Defending from GeoLocalization through Adversarial Road Trips

Niccolò Niccoli[1], Federico Becattini[2], and Lorenzo Seidenari[1]

[1] University of Florence, Florence, Italy
name.surname@unifi.it
[2] University of Siena, Siena, Italy
name.surname@unisi.it

Abstract. Retrieval-based image geolocalization has emerged as a powerful technique for determining the location of a query image by matching it against a large, geotagged database. The success of deep learning based approaches has raised concerns regarding privacy and safety. A way to protect users from geolocalization is to design adversarial attacks for such methods. In this paper, we introduce RoadTrip Attack (RTA), a novel and highly effective targeted adversarial attack for geolocalization. RTA conceptualizes the adversarial process as finding an optimal "distractor" journey to a specific, attacker-chosen location. It employs a beam search algorithm to iteratively construct a sequence of incorrect geographic locations that form a path to the target. At each step, the attack generates subtle perturbations to the query image, guiding the geolocalization model toward the next location in this deceptive path. We show that our method is also strong in black-box settings, obtaining highly transferable attacks with less perceptible image artifacts.

Keywords: Adversarial attack · Image geolocalization · Privacy

1 Introduction

Recent advances in image-based geolocalization [4, 11, 44] pose pressing privacy and safety issues. Retrieval-based geolocalization methods [16, 44] can accurately predict a location's country and even region from a single image. The public availability of such models allows malicious actors to invade user privacy and, in more critical cases, may even endanger their safety and well-being. Modern image-based geolocalization models loosely fall into either retrieval or classification approaches. Retrieval-based methods rely on a large set of geolocated samples and some feature learning procedure. Classification-based methods generally partition the coordinate space hierarchically through classifiers. While this second approach has shown a slight edge in terms of precision, retrieval-based methods can be scaled up to incorporate more knowledge without relearning the feature space. Moreover, performing location recognition via retrieval, which compares a query embedding against a database of geotagged image prototypes, is formally equivalent to a classification task in which the feature vectors of the prototypes serve as the weight vectors for their respective location classes in a linear classifier.

In this paper, we address the privacy issues stemming from the widespread availability of geolocalization models through the lens of adversarial machine learning. We propose to protect user privacy by attacking image-based localization models. Adversarial methods are designed to produce a subtle perturbation that, when added to an image, induces erroneous behavior in the targeted

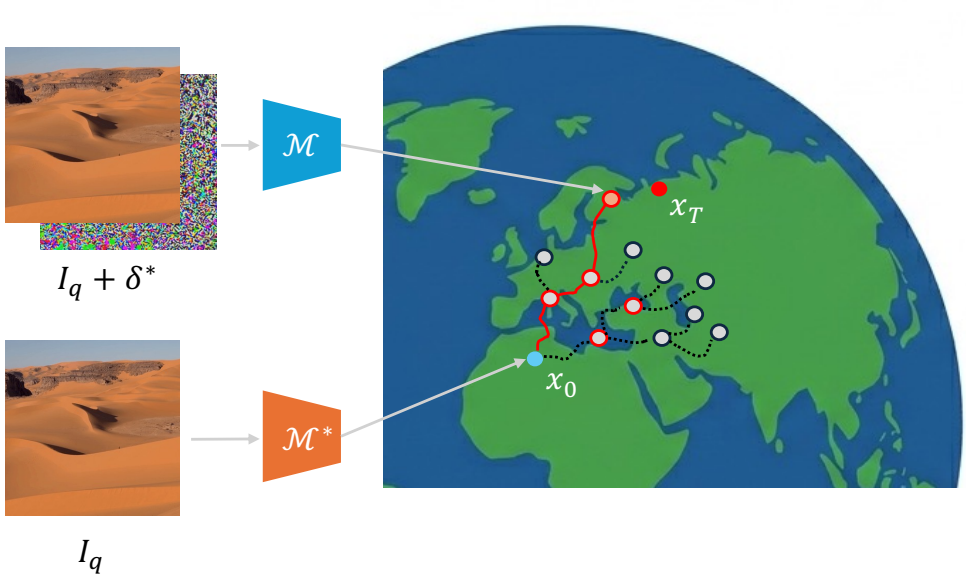


Fig. 1: RoadTrip Attack. Starting from an image I_q correctly localized at x_0 by the victim model $\mathcal{M}^*$, the attack crafts an adversarial image $I_q + \delta^*$. By leveraging a surrogate model $\mathcal{M}$, the algorithm explores a "beam" of black-box adversarial solutions (red dots). These are selected as the optimal subset of iteratively sampled geographic neighbors (gray dots). The final optimal "trip" (highlighted in red) demonstrates the gradual shift from the ground truth to the intended adversarial target location x_T.

model. Adversarial machine learning has been used in the past as a means of user privacy protection, for example in the field of text-based image editing [23, 34, 37, 43]. Here, perturbations are crafted to immunize user images from diffusion-based editing. Despite the extensive body of work on adversarial machine learning [1, 3, 6, 8, 25, 28, 40], robustness studies specifically targeting modern geolocalization models remain virtually non-existent, with GeoShield [24] representing the sole exception in the literature to date. We study the problem using different threat models: we first formulate a white-box approach, where we assume to have knowledge of the geolocalization model used by a malicious user; we then study the transferability of our approach in a black-box scenario, i.e. attacking a surrogate model, without having prior knowledge of the target geolocalization model.

First, we analyze the effectiveness of the state-of-the-art attacks such as Projected Gradient Descent (PGD) [25], Carlini & Wagner [5] and Fast Gradient Sign Method (FGSM) [10] in disrupting retrieval models. We then use this analysis to craft an even more effective method that leverages the geographic nature of the geolocalization task. Attacks like PGD have a significant advantage over simpler non-iterative attack methods, such as FGSM: the use of multiple steps and random restarts. This iterative process suggests that PGD could be made even stronger if multiple optimization paths were explored simultaneously during the attack.

A naive solution might be to sample multiple random perturbations in the image feature space at each iteration to broaden the search. However, a key issue arises: we can make few assumptions about the structure of this visual space. Recent work has shown that the feature spaces of contrastively-trained models can contain discontinuities among embeddings of the same modality [27, 49]. Furthermore, since perturbed images are out-of-distribution data, their behavior within

this space is unpredictable. Instead, we leverage the inherent geometrical structure of the problem. By performing targeted attacks, we can guide the model toward a set of intermediate geographic locations. We propose sampling these intermediate targets from a uniform distribution within a disk of a given radius around the target location. A second issue arises regarding the radius of this sampling disk. To ensure convergence, we employ a simple solution: scaling the sampling radius by the Haversine distance between the current location and the final attack target. Finally, to keep our approach computationally tractable, we adapt beam search to this problem. At each iteration, a batch of random geographic targets is used to generate adversarial candidates, but we only continue the optimization process for a small subset (the "beam") of the most successful ones. This approach finds non-obvious sequences of optimization paths, resulting in more effective adversarial noise. Because the sequence of intermediate targets resembles stops on a road trip from the original location to the adversarial one, we name our method the RoadTrip Attack. A visual sketch of our approach is presented in Fig. 1.

We evaluate our proposed attack by reporting white-box results on two popular geolocalization datasets. We also show that the attack effectively transfers to two other state-of-the-art models in a black-box fashion. Our results show that the RoadTrip Attack significantly outperforms standard adversarial attack baselines, as well as the VLM-based GeoShield attack, especially at extremely low perturbation budgets. Our contribution is threefold:

- We address a pressing privacy and safety concern regarding modern image-based geolocalization methods leveraging adversarial machine learning.
- We propose the RoadTrip Attack, a novel adversarial method specifically designed to exploit the inherent spatial structure of the geolocalization problem. RoadTrip improves over standard gradient-based methods by finding non-trivial optimization paths toward stronger attacks without requiring a higher noise budget.
- We empirically demonstrate that our specialized attack consistently and significantly outperforms strong, general-purpose baselines as well as geolocalization-based attacks, requiring a small perturbation budget both in white-box and black-box settings.

2 Related Work

Since in this work we seek to protect privacy from image-based geolocalization through adversarial machine learning, in the following we first review recent methods to estimate geographic coordinates from imagery and then we address adversarial machine learning literature.

2.1 Visual Worldwide Geolocalization

Visual worldwide geolocalization aims to predict the geographic coordinates of an image using only its visual content. Research in this area has evolved through three main methodological paradigms: classification-based [7, 11, 29, 33, 36, 41, 46], retrieval-based [4, 9, 12, 14, 32, 44, 45, 48], and generation-based approaches [16, 17, 51]. Some methods use a hybrid approach combining classification and regression [2, 11, 15, 19, 45].

Classification-based methods The earliest approaches conceptualized geolocalization as a multi-class classification task, partitioning Earth's surface into discrete geographic cells. Pioneering work such as PlaNet [46] leverages structured grid systems like Google's S2 geometry to create hierarchical

tessellations, where each cell represents a distinct class and the cell centroid serves as the estimated location [29, 36].

Researchers have explored diverse spatial partitioning methodologies from uniform regular grids [46] to adaptive partitions based on data density [7], semantic-driven divisions aligned with geographic features [41], and administrative boundary-based partitions [11, 33]. Contemporary advances like PIGEON [11] have elevated this paradigm by introducing semantically meaningful geocell construction with multi-task contrastive pretraining, achieving state-of-the-art performance while maintaining computational efficiency.

Retrieval-based methods Retrieval-based approaches reframe geolocalization as a nearest-neighbor search within high-dimensional feature spaces, assigning the coordinates of the most similar match as the predicted location [4, 48]. Early implementations relied on handcrafted descriptors including color histograms [13], gist features [32], and SIFT descriptors with SVM classifiers [14]. Deep learning architectures marked a transformative advancement [45], enabling end-to-end learning of robust visual representations with significantly enhanced matching accuracy. GeoCLIP [44] represents a notable breakthrough by introducing specialized location encoders using Random Fourier Features and Equal Earth projections, enabling direct image-to-GPS coordinate retrieval with hierarchical search capabilities.

Hybrid approaches Recognition of complementary strengths in classification and retrieval paradigms has motivated hybrid methodologies that strategically combine both approaches. These frameworks integrate discrete classification with continuous retrieval through ranking loss formulations [45] and contrastive learning objectives [19]. Notable implementations include classification-then-regression architectures that first predict coarse regions before applying fine-grained refinement [2, 11]. Advanced hybrid approaches estimate probability distributions over geographic space using spherical Gaussian formulations [15], providing uncertainty quantification alongside location estimates.

Generation-based methods The most recent paradigm leverages large-scale multimodal models to generate location predictions through sophisticated reasoning processes. Systems like Img2Loc [51], G3 [16], and GeoRanker [17] implement retrieval-augmented generation frameworks that synthesize visual features, GPS coordinates, and textual geographic descriptions within unified representational spaces. Unlike conventional retrieval methods relying solely on visual similarity, generation-based approaches incorporate explicit geo-alignment mechanisms into image representations, enabling sophisticated reasoning for heterogeneous queries that may differ substantially from reference data [16]. This paradigm demonstrates particular advantages when handling novel viewpoints, unusual lighting conditions, or rare geographic locations with limited training data.

Img2Loc [51] leverages a two-step RAG workflow for global geolocalization. By performing an initial retrieval of visually analogous images, the model extracts spatial metadata to serve as in-context evidence. This context is then injected into the prompt, allowing the VLLM to produce more precise geographic estimates. G3 [16] utilizes a three-stage framework to improve geolocalization accuracy. First, it merges images, text, and coordinates into a single searchable format. Then, using an RAG-based approach, it identifies a diverse set of possible locations. Finally, it filters those candidates to ensure the most reliable result. GeoRanker [17] introduces a distance-aware ranking framework for worldwide image geolocalization that goes beyond simple visual similarity. It uses large vision-language models to jointly encode interactions between a query image and candidate locations, learning both absolute and relative geographic distances through a multi-order

distance loss. This enables structured spatial reasoning over candidates and yields state-of-the-art performance on standard benchmarks.

2.2 Adversarial Machine Learning

Adversarial machine learning encompasses a variety of techniques designed to deliberately disrupt the performance of target models. These adversarial examples are crafted to be imperceptible to human observers while causing the model to produce significantly altered or incorrect outputs [5, 10, 25, 40].

The most common framework for generating these inputs is the additive threat model. Given an original image x, an adversarial counterpart is constructed as $x_{adv} = x + \delta$. To ensure the manipulation remains visually imperceptible, the perturbation δ is strictly constrained by a predefined budget ε, formulated as $||\delta||_\infty < \varepsilon$.

Generating this perturbation typically involves maximizing the target model's loss, with established methods approaching this through varying optimization strategies. The Fast Gradient Sign Method (FGSM) [10] generates the perturbation in a single step by moving in the direction of the gradient, while Projected Gradient Descent (PGD) [25] employs an iterative approach to maximize the loss while projecting back into the ε-budget constraint. In contrast, Carlini & Wagner (C&W) [5] diverges from strict budget bounding by minimizing the additive perturbation δ alongside a soft-constraint on the adversarial loss; here, a constant c acts as a hyperparameter to balance the priority of forcing misclassification against the goal of maintaining a minimal perturbation size. Several adversarial generation methodologies rely on stochastic meta-heuristics, such as differential evolution and evolutionary algorithms [30,39]. To enhance transferability, FTQ-PSO [22] introduces a population-based heuristic within the latent space, employing particle swarm optimization in a gray-box setting. In contrast to these purely stochastic or swarm-based approaches, our method integrates gradient-based local optimization with a systematic, heuristic-guided beam search. This hybrid strategy allows for the simultaneous exploration of multiple high-potential search paths while ensuring each candidate is locally refined for maximum adversarial impact.

Despite the rapid evolution of adversarial techniques in general computer vision, the intersection of adversarial machine learning and image geolocalization remains underexplored. GeoShield [24], which to our knowledge represents the only dedicated image-geolocalization-related adversarial method currently available in the literature, utilizes a feature disentanglement module to separate geographic from non-geographic information, alongside an exposure element identification module that pinpoints position-revealing regions within an image. By applying scale-adaptive adversarial perturbations to these specific areas, the model effectively misleads geolocation inference across various resolutions while maintaining the visual and semantic integrity of the original image.

3 Method

Our method leverages the intrinsic geometric structure of the image-based geolocalization problem. Similarly to other iterative [20, 25] adversarial attacks, we create a sequence of image perturbations to induce a distortion in the feature space in order to make localization inaccurate. Our method is based on alternating beam search and projected gradient descent [25] optimization. The main idea of attacking a geolocalization method is to target a location x_T that differs from the ground truth location x_0 and alter the image representation so that features are incorrectly placed close to the requested target.

3.1 Attacking Retrieval Based Localization

Our approach utilizes projected gradient descent [25] to attack retrieval-based geolocalizers. In this framework, the model leverages a gallery of pre-computed embeddings $\mathcal{D} : \{\mathbf{e}_0, \ldots, \mathbf{e}_k\}$. Depending on the architecture, these represent either a collection of geo-located reference images or a set of geographic coordinate embeddings. The system processes a query image I_q through an encoder to produce a query embedding. It then computes similarity scores s_i between the query and each gallery entry:

$$s_i = \langle \text{enc}(I_q), \mathbf{e}_i \rangle \tag{1}$$

The final geographic prediction is derived from the coordinates associated with the gallery embeddings that yield the highest similarity scores. Our method seeks to perturb I_q such that the similarity scores are redirected away from the ground-truth region toward distant, incorrect entries in the gallery.

Therefore, for a targeted attack against such methods, we minimize the cross-entropy loss with respect to the target embedding:

$$\mathcal{L}_{\text{CE}}\left(\text{enc}(I_q), \mathbf{e}_T\right) = -\sum_{i=1}^{C} q_i \log p_i, \tag{2}$$

where p_i are softmax outputs of scores s_i and $q_i = \mathbf{1}_{\{i=T\}}$. Here, C is the number of gallery images, and T is the index of the target embedding.

Each internal iteration of our method will update the image I_q according to:

$$I_q^{(t+1)} = \mathit{\Pi}_{\mathcal{B}_\epsilon(I_q)}\left(I_q^{(t)} - \alpha \,\text{sign}\left(\nabla_{I_q^{(t)}} \mathcal{L}_{\text{CE}}\left(\text{enc}\left(I_q^{(t)}\right), \mathbf{e}_T\right)\right)\right) \tag{3}$$

where α is the iteration step, ϵ is the attack budget, and $\mathit{\Pi}_{\mathcal{B}_\epsilon(I_q)}$ is the projection operator clipping the adversarial image in the l_∞ ball centered at the original query image I_q with radius ϵ.

3.2 RoadTrip Attack

The optimization landscape for adversarial attacks on geolocalization models is challenging. Visually similar features, such as comparable architecture or vegetation, can exist in geographically distant locations, creating a highly complex and non-convex feature space. A standard gradient-based optimization like PGD [25] follows a direct, greedy path toward the target location. This approach is susceptible to getting trapped in poor local minima, where the model's output is closer to the target but the required perturbation is unnecessarily large or sub-optimal.

To overcome this, the RoadTrip Attack reframes the problem. Instead of a direct attack toward the adversarial target, we conceptualize the attack as a journey. The intuition is that finding a sequence of easier, intermediate steps can lead to a more effective final perturbation. Our approach explores multiple branching paths in parallel, using beam search to prune unpromising routes. This method acts as a form of "geographic annealing", allowing the optimization to navigate the complex feature space more effectively and discover non-obvious, stronger adversarial solutions that a direct path would miss.

The main idea of the RoadTrip attack is to avoid a direct path toward the target but instead iteratively sample intermediate targets in a disk around the final target and leverage beam search to pick the most promising top-k candidates for the next step. Unlike PGD, which optimizes toward a fixed target, our method concurrently builds multiple paths toward the final destination. This

Algorithm 1 RoadTrip Attack.

Require: I_q,x_q, x_T, ϵ, α, $\mathcal{L}(\cdot)$, $\mathcal{G}(\cdot)$, η, N, J, K
$\delta_1 \sim \mathcal{U}_B$ *//Initialize perturbation*
$\mathcal{S}_1 \leftarrow \{(x_q, \delta_1)\}$ *//Initialize source set*
for $i = 1$ to N **do**
  $\mathcal{S}^*_{i+1} \leftarrow \emptyset$ *//Initialize next source candidates set*
  for $(x_{m,i}, \delta_{m,i}) \in \mathcal{S}_i$ **do**
    $\mathrm{R} \leftarrow \eta\mathcal{G}(x_{m,i}, x_T)$ *//Estimate sampling radius*
    $\mathcal{T}_i \leftarrow \{\tau^j_m | \tau^j_m \sim \mathcal{U}_{D_{\mathrm{R}}(x_{T,i})},\ \forall j \in [1 \dots J]\}$ *//Populate target set*
    for $\tau^j_m \in \mathcal{T}_i$ **do**
      $\delta^j_{m,i+1} \leftarrow \mathrm{P}(I_q, \tau^j_m, \delta_{m,i}, \epsilon, \alpha, \mathcal{L})$ *//Update prev. noises $\delta_{m,i}$ for targets τ^j_m*
      $x^j_{m,i+1} \leftarrow \mathcal{M}(I_q + \delta^j_{m,i+1})$
      $\mathcal{S}^*_{i+1} \leftarrow \mathcal{S}^*_{i+1} \cup \{(x^j_{m,i+1}, \delta^j_{m,i+1})\}$
    end for
  end for
  $\mathcal{S}_{i+1} \leftarrow \mathrm{top}_K(\mathcal{S}^*_{i+1}, x_T, \mathcal{G})$
end for
$(x^*, \delta^*) \leftarrow \mathrm{argmin}_{(x,\delta)} \mathcal{G}(x, x_T)$ s.t. $(x, \delta) \in \mathcal{S}_N$
return δ^*

iterative process of generating and pruning candidate solutions allows the attack to discover more effective adversarial paths. A detailed, step-by-step description of our method is provided in Alg. 1 and is explained in the remainder of this section.

In the following, we measure distances between coordinates on the earth using the Haversine distance

$$\mathcal{G} = 2R \sin^{-1}\left(\sqrt{\mathrm{h}(\phi_2 - \phi_1) + \cos(\phi_1)\cos(\phi_2)\ \mathrm{h}(\lambda_2 - \lambda_1)}\right) \tag{4}$$

which measures the shortest path between coordinates (ϕ_1, λ_1) and (ϕ_2, λ_2), expressed in latitude ϕ_i and longitude λ_i, given the average Earth radius R, and where $\mathrm{h}(\theta) = \sin^2\left(\frac{\theta}{2}\right)$.

Our method iteratively grows and prunes a set of candidate starting points $\mathcal{S}_i$ and perturbations $\mathcal{N}_i$. The state of our algorithm is retained by a set of image perturbation values $\mathcal{N}_i : \{\delta_{1,i} \dots \delta_{K,i}\}$ and a set of geographic coordinates $\mathcal{S}_i : \{x_{1,i} \dots x_{K,i}\}$.

At any given iteration i, for each starting point $x_{m,i} \in \mathcal{S}_i$ we uniformly sample a set of target coordinates τ^j_m in a disk $D_R(x_T)$ of radius R, centered in x_T. For each of the sampled locations τ^j_m, we create a perturbation via projected-gradient optimization, aiming at altering the image representation to make it geographically closer to location τ^j_m, according to some loss $\mathcal{L}$, which in its vanilla form is a cross-entropy loss $\mathcal{L}_{CE}$.

Let

$$\delta_{m,i+1} \leftarrow \mathrm{P}(I_q, \tau^j_m, \delta_{m,i}, \epsilon, \mathcal{L}) \tag{5}$$

be the projected gradient descent attack on query image I_q directed toward location τ^j_m which, at convergence, returns the best perturbation $\delta_{m,i+1}$. Such a perturbation, when applied to I_q, leads the model to yield the corresponding location $x_{m,i+1}$.

We initialize the noise $\delta_{m,i}$ by uniformly sampling in $B(\epsilon)$ where

$$B(\epsilon) = \{\delta \in \mathbb{R}^n : \|\delta\|_\infty \leq \epsilon\}, \tag{6}$$

We name $\mathcal{T}_{i+1}$ the set of all attacks converged to solutions $x_{m,i+1}$, and pick the top-K according to $\mathcal{G}$. Since we are targeting x_T we pick the K closest solutions to x_T as the subsequent iteration of starting points $\mathcal{S}_{i+1}$.

As in beam search, we iteratively drop less promising solutions and move toward our target set of candidates. After N steps, we pick the best perturbation δ_{i^*} as the one corresponding to $x_{i^*} \in \mathcal{S}_N$, i.e. the closest solution to x_T. An important implementation detail is the choice of the sampling radius R. A fixed radius presents a trade-off: a large radius would be inefficient for fine-grained adjustments when the solution is already close to the target, while a small radius would make the trip exceedingly slow and computationally expensive when far away. To resolve this, we employ an adaptive radius strategy. As outlined in Alg. 1, the radius R is dynamically scaled in proportion to the Haversine distance between the current candidate location $x_{m,i}$ and the final adversarial target x_T. When the current solution is geographically distant from the final target, the algorithm uses a larger radius to explore wider jumps, rapidly closing the distance. Conversely, as the attack path converges toward the target, the sampling radius automatically shrinks. This narrowing focus allows for a more fine-grained search, enabling the precise optimization required to reach the intended location. In our experiments, we set the scaling factor η to 0.5.

The target location x_T is sampled randomly from a gallery of GPS coordinates under the condition that the Haversine distance between x_0 and x_T is more than 2500km. Other than that, the GPS gallery is used only during noise optimization to compute cross-entropy loss (which is inherent to retrieval-based models). Thus, RoadTrip Attack does not require access to the full database for intermediate target selection.

In the black-box setting, we use GeoCLIP [44] as a surrogate for the attacked models. Notably, this approach requires access to neither the victim's internal architecture nor its gallery.

4 Experiments

4.1 Datasets and Metrics

We evaluate our proposed attack on two geolocalization benchmarks, Im2GPS3k and YFCC4k [45]. IM2GPS3k consists of around 3000 images from Im2GPS [13]. Note that this differs from the original Im2GPS test set [13]. YFCC4k comprises around 4000 random images from the YFCC100m [42] dataset. YFCC100m was not collected specifically for geolocalization but for generic computer vision tasks. Therefore, its distribution is different from Im2GPS and it is considered more challenging. To evaluate the effectiveness of our attack, we use three different metrics. Ground Truth Accuracy ACC_{GT} is the primary metric to assess the success of the attack. It measures the percentage of attacked images that are still localized within a distance threshold of their original true location. For a successful attack, this value should be as low as possible, demonstrating the model has been successfully fooled. Target Accuracy ACC_{Target} measures the percentage of attacked images whose predicted location falls within a certain Haversine distance threshold from the chosen adversarial target location x_T. A higher ACC_{Target} value indicates a more successful targeted attack. We evaluate the attack for different ACC_{GT} and ACC_{Target} distance thresholds, namely 1km, 25km, 200km, 750km, and 2500km, computed with the Haversine distance.

We vary the budget for all attacks, setting $\epsilon = \{2/255, 4/255\}$ for PGD, RTA and FGSM and $c = \{0.1, 1\}$ for Carlini-Wagner. Considering the high target accuracy reached by RTA at 1km we do not test larger budgets.

			ACC_{GT} (↓)					ACC_{Target} (↑)				
	ε	LPIPS (↓)	@1km	@25km	@200km	@750km	@2500km	@1km	@25km	@200km	@750km	@2500km
CW	c = 0.1	0.031	0.43%	1.63%	2.80%	5.44%	10.88%	53.79%	55.59%	59.26%	63.90%	74.74%
CW	c = 1.0	0.079	0.10%	0.43%	0.63%	1.40%	3.97%	81.95%	83.35%	84.35%	86.55%	90.52%
FGSM		0.007	6.64%	15.98%	23.69%	37.30%	55.39%	0.33%	0.33%	0.53%	1.77%	9.41%
PGD	2/255	0.007	0.40%	1.03%	2.10%	4.34%	8.34%	71.81%	73.97%	76.24%	79.71%	85.82%
RTA		0.027	**0.03%**	**0.03%**	**0.03%**	**0.10%**	**1.03%**	**93.43%**	**94.99%**	**95.70%**	**96.96%**	**98.50%**
FGSM		0.042	4.57%	12.08%	18.82%	30.93%	49.15%	0.40%	0.40%	0.93%	2.50%	11.64%
PGD	4/255	0.044	**0.00%**	0.07%	0.23%	0.33%	1.27%	96.63%	98.13%	98.47%	98.67%	99.03%
RTA		0.059	**0.00%**	**0.00%**	**0.00%**	**0.03%**	**0.73%**	**98.93%**	**99.73%**	**99.83%**	**99.87%**	**99.93%**

Table 1: Performance of RTA attacking GeoCLIP [44] on Im2GPS3k [45].

			ACC_{GT} (↓)					ACC_{Target} (↑)				
	ε	LPIPS (↓)	@1km	@25km	@200km	@750km	@2500km	@1km	@25km	@200km	@750km	@2500km
CW	c = 0.1	0.022	0.35%	0.77%	1.90%	4.41%	10.10%	47.38%	49.27%	52.40%	57.50%	69.49%
CW	c = 1.0	0.054	0.13%	0.26%	0.49%	1.41%	4.45%	76.85%	78.62%	80.42%	82.43%	87.57%
FGSM		0.008	4.14%	8.42%	15.70%	30.91%	49.80%	0.24%	0.24%	0.44%	1.74%	9.55%
PGD	2/255	0.009	0.15%	0.37%	0.82%	2.07%	4.23%	62.30%	64.30%	67.12%	71.62%	80.44%
RTA		0.031	**0.04%**	**0.07%**	**0.09%**	**0.24%**	**1.19%**	**88.60%**	**90.37%**	**92.15%**	**94.44%**	**97.49%**
FGSM		0.041	3.28%	6.33%	11.73%	25.62%	44.40%	0.26%	0.26%	0.57%	2.14%	11.44%
PGD	4/255	0.044	0.04%	0.13%	0.22%	0.42%	1.34%	94.80%	96.49%	97.00%	97.57%	98.48%
RTA		0.062	**0.02%**	**0.02%**	**0.02%**	**0.02%**	**0.68%**	**97.82%**	**99.01%**	**99.21%**	**99.51%**	**99.82%**

Table 2: Performance of RTA attacking GeoCLIP [44] on YFCC4k [45].

4.2 Implementation Details

All experiments have been run on a RTX 5000 Blackwell. For the models, we used the official codebases of GeoCLIP [44], G3 [16], and a custom implementation of Img2Loc [51] based on the one used as baseline by Jia et al. [16]; this was necessary since the official codebase of Img2Loc is not complete.

4.3 White-Box Results

We target a state-of-the-art geolocalization model, GeoCLIP [44], a retrieval method leveraging contrastively learned representations of images and coordinates. We benchmark our approach against widely used adversarial baselines, namely FGSM [10], PGD [25], and CW [5] on the Im2GPS3k and YFCC4k datasets [45], to ensure a comprehensive robustness evaluation.

Our experimental results, summarized in Tab. 1 and Tab. 2, demonstrate that the RoadTrip Attack consistently outperforms all baselines across all models and datasets. Among the baselines, PGD perform best. FGSM evidently suffers from the single-step attack and does not allow sufficiently complex perturbations to be crafted. CW has better performance with respect to FGSM, but still underperforms versus PGD. The benefits of RTA emerge more prominently in low-budget settings ($\epsilon = 2/255$). For instance, on the Im2GPS3k dataset (Tab. 1), with a minimal perturbation budget $\epsilon = 2/255$, the standard PGD attack only reduces the ACC_{GT}@2500km to 8.34%, whereas RTA in its vanilla formulation lowers it to 1.03%. At the same time, RTA achieves a much higher target success rate with the same budget, increasing ACC_{Target}@1km from 71.81% to 93.43%. This pattern is consistent across datasets.

	ACC_{GT} (↓)					ACC_{Target} (↑)				
	@1km	@25km	@200km	@750km	@2500km	@1km	@25km	@200km	@750km	@2500km
Clean	14.11%	34.47%	50.65%	69.67%	83.82%	–	–	–	–	–
PGD	0.40%	1.03%	2.10%	4.34%	8.34%	71.81%	73.97%	76.24%	79.71%	85.82%
Multi start	0.90%	0.90%	2.70%	3.60%	5.41%	72.07%	76.58%	79.28%	85.59%	89.19%
Random restart	0.45%	0.90%	2.52%	4.41%	8.55%	64.27%	66.70%	69.94%	75.16%	82.90%
Interpolated	0.99%	1.89%	3.60%	6.84%	13.77%	32.13%	33.57%	39.51%	47.79%	68.32%
RTA	**0.03%**	**0.03%**	**0.03%**	**0.10%**	**1.03%**	**93.43%**	**94.99%**	**95.70%**	**96.96%**	**98.50%**

Table 3: Comparison between PGD variants and RTA attacking GeoCLIP [44] on Im2GPS3k [45].

		ACC_{GT} (↓)				
Method	ε	@1km	@25km	@200km	@750km	@2500km
GeoShield	2/255	8.54%	24.42%	33.73%	47.45%	63.43%
RTA		**2.77%**	**7.61%**	**10.14%**	**16.08%**	**27.86%**
GeoShield	4/255	7.57%	19.99%	26.63%	38.14%	55.52%
RTA		**2.20%**	**6.14%**	**8.31%**	**13.78%**	**24.52%**
GeoShield	8/255	4.87%	13.41%	17.22%	27.29%	44.24%
RTA		**2.20%**	**5.74%**	**7.47%**	**12.61%**	**23.69%**
GeoShield	16/255	2.74%	8.31%	10.98%	17.62%	32.10%
RTA		**1.80%**	**4.94%**	**6.24%**	**11.04%**	**21.19%**

Table 4: Black-box transferability from GeoCLIP to Img2Loc on Im2GPS3k.

To highlight the contribution of the geographical framing of the RTA problem, we compare our results with several PGD-based variants under the same perturbation constraint, $\varepsilon = 2/255$ (Tab. 3).*Multi-start* runs five PGD optimizations in parallel and selects the best adversarial example. *Random restart* randomly perturbs the adversarial noise during optimization, introducing stochastic exploration. *Interpolated* uses five deterministic intermediate targets obtained by interpolating between the source and target locations. RTA consistently achieves lower ACC_{GT} and higher ACC_{Target} than all variants, indicating that its improvement is not only due to a larger search budget or random exploration, but to the adaptive selection of intermediate geographic targets.

4.4 Black-Box Results

To evaluate our method under a more realistic threat model, we tested RTA against image geolocalization models different from the one used for the attack. We focus on Img2Loc [51] and G3 [16], motivated by their state-of-the-art performance. We compare RTA to GeoShield [24], which to the best of our knowledge is the only available image geolocalization adversarial method existing in the literature. Note that the gallery used in the retrieval step of both these methods, MP16 [21], differs from the one on which we learn the adversarial noise, which is more than 40× smaller. In Tab. 4 and Tab. 5 we report ground truth accuracies for the attacks. We do not report target accuracy as GeoShield is untargeted. From the tables, it is clear that our method produces more disruptive adversarial perturbations. It is also very important to note that the adversarial noise added by RTA better preserves visual fidelity. To measure this, we compute the Learned Perceptual Image Patch Similarity (LPIPS) [50], which measures the visual quality and subtlety of the generated pertur-

		ACC_{GT} (↓)				
Method	ε	@1km	@25km	@200km	@750km	@2500km
GeoShield	2/255	12.38%	32.97%	44.98%	62.46%	78.14%
RTA		**1.17%**	**3.04%**	**4.54%**	**8.07%**	**19.12%**
GeoShield	4/255	9.71%	26.43%	36.07%	50.58%	68.67%
RTA		**0.80%**	**2.34%**	**3.17%**	**6.11%**	**15.42%**
GeoShield	8/255	6.67%	17.08%	22.96%	35.27%	54.99%
RTA		**0.77%**	**2.67%**	**3.67%**	**6.67%**	**16.15%**
GeoShield	16/255	3.74%	10.71%	14.58%	24.42%	45.01%
RTA		**0.73%**	**2.27%**	**3.50%**	**6.47%**	**16.08%**

Table 5: Black-box transferability from GeoCLIP to G3 on Im2GPS3k.

Method	ε	LPIPS (↓)	PSNR (↑)	ε	LPIPS (↓)	PSNR (↑)
GeoShield	2/255	0.047	35.26	8/255	0.183	32.96
RTA		**0.027**	**41.42**		**0.108**	**36.07**
GeoShield	4/255	0.097	34.60	16/255	0.305	29.94
RTA		**0.059**	**38.57**		**0.174**	**32.90**

Table 6: RTA adds less visible noise than GeoShield as highlighted by the lower LPIPS and the higher PSNR.

bations. LPIPS calculates the perceptual distance between the original and the perturbed image. Lower LPIPS values signify that the adversarial image is visually similar to the original, making the attack imperceptible to a human observer. As shown in Tab. 6, RTA alters the visual quality of the image in a much more subtle way compared to GeoShield, consistently across perturbation budgets. This difference is especially noticeable for high perturbation budgets, e.g. the LPIPS for GeoShield for $\varepsilon = 16/255$ is almost twice the LPIPS for RTA. This behavior can be clearly observed qualitatively in Fig. 2, where samples of attacked images are reported for different budgets. The adversarial noise for GeoShield is clearly visible to the human eye.

4.5 Runtime

RTA requires more computation, but unlike PGD, most of the computation can be parallelized. The comparison in Fig. 3a shows that RTA has a lower wall-clock time since the J intermediate targets (see Alg. 1) can be optimized in parallel. Recalling that in both cases we early-stop attacks when the target accuracy reaches 1km from the objective, this has a positive impact on our method, which, by searching a larger space, allows more early stops than PGD.We also report that GeoShield, in the same setting, runs at 20s/img, so 8x slower than RTA. All RTA variants in Tab. 1 and Tab. 2 use beam size 4, 5 parallel targets (respectively K and J in Alg.1), and for each of them 5 PGD steps. We use a step size of $\varepsilon/2$. Additionally, we report that PGD uses around 4GB of VRAM, RTA around 6GB, and GeoShield around 8.5GB.

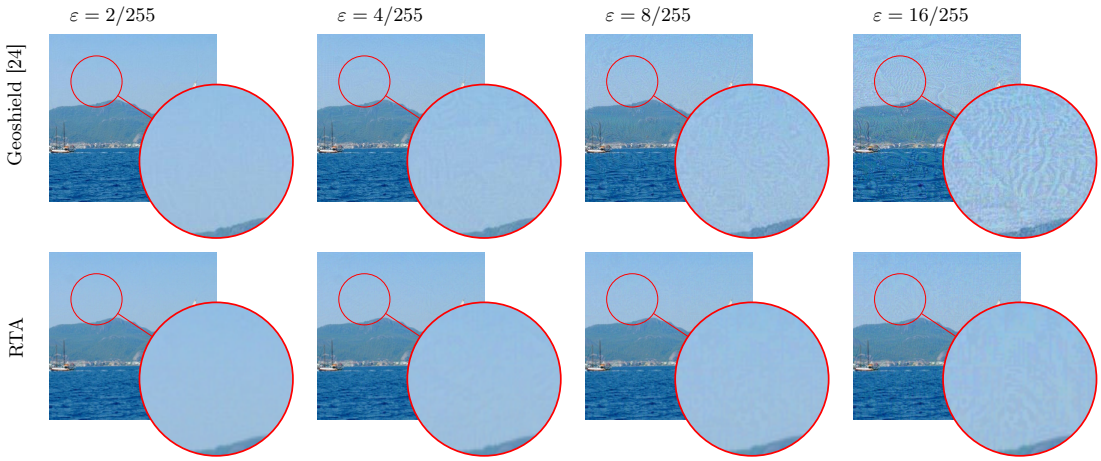


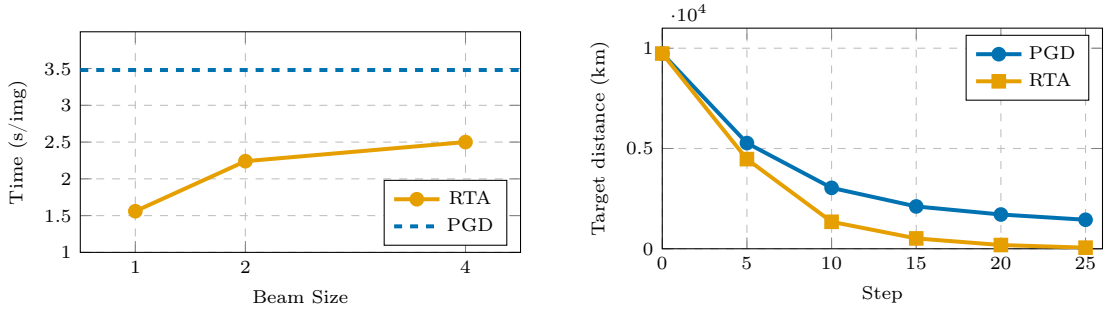
Fig. 2: Visual comparison of adversarial noise for four budgets $\varepsilon = 2/255, 4/255, 8/255, 16/255$ and two methods (Geoshield and ours). The magnified area highlights the fine-grained perturbation patterns for the attack. Best viewed in color on a screen.

4.6 Ablation studies

To better understand the behavior of RTA, we perform a series of ablation studies, focusing on the white-box setting. We report a convergence analysis in Fig. 3b against the strong PGD baseline. Here, we plot the distance from the target against the optimization step of the two methods. While PGD follows a more direct optimization path, RTA explores multiple paths, leading to a better solution. As the attack progresses, RTA's beam search strategy identifies a more effective trajectory, ultimately converging to a solution significantly closer to the adversarial target than PGD. This confirms that the iterative, multi-path approach is more effective at navigating the feature space to find strong adversarial solutions.

We then perform an ablation study on the beam size parameter (K in Alg. 1) of our proposed attack. We can observe from Fig. 4a a consistent trend showing that increasing the number of beams increases the attack's effectiveness. For the experiments reported throughout the paper we use K = 4 beams.

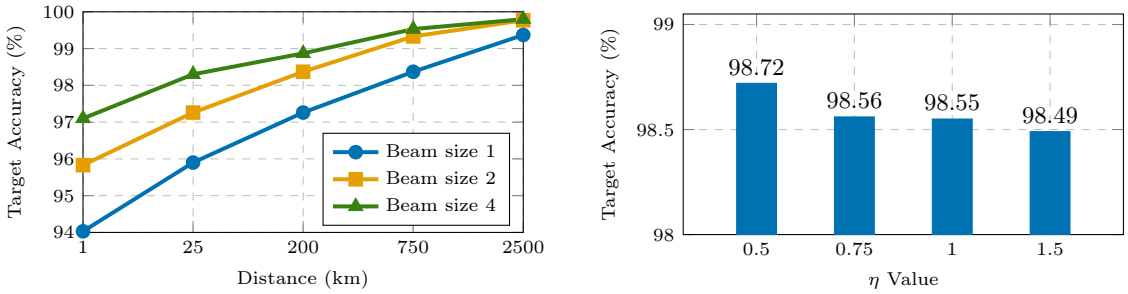
In addition to the beam size, we conducted a further ablation study to analyze the sensitivity of the RoadTrip Attack to the sampling radius scaling factor, η. This hyperparameter governs the exploration scope of the intermediate target sampling at each step of the attack. As illustrated in Fig. 4b, we evaluated the attack's performance with η values set to 0.5, 0.75, 1.0, and 1.5. The empirical results indicate that a value of $\eta = 0.5$ yields the highest target accuracy, particularly at the most challenging 1km and 25km thresholds. This suggests that a more conservative sampling strategy, which defines a tighter disk for intermediate targets relative to the remaining distance, allows the optimization to discover a more effective and precise adversarial path. Conversely, larger η values, such as 1.5, result in a performance degradation. This behavior implies that excessive spatial jumps may hinder the optimization process, preventing it from converging to the best adversarial solution.



(a) Processing time per image for RTA vs. PGD.

(b) Target distance at different attack steps.

Fig. 3: Ablation study comparing RTA with PGD on IM2GPS3k, comparing average processing time (a) and target distance (b).



(a) Impact of beam search size (K).

(b) Sensitivity to scaling factor (η).

Fig. 4: Experimental analysis of RTA parameters. (a) Larger beams increase target accuracy across distances. (b) Best average results are obtained with $\eta = 0.5$.

5 Conclusions

In this work, we introduced the RoadTrip Attack, a novel adversarial method to address the privacy risks of image geolocalization. Our approach formulates the attack as an optimal distractor journey to a set of intermediate locations, using a beam search algorithm to find effective perturbation paths. Extensive experiments demonstrate that our method significantly outperforms both standard adversarial attacks and the only prior existing attack tailored to geolocalization. RTA achieves superior success rates in both white-box and black-box settings, particularly in low-budget regimes where perturbations must remain subtle.

The success of this journey-based approach highlights a critical vulnerability: the complex optimization landscape of geolocalization models can be more effectively navigated through a sequence of intermediate steps rather than a direct path.

Future work will address the robustness of our attack against preprocessing and purification-based countermeasures. In particular, we will consider naive JPEG compression, input-transformation defenses [47], reconstruction-based purification [26,38], generative purification [31,35], and recovery perturbations [18].

Acknowledgements. This work was partially supported by the REG4AI project.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: Proc. of ECCV. pp. 484–501. Springer (2020)
2. Astruc, G., Dufour, N., Siglidis, I., Aronssohn, C., Bouia, N., Fu, S., Loiseau, R., Nguyen, V.N., Raude, C., Vincent, E., Xu, L., Zhou, H., Landrieu, L.: OpenStreetView-5M: The many roads to global visual geolocation. Proc. of CVPR (2024)
3. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Proc. of ICML. pp. 284–293. PMLR (2018)
4. Berton, G., Mereu, R., Trivigno, G., Masone, C., Csurka, G., Sattler, T., Caputo, B.: Deep visual geo-localization benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5396–5407 (2022)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
6. Chen, P.Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.J.: Ead: elastic-net attacks to deep neural networks via adversarial examples. In: Proc. of AAAI. vol. 32 (2018)
7. Clark, B., Kerrigan, A., Kulkarni, P.P., Cepeda, V.V., Shah, M.: 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. pp. 23182–23190 (2023)
8. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proc. of ICML. PMLR (2020)
9. Garg, S., Fischer, T., Milford, M.: Where is your place, visual place recognition? arXiv preprint arXiv:2103.06443 (2021)
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) 3rd Int'l Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6572
11. Haas, L., Skreta, M., Alberti, S., Finn, C.: Pigeon: Predicting image geolocations. In: Proc. of CVPR. pp. 12893–12902 (2024)
12. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: Proc. of CVPR. pp. 14141–14152 (2021)
13. Hays, J., Efros, A.A.: Im2gps: estimating geographic information from a single image. In: Proc. of CVPR. pp. 1–8. IEEE (2008)
14. Hays, J., Efros, A.A.: Large-scale image geolocalization. In: Choi, J., Friedland, G. (eds.) Multimodal Location Estimation of Videos and Images, pp. 41–62. Springer (2015). https://doi.org/10.1007/978-3-319-09861-6_3, https://doi.org/10.1007/978-3-319-09861-6_3
15. Izbicki, M., Papalexakis, E.E., Tsotras, V.J.: Exploiting the earth's spherical geometry to geolocate images. In: Proc. of ECML-PKDD. pp. 3–19. Springer (2019)
16. Jia, P., Liu, Y., Li, X., Zhao, X., Wang, Y., Du, Y., Han, X., Wei, X., Wang, S., Yin, D.: G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. Proc. of NeurIPS **37**, 53198–53221 (2024)
17. Jia, P., Park, S., Gao, S., Zhao, X., Li, S.: Georanker: Distance-aware ranking for worldwide image geolocalization. Proc. of NeurIPS **38**, 17673–17699 (2026)
18. Jiang, W., Diao, Y., Wang, H., Sun, J., Wang, M., Hong, R.: Unlearnable examples give a false sense of security: Piercing through unexploitable data with learnable examples. In: Proc. of ACM MM. pp. 8910–8921 (2023)
19. Kordopatis-Zilos, G., Galopoulos, P., Papadopoulos, S., Kompatsiaris, I.: Leveraging efficientnet and contrastive learning for accurate global-scale location estimation. In: Proc. of ICMR. pp. 155–163 (2021)
20. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial intelligence safety and security, pp. 99–112. Chapman and Hall/CRC (2018)
21. Larson, M., Soleymani, M., Gravier, G., Ionescu, B., Jones, G.J.: The benchmarking initiative for multimedia evaluation: Mediaeval 2016. IEEE MultiMedia **24**(1), 93–96 (2017)

22. Li, C., Jiang, T., Wang, H., Yao, W., Wang, D.: Optimizing latent variables in integrating transfer and query based attack framework. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(1), 161–171 (2025). https://doi.org/10.1109/TPAMI.2024.3461686
23. Liang, C., Wu, X.: Mist: Towards improved adversarial examples for diffusion models. arXiv preprint arXiv:2305.12683 (2023)
24. Liu, X., Jia, X., Xun, Y., Qin, S., Cao, X.: Geoshield: Safeguarding geolocation privacy from vision-language models via adversarial perturbations (2025), https://arxiv.org/abs/2508.03209
25. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th Int'l Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=rJzIBfZAb
26. Meng, D., Chen, H.: Magnet: a two-pronged defense against adversarial examples. In: Proc. of SIGSAC. pp. 135–147 (2017)
27. Mistretta, M., Baldrati, A., Agnolucci, L., Bertini, M., Bagdanov, A.D.: Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. In: Proc. of ICLR (2025)
28. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proc. of CVPR (2016)
29. Muller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: Proc. of ECCV. pp. 563–579 (2018)
30. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proc. of CVPR. pp. 427–436 (2015)
31. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proc. of ICML. vol. 162, pp. 16805–16827. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/nie22a.html
32. Oliva, A., Torralba, A.: Building the gist of a scene: the role of global image features in recognition. Progress in brain research **155**, 23–36 (2006), https://api.semanticscholar.org/CorpusID:2432623
33. Pramanick, S., Nowara, E.M., Gleason, J., Castillo, C.D., Chellappa, R.: Where in the world is this image? transformer-based geo-localization in the wild. In: Proc. of ECCV. pp. 196–215. Springer (2022)
34. Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., Madry, A.: Raising the cost of malicious ai-powered image editing (2023), https://arxiv.org/abs/2302.06588
35. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: Protecting classifiers against adversarial attacks using generative models. In: Proc. of ICLR (2018)
36. Seo, P.H., Weyand, T., Sim, J., Han, B.: Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In: Proc. of ECCV. pp. 536–551 (2018)
37. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by text-to-image models. In: 32nd USENIX Security Symp. (USENIX Security 23). pp. 2187–2204 (2023)
38. Song, Y., Kim, T., Nowozin, S., Ermon, S., Kushman, N.: Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In: Proc. of ICLR (2018)
39. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019)
40. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: Bengio, Y., LeCun, Y. (eds.) 2nd Int'l Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014), http://arxiv.org/abs/1312.6199
41. Theiner, J., Müller-Budack, E., Ewerth, R.: Interpretable semantic photo geolocation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 750–760 (2022)
42. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)

43. Trippodo, M., Becattini, F., Seidenari, L.: Immunizing images from text to image editing via adversarial cross-attention. p. 10535–10543. Proc. of ACM MM, Association for Computing Machinery, New York, NY, USA (2025)
44. Vivanco Cepeda, V., Nayak, G.K., Shah, M.: Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. Proc. of NeurIPS **36**, 8690–8701 (2023)
45. Vo, N., Jacobs, N., Hays, J.: Revisiting im2gps in the deep learning era. In: Proc. of CVPR. pp. 2621–2630 (2017)
46. Weyand, T., Kostrikov, I., Philbin, J.: PlaNet - Photo Geolocation with Convolutional Neural Networks, p. 37–55. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-46484-8_3, http://dx.doi.org/10.1007/978-3-319-46484-8_3
47. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)
48. Xu, Y., Shamsolmoali, P., Granger, E., Nicodeme, C., Gardes, L., Yang, J.: Transvlad: Multi-scale attention-based global descriptors for visual geo-localization. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2840–2849 (2023)
49. Yi, C., Ren, L., Zhan, D.C., Ye, H.J.: Leveraging cross-modal neighbor representation for improved clip classification. In: Proc. of CVPR. pp. 27402–27411 (2024)
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of CVPR. pp. 586–595 (2018)
51. Zhou, Z., Zhang, J., Guan, Z., Hu, M., Lao, N., Mu, L., Li, S., Mai, G.: Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In: Proc. of SIGIR. p. 2749–2754. SIGIR 2024, ACM (Jul 2024). https://doi.org/10.1145/3626772.3657673, http://dx.doi.org/10.1145/3626772.3657673

Supplementary Material

6 Adversarial Image Examples

In Fig. 2 of the main paper, we showed a single adversarial example due to space limitations; in Fig. 5, we show additional attacked images with zoomed-in crops to highlight the difference in noise patterns between our method and the competing method GeoShield [24].

7 Internal Representation Analysis

In this section, we analyze the internal representations of the black-box target models (Img2Loc [51] and G3 [16]) by examining the nearest neighbors retrieved during the retrieval stage. This qualitative assessment allows us to visualize how adversarial perturbations displace the query image within the latent manifold. Let $\Phi(I)$ represent the image encoding. We define the intra-modal similarity between a query image I_q and a gallery image I_g as:

$$\text{sim}_{I2I}(I_q, I_g) = \langle \Phi(I_q), \Phi(I_g) \rangle \tag{7}$$

In Figures 6, 7, 8, 9, 10, 11, 12, and 13 we display the top-5 retrieved images for each attack scenario. In the baseline case of unattacked images, the visual consistency of the internal representations is evident; the top-5 nearest neighbors frequently correspond to the exact same landmark or geographic location as the query. Interestingly, this visual coherence largely persists for images attacked with GeoShield, suggesting that the perturbation is insufficient to fully decouple the image from its original semantic cluster. In contrast, for images attacked with our proposed method, we observe a significant representational shift. Even at a small perturbation budget of $\varepsilon = 2$, the retrieved neighbors diverge sharply from the query's original location, successfully pushing the representation into visually distant regions of the embedding space.

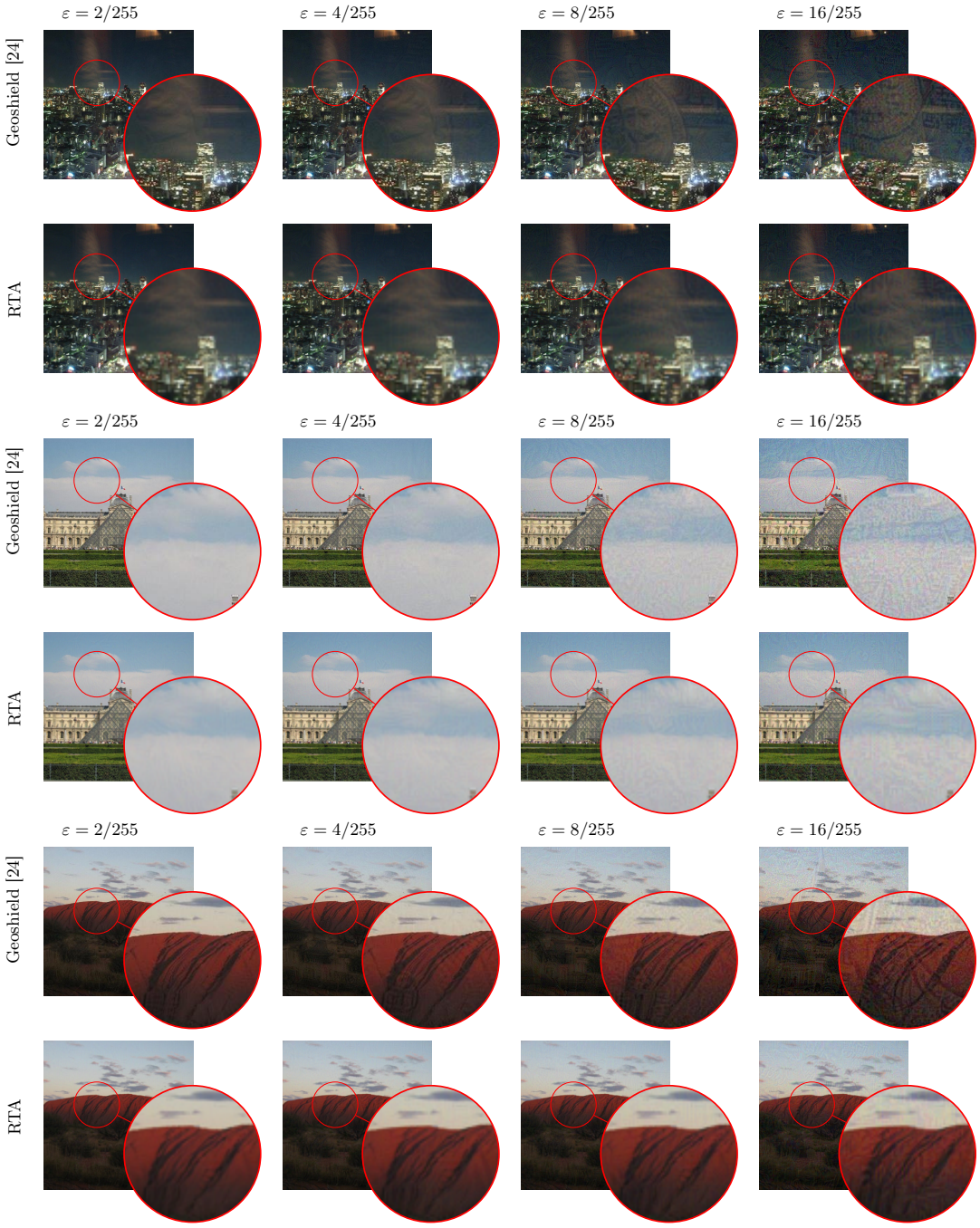


Fig. 5: Visual comparison of adversarial noise for four budgets $\varepsilon = 2/255, 4/255, 8/255, 16/255$ and two methods (Geoshield and ours). The magnified area highlights the fine-grained perturbation patterns for the attack. Best viewed in color on a screen.

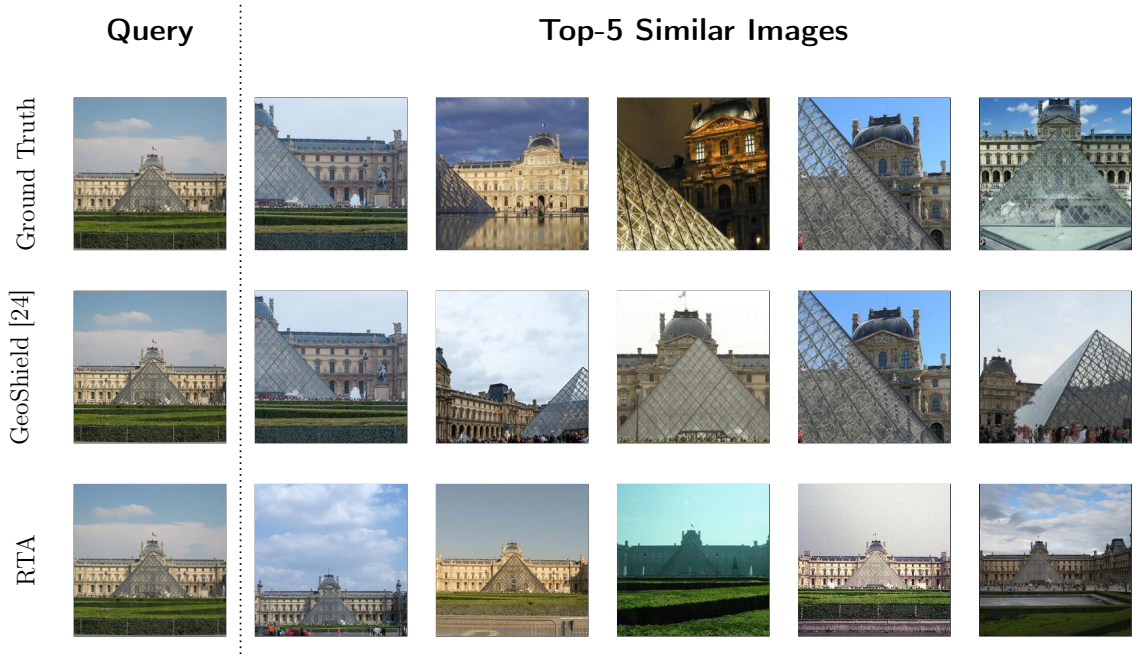


Fig. 6: Example of retrieved images by Img2Loc [51]. $\varepsilon = 2/255$

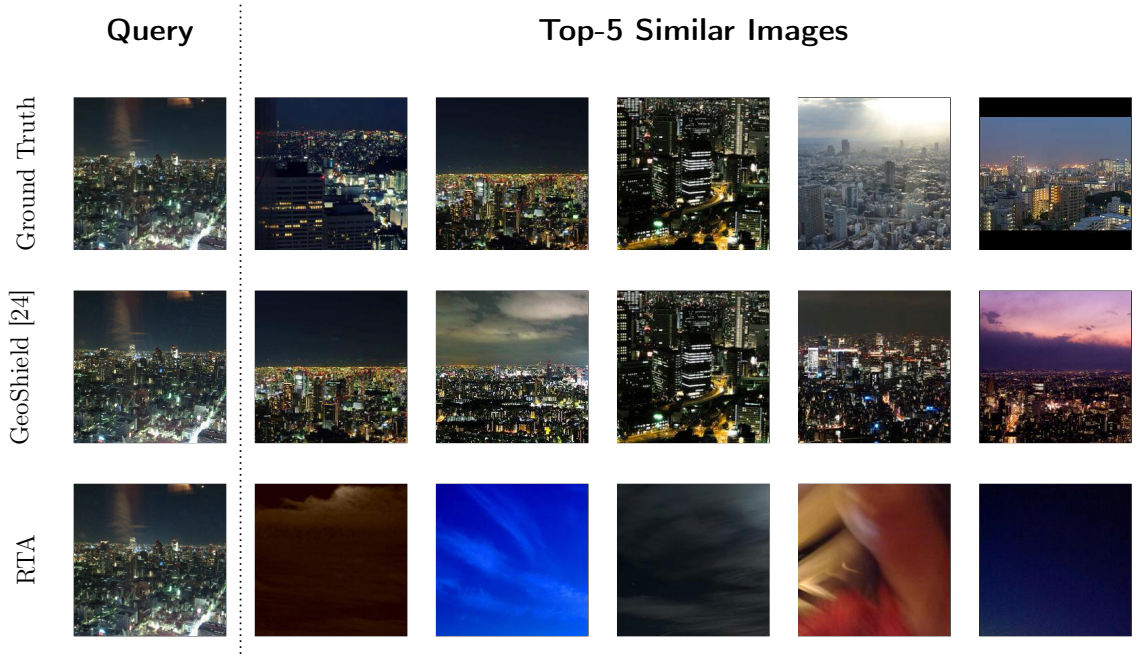


Fig. 7: Example of retrieved images by Img2Loc [51]. $\varepsilon = 4/255$

Fig. 8: Example of retrieved images by Img2Loc [51]. $\varepsilon = 8/255$

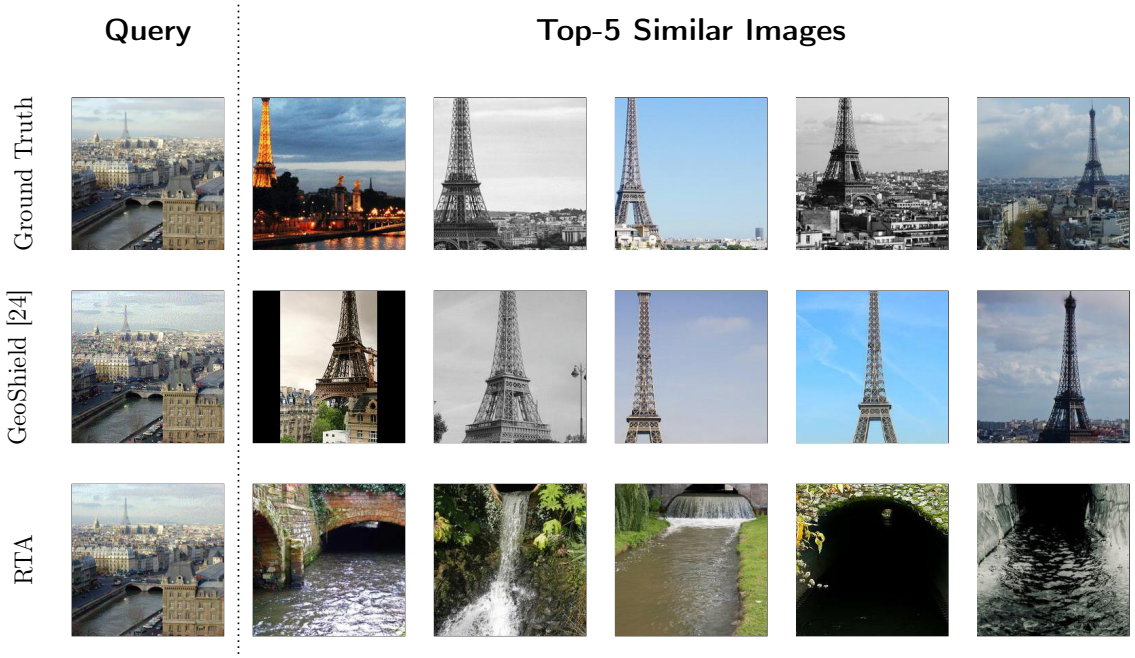


Fig. 9: Example of retrieved images by Img2Loc [51]. $\varepsilon = 16/255$

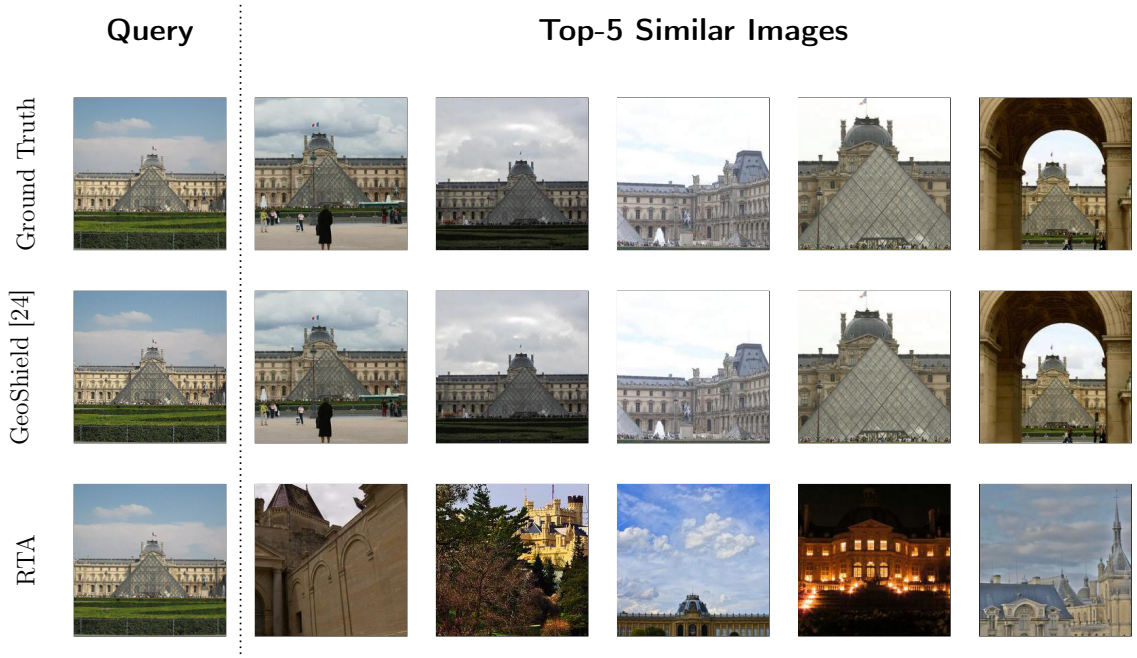


Fig. 10: Example of retrieved images by G3 [16]. $\varepsilon = 2/255$

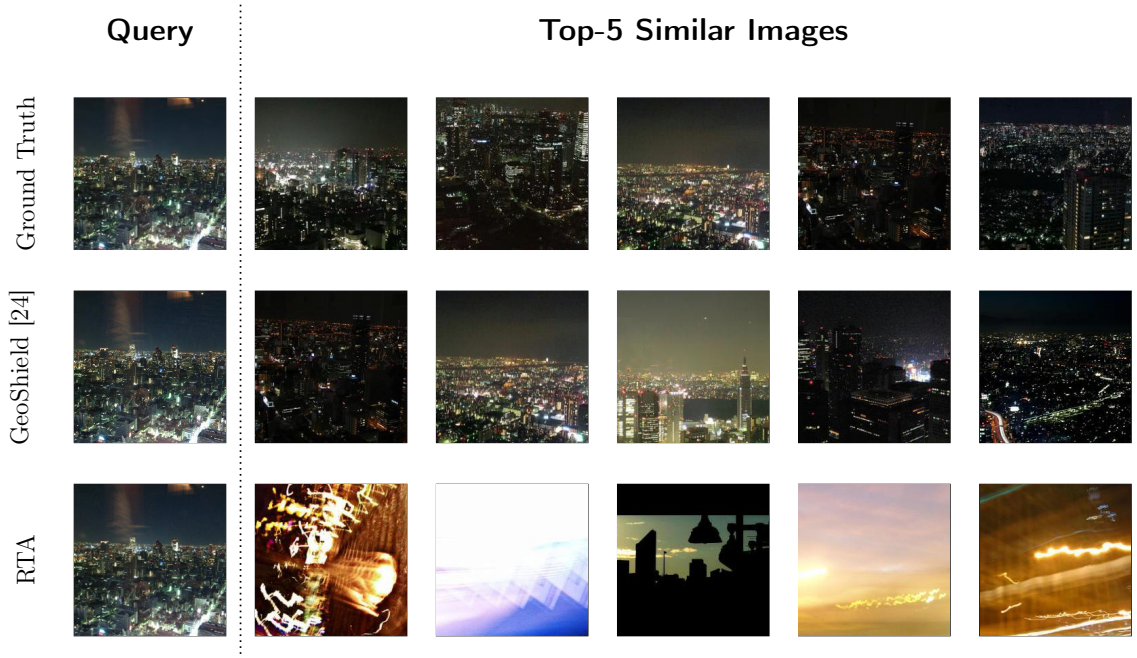


Fig. 11: Example of retrieved images by G3 [16]. $\varepsilon = 4/255$

Fig. 12: Example of retrieved images by G3 [16]. $\varepsilon = 8/255$

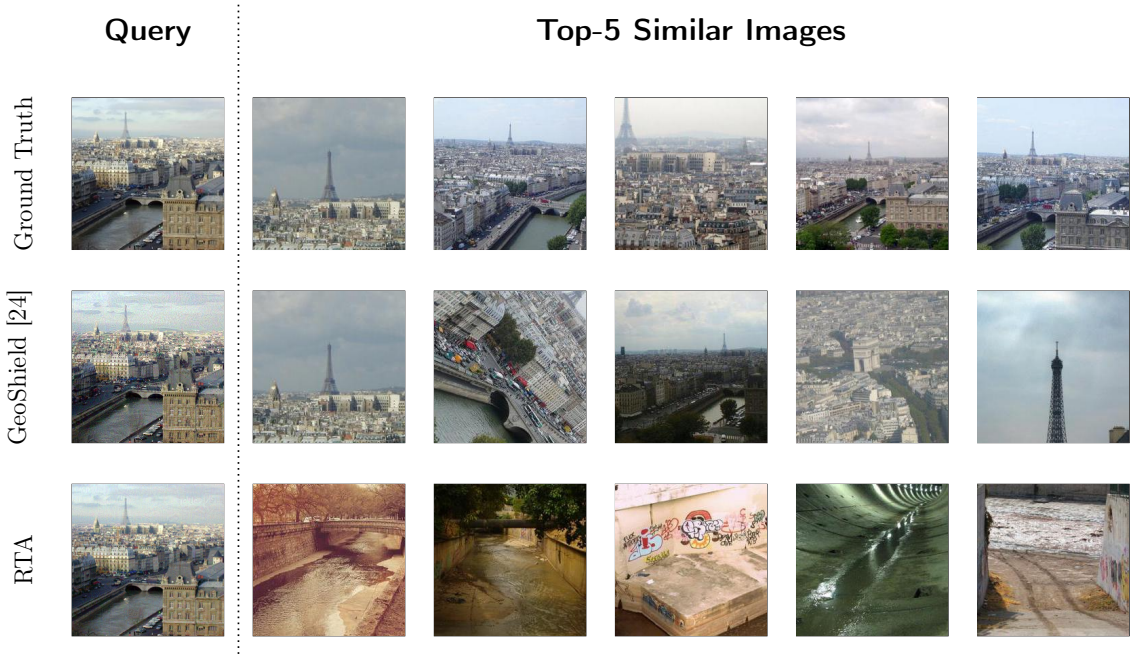


Fig. 13: Example of retrieved images by G3 [16]. $\varepsilon = 16/255$