IM2City: Image Geo-localization via Multi-modal Learning

Meiliu Wu
mwu233@wisc.edu
Spatial Computing and Data Mining Lab
University of Wisconsin-Madison
Madison, Wisconsin, USA

Qunying Huang
qhuang46@wisc.edu
Spatial Computing and Data Mining Lab
University of Wisconsin-Madison
Madison, Wisconsin, USA

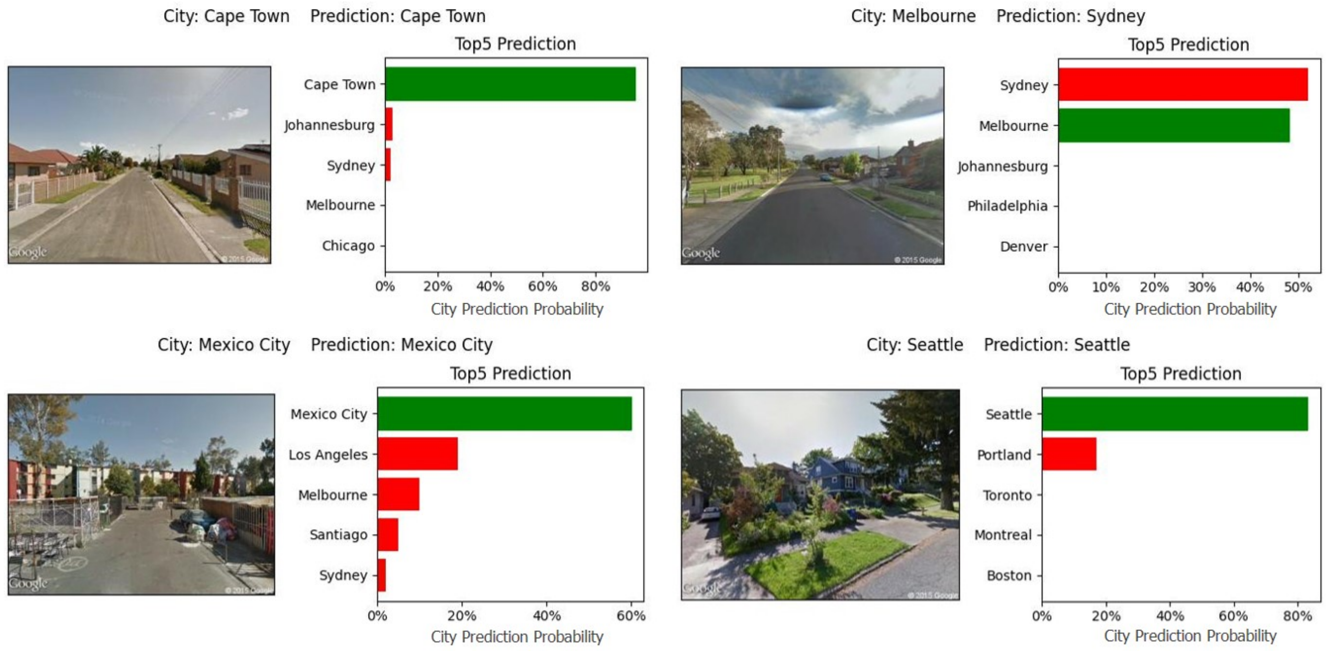

Figure 1: Traditionally image geo-localization is accomplished by two approaches: 1) image matching between the query image and the reference images, or 2) a classification-based scheme that assigns an image to a grid on the world map. Here we show that a street view image can be geo-localized at a city level in the global scale, based on multi-modal learning with natural language and computer vision.

ABSTRACT

This study investigated multi-modal learning as a stand-alone solution to image geo-localization problems. Based on the successful trials on the contrastive language-image pre-training (CLIP) model, we developed **GE**o-localization **M**ulti-modal (GEM) models, which not only learn the visual features from input images, but also integrate the labels with corresponding geo-location context to generate textual features, which in turn are fused with the visual features for image geo-localization. We demonstrated that simply utilizing the image itself and appropriate contextualized prompts (i.e., mechanisms to integrate labels with geo-location context as textural features) is effective for global image geo-localization, which traditionally requires large amounts of geo-tagged images for image matching. Moreover, due to the integration of natural language, our GEM models are able to learn spatial proximity of geo-contextualized labels (i.e., their spatial closeness), which is often neglected by classification-based geo-localization methods. In particular, the proposed Zero-shot GEM model (i.e., geo-contextualized prompt tuning on CLIP) outperforms the state-of-the-art model - Individual Scene Networks (ISN), obtaining 4.1% and 49.5% accuracy improvements on the two benchmark datasets, IM2GPS3k and Place Plus 2.0 (i.e., 22k street view images across 56 cities worldwide), respectively. In addition, our proposed Linear-probing GEM model (i.e., CLIP's image encoder linearly trained on street view images) outperforms ISN even more significantly, obtaining 16.8% and 71.0% accuracy improvements, respectively. By exploring optimal geographic scales (e.g., city-level vs. country-level), training datasets (street view images vs. random online images), and pre-trained models (e.g., ResNet vs. CLIP for linearly probing), this research sheds light on integrating textural features with visual features for image geo-localization and beyond.

GeoAI '22, November 1, 2022, Seattle, WA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9532-8/22/11…$15.00
https://doi.org/10.1145/3557918.3565868

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Scene understanding**.

KEYWORDS

image geo-localization, street view images, multi-modal learning, computer vision, natural language

ACM Reference Format:

Meiliu Wu and Qunying Huang. 2022. IM2City: Image Geo-localization via Multi-modal Learning. In *The 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI '22) (GeoAI '22), November 1, 2022, Seattle, WA, USA.* ACM, Seattle, WA, USA, 12 pages. https://doi.org/10.1145/3557918.3565868

1 INTRODUCTION

Location is the crucial contextual information for many image applications, such as social media geo-tagging and image searching. For example, if we are collecting photos of tourist attractions, then the pictures taken from popular tourism cities would be desired; if studying the characteristics of an urban landscape, the images from various sources (e.g., bird-viewed remote sensing and street-viewed open scenes) within the targeted urban area would be of interest. However, such location information is often missing in many image datasets. Therefore, image geo-localization has been an important task for decades in Computer Vision [3, 26]. Currently, there are two main approaches to address this task - one is image matching, and the other is classification-based. The former approach uses geo-tagged images as references and infers the geo-location of a query image based on its most similar reference images (e.g., [8, 9]), and the latter approach partitions the world map into grids and trains a model to classify a given query image to one of the grids (e.g., [21, 25]).

There are several limitations in the first approach, i.e., image matching. First, from operational perspectives, previous studies are tied with an undesirable prerequisite - a large, off-the-shelf database with geo-referenced images. Moreover, the geo-localization performance is largely limited by the spatial coverage of those available reference images. Second, the process of image matching involves handcrafting the similarity functions between the query images and reference images in order to retrieve the optimal matching, which often ended up to be a task-specific training due to spatial variability in scale as well as changes in viewpoint and lighting [13]. Third, this approach involves significant costs in terms of the indexing space and computational complexity, since both the query and reference images need to be processed for feature extraction and similarity measurement. In other words, the geo-localization accuracy of image matching methods is highly constrained by the availability of reference images as well as the computational capacity allocated.

Later, the second approach, i.e., classification-based methods, was proposed to solve the limitations of image matching. After the training stage, the classification model can directly perform the forward propagation to predict the geo-location of any query image, in which only the model parameters (i.e., weights) are stored in memory, largely saving computational space and time. However, classification-based methods also have deficiencies. Such methods typically neglect spatial relationships of the proximate grids. For instance, classifying an image of New York City to Philadelphia is treated equally wrong as classifying it to Tokyo. This is due to the fact that classification-based methods often consider geo-localization as a labeling process, in which "labels"/"classes" are categorical variables without any relationships or semantics implicitly representing spatial proximity. However, we argue that geo-localization is beyond "classification", and spatial proximity of the grids should be taken into consideration. To address this issue, we develop **GE**o-localization **M**ulti-modal (GEM) models, which use both natural language and computer vision techniques for image geo-localization tasks. Our first model, Zero-shot GEM, is built upon the Contrastive Language-Image Pre-training (CLIP) model [19], and further tuned by the geo-contextualized prompts (i.e., integrating labels with geo-location context) that are designed for image geo-localization. Next, also built upon CLIP, we developed the second model, Linear-probing GEM, by linearly training the image encoder with Google Street View images.

Our theoretical foundation is that geo-location text is often part of the description of visual features (e.g., language-image pairwise data), and similar visual features can help build the connection of similar geo-location text. Specifically, contrastive learning, a widely used deep learning paradigm where similar samples are pushed towards each other in the embedding space while those dissimilar samples are pulled against each other, can be applied to learn the linkage between visual features and their corresponding geo-contextualized text. The feasibility of contrastive learning has been preliminarily demonstrated by CLIP, in which an image encoder and a text encoder are trained simultaneously [19]. Meanwhile, we propose that a suitable geographic scale (e.g., city-level vs. country-level) and a desirable training dataset (e.g., Google Street View images vs. random online images) should also be considered for image geo-localization.

To sum up, this study contributes to four main aspects: (1) a methodological framework of how to integrate textural features with visual features for image geo-localization is proposed, featured by an in-depth investigation of the suitability of geographic scales, training datasets, and pre-trained models; (2) multiple experiments are conducted for prompt tuning and prompt ensembling, shedding light on how to incorporate appropriate prompts with geo-contextualized labels; (3) the effectiveness of combining textural features with visual features for image geo-localization is proven by the strongly boosted performances of our GEM models, compared with the state-of-the-art model - Individual Scene Networks (ISN) [16] - in different benchmarks; and (4) we demonstrate that this multi-modal learning fashion not only helps learn spatial relationships of the geo-contextualized labels, but also enriches the geo-localization results with fruitful semantics (e.g., administrative cities instead of arbitrary grids), which is more practical for downstream applications based on image geo-localization, e.g., searching images by geo-location text and supplementing geo-locations to image datasets as meta-information.

2 RELATED WORK

Image geo-localization can be generally defined as: given an image, where was it taken? This is a cutting-edge yet challenging problem in Computer Vision [3, 26], and is mainly addressed by two streams - image matching and classification-based methods. This section will first elaborate the strengths and weaknesses of previous work using image matching (Section 2.1) and classification-based methods (Section 2.2). Next, given the recent promising development of multi-modal learning based on language-image [19], we also look into the current progress in applying this technique to image geo-localization in Section 2.3.

2.1 Image Matching for Geo-localization

Image matching consists of two steps: (1) learning visual features from both query images and reference images, and (2) measuring the similarity of the features extracted from the two sets of images in order to find the best match. IM2GPS [8, 9] was the first attempt of applying image matching for global image geo-localization. It was trained on 6 million geo-tagged images collected from Flickr. To geo-localize a query image, IM2GPS performs the k-nearest neighbors algorithm in all reference images, and outputs the average GPS coordinates of the returned images. Later, IM2GPS-deep [24] significantly improved the results of the IM2GPS by using deep learning models to extract features. On the other hand, as reference images are stored in a large database, many studies also put efforts on constructing a database where the most similar reference images can be retrieved efficiently [1, 17, 27]. Nonetheless, this image matching approach is inevitably limited by its expensive cost in both space and time, making it unrealistic for large-scale image geo-localization in practice [11].

2.2 Classification-based Geo-localization

Classification-based methods, which formulate geo-localization as a classification task on worldwide grids, were introduced to overcome the limitation of image matching methods in space and time complexity. The original workflow is straightforward: first, the world map is partitioned into grids; second, a classification model is used to assign a query image to one of the grids. The predicted geo-location of a query image is the center of the predicted grid.

The first implementation of this classification-based idea is the PlaNet algorithm [25], in which the grid-based system is constructed recursively: (1) the world is initially partitioned into 6 grids; (2) the grid with the most images is further partitioned into smaller grids; and (3) step 2 is repeated until the total number of grids T (as a hyperparameter) is reached. Later, the CPlaNet algorithm [21] and the Individual Scene Networks (ISN) approach [16] were proposed to reduce the total number of grids T by applying multiple cross entropy output layers. However, the optimal T still requires careful tuning. Additionally, all these classification-based methods use the Cross-entropy Loss for training, which is not necessarily correlated with geo-localization accuracy [11]. Most importantly, treating image geo-localization as a simple "classification" task is conceptually problematic, as this process considers the grids as categorical variables without any relationships or semantics, leading to the failure of revealing spatial relationships of the proximate grids.

2.3 Multi-modal Learning for Image Geo-localization

Recently, multi-modal learning that combines natural language and computer vision has shown great potentials in solving the issues of classification-based methods. First, it is more flexible to scale natural language supervision compared with grid-based labeling for image geo-localization. For instance, city-level geo-localization can simply use city names as labels when natural language is integrated, instead of constructing another grid-based system that is suitable at a city level. This flexibility is also applicable for other scales such as country-level and continent-level in natural language supervision. Second, such multi-modal learning not only learns a visual representation, but also links that representation to corresponding geo-contextualized text, making it more feasible for the model to understand spatial relationships of geo-locations across different scales.

The first work that applied multi-modal learning for image geo-localization is well-known as the CLIP model [19], which was built on 400 million text-image pairwise data collected from the Internet. The text encoder and image encoder in the CLIP model are trained simultaneously via contrastive learning. Specifically, the paired dot products of the outputs from the two encoders would have the largest values (representing the largest probability), while the non-paired dot products would have lower values. For the downstream task of image geo-localization, the authors built a new dataset called "Country211", which is a subset of the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset [22], by filtering out the countries that have at least 300 images with GPS coordinates, and sampling 200 images for training and 100 images for testing, for each country. CLIP achieves 34.9% accuracy for zero-shot and 46.4% after linear-probing. To compare with the state-of-the-art [16] in the IM2GPS benchmark, the geo-location of a query image is predicted as the GPS coordinates of the nearest image in a set of reference images using CLIP's embedding space, and the performance is not as competitive as [16].

We emphasize that one of the most important contributions from CLIP is using natural language as a training signal for image geo-localization. As geo-contextualized text are embedded as representative vectors, those similar vectors can indicate their spatial proximity since they are likely to be linked to words that describe similar geo-contexts (e.g., similar street views). However, CLIP still suffers from a few limitations in this country-level geo-localization task. First, the data overlapping rate between the 400-million pre-training dataset and Country211 is 21.5%. Second, despite this large overlap, Zero-shot CLIP only exhibits a 0.2% increase in accuracy compared with a ResNet-50 trained on only YFCC100M. The speculation from the authors is that the pre-training text paired with an image is often not related to the country of the image. In fact, a better scale for geo-localization based on language-image learning is probably the city level, as city level is more fine-grained and informative, and therefore more likely to be concurrent or geo-tagged with online images. Third, during zero-shot prediction, the CLIP model used `"a photo i took in {label}"` as the only geo-contextualized prompt for country-level geo-localization. As Table 1 shows, more experiments should be conducted based on prompt engineering (i.e., the description of the task embedded together with the textual input) [4, 20], since using inappropriate geo-contextualized prompts may incur a decreased accuracy compared with no prompt used (Section 4.1). Unfortunately, the authors did not put more efforts on such experiments. Lastly, the suitability of an image used for geo-localization tasks should be assessed [18]. For instance, an image of an ordinary mug or cat from the Internet probably provides few useful geo-location cues. Therefore, the 400-million pre-training dataset should be further evaluated before used for image geo-localization.

3 METHODOLOGY

To address the current deficiencies of multi-modal learning, we design our methodology by considering three aspects: (1) the appropriate scale that should be applied for global image geo-localization based on language-image pairwise input (Section 3.1); (2) the desirable dataset for image geo-localization and model training (Section 3.2); and (3) the pre-trained models that can provide useful backbones for text tokenization as well as visual feature extraction (Section 3.3).

3.1 Scale

The first consideration for global image geo-localization is the scale. Traditionally, the scale involves street-level (1 km), city-level (25 km), region-level (200 km), country-level (750 km), and continent-level (2500 km) [16, 21, 25]. The evaluation metric is the percentage of images correctly localized within one of these given radii. In this work, we choose city-level because when language-image data are employed, city-level is possibly the most fine-grained scale that can provide more distinguishable geo-contextualized labels (i.e., administrative city names), while street-level may incur too many confusions globally (e.g., there are 10,866 "Second Street" in United States according to the 1993 U.S. Census report) and country-level can be too coarse-grained to be used in geo-tags and therefore less likely to be concurrent with images.

3.2 Data

As discussed in Section 2.3, we should assess the suitability of an image used for geo-localization tasks [18]. Specifically, not all images are worthy to be used for training or testing a model in performing global image geo-localization at a city level. For our experiments, we suggest that compared with random online images, street view images can provide more concrete and subtle visual features in urban areas, and therefore are more appropriate to be used to differentiate cities in the world. In addition, geo-localizing street view images has become a main research trend [2, 27], as these images have become largely available in public and can provide fine-grained and representative information for many downstream tasks in urban studies [28].

The **Place Plus 2.0** dataset [7], which consists of approximately 111 thousand Google Street View images from 56 major cities within 28 countries around the world captured between years 2007 and 2012, will be used to train our Linear-probing GEM model. These images were collected with latitude-longitude coordinates uniformly sampled from grids that are spatially intersected with city boundaries. Additionally, a subset of it (20%) will be used as a new benchmark to evaluate and compare the performances of our GEM models, the state-of-the-art model, and the baseline. The spatial distribution of the cities in the Place Plus 2.0 dataset is displayed in Figure 8.

Besides, we also employ another public benchmark dataset for model evaluation — IM2GPS3k [24], which contains 3,000 images originated from the IM2GPS dataset [8]. IM2GPS is a dataset of about 6.5 million images collected from the Internet, and these images were tagged with both GPS coordinates and geographic keywords, including every country and territory, the 200 most populated cities in the world, every US state, and popular tourist sites (e.g., “Pisa”, “Nikko”, “Orlando”).

3.3 Models

This work develops two GEM models, i.e., Zero-shot GEM and Linear-probing GEM, both of which are based on the pre-trained CLIP model using language-image pairwise data as input [19]. The former will be tuned by prompt engineering, and the latter will be trained on the Place Plus 2.0 dataset. For a further comparison, we also constructed a Linear-probing ResNet-152 model [10] pre-trained on ImageNet [5] as a baseline given its widespread adoption and robust performance on different benchmarks.

Model 1: Zero-shot GEM. Our implementation follows the practice in [19] with three steps as follows: 1) Text Encoding; 2) Image Encoding; and 3) Loss Calculation by Contrastive Learning.

Text Encoding. The text encoder is a Transformer [23] modified by [20]. It is a 12-layer, 768-width, and 12-attention-head model. The activations of the last layer of the Transformer are treated as the feature representation of the text input $T^{input}_{[n,d_t]}$, where n is the size of minibatch, and d_t is the feature dimensions of each text sample. This feature matrix is then layer-normalized and linearly projected into the multi-modal embedding space as $T^{output}_{[n,d_e]}$, where d_e is the dimensions of joint multi-modal embedding (Equation 1). Particularly, for image geo-localization, multiple geo-contextual prompts are constructed for prompt tuning (See Figure 2(a)) and the results are shown in Table 1.

$$T^{output}_{[n,d_e]} = \left\| T^{input}_{[n,d_t]} \times W^{T}_{[d_t,d_e]} \right\|^2 \tag{1}$$

Image Encoding. The image encoder is based on Vision Transformer ViT-L/14 [6], which we found performs best compared with other available image encoders (e.g., EfficientNet-style RN50x64, ViT-B/32, and ViT-B/16) in the CLIP model. It is a 24-layer, 1024-width, and 16-attention-head model. Similarly, the image feature matrix $I^{input}_{[n,d_i]}$, where n is the size of minibatch and d_i is the feature dimensions of each image sample, is then layer-normalized and linearly projected into the multi-modal embedding space as:

$$I^{output}_{[n,d_e]} = \left\| I^{input}_{[n,d_i]} \times W^{I}_{[d_i,d_e]} \right\|^2 \tag{2}$$

Loss Calculation by Contrastive Learning. After computing the feature embeddings of the images and the feature embeddings of the paired texts by the two encoders, the

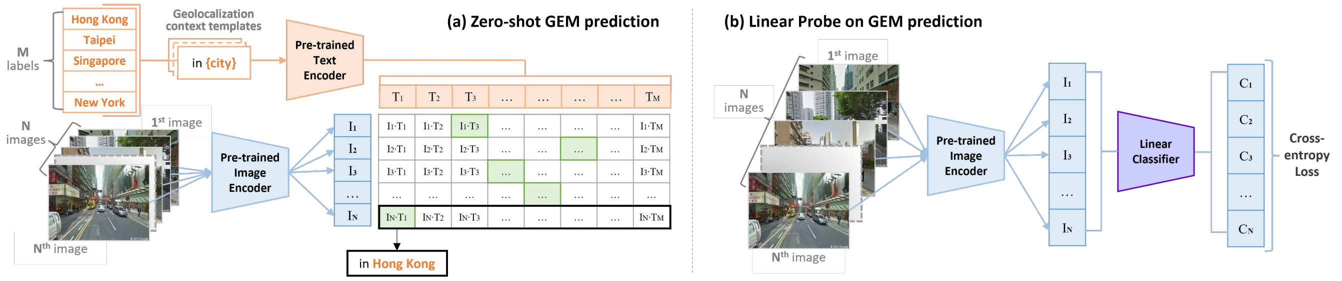
(a) Zero-shot GEM prediction
M labels
Hong Kong
Taipei
Singapore
...
New York
Geolocalization context templates
in {city}
Pre-trained Text Encoder
T1 T2 T3 … … … … TM
N images
1st image
Nth image
Pre-trained Image Encoder
I1 I2 I3 … IN
I1·T1 I1·T2 I1·T3 … … … … I1·TM
I2·T1 I2·T2 I2·T3 … … … … I2·TM
I3·T1 I3·T2 I3·T3 … … … … I3·TM
IN·T1 IN·T2 IN·T3 … … … … IN·TM
in Hong Kong
(b) Linear Probe on GEM prediction
N images
1st image
Nth image
Pre-trained Image Encoder
I1 I2 I3 … IN
Linear Classifier
C1 C2 C3 … CN
Cross-entropy Loss

Figure 2: Architectures of our models: a) Zero-shot GEM: it consists of two encoders - one is the image encoder that learns visual features from input images and outputs a representation vector for each image n, i.e., I_n, and the other is the text encoder that embeds the geo-contextualized labels and outputs a representation vector for each label m, i.e., T_m, For the prediction of each image n, the largest dot product among all $(I_n \cdot T_m)$ indicates the predicted label; b) Linear-Probing GEM: a linear classifier (i.e., Logistic Regression) is trained on the last-step feature maps from the image encoder in the supervised fashion.

cosine similarity of these two sets of embeddings is then calculated and scaled by a temperature parameter t (Equation 3). Finally, the scaled pairwise cosine similarities are normalized by softmax, and used to measure the symmetric cross entropy loss (Equation 4). For more details of the pre-training updates, the Adam optimizer [12] was used with decoupled weight decay regularization [15] for all weights (not gains or biases), and the learning rate (initialized as 4×10^{-4}) was decayed using a cosine schedule [14].

$$logits = S_c(T^{output}_{[n,d_e]}, (I^{output}_{[n,d_e]})^{\top}) \times e^t \tag{3}$$

where S_c is the cosine similarity and t is the temperature parameter. The symmetric loss function is defined as below:

$$\begin{aligned} loss_T &= -\sum_{c=1}^{M} y_{i,c} \log(logits^T_{i,c}) \\ loss_I &= -\sum_{c=1}^{M} y_{i,c} \log(logits^I_{i,c}) \\ loss &= (loss_T + loss_I)/2 \end{aligned} \tag{4}$$

where M is number of classes, log is the natural log, y is a binary indicator (0 or 1) if class label c (c = [1, ..., n]) is the correct classification for image i (i = [1, ..., n]), $logits^T_{i,c}$ is the predicted probability that image i is of class c using $logits$ in Equation 3 normalized along the Text axis via a softmax, and $logits^I_{i,c}$ is the predicted probability that image i is of class c using $logits$ normalized along the Image axis via a softmax.

Model 2: Linear-probing GEM. The image encoder from the CLIP-ViT-L/14 model (without the final classification layer) is used as the visual feature extractor. Specifically, for each image, a feature vector is output by this pre-trained image encoder, and then fed into a linear classifier for model training on the Place Plus 2.0 dataset (See Figure 2(b)). We train a logistic regression classifier based on the scikit-learn implementation, with Cross-entropy Loss calculated, L-BFGS as the optimization strategy, and maximum 1,000 iterations.

Model 3: Linear-probing ResNet-152. Likewise, a simple supervised baseline of a linear classifier is trained on top of ResNet-152 [10] features from the Place Plus 2.0 images, with Cross-entropy Loss calculated, Stochastic Gradient Descent (SGD) as the optimizer, and learning rate as 0.001. Note that all models are trained with 32 epochs.

Table 1: Prompt engineering can improve Zero-shot GEM accuracy.

Geo-contextualized Prompts	Top1 Accuracy
Prompt #1: "to {label}"	46.24%
Prompt #2: "city of {label}"	61.69%
no prompt	61.72%
Prompt #3: "at {label}"	61.77%
Prompt #4: "from {label}"	62.52%
Prompt #5: "in {label}"	63.23%
Prompt ensembling (softmax weighted)	**64.42%**

4 EXPERIMENTS

With our methodology described above, several experiments were conducted. First, prompt engineering techniques were applied to the Zero-shot GEM, e.g., the combination of different geo-contextualized prompts with the city names, and the ensemble of these combinations based on the softmax weighted averaging (Section 4.1). Second, few-shot learning was employed on the Linear-Probing GEM and ResNet-152 to compare their performances with the Zero-shot GEM's (Section 4.2). Lastly, the fully-supervised Linear-probing GEM and ResNet-152 were also built for a more in-depth comparison among these models (Section 4.3).

4.1 Prompt Engineering in Zero-Shot GEM

As discussed in Section 2.3, involving natural language in multi-modal learning should consider prompt engineering. Therefore, we built several geo-contextualized prompts (e.g., `"city of {label}"`, `"from {label}"`, and `"in {label}"`), and experimented them on the Place Plus 2.0 benchmark for city-level geo-localization. As results, Table 1 shows that combining city labels with prepositions that normally bind with city names (e.g., "at", "from", and "in") in the contextual prompt can fairly boost the prediction performance of

Table 2: Zero-shot GEM outperforms the state-of-the-art ISN on IM2GPS3k benchmark.

Model	Top1 Accuracy
ISN (city-level 25 km) [16]	28.0%
Zero-shot GEM (city-level)	**32.1%**

Zero-shot GEM. However, if the provided prompt contains strongly constrained words (e.g., "`city of {label}`") or prepositions (e.g., "to"), its prediction performance may be even worse than no prompt used.

In addition, we performed prompt ensembling based on weighted averaging, where the weight for each prompt is proportional to the accuracy on the training set using softmax:

$$\sigma(s_i) = \frac{e^{s_i}}{\sum_{j=1}^{K} e^{s_j}} \quad for\ i = 1, 2, \ldots, K \tag{5}$$

where K is the total number of prompts, and s_i is the top1 accuracy score of prompt i and used as the exponent of the base e (i.e., e^{s_i}).

Table 1 shows that prompt ensembling (softmax weighted) yields the best output, achieving an accuracy improvement by 2.70% compared with the case of no prompt. In the following sections, unless there is a specification, all results reported as "Zero-shot GEM" are based on this prompt ensembling as it performs the best.

After determining the prompt design, we then evaluate its prediction performance on city-level geo-localization globally, using the public benchmark dataset IM2GPS3k. Specifically, 461 unique cities are identified worldwide in IM2GPS3k, and Zero-shot GEM achieves 32.1% top1 accuracy and 57.6% top5 accuracy, obtaining a 4.1% top1 accuracy improvement compared with the state-of-the-art ISN [16] (Table 2). Note that top1 accuracy indicates whether the prediction (the one with the highest probability) is exactly the expected answer, and top5 accuracy indicates whether any of the top 5 predictions with the highest probabilities matches the expected answer. Meanwhile, it is also worth noting that the predicted city labels from Zero-shot GEM are generally more useful for a broader range of applications since they carry more semantics rather than just grid cells with numeric indices. In addition, we found that Zero-shot GEM is robust to geo-localizing other types of images such as city paintings of street views. More results are shown in Figure 9 in Appendix. This finding exhibits strong potential in using Zero-shot GEM for different image geo-localization tasks with various sources.

4.2 Zero-shot vs. Few-shot Learning

Next, we compared Zero-shot GEM performance with few-shot supervised models, i.e., Linear-probing GEM and ResNet-152, on the new benchmark dataset Place Plus 2.0. Note that 1-shot learning traditionally means using one image per class for training, assuming that an instance of each class is pictured in a single image. However, for city-level image geo-localization, it is inappropriate to define one image per city as 1-shot, as different cities are in various sizes and one single street view image cannot cover the spatial extent of a city. To rectify this issue, we define 1-shot learning for city-level geo-localization as one image per area unit for each city. In particular, for training on the Place Plus 2.0 dataset, we use 1 image per 74 square kilometers for each city as 1-shot, guaranteeing each city at least has one training sample.

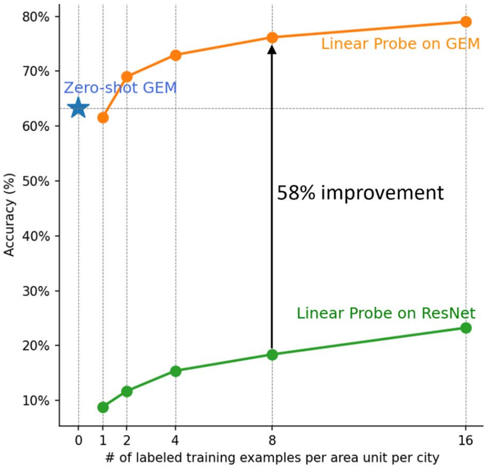

Figure 3: Zero-shot GEM significantly outperforms few-shot Linear-probing ResNet-152, and is marginally better than 1-shot Linear-probing GEM.

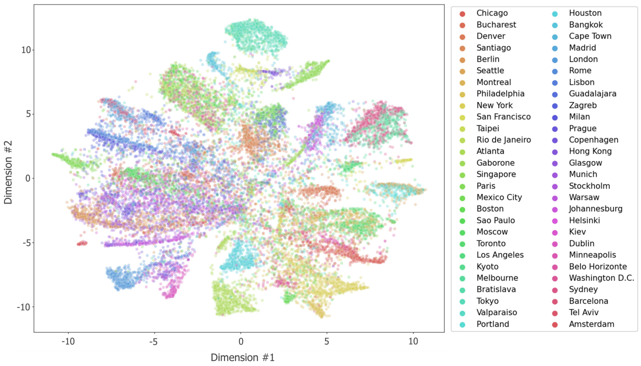

Figure 4: Zero-shot GEM transfer: Use t-SNE to visualize its visual feature space of the cities in the Place Plus 2.0 dataset.

Figure 3 visualizes the difference of prediction performances among Zero-shot GEM and few-shot Linear-probing GEM and ResNet-152 based on the testing set of the Place Plus 2.0 dataset. Surprisingly, Zero-shot GEM wins the other two 1-shot models with 64.42% top1 accuracy (and 93.66% top5 accuracy), achieving 2.82% and 55.56% top1 accuracy better than Linear-probing GEM and ResNet-152, respectively. Although it might be intuitive to expect zero-shot worse than few-shot, yet this unexpected outcome can be explained by the different training approaches between zero-shot and few-shot. First, Zero-shot GEM is self-supervised by natural language, which enables visual features to be directly described or specified based on large amounts of pre-training pairwise data. This process easily produces distinctive representations for each city label (Figure 4). By contrast, traditional supervised learning has to derive the visual features merely from input images, and such context-irrelevant image-based learning has the drawback that many different visual objects in one image can be linked to its class without knowing which object(s) should be primary, especially in the one-shot case, where the representative visual features of each class are not yet established. This finding proves that natural language significantly facilitates the reference of learnt visual features, enabling zero-shot transfer of the model to downstream geo-localization tasks.

Meanwhile, after 16-shot learning, Linear-probing GEM markedly improves its prediction performance with 79.01% top1 accuracy (and 97.41% top5 accuracy), achieving about 58% higher performance than the baseline Linear-probing ResNet-152. This result sturdily proves that language-image models are much more capable of geo-localizing street view images at a city level, compared with the traditional image-based classification methods.

4.3 Fully-Supervised Linear-probing

In addition to few-shot learning, we also build fully-supervised Linear-probing GEM and ResNet-152 based on the training set of the Place Plus 2.0 dataset (80%). As results, most cities have obtained substantial improvements on Zero-shot GEM (averagely 30.33% increase) and Linear-probing GEM (averagely 52.32% increase), compared with the baseline Linear-probing ResNet-152 (Figure 5).

Figure 8 displays the spatial distribution of the studied cities, with graduated colors referring to the top1 accuracy of Linear-probing GEM. It reveals that clustered cities are generally in lighter colors (e.g., European cities and Brazilian cities), implying that they may share similar visual feature spaces that confuse Linear-probing GEM, and therefore cannot receive high accuracies during testing.

Moreover, we discover a linearly correlated performance between Zero-Shot GEM and Linear-probing GEM on the studied cities, and most cities are geo-localized better by Linear-probing GEM (Figure 6). Specifically, although some cities cannot obtain ideal geo-localization outcomes from Zero-Shot GEM, such as Sao Paulo, Barcelona, Rio de Janeiro, Kiev, and others that are labeled in the bottom-right of Figure 6, yet they still can gain excellent results from Linear-probing GEM. This finding suggests that Zero-shot GEM may not be able to capture effective visual features for these cities due to the bias in the pre-training pairwise data (e.g., lack of data), but through a linear probe, these cities can be geo-localized desirably.

To further explore in detail how Linear-probing GEM performs city-level geo-localization based on input images, we visualize its attention map (i.e., weights of the last output layer) by averaging the weights over the 16 attention heads (Figure 7). Specifically, Figure 7(a) reveals that both buildings and vegetation can be main representations, while Figure 7(b) shows that the text in the image are strong signals (e.g., with characters in a specific language) and Figure 7(c) highlights highways and buildings. In the street views where there is no such features, vegetation alone can be a distinctive representation for image geo-localization (Figure 7(d)). These results demonstrate that Linear-probing GEM is able to leverage essential visual features for effective geo-localization.

We also investigate the cities that obtain <70% top1 accuracy in Linear-probing GEM, including Bratislava, Belo Horizonte, Kyoto, Prague, and Copenhagen (Table 3). For instance, the city Bratislava only obtains 43% top1 accuracy, as the model often wrongly geo-localizes its street view images as its neighboring cities (e.g., Prague, Warsaw, and Zagreb). This is probably due to the fact that these European cities have very similar street view scenes and therefore

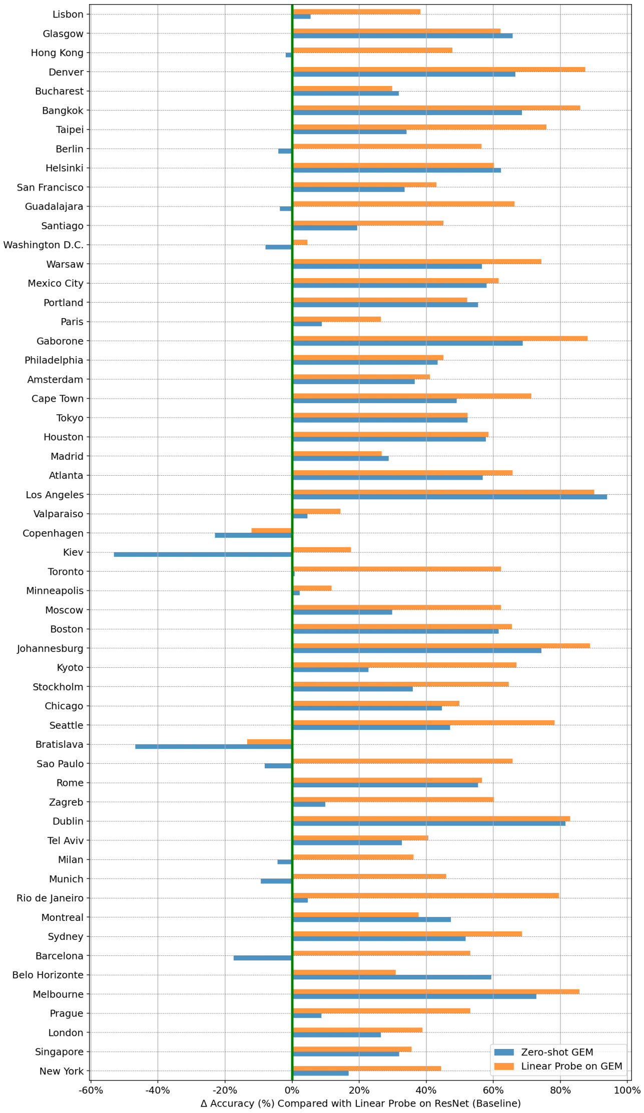

Figure 5: Zero-shot GEM and fully-supervised Linear-probing GEM outperform the fully-supervised baseline.

Table 3: Cities with <70% top1 accuracy in Linear-probing GEM are wrongly labeled as their spatially proximate ones.

City	Acc.	Continent	Top3 wrong labels
Bratislava	43%	Europe	Prague, Warsaw, Zagreb
Belo Horizonte	64%	South America	Rio de Janeiro, Sao Paulo, Guadalajara
Kyoto	67%	Asia	Tokyo, Taipei, Paris
Prague	67%	Europe	Warsaw, Bratislava, Zagreb
Copenhagen	67%	Europe	Stockholm, Helsinki, Amsterdam

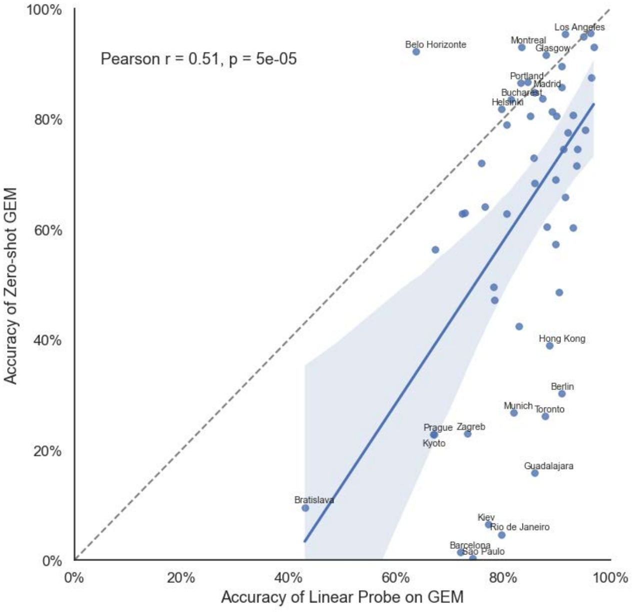

Figure 6: Zero-shot GEM performance is linearly correlated with Linear-probing GEM performance on the cities but mostly less competitive.

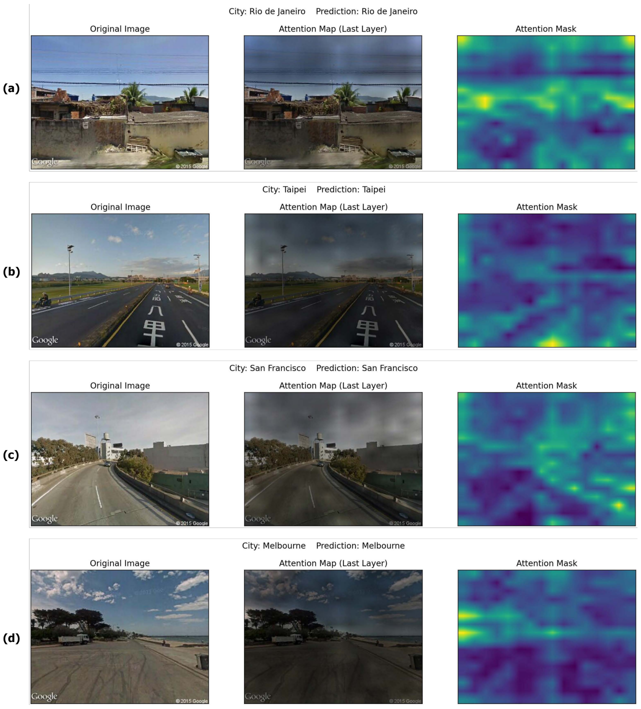
City: Rio de Janeiro Prediction: Rio de Janeiro
Original Image
Attention Map (Last Layer)
Attention Mask
(a)
City: Taipei Prediction: Taipei
Original Image
Attention Map (Last Layer)
Attention Mask
(b)
City: San Francisco Prediction: San Francisco
Original Image
Attention Map (Last Layer)
Attention Mask
(c)
City: Melbourne Prediction: Melbourne
Original Image
Attention Map (Last Layer)
Attention Mask
(d)

Figure 7: Examples of attention maps from Linear-probing GEM.

Table 4: Linear-probing GEM significantly outperforms all models on the Place Plus 2.0 dataset.

Model	Top1 Accuracy
ISN (city-level 25 km) [16]	14.92%
Linear-probing ResNet152	39.61%
Zero-shot GEM	64.42%
16-shot Linear-probing GEM	79.01%
Linear-probing GEM	**85.92%**

Table 5: GEM models outperform the state-of-the-art ISN on the IM2GPS3k benchmark (a subset within the cities in the Place Plus 2.0 dataset).

Model	Top1 Accuracy
ISN (city-level 25 km) [16]	45.05%
Zero-shot GEM ("in {label}")	**65.42%**
Linear-probing GEM	61.87%

Linear-probing GEM cannot successfully distinguish them from their homogeneous visual feature space. Interestingly, Bratislava is predicted best by the baseline Linear-probing ResNet-152 (See Figure 5), indicating that language-image models may not be useful for geo-localizing this city because of data bias. For more details about these low-accuracy cities, the frequencies of their predicted labels have been plotted in Figure 10 in Appendix. Likewise, they are mostly misinterpreted as their neighboring cities with similar street views. This suboptimal performance, however, still provides a positive indication that Linear-probing GEM is capable of identifying spatial relationships of spatially proximate cities.

Lastly, we evaluate GEM models, the baseline, and the state-of-the-art ISN [16] on two benchmark datasets. First, Table 4 shows the performances of all models on the Place Plus 2.0 benchmark (i.e., 22k street view images across 56 cities worldwide). Our Linear-probing GEM significantly outperforms ISN as well as the baseline by 71% and 46%, respectively. In fact, even 16-shot Linear-probing GEM can produce 79% top1 accuracy, meaning that using a small amount of training samples (i.e., street view images) can also achieve satisfying geo-localization results by applying Linear-probing GEM. More sampled results of Linear-probing GEM can be visualized in Figure 11 in Appendix.

Second, Table 5 demonstrates that both Zero-shot and Linear-probing GEM can significantly outperform the state-of-the-art ISN model on the IM2GPS3k benchmark for city-level geo-localization. Specifically, Zero-shot GEM is 20.37% ahead of ISN, while Linear-probing GEM is just 16.82% ahead. That is, Zero-shot GEM actually performs best in this benchmark subset. This outcome could be explained by the differences between this benchmark and the training data used by these models. Specifically, IM2GPS3k was collected from Flickr, whose distribution domain is similar to the pre-training pairwise data used by CLIP (e.g., YFCC100M) to a large extent. Therefore, Zero-shot GEM can straightforwardly learn the visual features of the IM2GPS3k images. By contrast, Linear-probing GEM is trained on Google Street View images from the Place Plus 2.0 dataset, whose domain distribution is largely different from online social media images. This inconsistency between training

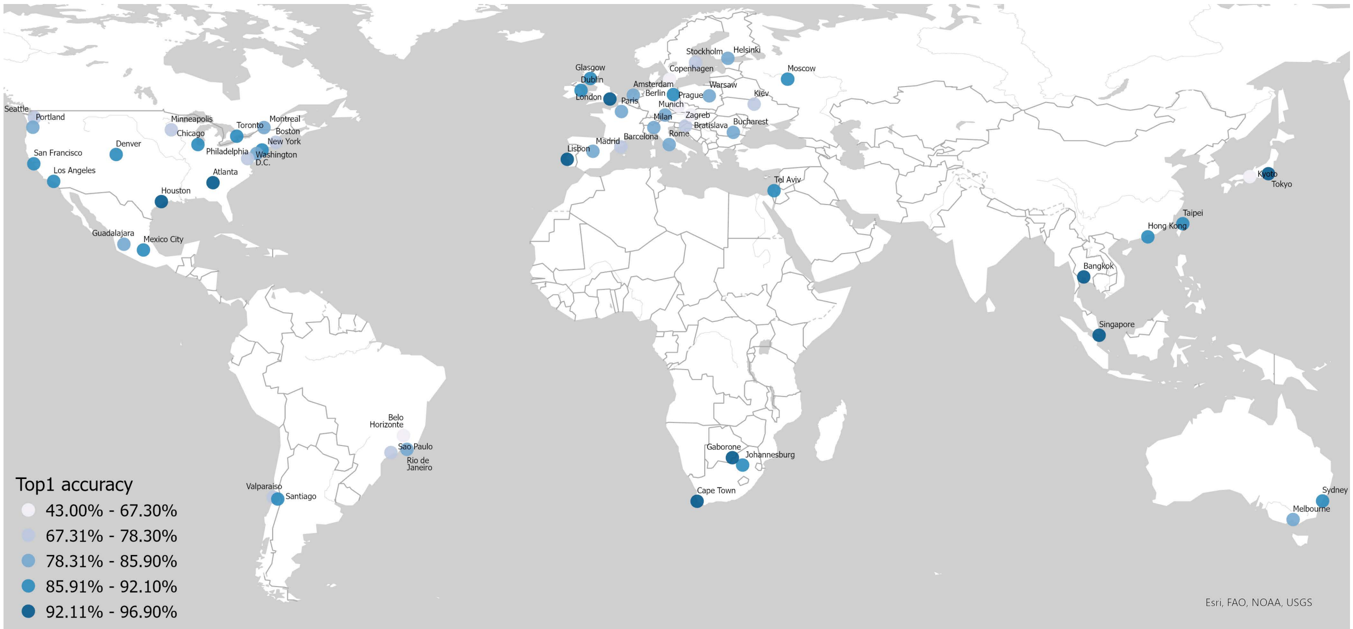

Figure 8: Spatial distribution of the cities in Place Plus 2.0, with graduated colors based on the top1 accuracy of Linear-probing GEM.

and testing domains leads to the less preferable performance of Linear-probing GEM compared with Zero-shot GEM. Nonetheless, the achievement of Linear-probing GEM is not significantly undermined, and it still achieves 16.82% accuracy improvement compared with ISN.

5 CONCLUSION AND DISCUSSIONS

In this work, we applied multi-modal learning to addressing image geo-localization problems, with the integration of textual and visual features. Based on the successful trials on CLIP, we developed Zero-shot GEM by prompt engineering and Linear-probing GEM that was trained on global street view images. The proposed GEM models are capable of connecting the geo-contextualized text to corresponding visual features for image geo-localization, while traditional methods only learn the visual features from input images. This language-image multi-modal learning achieves remarkably boosted performances compared with the baseline and the state-of-the-art model ISN [16]. Moreover, it helps GEMs learn the spatial relationships of the geo-contextualized labels, which is often neglected by traditional classification-based methods. As a bonus, the geo-localization results are enriched with intelligible semantics (e.g., administrative cities instead of arbitrary grids), which is more practical for downstream applications based on image geo-localization (e.g., searching images by geo-location text).

Another questionable common practice for large-scale image geo-localization via deep learning is to train models with massive random online images without considering the task context. Specifically, images that are irrelevant to a given task will be less likely to contribute to the model training (e.g., animal images are probably not useful for geo-localization). As such, we need to ponder on the effectiveness and validity of this practice. Do we need a model to geo-localize each single image in the world? Is it even practicable or reasonable? Is there any way we can improve data efficiency (e.g., using suitable input data such as street view images) instead of focusing on data quantity (e.g., using a large amount of input data without considering their suitability for the task)? We believe that our experiments of city-level geo-localization based on urban street view images rather than random online images at least pave the way to rethink these questions.

On the other hand, some limitations in our work can be addressed with further efforts. For example, although Place Plus 2.0 is one of the largest available street view image datasets that are globally distributed, yet the studied cities are still limited to those most populated ones in the world. Given the promising results of few-shot learning in Section 4.2, a broader scope of street view images can be harvested, in order to extend the applicable area for image geo-localization. Besides, the text encoder only supports natural language descriptions in English, and for the cities where English is not the official language, their geo-localization performance could be undermined.

For future directions, our work can be improved or expanded in the following three aspects: (1) exploring different types of images (e.g., remote sensing images) for geo-localization tasks. In our experiment (Figure 9), Zero-shot GEM has shown powerful and robust transferability in city paintings of street views, and therefore more experimental trials can be performed with other types of images; (2) improving the geo-localization accuracy by training the text encoder and image encoder from scratch based on specific (geo-context, spatial image) pairwise dataset, such as online articles/journals related to geographic places, and websites/mobile apps (e.g., Google Maps and Tripadvisor); and (3) using the spatial relationships of the top prediction labels to upscale image geo-localization tasks (e.g., geo-localizing an image at the region/country level instead of the city level). For example, while performing country-level image geo-localization, the country names can be directly used as input for the text encoder (i.e., using the same methodology in this paper). However, we can alternately use city names as the input and see if the top prediction labels are within the same country or the neighboring ones.

ACKNOWLEDGMENTS

This work was supported by the Vilas Associates Award from UW-Madison.

REFERENCES

[1] Artem Babenko and Victor Lempitsky. 2015. Aggregating deep convolutional features for image retrieval. *arXiv preprint arXiv:1510.07493* (2015).

[2] Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. 2011. Geo-localization of street views with aerial image databases. In *Proceedings of the 19th ACM international conference on Multimedia*. 1125–1128.

[3] Jan Brejcha and Martin Čadík. 2017. State-of-the-art in visual geo-localization. *Pattern Analysis and Applications* 20, 3 (2017), 613–637.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[7] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. In *European conference on computer vision*. Springer, 196–212.

[8] James Hays and Alexei A Efros. 2008. IM2GPS: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*. IEEE, 1–8.

[9] James Hays and Alexei A Efros. 2015. Large-scale image geolocalization. In *Multimodal location estimation of videos and images*. Springer, 41–62.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Mike Izbicki, Evangelos E Papalexakis, and Vassilis J Tsotras. 2019. Exploiting the earth's spherical geometry to geolocate images. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 3–19.

[12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[13] Jan Knopp, Josef Sivic, and Tomas Pajdla. 2010. Avoiding confusing features in place recognition. In *European Conference on Computer Vision*. Springer, 748–761.

[14] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[15] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[16] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 563–579.

[17] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*. 3456–3465.

[18] Apostolos Panagiotopoulos, Giorgos Kordopatis-Zilos, and Symeon Papadopoulos. 2022. Leveraging Selective Prediction for Reliable Image Geolocation. In *International Conference on Multimedia Modeling*. Springer, 369–381.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[21] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. 2018. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 536–551.

[22] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[24] Nam Vo, Nathan Jacobs, and James Hays. 2017. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*. 2621–2630.

[25] Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*. Springer, 37–55.

[26] Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. 2021. Visual and Object Geo-localization: A Comprehensive Survey. *arXiv preprint arXiv:2112.15202* (2021).

[27] Eyasu Zemene, Yonatan Tariku Tesfaye, Haroon Idrees, Andrea Prati, Marcello Pelillo, and Mubarak Shah. 2018. Large-scale image geo-localization using dominant sets. *IEEE transactions on pattern analysis and machine intelligence* 41, 1 (2018), 148–161.

[28] Fan Zhang, Bolei Zhou, Liu Liu, Yu Liu, Helene H Fung, Hui Lin, and Carlo Ratti. 2018. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning* 180 (2018), 148–160.

A APPENDIX

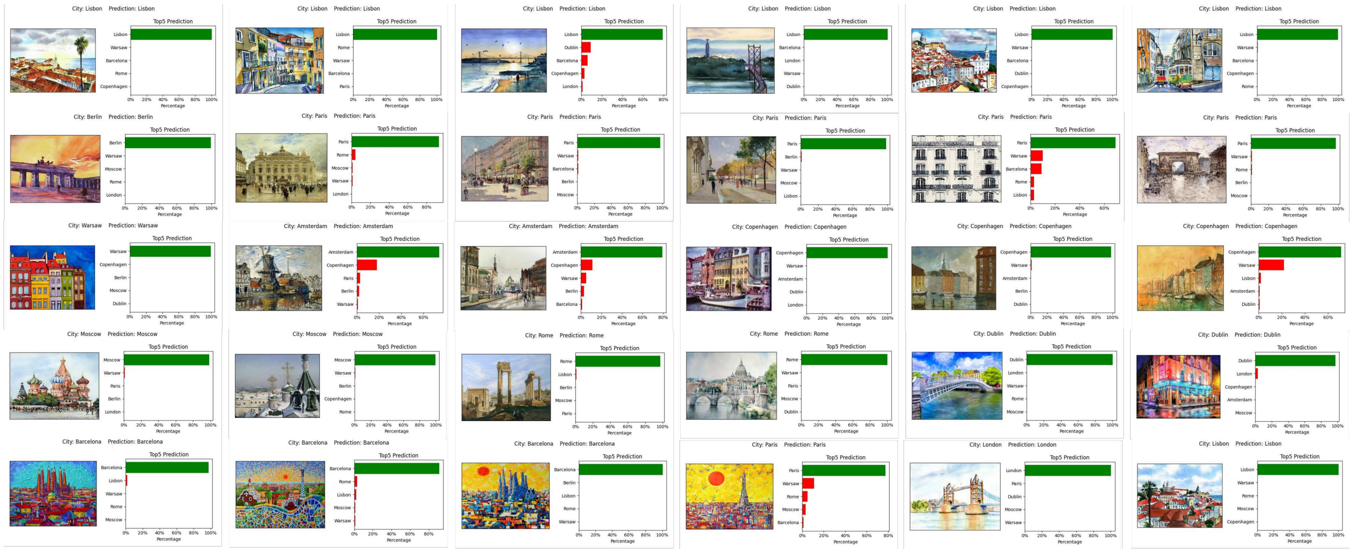
City: Lisbon Prediction: Lisbon
City: Lisbon Prediction: Lisbon
City: Lisbon Prediction: Lisbon
City: Lisbon Prediction: Lisbon
City: Lisbon Prediction: Lisbon
City: Lisbon Prediction: Lisbon
City: Berlin Prediction: Berlin
City: Paris Prediction: Paris
City: Paris Prediction: Paris
City: Paris Prediction: Paris
City: Paris Prediction: Paris
City: Paris Prediction: Paris
City: Warsaw Prediction: Warsaw
City: Amsterdam Prediction: Amsterdam
City: Amsterdam Prediction: Amsterdam
City: Copenhagen Prediction: Copenhagen
City: Copenhagen Prediction: Copenhagen
City: Copenhagen Prediction: Copenhagen
City: Moscow Prediction: Moscow
City: Moscow Prediction: Moscow
City: Rome Prediction: Rome
City: Rome Prediction: Rome
City: Dublin Prediction: Dublin
City: Dublin Prediction: Dublin
City: Barcelona Prediction: Barcelona
City: Barcelona Prediction: Barcelona
City: Barcelona Prediction: Barcelona
City: Paris Prediction: Paris
City: London Prediction: London
City: Lisbon Prediction: Lisbon
Top5 Prediction
Percentage

Figure 9: Zero-shot GEM is also robust to geo-localizing city paintings of street views, a small dataset with 30 images from 11 European cities (source).

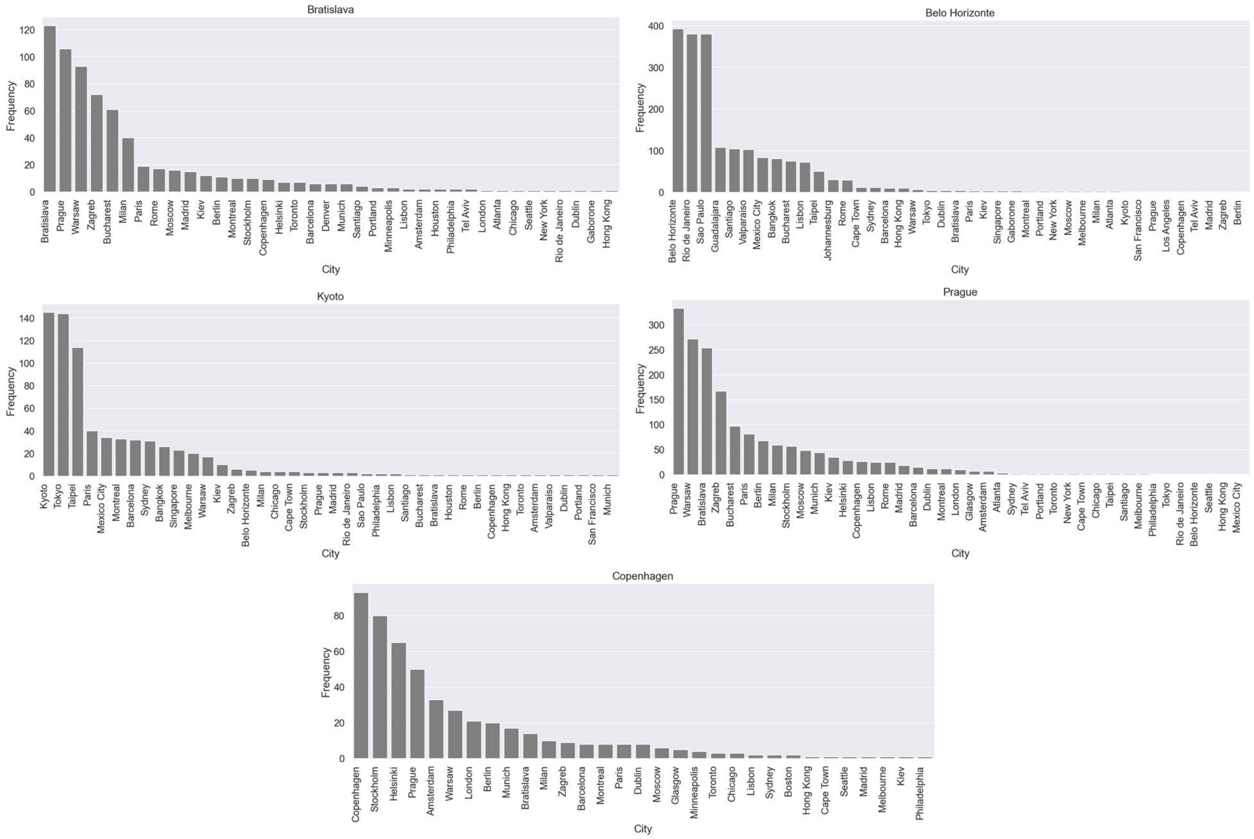

Figure 10: Frequencies of the predicted labels are plotted for the low-accuracy cities in Linear-probing GEM, i.e., Bratislava, Belo Horizonte, Kyoto, Prague, and Copenhagen.

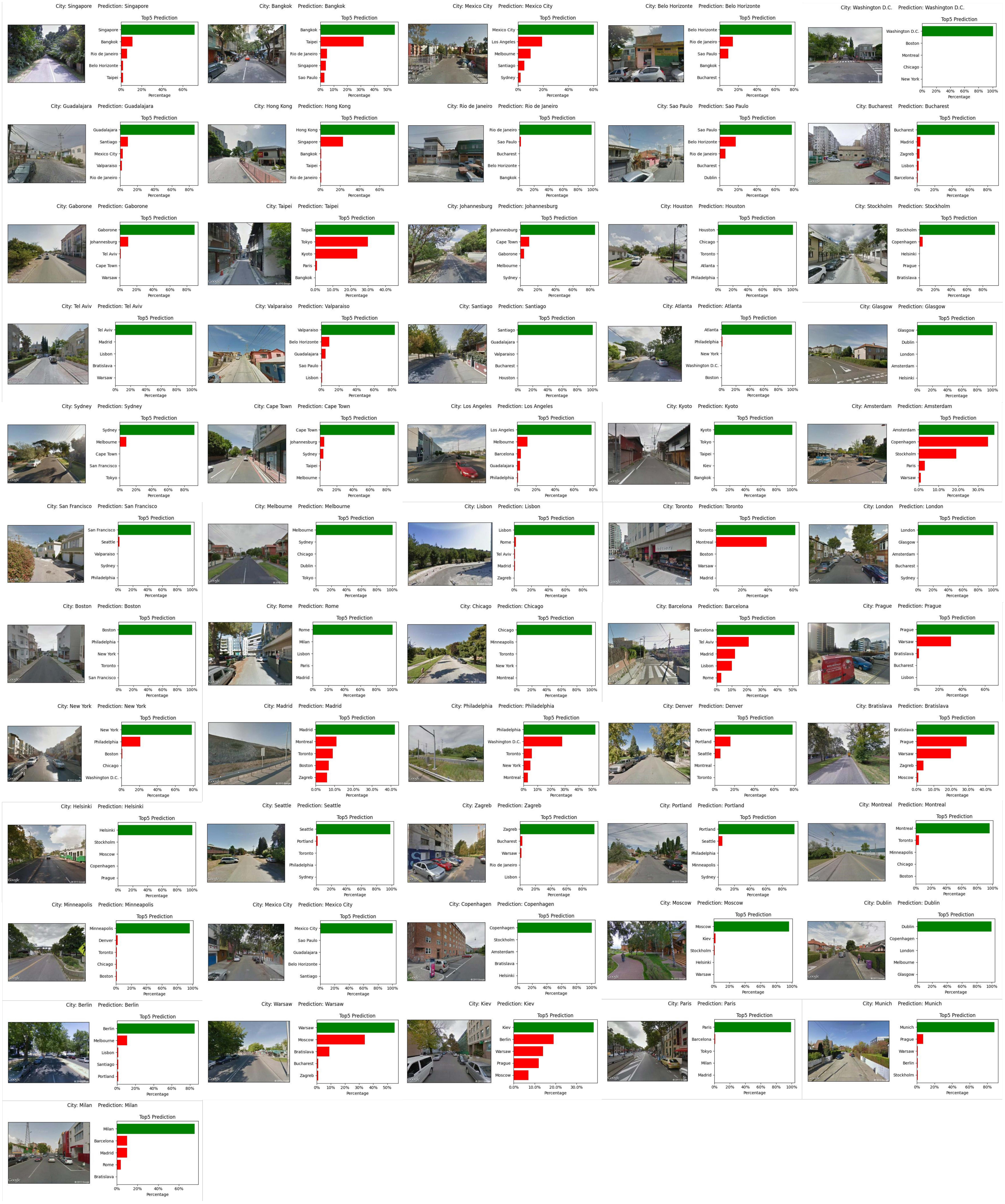

Figure 11: Geo-localization examples of the 56 cities in Place Plus 2.0 by Linear-probing GEM.