ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230255

A Siamese Based System for City Verification

Omran Alamayreh*, Jun Wang, Giovanna Maria Dimitri, Benedetta Tondi and Mauro Barni

Department of Information Engineering and Mathematics, University of Siena.

Abstract. Image geolocalization is receiving increasing attention due to its importance in several applications, such as image retrieval, criminal investigations and fact-checking. Previous works focused on several instances of image geolocalization including place recognition, GPS coordinates estimation and country recognition. In this paper, we tackle an even more challenging problem, which is recognizing the city where an image has been taken. Due to the vast number of cities in the world, we cast the problem as a verification problem, whereby the system has to decide whether a certain image has been taken in a given city or not. In particular, we present a system that given a query image and a small set of images taken in a target city, decides if the query image has been shot in the target city or not. To allow the system to handle the case of images, taken in cities that have not been used during training, we use a Siamese network based on Vision Transformer as a backbone. The experiments we run prove the validity of the proposed system which outperforms solutions based on state-of-the-art techniques, even in the challenging case of images shot in different cities of the same country.

1 Introduction

The ability to recognize the geographic location, where an image has been taken, is of crucial importance in several applications like image retrieval, criminal investigations, fact-checking, and to prevent the diffusion of fake news and fight misinformation campaigns. For instance, recognizing the location portrayed in a photo could be extremely useful to identify text-image inconsistencies in the news, and reveal if a certain image was actually taken in a different location with respect to the one referenced in the text. Most of the works carried out so far treat image geolocalization as an inference problem, in which the final goal is to estimate the geo-coordinates of the image scene obtaining the best possible accuracy [22,23,35]. More recently, some works have shown the possibility to set the image geolocalization task as a classification problem, where the goal is to identify the country where an image has been shot, by relying on the architectural, engineering and, so to say, social characteristics of the images, like house shapes, cars, shops and roads signs [4, 20].

In this work, we adopt yet a different perspective, consisting in identifying the city where an image has been taken. In most cases, in fact, recognizing the source city of an image is more important than providing a precise geo-coordinates estimate of the scene, while recognizing only the country where the image has been taken, does not provide a sufficiently precise localization. As a matter of fact, city recognition is an extremely challenging task, due to the huge number of cities worldwide and the high similarities between many urban environments belonging to the same country. To address this challenge, we shift from the inference and classification approaches adopted so far and set the city recognition problem as a verification task. In particular, we present a system based on a Siamese network, whose goal is to evaluate if a source image has been taken in a target city, assuming that a small set of images of the target city are available.

One may argue that once the geo-coordinates have been estimated, tracing back the city should be a trivial task. However, this is true only in an ideal case. In fact, in non-ideal conditions, a system which is asked to minimize the spatial localization error may result in a wrong identification of the city, if doing so allows a better estimate of the geo-coordinates.

With the above ideas in mind, the novel system we are proposing employs a Siamese network with a Vision Transformer (ViT) backbone. ViT architecture has been proven to be extremely successful for image classification and was widely applied in several fields. Our proposed framework takes as inputs both a query image and a small set of pictures taken in a target city and decides if the query image has been shot in the target city or not. The system is complemented in both the training and test phases by the GeoVIPP country classifier described in [4], to rule out the possibility that images taken in different countries are judged to belong to the same city. Moreover, we exploit the Places365-CNNs model [39] to measure the semantic similarity between the input image and the reference images of the target city, weighting more heavily the similarity (or dissimilarity) between images having similar semantic content.

To train and test our verification system, we constructed a dataset extracted from the VIPPGeo dataset described in [4]. Overall, the city verification dataset we built consists of $120,909$ high-quality urban images from 19 cities all around the world.

We present several experiments on the city verification dataset, to demonstrate the effectiveness of our system for both closed and open set scenarios. In the former case, the images used in the testing phase belong to cities that were also used in the training phase. In the open set scenario, instead, the to-be-verified images do not belong to any of the cities used in the training phase. In both cases, we also considered the difficult case of images taken in cities of the same country. We compared the performances of our system with those of a verification system, built on top of a geolocalization network, providing an estimate of the image geo-coordinates and using them to decide if the query image belongs to the target city or not. The results of the experiments show that the proposed system outperforms the state-of-the-art geo-coordinates estimation method [22] in both closed and open set scenarios.

The paper is organized as follows. In Section 2, we briefly review the relevant state of the art. In Section 3, we present the city verification dataset and the proposed ViT-based Siamese architecture. In Section 4, we describe the experimental setting and the results we got. Finally, we draw our conclusions and outline some perspectives for future research in Section 5.

* Corresponding Author. Email: omran@diism.unisi.it

Please check ArXiv or contact the authors for any appendices or supplementary material mentioned in the paper.

2 Related work

In this section, we start by reviewing the most common approaches for image geolocalization. Then, we will briefly review Vision Transformers (ViT) and Siamese Neural Networks.

2.1 Geolocalization of images

Geolocalization is a very challenging task due to the amount of variability in photos, in terms of scenery, environment and architecture. In the literature, several works have addressed such a challenge, adopting different strategies.

The first attempt, at planet-scale image geolocalization, was introduced in Im2GPS [13], in which a query image is matched against a database of six million geo-tagged images and the location is inferred from the retrieved set. Subsequently, in [14], the method in Im2GPS [13] was further improved by incorporating multi-class support vector machines, to refine the search process. Another noteworthy work is [32], where the authors proposed learning a feature representation with a convolutional neural network (CNN) to enhance the performance of Im2GPS [13].

In Planet [35], the authors formulated the problem of image geolocalization as a classification problem. The earth's surface was divided into thousands of multi-scale geographic cells and a convolutional neural networks (CNN) model was trained using millions of geotagged images. However, the granularity of the partitioning in image geolocalization is crucial, as larger cells may yield lower location accuracy, while smaller cells may result in reduced training examples per class, making the model susceptible to overfitting. To address this issue, CPlanet [26] proposed a combinatorial partitioning algorithm that generates a multitude of fine-grained output classes by intersecting multiple coarse-grained partitioning of the earth's surface.

In more recent studies, in [23], a selective prediction method was introduced to assess the suitability of an image for the geolocalization task, resulting in the removal of non-localizable images and thereby increasing the overall accuracy. In [16] Izbicki et al. introduced the Mixture of von Mises Fisher (MvMF) loss function which is able to exploit the spherical geometry of the Earth to improve geolocalization accuracy. In [18], the authors introduced a mixed classification and retrieval scheme, combining the strengths of both methods in a unified solution, achieving new state-of-the-art performance at fine granularity scales. Moreover, Pramanick et al. [25] recently introduced TransLocator, a unified dual-branch transformer network that achieves continent-level accuracy improvement over the existing state-of-the-art methods.

In the context of understanding geolocalization models, a recent study, by [28], introduced a novel semantic partitioning method, capable to enhance the interpretability of prediction results, still achieving state-of-the-art results in terms of geolocalization accuracy on benchmark test sets. One of the best-performing systems proposed to date is [22]. In this work, the authors proposed a classification system, in which the earth is subdivided into geographical cells. Images taken in various types of environments (urban outdoor, indoor or natural) are incorporated, so as to embed in the learning process specific features of several environmental settings. The deep learning architecture used is based on the ResNet network architecture [15]. Results obtained on benchmark datasets demonstrate the capability of this system, positioning [22] as a reference benchmark for image geolocalization. In our experiments, we adapted the framework proposed in [22] to be used as a city verifier, enabling a meaningful comparison with our results on the city verification task.

Most recently, a new direction of research focused exclusively on country recognition. In G^3 [20], the authors showed how language can be leveraged to improve image geolocalization. Their approach involves predicting the country of an image by exploiting a set of clues extracted from a textual guidebook for the GeoGuessr game. A new massive dataset for country recognition, the VIPPGeo dataset, was introduced in [4]. The dataset contains nearly 4 million high-quality urban images. In [4] the authors utilized the VIPPGeo dataset to train a ResNet-based classifier, which achieves state-of-the-art performance in the country recognition task [4]. In this paper, we leveraged the VIPPGeo dataset to construct a dataset containing images from 19 cities around the world.

2.2 Vision Transformers

Transformers were first described in the work by Vaswani et al. [30] based on attention mechanism in natural language processing tasks, e.g., machine translation and question answering. The basic building block of a transformer consists of the multi-head self-attention mechanism that exploits a deep relationship among the elements of embedding words. Vision Transformer (ViT), a variant of transformer targeting computer vision tasks, was first presented in [11] for image classification, by taking a sequence of image blocks as input. Thanks to their outstanding performance, more and more researchers are proposing transformer-based models for improving a wide range of visual tasks, including object detection [38], semantic segmentation [27, 37], image processing [9], and video understanding [5]. Traditional CNNs have gradually been substituted by transformers as the preferred model in the field of computer vision, with several models proposed such as Swinformer, BERT [17], and BEVT [34]. In contrast to recurrent neural networks, transformers are able to focus on the whole sequence, not focusing mainly on short-term dependencies. Moreover, transformers are purely based on the attention mechanisms and their uniqueness consists in an implementation which is optimized for parallelization purposes [8]. As opposed to other approaches, like hard attention [31], which is stochastic in nature and needs Monte Carlo sampling for attention location sampling, transformers scale well to high-complexity models and large-scale datasets. Additionally, pre-trained transformers trained using pretext tasks on large-scale (unlabelled) datasets [17,30] are adopted as starting point of the training procedure, thus significantly reducing the cost of manual annotations.

2.3 Siamese Networks

The Siamese network framework was first proposed by Bromley et al. [7] in 1993 for verification tasks. A basic Siamese network adopts two subnetworks with shared weights as feature extractors. The final decision, then, is made by comparing the outputs of the two subnetworks [19, 33]. Learning knowledge by comparing the features extracted by the two branches instead of directly using labels gives the possibility to learn with unlabeled data and plays an important role in overcoming the limited label issue in real-life applications. In recent years, Siamese architectures have attracted increasing attention in addressing various matching problems, such as object tracking [12], image matching [21], image identification [29] and image change detection [6, 36].

In this work, we combined a Vision Transformer (ViT) model [11] with a Siamese architecture [7] for the city verification task. Specifically, we constructed a two-branch network, with ViT as the backbone. The network is then followed by fully connected layers with ReLU activation functions. The final layer of the network uses a Sigmoid activation function to make the final decision.

3 Methodology

In this section, we describe the city verification dataset and the proposed city verification system.

3.1 City Verification Dataset

The city verification dataset was constructed starting from the VIPPGeo [4] dataset. The VIPPGeo dataset has been built by using three publicly available data sources: Flickr [1], Mapillary [2] and Unsplash [3]. The dataset includes urban pictures with different characteristics, shot with various cameras, from a wide range of different photographers and largely diverse views of world areas. The images were crawled by using the APIs released from each of the 3 data sources. In total, the VIPPGeo dataset contains $3,813,651$ geotagged images.

The city verification dataset consists of $120,909$ images from 19 cities worldwide (refer to Table 2 and 3 for the cities' names.). To obtain the images of each city, we retrieved all the images shot within a circular area around the city centre. The diameter of the circular area varies from 5 Km to 10 Km, depending on the availability of images in the VIPPGeo dataset for that particular city. We used a flexible diameter length to ensure that a fixed number of images were collected for each city in our dataset. Our city verification dataset is partitioned into two sets: Closed and Open sets. We have collected $12,000$ images for each city class in the Closed set and 101 images for each city class in the Open set.

The VIPPGeo dataset has been built by filtering the images to ensure that the dataset contains urban images [4]. To build the city verification dataset, we added additional constraints to maximize the relevance of the dataset for the city verification task. In particular, we employed various filtering strategies to eliminate unsuitable or irrelevant images, like, for example, images containing only faces, natural images, and indoor images. Similar to [4], we used the Places365-CNNs model [39] to implement the above filtering strategy.

The Closed set portion of the city verification dataset consists of 10 city classes, with two pairs of cities belonging to the same country. This subset contains a total of $120,000$ images, $100,000$ out of which were used to train the proposed ViT Siamese network, $10,000$ for validation, $10,000$ for testing, and $1,010$ images for verification. The verification subset dimension is reduced by applying restrictive filtering on the test set and we have used it to evaluate the performance of the overall city verification system, which includes not only the Siamese network but also other auxiliary components. In contrast, we used the testing subset exclusively to assess the performance of the Siamese network. Further details are given in the following sections.

The Open set portion of the dataset consists of 10 city classes, including 9 new cities that were not present in the Closed set, and one city shared with the Closed set. The Open set contains two pairs of cities coming from the same country, with each class containing 101 images. The images in the Open set are much smaller than that of the Closed set since these images were not used for training. The Open set was used exclusively for the verification task. Table 1 summarizes the characteristics of the publicly available city verification dataset.

Table 1: The number of images in the City Verification Dataset.

	Training	Validation	Siamese-testing	Verification
Closed Set	100000	10000	10000	1010
Open Set	0	0	0	1010

3.2 Proposed City Verification System

Figure 1 shows the overall architecture of the proposed city verification system. The core of the system is formed by a Siamese Network [7] with a Vision Transformer backbone (ViT) [11]. The other building blocks include the country classifier (GeoVIPP) introduced in [4], and the Places365-CNNs model [39] to measure the semantic similarity between images. The verification pipeline of our system is outlined in the following:

1. Given an input image (hereafter referred to as query image) and a claim on the city where the image has been taken (hereafter referred to as claimed city), the image is passed through the GeoVIPP [4] country classifier to obtain the Top-2 country predictions.
2. If the country of the claimed city does not appear in the Top-2 country predictions, the image city claim is not verified, and the image is not passed to the subsequent steps of the system.
3. If the country of the claimed city appears in the Top-2 country predictions, the query image is paired with m reference images taken in the claimed city for verification.
4. The image pairs (composed by the query image and the m reference images from the claimed city) are fed to the Siamese network.
5. The semantic similarity between the two images in each pair is calculated using the Places365-CNNs model [39] (refer to Section 3.2.4 for the details).
6. The Siamese network outputs a probability between 0 and 1 for each image pair, with 0 meaning that the images belong to the same city, and 1 that they belong to different cities.
7. The scores given by the Siamese network are weighted according to the image similarities evaluated by the Places365-CNN network and summed together. Eventually, the weighted score is thresholded to verify if the query city has been shot in the claimed city or not.

In the following sections, all the components of the verification system are described in detail, starting with the Siamese network which represents the backbone of the system.

3.2.1 ViT-based Siamese Network

In the proposed framework, for each branch of the Siamese network we have used the ViT-L/16 variant of the Vision Transformer model provided by [11]. We have set the last layer of the ViT model to have 64 output units. Afterwards, we concatenated the output of both networks to form a single-layer feed-forward network with a size of 128 units. The concatenated output is then passed through a Rectified Linear Unit (ReLU) activation function. Then, the output is forwarded to a layer with 64 units, and finally to a Sigmoid activation function. A Sigmoid output equal to or close to 0 indicates that the image pairs come from the same city, while a value close to 1 indicates that the image pairs come from different cities.

Given that the size of the ViT input is fixed ($224 \times 224, 3 - band$ images with 16 patches), and given that the images to be verified have very different dimensions, we adopted a strategy to analyse the entire image content without changing the aspect ratio of the images, since this could affect the performance of the system. During the validation, testing and verification phases, the query image is first resized in such a way that the lower dimension (either width or height) is equal to 256. Then, a crop of size 224×224 is taken from the resized image. The prepossessing step ensures that all images fed to the Siamese network are consistent in size and orientation.

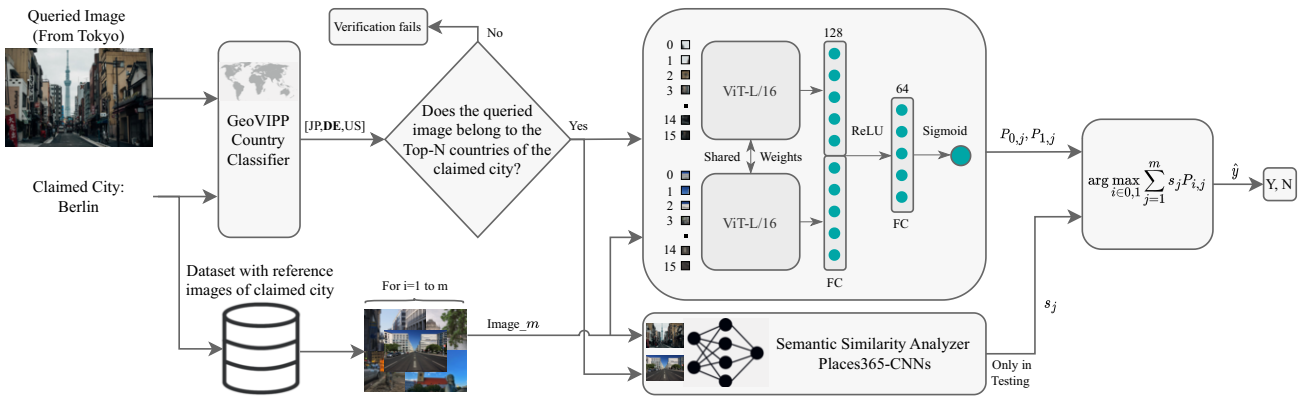

Figure 1: The figure illustrates the overall workflow of the proposed architecture. For more verification examples see the supplementary material. The code and the dataset are publicly available at the following link: https://github.com/alamayreh/city_verifier .

Furthermore, during the training phase, we applied geometric augmentation to prevent overfitting and promote the learning of robust features. The images are first resized so that the lower dimension is equal to 256. Then, we randomly applied horizontal flipping to the training images. Subsequently, we took a random crop covering at least $2/3$ of the image area. Eventually, we resized the crop to 224×224.

In all phases (i.e., training, validation, testing, and verification), the images were normalized to bring the mean pixel values and the standard deviation equal to those of the Imagenet dataset[1].

3.2.2 ViT Siamese Training

Based on the number of images per class in the city verification dataset, which is equal to $10,000$ images, we can form a very large number of possible image pairs that we can use to train the Siamese network. Specifically, the total number of image pairs that we can form is in the order of 5×10^9.

To account for this large number of image pairs and to force the Siamese network to learn robust and discriminative features, without being biased towards any specific city class, we adopted a proper sampling strategy to be used during training. The strategy ensures that for each training epoch, we have a balanced and representative number of positive and negative image pairs. In addition, the sampling strategy ensures that the negative pairs are evenly distributed across different cities, preventing any bias towards specific cities. During each epoch, a total of 16384 image pairs are sampled to train the Siamese network. The large number of image pairs sampled at each epoch provides a diverse and representative sample of the city verification dataset.

The Siamese network was trained for a total of 50 epochs, starting from a ViT model pre-trained on ImageNet [10]. The batch size was set to 32 image pairs. The network was trained using Binary Cross Entropy loss (BCE) and the Stochastic Gradient Descend (SGD) optimizer with a learning rate of $1 \cdot 10^{-2}$ a momentum of 0.9, and a weight decay of $1 \cdot 10^{-4}$. We used the loss function value on the validation set to select the best model for testing and verification. All the experiments have been performed using PyTorch [24] as a Deep Learning framework on a workstation equipped with one Intel(R) Xeon(R) E5-2620 24-Core CPU and four NVIDIA Quadro M6000 12GB GPU.

[1] This is common practice when using models pre-trained on Imagenet.

3.2.3 GeoVIPP Country Classifier

We exploited the country classifier (GeoVIPP) provided in [4] for the city verification task during both training and testing. During training, while sampling the negative pairs to be used within each epoch, we imposed a specific, country plausibility, condition to hold for 50% of the pairs. The condition can be expressed as follow. Suppose we have sampled Image *A* and Image *B* as candidate negative pairs. Before feeding the pair to train the Siamese network, we pass each image through the GeoVIPP country classifier, and we consider the negative image pair *(A, B)* valid if at least one of the Top-2 country predictions of Image *A* is equal to one of the Top-2 country predictions of Image *B*.

Adding this condition helps to ensure that the negative pairs used to train the Siamese network consist of challenging pairs to verify, in the sense that these images are coming from different cities but according to the country classifier, these images may potentially be from the same country. In this way, we also mimic the operating conditions enforced at test time, when the verification automatically fails if the country of the claimed city is not in the Top-2 countries estimated by the GeoVIPP country classifier on the query image.

In the testing phase, the role of the GeoVIPP country classifier is to let the verification fail if the country of the claimed city does not appear in the Top-2 country predictions made by GeoVIPP on the query image.

3.2.4 Semantic Similarity Analyzer

As shown in Figure 1, the final decision is made by accumulating the output of the Siamese network on all the image pairs formed by the query image and the reference image of the claimed city. The output of the Siamese network is weighted according to the semantic similarity between the two images in the pair. The rationale behind this choice is that the output of the Siamese network is more reliable when the input images represent similar scenes. To do so, we exploited the Places365-CNNs classification model [39].

Given an image, Places365 classifies it into one of 365 scenery categories, for example, house, street, river, etc. We reduced the output vector dimensions of the Places365 classification system from 365 to 16 by utilizing the scene hierarchy matrix provided in the same work. The hierarchy categories contain 16 major classes like transportation, forest, industrial, etc. The reduction results in a 16-long vector, whose components give the probability that the scene shown in the input image belongs to the 16 hierarchy scene classes consid-

Table 2: Results on the Closed Set. In each cell, the value on the right shows the total number of images that have successfully passed the GeoVIPP classifier test. The value on the left indicates the fraction of images that passed the overall verification procedure. The overall accuracy is 0.82 for positive pairs (True Positive rate) TP and 0.96 for the negative pairs (off-diagonal elements, True Negative rate $TN = 1 - FP$). The blue cells refer to images from the same cities, while the grey cells indicate the results obtained on different cities of the same country.

City	Amsterdam	Barcelona	Berlin	London	New York	L.Angeles	Rome	Milan	Paris	Tokyo
Amsterdam	0.89 - 98	0.00 - 00	0.02 - 15	0.07 - 24	0.00 - 04	0.00 - 04	0.00 - 01	0.00 - 01	0.01 - 09	0.00 - 00
Barcelona	0.00 - 01	0.83 - 90	0.03 - 06	0.04 - 05	0.01 - 17	0.14 - 17	0.00 - 19	0.15 - 19	0.21 - 26	0.00 - 02
Berlin	0.00 - 04	0.00 - 03	0.89 - 93	0.04 - 07	0.00 - 07	0.04 - 07	0.00 - 02	0.01 - 02	0.11 - 12	0.00 - 03
London	0.04 - 07	0.01 - 02	0.03 - 08	0.71 - 96	0.06 - 24	0.13 - 24	0.00 - 03	0.02 - 03	0.15 - 19	0.01 - 03
NewYork	0.00 - 01	0.00 - 00	0.00 - 05	0.00 - 06	0.91 - 99	0.04 - 99	0.00 - 01	0.00 - 01	0.00 - 02	0.03 - 04
L.Angeles	0.00 - 02	0.12 - 14	0.00 - 01	0.00 - 01	0.10 - 88	0.74 - 88	0.00 - 01	0.00 - 01	0.00 - 02	0.03 - 06
Rome	0.00 - 00	0.05 - 32	0.00 - 01	0.00 - 02	0.00 - 02	0.00 - 02	0.85 - 98	0.09 - 98	0.00 - 23	0.00 - 00
Milan	0.00 - 03	0.15 - 19	0.07 - 11	0.04 - 06	0.03 - 13	0.08 - 13	0.00 - 78	0.70 - 78	0.10 - 12	0.00 - 01
Paris	0.00 - 01	0.16 - 18	0.08 - 10	0.18 - 22	0.00 - 05	0.02 - 05	0.00 - 07	0.06 - 07	0.95 - 99	0.00 - 00
Tokyo	0.00 - 01	0.00 - 02	0.00 - 04	0.01 - 04	0.13 - 17	0.01 - 17	0.00 - 00	0.00 - 00	0.00 - 01	0.80 - 96

ered by Places365-CNN. Given such a vector, the semantic similarity s between the images of each image pair is computed as follows:

$$s = \frac{u \cdot v}{\|u\|_2 \|v\|_2} \tag{1}$$

where u and v are the output probability vectors provided by Places365 model for each image. Then, we used the similarity to weigh the output of the Siamese network. In particular, the final decision is made as:

$$\hat{y} = \arg\max_{i \in 0,1} \sum_{j=1}^{m} s_j P_{i,j} \tag{2}$$

where s is the semantic similarity between the two images in each pair, m is the total number of image pairs, P_0 is the probability that the images pair are coming from the same city and P_1 are coming from a different city. P_1 is equal to the Sigmoid output of the Siamese Network, while P_0 is equal to $(1 - P_1)$.

4 Experiments and Results

We conducted several experiments to evaluate the performance of the proposed City Verifier. We considered different scenarios, including Open and Closed datasets. We also conducted an ablation study to understand the impact of the GeoVIPP classifier on the accuracy of the overall system. Finally, we have compared our results with a verification system built on top of a state-of-the-art image geolocalization system.

4.1 Siamese Training

To start with, we trained the Siamese ViT model using image pairs sampled from the $100,000$ images of the training set. The model is continuously evaluated on $64,000$ image pairs sampled from the $10,000$ images of the validation set. We selected the best model based on the lowest loss value achieved by the trained models on the validation set. Then, The best model is used in the test and verification phases.

Upon training, the performance of the Siamese ViT model was evaluated on 64000 image pairs sampled from the $10,000$ images of the Siamese-testing set (see Table 1 for the dataset details). In the following, we present the results of the evaluation in the form of a confusion matrix (see Figure 2). The matrix indicates that the Siamese Network has an adequate capability to detect if two images belong to the same city. Even if the performance of the Siamese network on single image pairs may appear insufficient, this is good enough to let the overall verification system work, when the Siamese network is applied to all the images in the reference dataset of the claimed city, and when it is used in conjunction with the country classifier which contributes significantly to discard negative image pairs, when they belong to cities of different countries. This aligns with our goal, to maximize the Siamese network's ability to detect images from the same city and to rely on the GeoVIPP classifier to discard images that do not belong to the country of the claimed city.

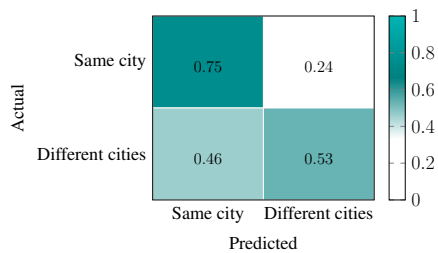

Figure 2: Confusion matrix showing the performance of the Siamese network in determining positive and negative pairs on 64000 image pairs sampled from the Siamese-testing set.

4.2 Verification Results - Closed Set

With regard to the verification task, which we remind is the final goal of our system, the results we got on images representing cities belonging to the Closed set are shown in Table 2. The experiments were conducted on the verification set, including 101 images per city class. As mentioned in the procedure outlined in Section 3.2, we paired each image from the claimed city with the 101 images of the reference dataset of the claimed city. The claimed cities are given in the first column of the table, while the ground truth of the query cities is shown in the first row of the table. For cases where the claimed city and the ground truth of the queried cities are the same (on the diagonal), we held out one image from the 101 images of the claimed city and paired it with the remaining 100 images. Then, we pass the pairs to the verification system. The process is repeated for each image of the 101 images. The fraction of images that passed the entire verification process is reported in the table.

Table 3: Verification accuracy results in the open set scenario. The meaning of the values is the same as in Table 2. The overall True Positive and True Negative rates are 0.71 and 0.98, respectively.

City	Amman	Istanbul	Mexico.C	Singapore	Quebec	Vancouver	Florence	Rome	R.Janeiro	Delhi
Amman	0.79 - 97	0.02 - 03	0.00 - 00	0.00 - 00	0.00 - 01	0.00 - 01	0.00 - 04	0.02 - 04	0.00 - 01	0.03 - 05
Istanbul	0.06 - 07	0.79 - 92	0.00 - 01	0.00 - 00	0.00 - 00	0.00 - 00	0.11 - 13	0.00 - 13	0.00 - 01	0.03 - 04
Mexico.C	0.00 - 00	0.00 - 00	0.55 - 71	0.00 - 00	0.06 - 07	0.00 - 07	0.02 - 04	0.00 - 04	0.01 - 06	0.00 - 01
Singapore	0.00 - 00	0.00 - 00	0.00 - 00	0.66 - 100	0.00 - 00	0.00 - 00	0.00 - 00	0.00 - 00	0.00 - 02	0.01 - 04
Quebec	0.00 - 00	0.00 - 00	0.00 - 01	0.00 - 00	0.70 - 88	0.20 - 88	0.00 - 00	0.00 - 00	0.00 - 00	0.00 - 01
Vancouver	0.00 - 00	0.00 - 00	0.00 - 00	0.00 - 01	0.38 - 86	0.64 - 86	0.00 - 00	0.00 - 00	0.00 - 02	0.00 - 00
Florence	0.00 - 01	0.00 - 01	0.00 - 00	0.00 - 00	0.00 - 00	0.00 - 00	0.77 - 99	0.21 - 99	0.00 - 00	0.01 - 02
Rome	0.00 - 00	0.00 - 02	0.00 - 01	0.00 - 00	0.00 - 00	0.00 - 00	0.20 - 98	0.85 - 98	0.00 - 00	0.00 - 02
R.Janeiro	0.00 - 00	0.00 - 01	0.04 - 08	0.00 - 00	0.00 - 00	0.00 - 00	0.00 - 01	0.00 - 01	0.57 - 81	0.03 - 05
Delhi	0.06 - 09	0.01 - 02	0.00 - 01	0.00 - 00	0.00 - 00	0.00 - 00	0.00 - 01	0.00 - 01	0.00 - 00	0.78 - 99

Each cell of the table reports two values. The value on the right shows the total number of images that have successfully passed the GeoVIPP classifier test, that is, the claimed country belongs to the Top-2 countries predicted by the classifier on the query image. The number on the left, instead, shows the overall verification accuracy, that is the fraction of images that passed the verification procedure. The blue cells (on the diagonal) represent the verification results when the claimed city is correct, while the grey cells indicate results when the query and claimed cities are different, but from the same country.

As shown by the table, the verification system demonstrates very good performance, even in the difficult case when the claimed and query cities belong to the same country[2]. As an overall performance metric, we considered the True Positive (TP) and False Positive (FP) rates, defined as:

$$TP = \frac{1}{C}\sum_{i=1}^{C} a_{i,i} \tag{3}$$

$$FP = \frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C} a_{i,j}, \tag{4}$$

where C is the total number of cities and $a_{i,j}$ are the left values reported in Table 2. In particular, TP in Equation (3) reports the accuracy of the verification when the query city corresponds to the claimed one (the larger the better), while FP in Equation (4) gives the probability that a false claim is verified (the lower the better). The values of TP and FP calculated from Table 2 are 0.82 and 0.03, respectively.

4.3 Verification Results - Open Set

We also evaluated the performance of the verification in an open set scenario where both the query and claimed cities do not belong to the set of cities used during training. The Open set portion of the dataset consists of 10 cities, each city class containing 101 images. The only exception to the open set rule is represented by the images of Rome, given that Rome was also included in the Closed set portion of the dataset. The reason for this exception is that we wanted to evaluate the accuracy of the system in a mixed scenario where the query (res. claimed) city has been seen during training and the claimed (res. query) city has not.

[2] In this case, in fact, the country classifier is of no help.

As illustrated in Table 3, the city verifier maintains very good performance also in the open set scenario, especially for cities belonging to different countries. A certain performance drop can be observed when the claimed and query cities are different but belong to the same country, however, even in this difficult case, the verifier maintains a certain capability to recognize if the query image depicts the claimed city or not. Applying Equations (3) and (4) to Table 3,. we now get $TP = 0.72$ and $FP = 0.02$.

4.4 Ablation Study

As a further investigation, we conducted an ablation study that involves systematically removing and tuning the GeoVIPP country classifier [4] at different stages of the process and evaluating its impact on the overall performance. By carefully analyzing the results, we were able to gain insights into the contribution and importance of the GeoVIPP classifier in the proposed city verification framework.

Table 4: Impact of GeoVIPP country classifier on n verification accuracy at different stages of the verification pipeline.

GeoVIPP		No use	Sampling	Verification	Sampling and Verification
Closed Set	(TP)	0.9079	0.8782	0.8544	**0.8287**
	(FP)	0.2639	0.3410	0.0391	**0.0388**
Open Set	(TP)	0.7772	0.7851	0.7118	**0.7128**
	(FP)	0.3785	0.4832	0.0242	**0.0209**

We started first by removing the GeoVIPP [4] classifier from the system. Then, we measured the verification accuracy on the Open and Closed city verification datasets using Equations (3) and (4). The results are reported in Table 4. The column labelled *No use* refers to a case where the GeoVIPP classifier is not used at all, while the *Sampling* column shows the results when the GeoVIPP is only used in the sampling stage of pairs during training of the Siamese network. The *Verification* column represents the case where the GeoVIPP is not used during training, but only during verification. Finally, the last column refers to the case where the GeoVIPP is used in all stages. Upon inspection of the Table, we can see that incorporating the GeoVIPP country classifier in different stages of the city verification leads to a performance increase in both open and closed sets scenarios.

We also conducted an experiment to investigate the effect of changing the Top-N prediction of the GeoVIPP classifier used in the verification process. Increasing the value of N means allowing more

images to be passed to the city verification system, with an expected positive effect on the True Positive rate (diagonal values in Tables 2 and 3) and a negative effect on on the False Positive rate (off-diagonal values in Tables 2 and 3). As we can see from Figure 3, choosing N involves finding a trade-off between TP and FP. According to our experiments, a good trade-off is obtained by letting $N = 2$.

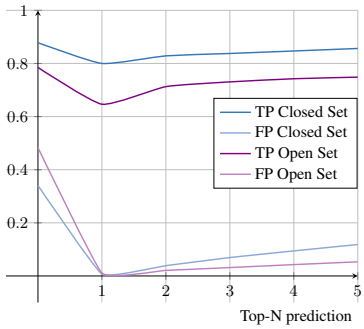

Figure 3: Verification accuracy using different Top-N predictions by GeoVIPP country classifier (see the supplementary material for the full tables obtained by using Top-1 and Top-3 results.

4.5 *Comparisons with State-of-the-Art*

As a last set of experiments, we compared the results of our system with those obtained by relying on the state-of-the-art ISN method described in [22]. Such a method provides an estimation of the GPS location where the image was captured. As to comparing our results with the state-of-the-art, we can only compare with methods estimating image GPS location. However, the source code of such methods is only rarely available and the results reported in the papers refer to different kinds of problems. The code of the method described in ISN [22] is available, making it suitable for our comparison, all the more that such a system is a well-recognized benchmark in the field.

To turn the system in [22] into a city verifier, we defined circles around the city centres with different diameters, precisely 25 Km, 50 Km, and Flexible. A city claim is positively verified if the GPS estimation of the query city falls within the predefined radius of the claimed city. The Flexible diameter was adjusted until we achieved the same False Positive rate FP of our system. This yielded a radius equal to 463 Km and 711 Km for the open and closed set cases, respectively. In particular, we used the Haversine[3] formula to measure the distance between the city centre coordinates and the estimated geo-coordinates [22]. By using the procedure outlined in Section 3.2, we constructed tables similar to Table 2 and Table 3 for both the Closed and Open sets (such tables are available as supplementary material). Then, we applied Equation (3) and Equation (4) to calculate TP and FP.

Figure 4 illustrates the TP rates obtained when the claimed and the query cities are the same. As we can see from the figure, the proposed method outperforms ISN even when we relax the diameter constraint and enlarge the predefined circle around the city centres. In Figure 4, the slightly higher verification accuracy obtained on the closed set compared to the open set case, can be explained by the fact that some cities in the Open set have fewer images in the original dataset on which ISN has been trained [22]. With regard to the False Positive rate, we got the results reported in Table 5. As shown in the table, both our system and ISN achieve good performance.

[3] The Haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes.

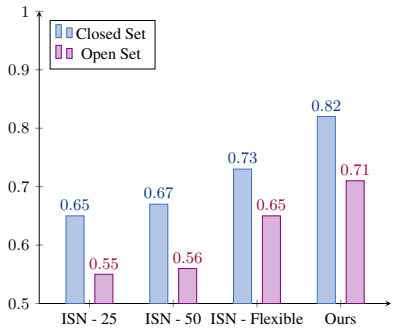

Figure 4: True Positive rate comparison with SOTA [22] for various values of proximity circle diameters (from left to right, 25 Km, 50 Km, and Flexible).

Table 5: False Positive rate comparison with SOTA [22].

	ISN-25	ISN-50	ISN-Flexible	Ours
Closed Set	0.0049	0.0051	0.0388	0.0388
Open Set	0.0009	0.0009	0.0209	0.0209

5 Conclusion

We have introduced the city verification task, a new instance in the class of image geolocalization problems and presented a novel system to address it. The proposed system uses a Siamese network backboned with Vision Transformer, coupled with a country classifier and Semantic similarity analyser. Given a query image and a small set of images taken in a target city, the system accurately determines whether the query image was taken in the target city or not.

The city verification problem involves two key factors: i) the vast number of cities worldwide, and ii) the variability in scenes from the same city. Our Siamese-based verifier effectively addresses both challenges. Firstly, it is not necessary to retrain the system to handle images belonging to cities that have not been used in the training set. Secondly, our system gives more voting power to the reference images that are semantically similar to the query image. Moreover, our system compares the query image with several reference images of the claimed city. We argue that if the number of reference images is large enough, then the verifier can handle the variability of images from the same city properly.

While our system shows promising results, we acknowledge that the experiments we carried out can only prove the plausibility of the arguments our system relies on. Thus, future research should focus on improving the generality and scalability of our approach. Additionally, optimizing the choice of images in the reference dataset and adopting more sophisticated fusion strategies could further enhance the system's performance.

Acknowledgements

This work has been partially supported by the PREMIER project under contract PRIN 2017 2017Z595XS-001, funded by the Italian Ministry of University and Research, and by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

[1] Flickr website. flickr.com.
[2] Mapillary website. mapillary.com.
[3] Unsplash website. unsplash.com.
[4] Omran Alamayreh, Giovanna Maria Dimitri, Jun Wang, Benedetta Tondi, and Mauro Barni, 'Which country is this picture from? new data and methods for dnn-based country recognition', in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, (2023).
[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, 'Vivit: A video vision transformer', in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, (2021).
[6] Wele Gedara Chaminda Bandara and Vishal M Patel, 'A transformer-based siamese network for change detection', in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 207–210. IEEE, (2022).
[7] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah, 'Signature verification using a" siamese" time delay neural network', *Advances in neural information processing systems*, **6**, (1993).
[8] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath, 'An attentive survey of attention models', *ACM Transactions on Intelligent Systems and Technology (TIST)*, **12**(5), 1–32, (2021).
[9] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao, 'Pre-trained image processing transformer', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12299–12310, (2021).
[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'Imagenet: A large-scale hierarchical image database', in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, (2009).
[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., 'An image is worth 16x16 words: Transformers for image recognition at scale', *arXiv preprint arXiv:2010.11929*, (2020).
[12] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang, 'Learning dynamic siamese network for visual object tracking', in *Proceedings of the IEEE international conference on computer vision*, pp. 1763–1771, (2017).
[13] James Hays and Alexei A Efros, 'Im2gps: estimating geographic information from a single image', in *2008 ieee conference on computer vision and pattern recognition*, pp. 1–8. IEEE, (2008).
[14] James Hays and Alexei A Efros, 'Large-scale image geolocalization', *Multimodal location estimation of videos and images*, 41–62, (2015).
[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).
[16] Mike Izbicki, Evangelos E Papalexakis, and Vassilis J Tsotras, 'Exploiting the earth's spherical geometry to geolocate images', in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pp. 3–19. Springer, (2020).
[17] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of NAACL-HLT*, pp. 4171–4186, (2019).
[18] Giorgos Kordopatis-Zilos, Panagiotis Galopoulos, Symeon Papadopoulos, and Ioannis Kompatsiaris, 'Leveraging efficientnet and contrastive learning for accurate global-scale location estimation', in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pp. 155–163, (2021).
[19] Yikai Li, CL Philip Chen, and Tong Zhang, 'A survey on siamese network: Methodologies, applications, and opportunities', *IEEE Transactions on Artificial Intelligence*, **3**(6), 994–1014, (2022).
[20] Grace Luo, Giscard Biamby, Trevor Darrell, Daniel Fried, and Anna Rohrbach, 'Gˆ 3: Geolocation via guidebook grounding', *arXiv preprint arXiv:2211.15521*, (2022).
[21] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu, 'Siamese network features for image matching', in *2016 23rd international conference on pattern recognition (ICPR)*, pp. 378–383. IEEE, (2016).
[22] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth, 'Geolocation estimation of photos using a hierarchical model and scene classification', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 563–579, (2018).
[23] Apostolos Panagiotopoulos, Giorgos Kordopatis-Zilos, and Symeon Papadopoulos, 'Leveraging selective prediction for reliable image geolocation', in *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II*, pp. 369–381. Springer, (2022).
[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., 'Pytorch: An imperative style, high-performance deep learning library', in *Advances in Neural Information Processing Systems*, (2019).
[25] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa, 'Where in the world is this image? transformer-based geo-localization in the wild', in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pp. 196–215. Springer, (2022).
[26] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han, 'Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 536–551, (2018).
[27] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid, 'Segmenter: Transformer for semantic segmentation', in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, (2021).
[28] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth, 'Interpretable semantic photo geolocation', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 750–760, (2022).
[29] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang, 'A siamese long short-term memory architecture for human re-identification', in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pp. 135–153. Springer, (2016).
[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', *Advances in neural information processing systems*, **30**, (2017).
[31] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, 'Show and tell: A neural image caption generator', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, (2015).
[32] Nam Vo, Nathan Jacobs, and James Hays, 'Revisiting im2gps in the deep learning era', in *Proceedings of the IEEE international conference on computer vision*, pp. 2621–2630, (2017).
[33] J Wang, B Tondi, and M Barni, 'An eyes-based siamese neural network for the detection of gan-generated face images', *Front. Sig. Proc. 2: 918725. doi: 10.3389/frsip*, (2022).
[34] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan, 'Bevt: Bert pretraining of video transformers', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14733–14743, (2022).
[35] Tobias Weyand, Ilya Kostrikov, and James Philbin, 'Planet-photo geolocation with convolutional neural networks', in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pp. 37–55. Springer, (2016).
[36] Panli Yuan, Qingzhan Zhao, Xingbiao Zhao, Xuewen Wang, Xuefeng Long, and Yuchen Zheng, 'A transformer-based siamese network and an open optical dataset for semantic change detection of remote sensing images', *International Journal of Digital Earth*, **15**(1), (2022).
[37] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, et al., 'Segvit: Semantic segmentation with plain vision transformers', *Advances in Neural Information Processing Systems*, **35**, 4971–4982, (2022).
[38] Zixiao Zhang, Xiaoqiang Lu, Guojin Cao, Yuting Yang, Licheng Jiao, and Fang Liu, 'Vit-yolo: Transformer-based yolo for object detection', in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2799–2808, (2021).
[39] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, 'Places: A 10 million image database for scene recognition', *IEEE transactions on pattern analysis and machine intelligence*, **40**(6), 1452–1464, (2017).