GeoSURGE: Geolocalization using Semantic Fusion with Hierarchy of Geographic Embeddings

Angel Daruna, Nicholas Meegan, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar
{angel.daruna,nicholas.meegan,han-pang.chiu,supun.samarasekera,rakesh.kumar}@sri.com
SRI International
Princeton, NJ, USA

Preprint

***Abstract*— Worldwide visual geo-localization seeks to determine the geographic location of an image anywhere on Earth using only its visual content. Learned representations of geography for visual geo-localization remain an active research topic despite much progress. We formulate geo-localization as aligning the visual representation of the query image with a learned geographic representation. Our novel geographic representation explicitly models the world as a hierarchy of geographic embeddings. Additionally, we introduce an approach to efficiently fuse the appearance features of the query image with its semantic segmentation map, forming a robust visual representation. Our main experiments demonstrate improved all-time bests in 22 out of 25 metrics measured across five benchmark datasets compared to prior state-of-the-art (SOTA) methods and recent Large Vision-Language Models (LVLMs). Additional ablation studies support the claim that these gains are primarily driven by the combination of geographic and visual representations.**

I. Introduction

Visual geo-localization supports critical applications across autonomous systems, emergency response, and personalized digital experiences by accurately estimating geographic locations from images. Many variations of the visual geo-localization problem exist including city scale [1], cross-view [2], and global [3]. Among these variations, global geo-localization remains a challenging problem due to the complexity and diversity of images from the world. In this work, we address the global visual geo-localization problem where only one image of the scene is provided.

Most prior global visual geo-localization approaches can be categorized as retrieval based [4], [5] and classification based [6]. The former compares the query image with references from a world-scale database to estimate the geo-location of the input image. Classification-based methods discretize the world into a grid of cells (geocells), and then train a classifier to assign the query image to the geocell it is in. Both approaches have limitations. Retrieval-based methods require more computation during inference given a large reference database while classification-based methods tradeoff between the granularity and coverage of geocells. Recent SOTA includes GeCLIP [7], Img2Loc [8], and G3 [9]. GeoCLIP used GPS coordinates as the reference database instead of geotagged images and introduced several specialized components, such as Random Fourier Features, to mitigate information loss when summarizing scene information into two dimensions. Img2Loc leveraged LVLMs with geotagged image references, providing substantial performance gains. G3 followed combining ideas from GeoCLIP and Img2Loc. Despite much progress, learned representations of geography are challenging to design due to their low-dimensional nature.

In this paper, we propose a new approach, GeoSURGE (Geo-localization using Semantic Fusion with Hierarchy of Geographic Embeddings), which aims to combine the strengths of retrieval and classification approaches. GeoSURGE explicitly models the world as a hierarchy of geographic embeddings, which is a hierarchical and distributed representation. Similar to classification-based methods, the Earth's surface is partitioned into geocells. However, unlike prior works, GeoSURGE explicitly represents each geocell as a learned feature vector. By learning and matching geocell representations to visual representations of query scenes, geo-localization is realized as correctly matching visual and geographic features.

Another innovation from GeoSURGE is a semantic fusion process that combines both appearance and semantic segmentation features to efficiently enrich the visual representation. Specifically, we fuse CLIP features from an RGB image [10], [7] with its semantic segmentation map using latent cross-attention. While appearance-based features identify fine-grained visual cues from images, semantic segmentation yields more invariant features to large changes in imaged conditions. Semantics also help determine which regions of an image may be unreliable for localization, such as humans and vehicles.

We rigorously evaluated GeoSURGE through benchmark experiments following protocols from prior visual geo-localization studies and ablation studies. Our benchmark experiments included five widely used visual geo-localization datasets: IM2GPS [4], IM2GPS3k [11], YFCC4k [11], YFCC26k [12], and GWS15k [13]. In addition to SOTA planet-scale geo-localization methods, we compared GeoSURGE with LVLMs baselines due to their recent emergence to visual geo-localization. Empirically, our method achieved new bests on 22 out of 25 results measured across the five benchmark datasets. We followed benchmark experiments with several ablation studies to quantify the impact of each design choice on overall performance. The ablation studies revealed that these gains are primarily driven by our geographic and visual representations. These experiments

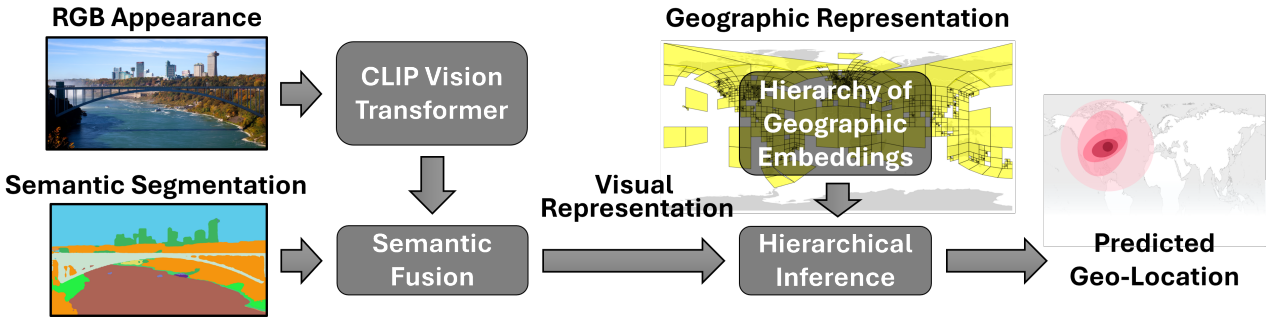

Fig. 1. GeoSURGE Approach Overview: The location of an input image is predicted via hierarchical inference, by matching the visual representation of the image against the geographic representation, which is learned beforehand. The visual representation is generated from the semantic fusion module, which enriches appearance features with semantic segmentation.

confirm that models trained specifically for visual geo-localization, like ours, continue to hold substantial advantages for advancing geo-localization accuracy.

In summary, GeoSURGE has the following key contributions to planet-scale image-based geo-localization:

1) It learns a novel representation that explicitly models the world as a hierarchy of geographic embeddings.
2) It uses a new semantic fusion module with latent cross-attention to integrate a semantic segmentation map and appearance features within its visual representation.
3) It achieves new bests on 22 out of 25 metrics measured against SOTA image geo-localization methods and LVLM-based methods across five widely-used public benchmark datasets.

II. Related Works

Planet-scale image geo-localization approaches can be mainly categorized into two types: retrieval-based and classification-based methods, with some works combining the two approaches for improved performance.

Classification-Based Methods divide the world into geocells and determine which geocell contains a query image to infer its location [6]. The seminal work of Planet partitioned Earth into geocells with respect to the distribution of training images using Google's S2 library and trained a convolutional neural network to assign a query image to a geocell [6]. Many later works explored other partitioning methods including hierarchies [14], [13], [11], [12], [15], [3], semantic knowledge [16], [10], or a combination of approaches (e.g., hierarchy and semantic knowledge [10]).

In addition to partitioning, many classification-based methods explored how to improve performance by incorporating scene knowledge or semantics. Individual Scene Networks (ISNs) were used to diversify learned features across indoor, urban, and natural scenes [12]. GeoDecoder further diversified features across both geographic levels and 16 visual scene categories [13]. IM2City [17] introduced CLIP and visual-language grounding, by training with images paired with their city names. In G^3, image embeddings and clue embeddings from human-written GeoGuessr guidebooks were combined to predict the country of the query image [18]. TransLocator [19] trains a dual-branch transformer with query image and corresponding semantic segmentation maps.

GeoSURGE also develops a semantic fusion module but uses recent advances in CLIP-based features and latent multi-headed attention to promote memory efficiency [20]. While we use many concepts from classification-based methods such as geocells and hierarchy, we treat visual geo-localization like a retrieval problem wherein we match visual and geographic representations.

Retrieval-based methods typically compare a query image to a large database of geotagged images to infer the location of the query image [4], [5]. The seminal work of IM2GPS first investigated planet-scale geo-localization via retrieval by computing similarities between hand-crafted features of the query image and a geotagged image database over 6 million images [4]. Recently Img2Loc [8] leverages LVLMs, such as GPT-4V, extending the retrieval-based approach with retrieval augmented generation (RAG). After retrieving the most and least similar reference images to the query image, Img2Loc prompts an LVLM to geotag the query image. In this way, Img2Loc benefits from the Internet-scale multimodal corpus used to train LVLMs to infer image coordinates from visual cues.

GeoCLIP [7] reformulated the retrieval-based approach as comparing the query image to GPS coordinates using CLIP-style contrastive learning, effectively opting for a reference database of GPS coordinates instead of geotagged images. Their design choice allows the network to accumulate visual information from multiple scenes associated with the same coordinates, in essence learning a representation of the location. Similar in spirit to Img2Loc, G3 [9] extends GeoCLIP's retrieval-based approach with RAG using LVMLs, such as GPT-4V.

In GeoSURGE, we also used CLIP-style contrastive learning but propose to learn representations of geocells instead of GPS coordinates. This crucial design choice allows for more effective learning of geographies due to the low-dimensional nature of GPS, necessitating specialized learning components like Random Fourier Features in GeoCLIP. Our insight is substantiated by the benchmark performance gains over GeoCLIP while using the same resources like network backbones and data.

Hybrid Methods aim to merge retrieval and classification to improve performance. [L]kNN use features derived from networks trained with a classification learning objective in their retrieval-based inference system [11]. In [21] and

[3], a retrieval-within-geocell scheme is used to refine the predicted location of the query image inside the S2 geocell, using features trained via classification. PIGEOTTO [10] refines the predicted location of the query image through a hierarchical retrieval mechanism that examines both top-K geocells and clusters within these geocells. GeoSURGE can also be considered a hybrid as it is trained to match scene features with geographic features (i.e., retrieval) that represent geocells (i.e., classification).

III. Approach

GeoSURGE is trained to geolocate images through contrastive learning by aligning image features with geographic features that represent regions of Earth as in Figure 1. The visual representation of an image is derived from both the RGB image and its semantic segmentation map, which are assumed inputs. For training, we also assume access to a dataset of geotagged images. The geographic representation is a learned hierarchy of embeddings that correspond to regions of Earth. Geo-localization is implemented by outputting the location of the geographic representation that contains a query image, which is unknown. Therefore, GeoSURGE learns a function to maximize the similarity between query image features and the geographic features corresponding to the region of Earth that contains that image. Below we detail the design of these representations, our training procedure, and how we use these two representations for inference in a hierarchical manner.

A. Geographic Representation

GeoSURGE explicitly models geography as a hierarchical and distributed representation. Earth's surface is partitioned into geographic cells (geocells), similar to what typical classification-based methods do. However, unlike prior works, GeoSURGE explicitly learns a unique feature vector to represent each geocell using all training data samples within this geocell. Together, these vectors from all geocells can be learned to form a distributed representation (an embedding space) of the geographic regions covering Earth, a partition. Inspired by the hierarchical representations of prior geo-localization works [11], [12], GeoSURGE repeats this partitioning process by varying the granularity of the geocells. In this way, the geocells covering Earth form a partition hierarchy. Mathematically, this partition hierarchy is represented as a hierarchy of embeddings, each with a unique embedding space that is learned.

Specifically, GeoSURGE uses Google's S2 Library to divide Earth's surface into Schneider 2 (S2) geocells [12]. The process begins by projecting the Earth onto the six faces of a cube, resulting in six initial S2 geocells containing all training samples from a dataset. To balance the number of images within each geocell, any geocell containing more than τ_{max} samples is recursively subdivided. Geocells with fewer than τ_{min} samples are excluded to ensure that each geocell has a sufficient number of samples. This recursive splitting continues until all geocells contain a number of samples between τ_{min} and τ_{max}. This way ultimately produces a partition ρ of Earth's surface where each geocell contains a balanced number of training samples, $\rho(\tau_{\text{min}}, \tau_{\text{max}})$. GeoSURGE then explicitly models each geocell as a unique vector that represents all training samples within. Together, these vectors form the embedding space, $\mathcal{E}$, representing the partition of Earth's surface $\rho(\tau_{\text{min}}, \tau_{\text{max}})$ that is learned during training. Note, GeoSURGE is agnostic to different partitioning schemes. GeoSURGE currently uses this S2 partitioning [12] because our contributions are orthogonal to the partitioning method.

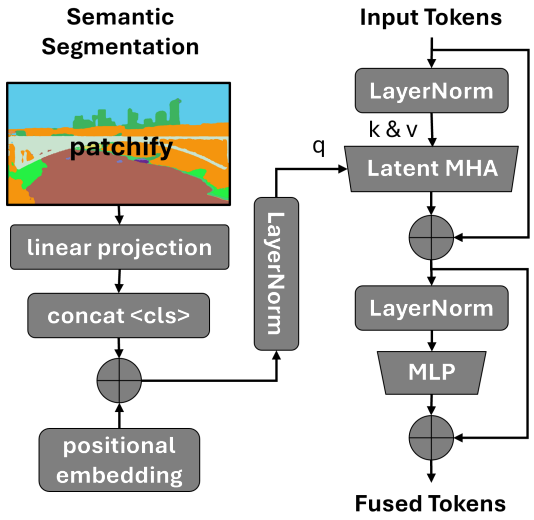

Fig. 2. Diagram of GeoSURGE's semantic fusion blocks.

GeoSURGE then forms a partition hierarchy of geocells covering Earth, by repeating the above partitioning process with varied granularity of the geocells, representing each partition as an embedding. GeoSURGE treats the number of partitions, the dimensionality of the embedding space representing a partition, and $(\tau_{\text{min}}, \tau_{\text{max}})$ for each partition as hyperparameters. During repeated partitioning, the value of τ_{max} is varied while the same value of τ_{min} is applied across all hierarchy partitions. This approach ensures that each geocell in the finest partition can be uniquely linked to its coarser parent partitions. Therefore, at inference time a hierarchical prediction is computed as the product of similarities between features of a query image and all containing geocells of the partition hierarchy (i.e., embeddings of corresponding geocells). However, during training, the embedding representing each partition is learned separately from other partitions to promote diversity among geographical features at different scales.

B. Visual Representation

GeoSURGE extracts visual features from both appearance (RGB image) and semantics (segmentation map) to match with learned features from the geographic representation. GeoSURGE uses CLIP [22] to extract features from the RGB image as in [7], [10], [9]. GeoSURGE then uses OneFormer [23] to get a semantic segmentation map corresponding to this RGB image. This semantic segmentation map is fused with the CLIP-extracted features through latent cross-attention within a semantic fusion module, as shown in Figure 2. The fused visual representation output from this semantic fusion

module is then matched to the learned geographic features for predicting the image's location.

The semantic fusion module combines features extracted from the RGB image and its semantic segmentation map into a single visual representation. GeoSURGE first uses the CLIP vision transformer (i.e., without projection and output normalization) to extract RGB appearance features as a sequence of tokens representing all patches of the RGB image plus a CLS token. All but the last CLIP Vision Transformer encoder blocks are kept frozen during training. The semantic fusion module enhances these RGB tokens, by fusing them with a semantic segmentation map produced from OneFormer [23] on the RGB image. The semantic segmentation map has the same width and height as the RGB image and semantic classes are defined based on ADE20K [24].

Shown in the left of Figure 2, the semantic fusion module extracts tokens from the semantic segmentation map, by linearly projecting all patches from the segmentation map, concatenating a CLS token, and adding positional embeddings. These semantic tokens are fused with the RGB tokens using latent multi-headed attention to promote memory efficiency [20]. RGB image tokens are used as the keys and values, while semantic segmentation map tokens are queries. The output of this attention module is then passed through an MLP, producing fused tokens for every patch of the image. Additionally, the semantic fusion module includes residual connections [25] and layer normalizations common in transformer architectures, as shown in Figure 2. Although Figure 2 shows only a fusion block, the semantic fusion module serially repeats these to promote hierarchies in fused features. As final steps, we extract the CLS of the fused tokens, then perform layer normalization and linear projection of the fused CLS token to produce the final visual representation.

C. Training

GeoSURGE seeks to correctly match the visual features from the query image to the learned geographic features at the correspondent location of this image. The geographic features are embeddings representing a partition hierarchy of Earth's surface. The visual features are extracted from the input RGB image and its semantic segmentation map. Our training process uses contrastive learning techniques [22] to maximize the similarity of correct pairs of visual and geographic features from the training dataset, while minimizing the similarity of incorrect pairs.

Specifically, GeoSURGE is trained to select the correct geographic location for a query image from a batch of random locations, by computing the similarity between the features of each location and the query image. Our training batches are constructed from sampling a dataset $\mathcal{D}$ of geotagged images (x, y). Given a training sample, we extract the fused CLS token as a feature vector $\mathbf{v}$, summarizing the the RGB and the semantic information of the sample x. We then use the sample's true location y to extract a second feature vector $\mathbf{g}$ from our geographic representation. We construct $\mathbf{g}$, by indexing the partition $\rho(\tau_{\text{min}}, \tau_{\text{max}})$ for the geocell containing y. The corresponding normalized vector of the geocell within the embedding $\mathcal{E}$ representing $\rho(\tau_{\text{min}}, \tau_{\text{max}})$ is $\mathbf{g}$.

Given a batch of correct pairs of visual and geographic features $(\mathbf{v}, \mathbf{g})$ of size B, we learn a function that maximizes the cosine similarity of the B correct pairs and minimizes the cosine similarity of the $B^2 - B$ incorrect pairs. Our learning objective is the InfoNCE loss function [26] shown in Equation 1 for a sample i in a batch. Following the above procedure, we extract feature vectors for each partition within our hierarchical geographic representation. The complete learning objective is the sum of losses for the partition hierarchy.

$$\mathcal{L}_i = -\log \frac{\exp(\mathbf{v}_i^\intercal \mathbf{g}_i / \tau)}{\exp(\mathbf{v}_i^\intercal \mathbf{g}_i / \tau) + \sum_{j \neq i} \exp(\mathbf{v}_i^\intercal \mathbf{g}_j / \tau)} \tag{1}$$

D. Hierarchical Inference

During inference (Figure 1), GeoSURGE matches the query image's visual features with all features of the learned geographic representation. GeoSURGE uses a hierarchical inference process to produce a single prediction that integrates the complete partition hierarchy of Earth's surface. In other words, the location of the query image is predicted via this hierarchical inference process as the location in the partition hierarchy of Earth's surface with the highest similarity to the query image's visual features.

Specifically, given a query image x with unknown location, GeoSURGE extracts the fused CLS token of its visual representation as a feature vector $\mathbf{v}$. The extraction process is described in previous sub-sections. GeoSURGE then computes the cosine similarity between $\mathbf{v}$ and the vectors of each embedding $\mathcal{E}$, which represent the partitions $\rho(\tau_{\text{min}}, \tau_{\text{max}})$ dividing Earth's surface in a hierarchical manner. For each geocell r in the finest partition, we compute the product of its similarity by integrating the similarities for each parent geocell r' that contains r. This way integrates similarities across all hierarchy levels, realizing hierarchical inference.

IV. Experiments

We perform experiments on publicly available benchmark datasets to have a fair comparison with existing works. We provide additional ablation studies to gauge the contribution of our design choices to overall performance. We observed that GeoSURGE provides new all-time bests in 22 / 25 metrics measured across five benchmark datasets. Ablation studies indicate that these improvements are mostly due to our geographic and visual representations.

Benchmark Evaluation Protocol follows the precedent established in previous works to maintain fair comparisons. We train GeoSURGE on the MediaEval Placing Tasks 2016 (MP-16) dataset containing more than 4 million images, holding 1% for validation and the remainder for training. We tested GeoSURGE on five benchmark datasets: IM2GPS [4], IM2GPS3k [11], YFCC4k [11] and YFCC26k [12], and GWS15k [13]. Note, while the GWS15k image dataset has not been publicly released, we use the same set of panorama IDs from the authors that uniquely identify the Google Street

View image panoramas from where the original GWS15k was constructed. In this way, our results on GWS15k can be directly compared with [13], [7] and avoid differences in results reported from [10] that may be introduced due to random sampling. Same as [13], we average the prediction of the Ten Crop method to provide a single prediction for the entire image. As in prior work, we report results using a threshold metric. We computed the great circle distance (GCD) from each predicted location to the ground truth location using the Haversine distance. After computing the GCDs for all predicted coordinates with respect to the ground truth, we compute the percentage of predictions within five error thresholds: 1km, 5km, 25km, 750km, 2500km. These thresholds correspond to 5 levels of localization: street, city, region, country, continent.

Implementation details are summarized as follows along with full details in supplementary materials to support recreation of results. We use Clip-ViT-Large-Patch14-336 as our visual backbone. The semantic fusion module extracts 128-dimensional tokens from the semantic segmentation maps by linearly projecting each 14 by 14 patch and uses 3 repeated fusion blocks with query dimensionalities of 128. Remaining dimensionalities of the semantic fusion module match the hidden dimensionality of visual backbone, 1024, to perform latent cross-attention. The latent dimension of 64 was used for cross-attention. We used 7 partitioning levels to divide Earth where τ_{min} is 50 and τ_{max} is 25000, 10000, 5000, 2000, 1000 750, 500 from coarsest to finest partitioning. We initialized the learnable temperature parameters in Equation 1 to 0.07 for each partition level with geographic embeddings having a dimensionality of 768 to match the visual backbone's projection dimensionality. We trained using the AdamW optimizer with initial learning rate 0.0001, weight decay 0.0001, and effective batch size of 1024. We used a step learning rate decay schedule with a gamma of 0.5 at each epoch. We used early stopping to finish training when the performance on the validation dataset did not improve for 4 epochs. GeoSURGE was trained in 21 hours on 8 NVIDIA A6000 GPUs.

A. Experimental Results

Tables I through V summarize our performance on five benchmark datasets. We bold the best result and underline the second best result for each distance threshold in a dataset. Compared prior global visual geo-localization works include [L]kNN [11], PlaNet [6], CPlaNet [14], ISNs [12], Translocator [19], GeoDecoder [13], GeoCLIP [7], and PIGEOTTO [10], Img2Loc [8], G3 [9], and RFM$_{10M}\mathcal{S}_2$ [27]. We place "-" in the table, if a prior method does not report its results on this dataset. If a method was published before a benchmark dataset was released, we do not include it in the table correspondent to that benchmark dataset.

Quantitative results show GeoSURGE achieves new all-time bests in 22 / 25 metrics across the five distance thresholds and five benchmark datasets. Excluding methods that use LVLMs, GeoSURGE performs better in all 25 metrics measured across five benchmark datasets. The consistent performance gains over SOTA GeoCLIP method highlight the importance of GeoSURGE's geographic representation design as GeoCLIP uses the same visual backbone with a GPS-based geographic representation. We further probe the performance attributable to GeoSURGE's geographic representation in ablation studies.

When considering recent LVLM-based methods, GeoSURGE outperforms both Img2Loc and G3 in 7 / 10 metrics across five distance thresholds and two benchmark datasets reported in [8], [9]. Specifically, Img2Loc or G3 in some cases perform better than GeoSURGE with smaller distance thresholds (street and city-level) for evaluation. Due to recent interests in utilizing LVLMs to geo-localization, we also evaluated three LVLM models which are newer than GPT-4V to better understand the LVLM contribution without RAG. These LVLM models included Qwen2-VL-72B [28], InternVL2-LLAMA3-76B [29], and NVLM-D-72B [30], which have impressive performance on various vision-language benchmarks. We used the chain-of-thought reasoning prompt to these LVLMs to geotag query images. For each evaluated LVLM, we generate a maximum of 500 tokens for the LVLM to reason about the location of the query image. We found performance from these three LVLMs using chain-of-thought prompting is worse than [8], [9], highlighting the importance of computationally intensive RAG to achieve superior geo-localization performance as in Img2Loc and G3.

Consequently, our findings indicate that methods trained specifically for visual geo-localization still hold significant value to global visual geo-localization, although efficient integration with LVLMs may be a promising future direction.

Qualitative results of sample success cases are shown in Figure 3. We show the predicted GPS, ground truth GPS, as well as the closest reference image to the predicted location within the ground truth geocell. We show that GeoSURGE is robust in location prediction, by showcasing an example with the Eiffel Tower (the bottom-left example in Figure 3): the original in Paris and replica in Las Vegas. Distinctive features derived from the semantic information assists the model in differentiating these two locations, such as the water and large building located close to the replica that are absent at the original location. Dynamic entities, such as the person in the top-left reference image and two people in the bottom-right reference image in Figure 3, are also accurately identified in semantic segmentation maps. This way implicitly avoids the utilization of unreliable imaged regions to geo-localization, via our semantic fusion module.

We also report some failure cases from GeoSURGE in Figure 4. Here, we show predicted reference-image pairs, denoted by color. For the pair denoted in red, while within country-level distance between the predicted location and ground truth, visual ambiguity with the RGB image and semantic image may lead to confusion in the prediction. For the pair denoted in blue, the large disparity in predicted latitude and longitude versus ground truth can be attributed to an insufficient number of training images within the ground truth cell. The closest reference image to the ground truth is approximately 806 kilometers away.

TABLE I

IM2GPS GCD ACCURACY; HIGHER IS BETTER.

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
[L]kNN	14.4	33.3	47.7	61.6	73.4
PlaNet	8.4	24.5	37.6	53.6	71.3
CPlaNet	16.5	37.1	46.4	62.0	78.5
ISNs	16.9	43.0	51.9	66.7	80.2
Translocator	19.9	48.1	64.6	75.6	86.7
GeoDecoder	22.1	50.2	69.0	80.0	89.1
GeoCLIP	-	-	-	-	-
PIGEOTTO	11.8	38.8	63.7	80.5	91.1
$\text{RFM}_{10\text{M}}\mathcal{S}_2$	-	-	-	-	-
Img2Loc/GPT-4V	-	-	-	-	-
G3/GPT-4V	-	-	-	-	-
Qwen2-72B	11.8	41.2	60.8	76.4	86.5
InternVL2-76B	12.7	37.1	51.9	66.7	81.0
NVLM-D-72B	8.9	29.1	42.6	59.1	69.6
GeoSURGE (Ours)	**27.0**	**54.4**	**70.0**	**84.4**	**93.2**

TABLE II

IM2GPS3K GCD ACCURACY; HIGHER IS BETTER.

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
[L]kNN	7.2	19.4	26.9	38.9	55.9
PlaNet	8.5	24.8	34.3	48.4	64.6
CPlaNet	10.2	26.5	34.6	48.6	64.6
ISNs	10.5	28.0	36.6	49.7	66.0
Translocator	11.8	31.1	46.7	58.9	80.1
GeoDecoder	12.8	33.5	45.9	61.0	76.1
GeoCLIP	14.1	34.5	50.7	69.8	83.8
PIGEOTTO	10.9	35.8	52.4	70.7	84.4
$\text{RFM}_{10\text{M}}\mathcal{S}_2$	-	-	-	-	-
Img2Loc/GPT-4V	17.1	**45.1**	57.9	72.9	84.7
G3/GPT-4V	16.6	40.9	55.6	71.2	84.7
Qwen2-72B	8.5	34.4	49.4	62.0	72.1
InternVL2-76B	10.0	30.4	43.5	59.3	72.4
NVLM-D-72B	7.2	23.3	32.3	44.5	55.3
GeoSURGE (Ours)	**17.2**	42.5	**58.1**	**74.6**	**87.6**

TABLE III

YFCC4K GCD ACCURACY; HIGHER IS BETTER.

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
[L]kNN	2.3	5.7	11.0	23.5	42.0
PlaNet	5.6	14.3	22.2	36.4	55.8
CPlaNet	7.9	14.8	21.9	36.4	55.5
ISNs	6.7	16.5	24.2	37.5	54.9
Translocator	8.4	18.6	27.0	41.1	60.4
GeoDecoder	10.4	24.4	33.9	50.0	68.7
GeoCLIP	-	-	-	-	-
PIGEOTTO	9.5	22.5	38.8	60.7	76.9
$\text{RFM}_{10\text{M}}\mathcal{S}_2$	-	33.5	45.3	61.1	77.7
Img2Loc/GPT-4V	14.1	29.6	41.4	59.3	76.9
G3/GPT-4V	**24.0**	**35.9**	47.0	64.3	78.1
Qwen2-72B	4.0	16.2	26.9	41.9	52.9
InternVL2-76B	5.1	14.4	22.6	35.2	46.5
NVLM-D-72B	2.4	7.4	9.7	12.8	15.1
GeoSURGE (Ours)	19.9	33.6	**48.7**	**67.4**	**82.0**

TABLE IV

YFCC26K GCD ACCURACY; HIGHER IS BETTER.

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
PlaNet	4.4	11.0	16.9	28.5	47.7
ISNs	5.3	12.3	19.0	31.9	50.7
Translocator	7.2	17.8	28.0	41.3	60.6
GeoDecoder	10.1	23.9	34.1	49.6	69.0
GeoCLIP	11.6	22.2	36.7	57.5	76.0
PIGEOTTO	10.1	24.6	41.3	62.6	78.7
$\text{RFM}_{10\text{M}}\mathcal{S}_2$	-	-	-	-	-
Img2Loc/GPT-4V	-	-	-	-	-
G3/GPT-4V	-	-	-	-	-
Qwen2-72B	4.3	16.8	27.5	39.8	50.3
InternVL2-76B	6.2	16.3	25.4	38.9	51.7
NVLM-D-72B	3.6	11.7	18.1	27.1	34.2
GeoSURGE (Ours)	**17.8**	**31.5**	**45.1**	**64.3**	**79.3**

TABLE V

GWS15K GCD ACCURACY; HIGHER IS BETTER.

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
GeoDecoder	0.7	1.5	8.7	26.9	50.5
GeoCLIP	0.6	3.1	16.9	45.7	74.1
PIGEOTTO	-	-	-	-	-
$\text{RFM}_{10\text{M}}\mathcal{S}_2$	-	-	-	-	-
Img2Loc/GPT-4V	-	-	-	-	-
G3/GPT-4V	-	-	-	-	-
Qwen2-72B	0.4	4.1	20.5	47.2	69.1
InternVL2-76B	0.6	2.3	10.1	28.2	49.9
NVLM-D-72B	0.1	1.3	7.1	22.5	39.0
GeoSURGE (Ours)	**1.0**	**4.6**	**21.9**	**54.7**	**80.8**

B. Ablation Studies

We conducted ablation studies using the IM2GPS dataset with different settings to our design choices in GeoSURGE. Tables VI through VIII show the impact of these design choices to the overall performance.

Geographic representation ablations are shown in Table VI, where we toggle whether GeoSURGE is trained with a retrieval learning objective that uses our geographic embeddings or a classification learning objective that does not. All other variables are set to defaults and held constant across the GeoSURGE ablations. We see the performance gains from our geographic representation particularly for Street distances, highlighting the contribution of the geographic representation.

Hierarchy Depth Ablations are shown in Table VII, where we vary the depth of the geographic hierarchy GeoSURGE uses during training and inference. All other variables are set to defaults and held constant across the GeoSURGE ablations. We see that deeper hierarchies improve performance, as observed in prior works. This stems from having smaller geocells with more precise locations as well as more diversity among geographical features at different scales.

Semantic fusion ablations are shown in Table VIII where we provide the number of increased percentage points observed by activating the semantic fusion module with varying number of fusion blocks compared to disabling semantic fusion altogether. All other variables are set to

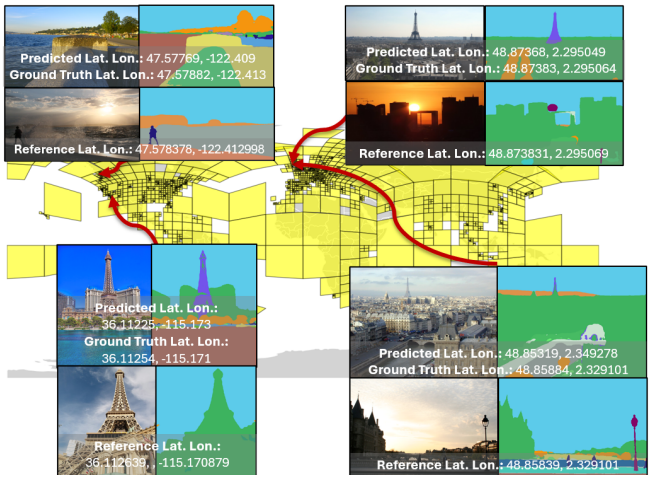

Fig. 3. Sample successful GeoSURGE predictions. Best viewed when zoomed.

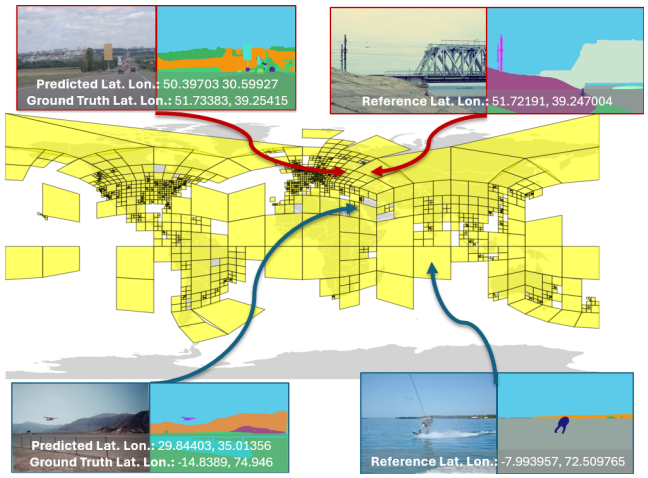

Fig. 4. Sample unsuccessful GeoSURGE predictions. Best viewed when zoomed.

TABLE VI
GEOGRAPHIC REPRESENTATION ABLATIONS

Geographic Embeddings?	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Yes	27.0	54.4	70.0	84.4	93.2
No	22.8	54.0	71.3	83.5	92.8

defaults and held constant across the GeoSURGE ablations. In the first row we see the largest performance improvement, with the most gains on the shorter Street and City distance metrics highlighting the importance of semantics for fine grained visual geo-localization.

V. CONCLUSION

In conclusion, GeoSURGE combines the strengths from classification and retrieval approaches, by learning a novel geographic representation that explicitly models the world as a hierarchy of geographic embeddings. Based on this geographic representation, GeoSURGE formulates the geo-localization problem as matching the visual representation of the query image against the feature vectors from the geographic representation. In addition, GeoSURGE efficiently enriches the visual representation of the image by using latent cross-attention to fuse appearance features with semantic segmentation.

TABLE VII
HIERARCHY DEPTH ABLATIONS

Hierarchy Levels	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
7	27.0	54.4	70.0	84.4	93.2
5	24.1	53.2	68.8	82.3	92.4
3	24.0	51.5	67.5	82.2	92.4
1	17.3	45.5	65.4	82.3	91.6

TABLE VIII
SEMANTIC FUSION ABLATIONS

Fusion Blocks	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
3	+ 3.9%	+ 2.9%	+ 1.7%	+ 0.9%	+ 0%
2	+ 2.6%	+ 0%	+ 0.5%	- 0.4%	+ 0%
1	+ 1.3%	+ 1.2%	- 0.8%	+ 0%	- 1.2%

GeoSURGE demonstrated new SOTA results across 22 of 25 metrics on five benchmark datasets. The consistent improvements over the most similar approaches, like GeoCLIP, underscore the importance of the geographic representation. Ablation studies further highlighted that our novel geographic and visual representations were central to these performance improvements from GeoSURGE. Specifically, geographic embeddings improved across all distance metrics while semantic segmentation provided the strongest gains for fine grained geo-localization. While these findings affirm that domain-specific models remain critical for global geo-localization tasks, future advancements might be drawn by efficiently combining the benefits of LVLMs.

REFERENCES

[1] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021.

[2] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013.

[3] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, et al. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21967–21977, 2024.

[4] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

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

[6] 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*, pages 37–55. Springer, 2016.

[7] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2749–2754, 2024.

[9] Pengyue Jia, Yiding Liu, Xiaopeng Li, Xiangyu Zhao, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, and Dawei Yin. G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. *Advances in Neural Information Processing Systems*, 37:53198–53221, 2024.

[10] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12893–12902, 2024.

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

[12] 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)*, pages 563–579, 2018.

[13] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23182–23190, 2023.

[14] 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)*, pages 536–551, 2018.

[15] 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*, pages 3–19. Springer, 2020.

[16] 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*, pages 750–760, 2022.

[17] Meiliu Wu and Qunying Huang. Im2city: image geo-localization via multi-modal learning. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 50–61, 2022.

[18] Grace Luo, Giscard Biamby, Trevor Darrell, Daniel Fried, and Anna Rohrbach. G3: Geolocation via guidebook grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5841–5853, 2022.

[19] 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 *European Conference on Computer Vision*, pages 196–215. Springer, 2022.

[20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[21] 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*, pages 155–163, 2021.

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

[23] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023.

[24] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[25] 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*, pages 770–778, 2016.

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[27] Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. Around the world in 80 timesteps: A generative approach to global visual geolocation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23016–23026, 2025.

[28] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[29] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[30] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024.