Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (37.th cycle)

Towards Robust Visual Geo-Localization: Cross-Domain, Sequential, and Fine-Grained Approaches for Place Recognition

Gabriele Trivigno

* * * * *

Supervisors
Prof. Barbara Caputo
Prof. Carlo Masone

Politecnico di Torino
October X, 2025

Summary

Determining the geographic origin of an image, *i.e.* answering the fundamental question "Where was this picture taken?" constitutes a foundational challenge in computer vision and robotics with far-reaching implications. This capability to localize visual content across diverse environments underpins critical applications spanning from large-scale place recognition (with meter-level precision) to fine-grained visual localization (achieving centimeter-level accuracy). The advancement of these fields has been propelled by the proliferation of camera-equipped devices, the growing availability of geotagged imagery, and increasing demands for autonomous systems operating in unstructured environments. Applications range from consumer technologies like augmented reality and intelligent photo organization to robotic navigation for autonomous vehicles, assistive tools for the visually impaired, and large-scale geospatial analysis. Within this landscape, the literature has evolved to address two principal instantiations of the problem: Visual Place Recognition (VPR), which identifies coarse locations through image retrieval, and Visual Localization, which estimates precise 6-degrees-of-freedom camera poses, often within pre-built 3D maps.

This thesis makes several contributions advancing both paradigms. We first establish a comprehensive benchmarking framework for VPR, analyzing architectural choices, feature aggregation methods, and training strategies to derive practical insights for real-world deployment. Our findings reveal that lightweight CNNs often outperform complex models when optimized efficiently, while careful pipeline design can drastically reduce computational costs without sacrificing accuracy.

Recognizing the fundamental limitations of single-image retrieval approaches in Visual Place Recognition, we conduct a systematic investigation of sequence-based methods that leverage temporal information for improved localization robustness. While conventional VPR systems process frames independently, we demonstrate that sequential analysis of image streams yields significant performance gains, particularly in challenging scenarios affected by perceptual aliasing or viewpoint variations. To this end, we establish a comprehensive taxonomy of sequential descriptor architectures, analyzing their frame aggregation mechanisms and inherent design trade-offs, extending beyond traditional metrics to assess practical deployment factors including computational efficiency and memory requirements across diverse

datasets. Secondly, we introduce SeqVLAD, a novel sequential descriptor that holistically encodes multi-frame inputs through an innovative spatiotemporal aggregation layer. The proposed method addresses critical limitations of existing sequence matching approaches, mainly their quadratic complexity scaling and sensitivity to motion assumptions, while maintaining computational efficiency. We further demonstrate that local-feature based re-ranking, often overlooked in VPR, is decisive in overcoming domain shifts (e.g., day-night variations) and occlusions. To support our analysis, we introduce two challenging datasets that purposefully contain heavy domain shifts, and severe occlusion in the crowdsourced queries that we colledted. These new datasets (SF test-night and SF test-occlusions) remain to this day unsolved by state-of-the-art methods.

Furthermore, having underlined the main bottleneck in scalability of conventional paradigms, we challenge them by proposing a scalable classification framework for VPR that replaces contrastive learning with an Additive Angular Margin Classifier, enabling fast, database-size-agnostic inference while maintaining fine-grained precision.

Finally, for precise 6-DoF localization, we develop a training-free pose refiner that generalizes across scene representations (e.g., point clouds, neural radiance fields) by leveraging pre-trained deep features in a render-and-compare paradigm, achieving state-of-the-art accuracy without per-scene optimization.

Together, these contributions present a holistic advancement across the spectrum of image localization capabilities. The thesis systematically addresses various facets of the localization spectrum from single-image place recognition to sequence-based methods and finally precise 6-DoF pose estimation, establishing new state-of-the-art performance at each level of spatial granularity. By unifying innovations in representation learning, temporal modeling, and geometric verification within a coherent framework, this work provides both theoretical insights and practical solutions for real-world visual localization systems operating under varying precision requirements and environmental constraints.