Section 4.4), using the labeled source and augmented images as well as the few unlabeled queries from the target domain.

Both these parts necessitate only a few images from the target domain, thus making AdAGeo and AdAGeo-Lite truly few-shot architectures for VG across domains. Formally, we consider the unsupervised domain adaptation scenario in which we have a labeled dataset $X^s = \{(x_i^s, y_i^s)\}_{i=1}^{n^s}$, where x_i^s is an image from the source domain D_s and y_i^s is its geo-tag of the image and an unlabeled target dataset $X^t = \{(x_j^t)\}_{j=1}^{n^t}$ made of samples from target domain D_t, with $n^t \ll n^s$.

4.1. Domain-Driven Data Augmentation

The purpose of the domain-driven data augmentation (DDDA) module is to find a mapping $D^s \mapsto D^{pt}$ from the source domain to a pseudo-target domain that better approximates the target domain, i.e., $D^{pt} \approx D^t$. This mapping can then be applied to the source dataset X^s to generate a new labeled dataset with pseudo-target images $X^{pt} = \{(x_i^{pt}, y_i^{pt})\}_{i=1}^{n^{pt}}$. Although X^{pt} has the same dimensionality and labels as X^s, it has a smaller discrepancy w.r.t. the target domain D^t.

The creation of the pseudo-target dataset is a data augmentation technique performed only once, offline. Afterward, we use both X^s and X^{pt} to train the descriptor extraction network, which not only leads to a more robust model but it also results in faster convergence of the training. In the rest of this section, we present two different implementations of the DDDA module: an approach that uses two autoencoders to learn a bi-directional mapping between D^s and D^{pt} (c.f. Section 4.1.1), and a non learnable method based on a Fourier transform (c.f. Section 4.1.2). These two DDDA methods are the distinctive difference between AdAGeo and AdAGeo-Lite.

4.1.1. Learned DDDA for AdAGeo

As a first solution for the DDDA module, we take inspiration from Cohen and Wolf (2019) who propose an approach for the problem of learning a bi-directional mapping between two domains, one (i.e., D^s) having many samples and the other (i.e., D^t) with only one sample available. This solution fits the VG scenario as it has already been demonstrated (Cohen and Wolf, 2019) to be effective on images depicting streets and urban places. Moreover, it does not requires massive resources and long training times. The main intuition of this method is to use two parallel autoencoders, named $Ae_S = Dec_S(Enc_S(x))$ and $Ae_T = Dec_T(Enc_T(x))$, respectively, and with Enc_S and Enc_T denoting encoders and Dec_S and Dec_T decoders. The goal is to minimize the distance between the distributions of the latent spaces of the two autoencoders, forcing the encoders to produce domain-invariant embeddings, while at the same time each decoder should be able to translate the embeddings to an image in its own domain. This is achieved by minimizing a reconstruction loss on both autoencoders:

$$L_{REC} = \sum_{s \in S} \|Ae_S(s)) - s\|_1 + \sum_{t \in T} \|Ae_T(t)) - t\|_1 \tag{1}$$

as well as cycle-consistency losses:

$$\begin{aligned} L_{sts-cycle} &= \sum_{s \in S} \|Dec_S(\overline{Enc_T}(\overline{Dec_T}(Enc_S(s)))) - s\|_1 \\ L_{tst-cycle} &= \sum_{t \in T} \|Dec_T(\overline{Enc_S}(\overline{Dec_S}(Enc_T(t)))) - t\|_1 \end{aligned} \tag{2}$$

Ineq. (2), the bar above a module means that its weights are frozen during the backpropagation of the loss. Moreover, as in Cohen and Wolf (2019), it is important that the embeddings approximate a Gaussian distribution, which helps the two domains to better align and can be achieved through a variational

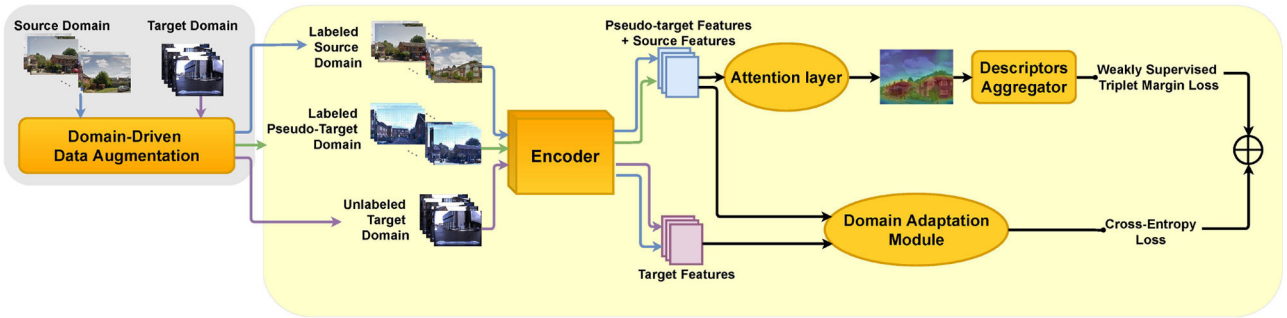

FIGURE 3 | The proposed architecture. First, a domain-driven data augmentation method is used to generate a labeled pseudo-target dataset from the source dataset and just 5 unlabeled target images. Then, the source dataset, pseudo-target dataset, and the 5 unlabeled target images are used to train the network that extracts the image descriptors. This network leverages an attention layer and a domain adaptation module to provide robustness to domain shifts. The difference between AdAGeo and AdAGeo-Lite is in the implementation of the domain-driven data augmentation.

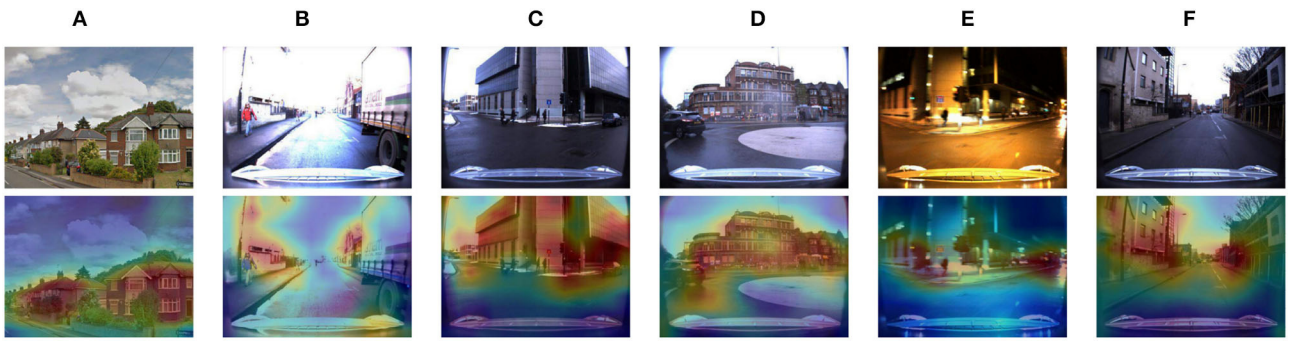

FIGURE 4 | Visualization of attention score maps on source domain **(A)** and target domains **(B–F)** unseen by the attention module at training time.

loss on both encoders:

$$\begin{aligned} L_{VEnc_S} &= \sum_{s \in S} KL(\{Enc_S(s) | s \in S\} \| \mathcal{N}(0, I)) \\ L_{VEnc_T} &= \sum_{t \in T} KL(\{Enc_T(t) | t \in T\} \| \mathcal{N}(0, I)) \end{aligned} \tag{3}$$

The complete loss is finally:

$$\begin{aligned} L_{final} = L_{REC} + L_{sts-cycle} + L_{tst-cycle} + \\ 0.001 L_{VEnc_S} + 0.001 L_{VEnc_T} \end{aligned} \tag{4}$$

Once the training process is finished, the pseudo-target images are generated as $x_i^{pt} = Dec_T(Enc_S(x_i^s))$, with their corresponding labels being $y_i^{pt} = y_i^s$.

4.1.2. Fourier-Based DDDA for AdAGeo-Lite

The DDDA solution based on two autoencoders discussed in Section 4.1.1, and originally presented in Berton et al. (2021b), is effective for few-shot domain-driven data augmentation, but it has some drawbacks: i) it requires a computationally heavy architecture, and ii) it introduces an additional training step to be performed before the descriptor extractor model can be trained. Moreover, it is arguable whether such a complex generative method is needed when the appearance discrepancies commonly encountered in VG due to environmental changes are mostly in the form of variations to the global photometric statistics. Indeed, as discussed by Yang and Soatto (2020), such global illumination discrepancies can be eliminated without having to be learned. Inspired by Yang and Soatto (2020), we propose a second DDDA method based on the lightweight Fourier Domain Adaptation (FDA). Formally, let $\mathcal{F}^A, \mathcal{F}^P : \mathbb{R}^{3 \times H \times W} \rightarrow \mathbb{R}^{3 \times H \times W}$ be the amplitude and phase components of the Fourier transform $\mathcal{F}$. Let us also denote with M_β a mask whose value is zero except for the center region defined by $\beta \in (0, 1)$:

$$M_\beta(h, w) = \mathbb{1}_{(h,w) \in [-\beta H : \beta H, -\beta W : \beta W]}. \tag{5}$$

where the center of the image is assumed to be in the position $(0, 0)$. Then, given two randomly sampled images $x^s \sim D^s$, $x^t \sim D^t$, the FDA-DDDA is formalized as:

$$x^{pt} = \mathcal{F}^{-1}([M_\beta \circ \mathcal{F}^A(x^t) + (1 - M_\beta) \circ \mathcal{F}^A(x^s), \mathcal{F}^P(x^s)]) \tag{6}$$

Ineq. (6), first, the low frequency part of the amplitude of the source image $\mathcal{F}^A(x^s)$ is replaced by that of the target image. Then, the modified spectral representation of x^s is mapped back to the

FIGURE 5 | Qualitative comparison between AdAGeo and AdAGeo-Lite data augmentation methods: on the first row, the RobotCar images are augmented with the DDDA method from AdAGeo-Lite, on the second row the augmentation is performed with AdAGeo. The columns, from left to right, refer to the following target domains: Snow, Rain, Sun, Night, and Overcast.

image $x^{s\to t}$, whose content is the same as x^s but will resemble the appearance of a sample from D^t. Usingeq. (6), we generate the pseudo-target dataset X^{pt}, without training.

4.2. Attention

The descriptor extraction network is tasked to produce the image representations that are used to perform the retrieval of similar places. In order to guide this network to focus on the most relevant features' areas for the retrieval task, we introduce an attention layer after the encoder. To this end, we took inspiration from the class activation map (CAM) (Zhou et al., 2016) which tries to focus on discriminative image areas that are the most useful to produce the class output in the image classification task, exploiting the final average pooling layer present in recent networks such as the ResNet (He et al., 2016). Let us consider for a given image of dimension $3 \times H \times W$ the extracted feature representation f of shape $D \times H_1 \times W_1$ where D is the number of kernels from the last convolutional layer in the encoder. Furthermore, consider also the backbone classifier block, which contains a fully connected layer with $D \times C$ weights w_{cd} with d values respectively for each class c. The attention map AM_c for a given class c is obtained by the following linear combination:

$$AM_c = \sigma\left(\sum_d f_d \cdot w_{cd}\right) \tag{7}$$

where σ is the softmax function and whose result has dimension $H_1 \times W_1$.

Finally, AM_c is upsampled to $H \times W$, and it is applied over the input image, to visualize the most relevant regions for that class c.

In our architecture, we use the fully connected layer of a CNN pretrained on Places365 (Zhou et al., 2017), which contains $C = 365$ classes, to produce the AM_c. The idea stems from the fact that the classes in Places365 (such as house, building, market) are inherently relevant to the task of geo-localization in urban environments. The images are passed to the whole backbone extracting the local features representation f from the last convolutional layer and producing the $AM_{c_{max}}$ for the category c_{max} with the highest probability predicted by the fully connected layer. Then, the features are spatially weighted with the scores calculated before:

$$f^w = f \cdot AM_{c_{max}} \tag{8}$$

producing new weighted features f^w with the same dimensions as f.

We demonstrate that the attention mechanism is useful also for the target images since the salience regions can help to distinguish also the elements across different domains. **Figure 4** shows the results obtained applying the attention mechanism over all domains at test-time, which shows significant visual results also over target domains unseen by the attention module.

4.3. Weakly Supervised Descriptors Aggregation

In order to transform the attentive embeddings into vectorized representations of each image we use a NetVLAD layer (Arandjelovic et al., 2018), a common descriptor aggregator for VG (Kim et al., 2017; Arandjelovic et al., 2018; Wang et al., 2019). To use NetVLAD, we first perform K-means clustering over 500 randomly sampled embeddings of images from the source and pseudo-target domains to compute K centroids. Then, given the embeddings f^w of dimension $D \times H_1 \times W_1$, reshaped with dimensions $D \times R$ where $R = H_1 \times W_1$, the element (j, k) of the VLAD representation V (Jégou et al., 2010) is computed as

$$V(j,k) = \sum_{i=1}^{R} \frac{e^{-\|f_i^w - c_k\|^2}}{\sum_{k'} e^{-\|f_i^w - c_{k'}\|^2}} \cdot \left(f_i^w(j) - c_k(j)\right) \tag{9}$$

where $f_i^w(j)$ and $c_k(j)$ are the j-th dimensions of the i-th embedding and k-th centroids, respectively. The fraction ineq. (9) is the soft-assignment of descriptor f_i^w to centroid k-th. Given the intrinsic nature of VG data, where the label for each image is represented solely by its position, it is not possible to use standard supervised losses to drive the training, because two photos taken in the same position (therefore, with the same label) but opposite

directions would depict different locations. To overcome this, we use a weakly supervised triplet margin loss as in Arandjelovic et al. (2018). For each query q, this loss is defined as

$$\mathcal{L}_{triplet} = \sum_{y}^{Y} h(\min_{i} d^2(F(q), F(p_i^q)) + m - d^2(F(q), F(n_y^q))) \quad (10)$$

where $d(\cdot, \cdot)$ represents the Euclidean distance, $F(x)$ is the features representation for image x, $\{p_i^q\}$ is the set of potential hard positives (images within 10 m from the query q), $\{n_y^q\}$ is the set of Y negatives (further than 25 m), h is the hinge loss, and m is a constant parameter chosen as margin.

4.4. Domain Adaptation

In order for the retrieval to work well across domains it is important that the embeddings produced by the attention module are domain agnostic, i.e., they do not encode domain-specific information. We achieve this by using a domain discriminator which receives embeddings from the three domains D_s, D_{pt}, and D_t. The discriminator is composed of two fully connected layers, and its goal is to classify the domain to which the embeddings belong. Just before the discriminator, there is a gradient reversal layer (Ganin and Lempitsky, 2015) that in the forward pass acts as an identity transform, while in the backward pass multiplies the gradient by -λ, where $\lambda > 0$. The use of this layer effectively sets up a min-max game, where the discriminator tries to minimize the domain classification loss, that is a cross-entropy loss $\mathcal{L}_{CE}$, while the feature extractor learns to produce domain-invariant embeddings, acting as an adversary to the discriminator.

5. EXPERIMENTS

In this section, we first explain the training details for AdAGeo and AdAGeo-Lite as well as the experimental protocol and the methods considered for comparisons. With this setting, we then report results from extensive experiments aimed at assessing both the domain adaptation and domain generalization capabilities of our methods. Finally, we include an ablation study to investigate the effect of the various components of our architecture.

5.1. Training Details

Prior to training the descriptor extraction network, we need to generate the pseudo-target dataset. For the learned DDDA method (c.f. Section 4.1.1), we adopt the architecture of Cohen and Wolf (2019) which consists of two encoders, made of two convolutional layers and four residual blocks, and two symmetric decoders, made of four residual blocks and two deconvolutional layers. To train this architecture we use the Adam optimizer with a learning rate of 0.0002 and batch size 1. For the non-learned DDDA method (c.f. Section 4.1.2), we use the implementation of the Fourier Domain Adaptation from Yang and Soatto (2020), and set the parameter β to 0.001. For both the DDDA methods and the training of the descriptor extractor network we use only 5 unlabeled images from the target domain.

Once the pseudo-target dataset is created, we use it together with the source dataset and the few target images used in the previous step to train the descriptor extraction network. For this network, we use a ResNet-18 (He et al., 2016) pre-trained on Places365 (Zhou et al., 2017) as the backbone, and we fine-tune it from the last two convolutional blocks to the end. The features extracted at the last convolutional layer, before ReLU, are passed to the attention and the domain adaptation modules. As an optimizer, we use Adam with a learning rate of 0.00001, and for each iteration, we use 4 tuples, each consisting of 1 query image, the best positive, and 10 negative samples. The negative samples are chosen following the standard procedure described by Arandjelovic et al. (2018), in order to increase the likelihood that $\mathcal{L}_{triplet} > 0$, by making sure that each negative is similar enough to the positive. The two losses are combined as $\mathcal{L}_{triplet} + \alpha \cdot \mathcal{L}_{CE}$ where $\alpha = 0.1$. Finally, unlike most domain adaptation methods which train the network for a constant number of epochs, or perform validation and early stopping on the source validation set, we perform validation and early stopping on the generated pseudo-target validation set, having a similar distribution to the target set, helps to optimally stop the training.

5.2. Datasets for Domain Generalization

Among the available multi-domain geolocalization datasets, we use Mapillary Street Level Sequences (MSLS) (Warburg et al., 2020) and St Lucia (Milford and Wyeth, 2008) to test the capability of our solution to generalize to unseen domains, including different cities. MSLS is composed of a collection of sequences from 30 different cities, and it encompasses a large variety of domain changes, such as night/day, long-term, and sun/rain. For our experiments, we chose 5 cities with the goal of having different degrees of similarity with the source dataset: Copenhagen, San Francisco, Nairobi, Tokyo, and São Paulo. For further assessment of the various methods, we also test on the St Lucia dataset, which was collected in a suburb of Brisbane and presents no heavy temporal changes between the database and queries. The size of each dataset used for domain generalization is shown in **Table 2**.

5.3. Comparisons With Other Methods

To evaluate AdAGeo and AdAGeo-Lite we compare them with other methods. To the best of our knowledge, the only

TABLE 2 | Sizes of cities from MSLS (Warburg et al., 2020) and St Lucia (Milford and Wyeth, 2008) datasets.

	Copenhagen	San Francisco	Nairobi	Tokyo	São Paulo	St Lucia
Gallery	12601	437	35096	6315	34823	1549
Queries	6595	427	18989	4525	26310	1464

TABLE 3 | Domain adaptation results: comparison between all methods, shown as recall@1 (R@1) on each target domain.

Method	#T	Snow	Rain	Sun	Night	Overcast	Avg
		R@1	R@1	R@1	R@1	R@1	
NetVLAD	0	50.1 ± 1.3	36.5 ± 0.6	17.7 ± 0.9	1.6 ± 0.4	60.0 ± 0.7	33.2
Wang	all	23.8 ± 6.2	11.2 ± 1.4	5.7 ± 0.5	0.9 ± 0.5	37.6 ± 8.2	15.8
NetVLAD+SAFN	all	57.3 ± 2.5	43.6 ± 0.4	19.1 ± 2.0	2.2 ± 0.7	68.3 ± 1.2	38.1
NetVLAD+DCORAL	all	60.2 ± 2.0	33.5 ± 1.1	14.1 ± 0.6	2.1 ± 0.8	61.2 ± 3.6	34.2
NetVLAD+GRL	all	68.9 ± 2.5	50.9 ± 2.0	27.1 ± 4.8	4.6 ± 1.2	76.9 ± 0.7	45.7
AdAGeo	5	73.3 ± 2.2	55.7 ± 1.8	29.6 ± 1.0	10.5 ± 1.9	**80.1** ± 1.5	49.8
AdAGeo-Lite	5	**73.8** ± 0.3	**57.6** ± 2.5	**30.6** ± 4.4	**11.1** ± 0.5	78.8 ± 1.9	**50.4**

#T shows the number of target images used at training time. Snow, Rain, Sun, Night, and Overcast are the 5 target domains of the SVOX+RobotCar dataset. The last column shows the average recall@1 across all domains. In bold is the best result for each target domain.

TABLE 4 | Domain generalization results: comparison between all methods, shown as recall@1 (R@1) on each target domain.

	Method	#T	Snow	Rain	Sun	Night	Overcast	Cph	Nairobi	Tokyo	SF	Sao Paulo	St Lucia
	NetVLAD	0	50.2	36.6	17.8	1.6	60.1	51.7	38.0	28.1	46.4	23.6	89.4
Snow	Wang	all	23.8	7,9	2,7	0,9	11,3	25,5	4,9	13,7	22,2	8,0	41,9
	NetVLAD+SAFN	all	57.4	37.7	16.3	1.7	67.7	53.3	40.4	29.0	47.6	24.0	86.1
	NetVLAD+DCORAL	all	60.3	34.9	15.6	3.4	74.0	42.5	30.5	24.6	37.4	19.9	83.8
	NetVLAD+GRL	all	68.9	48.7	21.9	2.3	76.4	52.3	40.4	28.2	46.6	23.4	85.1
	AdAGeo-Lite	5	**73.5**	**57.6**	**28.6**	**4.0**	**77.1**	**65.0**	**51.0**	**36.4**	**63.8**	**34.3**	**95.1**
Rain	Wang	all	8.0	11.2	0.1	0.4	17.3	21.5	4.7	12.4	22.0	6.5	39.7
	NetVLAD+SAFN	all	59.1	43.7	17.5	1.6	67.8	52.8	39.0	28.8	46.5	23.9	87.2
	NetVLAD+DCORAL	all	59.3	33.6	15.8	**3.5**	71.6	44.1	30.7	26.9	39.9	21.3	84.9
	NetVLAD+GRL	all	66.8	50.9	23.5	1.0	73.0	50.8	39.8	27.5	44.2	22.3	85.1
	AdAGeo-Lite	5	**73.8**	**55.4**	**27.7**	2.7	**79.2**	**65.8**	**55.6**	**36.2**	**63.7**	**35.3**	**95.8**
Sun	Wang	all	7.3	8.7	5.7	0.4	11.4	23.6	5.5	13.1	20.8	6.1	32.4
	NetVLAD+SAFN	all	58.9	42.4	19.1	1.9	71.1	51.4	41.0	27.9	46.7	23.6	86.2
	NetVLAD+DCORAL	all	53.1	35.8	14.1	2.4	64.8	46.4	33.6	24.7	40.1	21.1	82.6
	NetVLAD+GRL	all	64.7	45.7	27.1	1.0	72.8	52.3	41.4	28.0	46.2	24.0	85.0
	AdAGeo-Lite	5	**70.3**	**56.9**	**30.6**	**2.7**	**75.3**	**64.5**	**52.2**	**35.8**	**63.1**	**34.2**	**94.1**
Night	Wang	all	7.5	15.4	2.5	0.9	19.0	26.7	9.4	14.9	23.3	8.7	43.7
	NetVLAD+SAFN	all	57.2	37.5	16.3	2.2	67.6	49.2	40.1	27.5	45.2	23.1	84.2
	NetVLAD+DCORAL	all	55.0	34.8	16.4	2.1	69.1	47.7	35.2	24.8	41.4	22.0	83.5
	NetVLAD+GRL	all	57.9	40.3	19.1	4.6	67.9	53.7	39.3	28.9	49.3	25.5	88.0
	AdAGeo-Lite	5	**68.1**	**49.0**	**23.1**	**11.1**	**74.0**	**65.1**	**56.4**	**36.4**	**63.7**	**34.5**	**94.6**
Overcast	Wang	all	6.3	14.7	1.6	0.4	37.6	26.4	6.6	14.4	22.3	7.8	39.8
	NetVLAD+SAFN	all	56.8	39.3	16.2	1.5	68.3	51.6	37.0	28.0	45.8	23.6	84.2
	NetVLAD+DCORAL	all	49.6	34.7	14.5	1.5	61.3	44.7	35.3	22.8	38.5	20.7	83.6
	NetVLAD+GRL	all	66.5	43.7	21.0	2.0	77.0	49.1	37.3	26.3	42.8	22.0	84.7
	AdAGeo-Lite	5	**72.6**	**52.8**	**27.0**	**2.7**	**78.8**	**65.7**	**53.6**	**36.5**	**64.4**	**35.0**	**95.0**

The RobotCar scenario on the left indicates which one is used at training time for domain adaptation. #T shows the number of target images used at training time. In bold is the best result for each pair of train-test target domains.

other VG method built for domain adaptation is the one proposed by Wang et al. (2019), which uses an attention module and MK-MMD (Gretton et al., 2012). In particular, we use the code provided by the authors, in both its variants: the first one with just the attention mechanism (Wang: Att) and the second one with also the DA branch (Wang: Att+DA). Additionally, we also compare NetVLAD (Arandjelovic et al., 2018), one of the most used and well-established methods for VG. Given the focus of our research, we implement NetVLAD with various domain adaptation techniques, namely GRL (Ganin and Lempitsky, 2015), DeepCORAL (Sun and Saenko, 2016), and SAFN (Xu et al., 2019). For SAFN, we compute the features norm from the embeddings produced by the last convolutional layer of the backbone, using the code provided by the authors. For fairness of comparisons, we compare the methods using as backbones ResNet-18 (He et al., 2016) pretrained on Places365 (Zhou et al., 2017) and cropped at the last convolutional layer.

5.4. Qualitative DDDA Comparison Between AdAGeo and AdAGeo-Lite

The DDDA method used in AdAGeo has the downside that it requires a training phase that lasts roughly 12 h on a V100 GPU. The lighter architecture AdAGeo-Lite was developed to remove this additional training phase and to streamline the pipeline, using a simpler Fourier Domain Adaptation. We show in **Figure 5** some qualitative results obtained with the two DDDA methods. As we can see, there is little to no difference in the generated pseudo-targets.

5.5. Results: Domain Adaptation With AdAGeo

The methods are trained using SVOX as the source domain and Oxford RobotCar (Maddern et al., 2017) (around 800 images, depending on the domain, refer to **Table 1**) as the target. Depending on the method, a different number of target images has been used: NetVLAD uses none (i.e., no domain adaptation); for AdAGeo, we use 5 images (i.e., 5-shot scenario); and for other methods, we use the whole target set (about 800 images, depending on the domain **Table 1**). At test time, we use the SVOX test gallery and queries from Oxford RobotCar (Maddern et al., 2017), thus ensuring that the test-time photos come from a geographically disjoint area than the train/DA set. For methods with DA, training is performed separately for each of the 5 target domains (Snow, Rain, Sun, Night, and Overcast). As a metric, we use the commonly used recall@N (R@N), which measures the percentage of queries for which at least one of the first N predictions is located within a given threshold distance from the relative query. Following standard procedure, the threshold distance is set at 25 m (Kim et al., 2017; Arandjelovic et al., 2018; Zhu et al., 2018; Liu et al., 2019a; Wang et al., 2019; Warburg et al., 2020; Berton et al., 2021a,b; Hausler et al., 2021). All the results shown are the outcome of three experiments, and we report the mean and SD, to ensure further reliability.

In **Table 3** we show the results for each method. We find that AdAGeo and AdAGeo-Lite outperform all other approaches while using two orders of magnitude fewer target domain images. The comparable results between the two are in line with the qualitative similarity of their two pseudo-targets, as seen in **Figure 5**. This supports the intuition that the domain shift caused by the environmental changes (e.g., illumination and weather) have, for the most part, the characteristics of global photometric variations. Actually, AdAGeo-Lite gives slightly better results, perhaps due to the fact that finding the optimal configuration for the many hyperparameters of the learned DDDA is not trivial, whereas the FDA method used in AdAGeo-Lite only has one hyperparameter to tune.

From this results, it stands out that all models perform poorly on the Night target domain. This can be expected, as the shift between this domain and the source domain (StreetView) is extreme, with very dark images and with a strong yellow tones. Even though AdAGeo and AdAGeo-Lite surpass all competitors by at least 5.9 points, the results are still far from the other cases. We observe that the pseudo-target images generated for the Night target by AdAGeo and AdAGeo-Lite indeed present an accentuated yellow hue (cf. **Figure 5**), and indeed

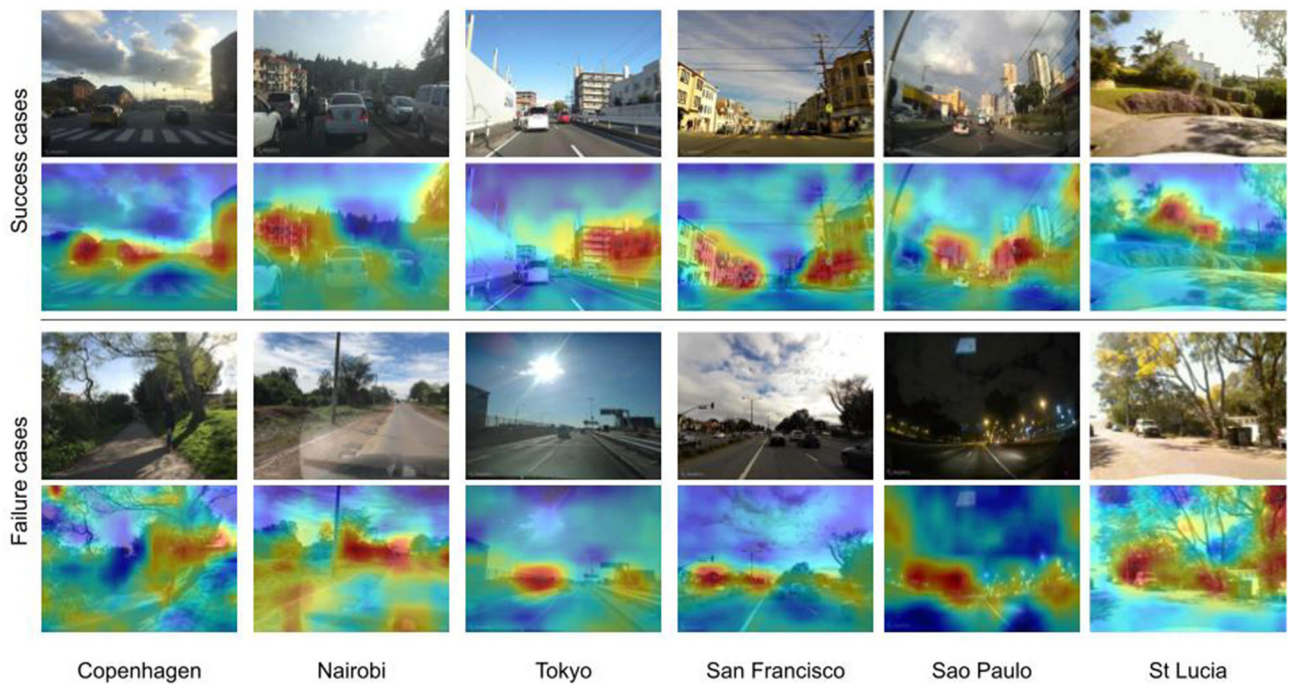

FIGURE 6 | Domain generalization qualitative results: top-half rows depict the queries and their associated attention map, which are successfully geolocalized in the first prediction. Bottom-half rows show failure cases which are not retrieved even after 20 predictions.

the ablation results presented later in Section 5.7 demonstrate that they improve the results. On the other hand, looking at the visualization of the attention score maps (refer to **Figure 4**) we note that the model is much less focused, particularly on the buildings as they are not well discernible. This is also confirmed by the ablation study that is presented later in Section 5.7.

5.6. Results: Domain Generalization

In domain generalization, we use a model trained on the source and pseudo target datasets X^s, X^{pt}, with pseudo targets adapted from D^t, to make out-of-distribution inference on a target domain $D^{t'} \neq D^t$ unseen at training time. For each scenario in the RobotCar dataset, we trained a model and tested its generalization abilities. For this set of experiments, we only use AdAGeo-Lite, since it outperformed AdAGeo in the domain adaptation experiments.

From **Table 4**, we observe that AdAGeo-Lite generalizes largely better than all other models, not only on the different RobotCar scenarios but also when considering sample images coming from new cities, not in the training dataset. These results extensively show that, as a side effect, our style transfer techniques are beneficial in terms of generalization, which is a highly desirable property in visual geo-localization. This is true not only for shifts between related domains such as images under different light conditions, but also for queries of dissimilar nature such as cities belonging to many geographical continents. Some visualizations of attention score maps on these datasets is shown in **Figure 6**.

The results also show that all models, including AdAGeo and AdAGeo-Lite, struggle to generalize to Tokyo and to São Paulo. We conjecture that this is due to the fact that both these scenarios present a large gallery (see **Table 2**), which increases the number of negatives per query, making geolocalization on these cities inherently more challenging.

TABLE 5 | Ablation table of AdAGeo on the SVOX+RobotCar dataset in a 5-shot setting with ResNet18 as the encoder.

Method	Snow	Rain	Sun	Night	Overcast	Avg
	R@1	R@1	R@1	R@1	R@1	
Baseline	50.1	36.5	17.7	1.6	60.0	33.2
Baseline+DDDA	61.3	45.3	23.3	6.1	71.1	41.4
Baseline+Att	49.4	39.9	24.5	3.3	64.0	36.2
Baseline+DA	65.3	49.7	25.4	6.0	75.2	44.3
Baseline+DDDA+Att	66.6	54.5	27.3	5.5	72.2	45.2
Baseline+DDDA+DA	67.2	51.5	24.8	9.4	78.4	46.3
Baseline+Att+DA	66.0	49.1	24.8	3.2	76.1	43.8
AdAGeo	73.3	55.7	29.6	f10.5	80.1	49.8

R@1, recall@1; DDDA, Domain-Driven Data Augmentation; Att, Attention layer; DA, Domain adaptation layer.

5.7. Ablation

We evaluate the components of AdAGeo by conducting an extensive ablation study over each target domain of SVOX+RobotCar. The results are shown in **Table 5**, where all experiments have been run in a 5-shot environment (except for the experiments where the target domain is not used) and all the modules combination are tried. As a baseline, we use a ResNet-18 encoder (cropped at the last convolutional layer) followed by a NetVLAD (Arandjelovic et al., 2018) layer. Then, we try all possible combinations of the three proposed components: Domain-driven data augmentation module (DDDA), Attention

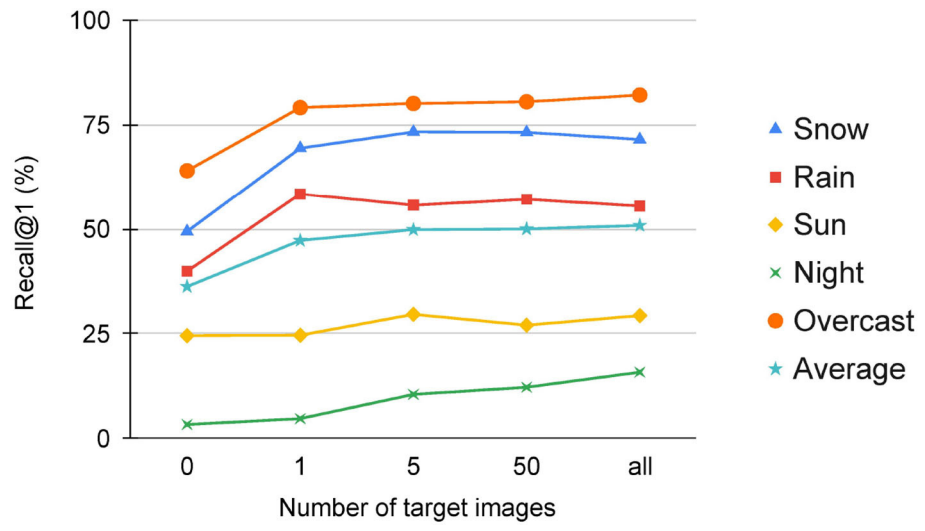

FIGURE 7 | Results of experiments with AdAGeo with 0-shots, 1-shot, 5-shots, 50-shots, and all-shots. With easier domains (Snow, Rain, Overcast), AdAGeo shows a good improvement in accuracy with just 1 target domain image, while with more challenging domains (Sun, Night), AdAGeo requires a higher amount of images to perform significant improvements.

module (Att), and Domain adaptation module (DA). The use of all components results in AdAGeo. As shown in **Table 5**, each module produces an improvement with respect to the baseline. The ablation study also proves that the modules are orthogonal to each other, giving consistent improvements when used alone as when used together. In particular, the attention module yields a 3% improvement on the baseline, and 3.5% on the final model, although it does not see the target domain at training time. Finally, the three modules together show an improvement of more than 16% on average over the baseline.

To better understand how the number of target domain images influences the outcome, we also conduct extensive experiments on this matter. Results are in **Figure 7**, and they show that the model can benefit from a high amount of target images on challenging domains, while for easier domains the model quickly saturates even with as few as 5 images.

6. CONCLUSIONS

In this study, we have proposed a method to tackle the problem of cross-domain visual geo-localization using only few unlabeled target images. The foundation of our architecture is to use two orthogonal domain adaptation techniques as well as an attention mechanism. In particular, we present two variations of our method, AdAGeo and AdAGeo-Lite, that differ in the domain-driven data augmentation module. Given the modularity of our method, in the future, we plan to extend it to use different style transfer solutions, e.g., TUNIT (Baek et al., 2021) for the case, when many unlabeled target images are available, or FUNIT (Liu et al., 2019b) for the case, when the training can rely on more considerable resources. Both AdAGeo and AdAGeo-Lite are able to outperform current state-of-the-art solutions while using two orders of magnitude fewer target images during training. Remarkably, our experiments show that the simpler domain-driven data augmentation method used in AdAGeo-Lite yields comparable or better results in comparison to the learned DDDA module used in the first AdAGeo version. Finally, we propose a new dataset, called SVOX, which, extends Oxford RobotCar and can be used as a large scale multi-domain dataset for visual place recognition, presenting a realistic scenario for future research in the field.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

VP, GB, CM, and BC contributed to conception and design of the project. VP and GB designed and implemented the core of the software. FM implemented the DDDA algorithm for AdAGeo-Lite. VP, GB, and FM performed the experiments. All the authors contributed to writing the manuscript.

ACKNOWLEDGMENTS

The content of this manuscript has been presented in part at the IEEE Winter Conference on Applications of Computer Vision (WACV) 2021, (Berton et al., 2021b). This work was supported by CINI.

REFERENCES

Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., and Gool, L. V. (2019). "Night-to-day image translation for retrieval-based localization," in *2019 International Conference on Robotics and Automation (ICRA)* (Montreal), 5958–5964.

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2018). NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 1437–1451. doi: 10.1109/TPAMI.2017.2711011

Aubry, M., Russell, B. C., and Sivic, J. (2014). Painting-to-3d model alignment via discriminative visual elements. *ACM Trans. Graph.* 33, 1–14. doi: 10.1145/2591009

Baek, K., Choi, Y., Uh, Y., Yoo, J., and Shim, H. (2021). "Rethinking the truly unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 14154–14163.

Benaim, S., and Wolf, L. (2017). "One-sided unsupervised domain mapping," in *Advances in Neural Information Processing Systems 30* (Long Beach, CA: Curran Associates, Inc.), 752–762.

Berton, G., Masone, C., Paolicelli, V., and Caputo, B. (2021a). "Viewpoint invariant dense matching for visual geolocalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal), 12169–12178.

Berton, G. M., Paolicelli, V., Masone, C., and Caputo, B. (2021b). "Adaptive-attentive geolocalization from few queries: a hybrid approach," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision* (Hawaii), 2918–2927.

Bolte, J.-A., Kamp, M., Breuer, A., Homoceanu, S., Schlicht, P., Hüger, F., et al. (2019). "Unsupervised domain adaptation to improve image segmentation quality both in the source and target domain," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Long Beach, CA: IEEE), 1404-1413.

Cao, B., Araujo, A., and Sim, J. (2020). "Unifying deep local and global features for image search," in *European Conference on Computer Vision-2020*, eds A. Vedaldi, H. Bischof, T. Brox, and J. -M. Frahm (Cham: Springer International Publishing), 726–743.

Chen, M., Kira, Z., Alregib, G., Yoo, J., Chen, R., and Zheng, J. (2019). "Temporal attentive alignment for large-scale video domain adaptation," in *ICCV* (Seoul), 6320–6329.

Chen, Z., Jacobson, A., Sünderhauf, N., Upcroft, B., Liu, L., Shen, C., et al. (2017a). "Deep learning features at scale for visual place recognition," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (Singapore: IEEE), 3223–3230.

Chen, Z., Maffra, F., Sa, I., and Chli, M. (2017b). "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC: IEEE), 9–16.

Cheng, R., Wang, K., Bai, J., and Xu, Z. (2020). Unifying visual localization and scene recognition for people with visual impairment. *IEEE Access* 8, 64284–64296. doi: 10.1109/ACCESS.2020.2984718

Cohen, T., and Wolf, L. (2019). "Bidirectional one-shot unsupervised domain mapping," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul), 1784–1792.

Cummins, M., and Newman, P. (2008). FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Rob. Res.* 27, 647–665. doi: 10.1177/0278364908090961

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., and Tao, D. (2019). Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern. Recognit.* 2019, 2422–2431. doi: 10.1109/cvpr.2019.00253

Ganin, Y., and Lempitsky, V. (2015). "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research*, eds F. Bach and D. Blei (Lille: PMLR), 1180–1189.

Garg, S., Suenderhauf, N., and Milford, M. (2018a). "Don't look back: robustifying place categorization for viewpoint- and condition-invariant place recognition," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (Brisbane), 3645–3652.

Garg, S., Suenderhauf, N., and Milford, M. (2018b). "Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics," in *Proceedings of Robotics: Science and Systems, Pittsburgh, Pennsylvania*.

Gordo, A., Almazán, J., Revaud, J., and Larlus, D. (2017). End-to-end learning of deep visual representations for image retrieval. *Int. J. Comput. Vis.* 124, 237–254. doi: 10.1007/s11263-017-1016-8

Gretton, A., Sriperumbudur, B. K., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., et al. (2012). "Optimal kernel choice for large-scale two-sample tests," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012*, eds P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Lake Tahoe, NV), 1214–1222.

Hausler, S., Garg, S., Xu, M., Milford, M., and Fischer, T. (2021). "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition* (Seattle), 14141–14152.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern. Anal. Mach. Intell.* 37, 1904–1916. doi: 10.1109/TPAMI.2015.2389824

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *CVPR*, 770–778.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., et al. (2018). "CyCADA: cycle-consistent adversarial domain adaptation," in *Proceedings of the 35th International Conference on Machine Learning* (Stockholm), Vol. 80, 1989–1998.

Hong, W., Wang, Z., Yang, M., and Yuan, J. (2018). "Conditional generative adversarial network for structured domain adaptation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 1335–1344.

Hu, H., Qiao, Z., Cheng, M., Liu, Z., and Wang, H. (2021). DASGIL: Domain adaptation for semantic and geometric-aware image-based localization. *IEEE Trans. Image Process.* 30, 1342–1353. doi: 10.1109/TIP.2020.3043875

Huang, S.-W., Lin, C.-T., Chen, S.-P., Wu, Y.-Y., Hsu, P.-H., and Lai, S.-H. (2018). "AugGAN: cross domain adaptation with gan-based data augmentation," in *The European Conference on Computer Vision (ECCV)* (Munich), 731–744.

Jégou, H., Douze, M., Schmid, C., and P?rez, P. (2010). "Aggregating local descriptors into a compact image representation," in *Conference on Computer Vision and Pattern Recognition (CVPR)* (San Francisco, CA).

Johns, E., and Guang-Zhong, Y. (2011). "From images to scenes: Compressing an image cluster into a single scene model for place recognition," in *IEEE International Conference on Computer Vision* (Barcelona: IEEE), 874–881.

Kim, H. J., Dunn, E., and Frahm, J. (2015). "Predicting good features for image geo-localization using per-bundle VLAD," in *IEEE International Conference on Computer Vision* (Santiago,: IEEE), 1170–1178.

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

Liu, L., Li, H., and Dai, Y. (2019a). "Stochastic attraction-repulsion embedding for large scale image localization," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 2570–2579.

Liu, M.-Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., et al. (2019b). "Few-shot unsupervised image-to-image translation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 10550–10559.

Lou, Y., Bai, Y., Wang, S., and Duan, L.-Y. (2018). "Multi-scale context attention network for image retrieval," in *Proceedings of the 26th ACM International Conference on Multimedia, MM '18* (New York, NY: Association for Computing Machinery), 1128–1136.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., et al. (2016). Visual place recognition: a survey. *IEEE Trans. Rob.* 32, 1–19. doi: 10.1109/TRO.2015.2496823

Maddern, W., Pascoe, G., Linegar, C., and Newman, P. (2017). "1 Year, 1000km: the oxford robotcar dataset," in *The International Journal of Robotics Research (IJRR)*.

Masone, C., and Caputo, B. (2021). A survey on deep visual place recognition. *IEEE Access* 9, 19516–19547. doi: 10.1109/ACCESS.2021.3054937

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

Middelberg, S., Sattler, T., Untzelmann, O., and Kobbelt, L. (2014). "Scalable 6-dof localization on mobile devices," in *European Conference on Computer Vision-2014*, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Cham: Springer International Publishing), 268–283.

Milford, M., and Wyeth, G. (2008). Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Trans. Rob.* 24:1038–1053. doi: 10.1109/TRO.2008.2004520

Milford, M. J., and Wyeth, G. F. (2012). "Seqslam: visual route-based navigation for sunny summer days and stormy winter nights," in *2012 IEEE International Conference on Robotics and Automation* (Saint Paul, MN: IEEE), 1643–1649.

Murillo, A. C., and Kosecka, J. (2009). "Experiments in place recognition using gist panoramas," in *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (Kyoto: IEEE), 2196–2203.

Nakka, K. K., and Salzmann, M. (2018). "Deep attentional structured representation learning for visual recognition," in *BMVC* (Newcastle).

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

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

Oertel, A., Cieslewski, T., and Scaramuzza, D. (2020). Augmenting visual place recognition with structural cues. *IEEE Rob. Autom. Lett.* 5, 5534–5541. doi: 10.1109/LRA.2020.3009077

Peng, G., Yue, Y., Zhang, J., Wu, Z., Tang, X., and Wang, D. (2021a). "Semantic reinforced attention learning for visual place recognition," in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (Xi'an: IEEE), 13415–13422.

Peng, G., Zhang, J., Li, H., and Wang, D. (2021b). "Attentional pyramid pooling of salient visual residuals for place recognition," in *IEEE International Conference on Computer Vision* (Montreal, QC: IEEE), 885–894.

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

Pion, N., Humenberger, M., Csurka, G., Cabon, Y., and Sattler, T. (2020). "Benchmarking image retrieval for visual localization," in *2020 International Conference on 3D Vision (3DV)*, 483–494.

Porav, H., Maddern, W., and Newman, P. (2018). "Adversarial training for adverse conditions: Robust metric localisation using appearance transfer," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (Brisbane, QLD: IEEE), 1011–1018.

Radenovi,c, F., Tolias, G., and Chum, O. (2019). Fine-tuning cnn image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1655–1668. doi: 10.1109/TPAMI.2018.2846566

Russo, P., Carlucci, F. M., Tommasi, T., and Caputo, B. (2018). "From source to target and back: symmetric bi-directional adaptive GAN," in *2018 IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, IEEE), 8099–8108.

S?nderhauf, N., and Protzel, P. (2011). "BRIEF-Gist - closing the loop by simple means," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (San Francisco, CA: IEEE), 1234–1241.

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., et al. (2018). "Benchmarking 6dof outdoor visual localization in changing conditions," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 8601–8610.

Shang, C., Palmer, A., Sun, J., Chen, K., Lu, J., and Bi, J. (2017). "Vigan: missing view imputation with generative adversarial networks," in *2017 IEEE International Conference on Big Data (Big Data)* (Boston, MA: IEEE), 766–775.

Sun, B., and Saenko, K. (2016). "Deep CORAL: correlation alignment for deep domain adaptation," in *Computer Vision-ECCV 2016 Workshops*, eds G. Hua and H. Jégou (Cham: Springer International Publishing), 443–450.

Sünderhauf, N., Neubert, P., and Protzel, P. (2013). "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *Proceedings of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)* (Karlsruhe), 2013.

Tomit,a, M.-A., Zaffar, M., Milford, M. J., McDonald-Maier, K. D., and Ehsan, S. (2021). ConvSequential-SLAM: a sequence-based, training-less visual place recognition technique for changing environments. *IEEE Access* 9, 118673–118683. doi: 10.1109/ACCESS.2021.3107778

Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., and Pajdla, T. (2018). 24/7 place recognition by view synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 257–271. doi: 10.1109/TPAMI.2017.2667665

Wang, Z., Li, J., Khademi, S., and van Gemert, J. (2019). "Attention-aware age-agnostic visual place recognition," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (Seoul: IEEE), 1437–1446.

Warburg, F., Hauberg, S., López-Antequera, M., Gargallo, P., Kuang, Y., and Civera, J. (2020). "Mapillary street-level sequences: a dataset for lifelong place recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: IEEE), 2623–2632.

Xu, R., Li, G., Yang, J., and Lin, L. (2019). "Larger norm more transferable: an adaptive feature norm approach for unsupervised domain adaptation," in *ICCV* (Seoul), 1426–1435.

Yang, Y., and Soatto, S. (2020). "FDA: fourier domain adaptation for semantic segmentation," in *2020 IEEE Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 4084–4094.

Zaffar, M., Khaliq, A., Ehsan, S., Milford, M., and McDonald-Maier, K. (2019). "Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing condition," in *IEEE International Conference on Robotics and Automation Workshop* (IEEE), 1–8.

Zhang, X., Wang, L., and Su, Y. (2021). Visual place recognition: a survey from deep learning perspective. *Pattern Recog.* 113, 107760. doi: 10.1016/j.patcog.2020.107760

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 2921–2929.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: a 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 1452–1464. doi: 10.1109/TPAMI.2017.2723009

Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE), 2242–2251.

Zhu, Y., Wang, J., Xie, L., and Zheng, L. (2018). "Attention-based pyramid aggregation network for visual place recognition," in *Proceedings of the 26th ACM International Conference on Multimedia, MM '18* (New York, NY: Association for Computing Machinery), 99–107.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Paolicelli, Berton, Montagna, Masone and Caputo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.