Université Paris Cité

generative geolocalization
Interpreting PLONK: Geographic Knowledge in Vision Encoders

BOUKHARI Imed Eddine

19/01/2026

Table of Contents

Introduction & Motivation

Outline

Research Focus: Understanding geographic knowledge in PLONK's encoder

Two Experiments:

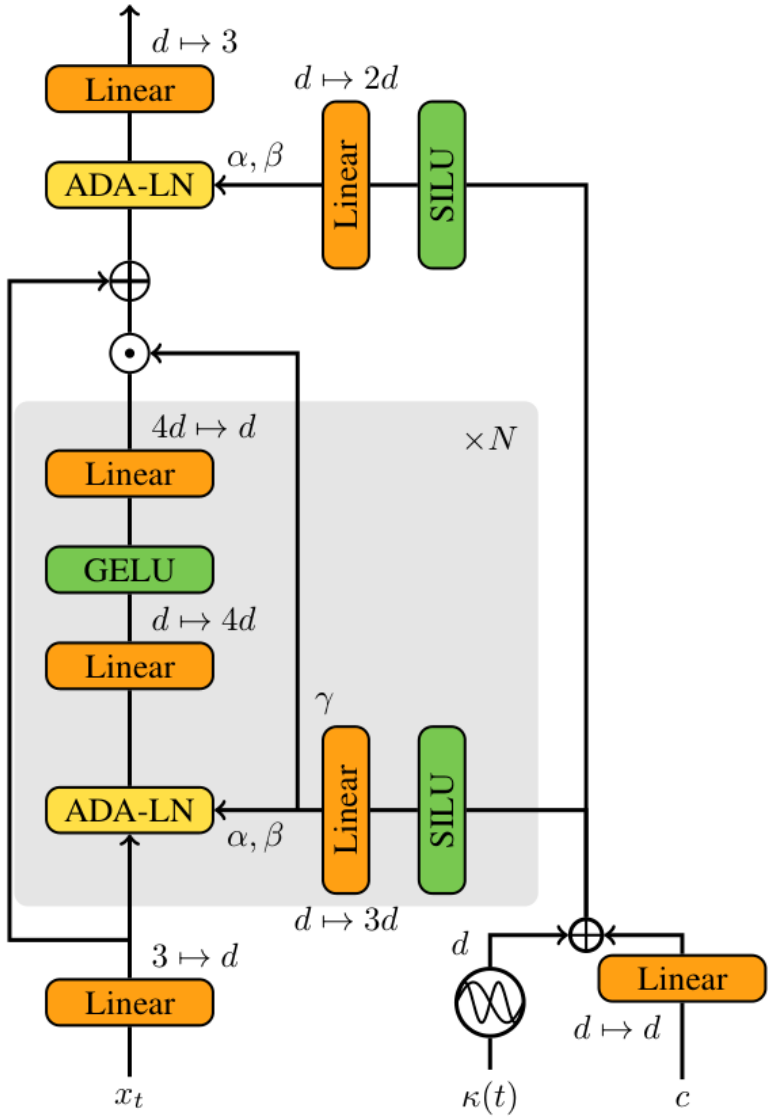
1. **Linear Probing**
   - How much does the frozen encoder already know?
   - Train simple classifier on StreetCLIP features
   - Test: Country, Region, City classification

2. **Attention Analysis**
   - What visual features does the model use?
   - Extract attention maps from 50,000 images
   - Analyze: Entropy, concentration, spatial patterns

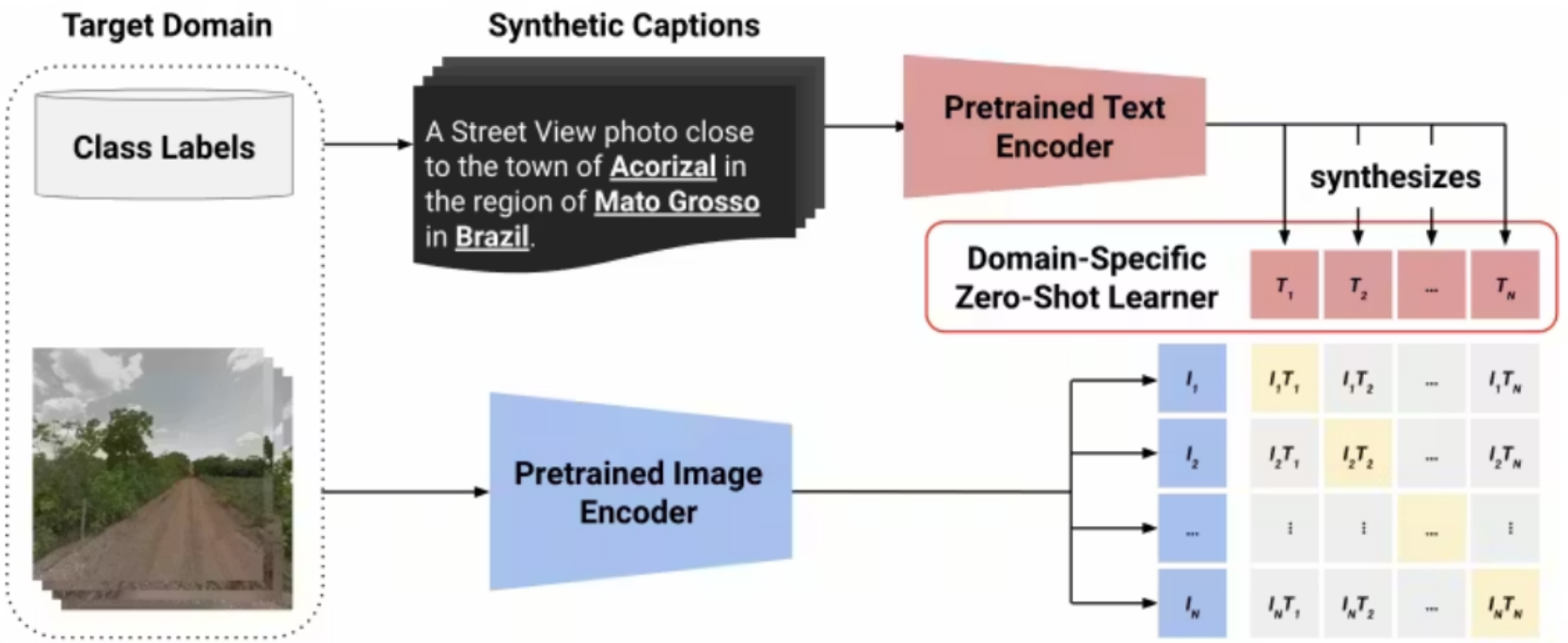
Goal: Decompose PLONK to understand where geographic knowledge comes from

PLONK Overview

PLONK: Generative Geolocation on Earth's Sphere



StreetCLIP: Vision Transformer Encoder



Linear Probing Experiments

Linear Probing: Where Does Geographic Knowledge Come From?

Question: Does StreetCLIP already know geography?

Method:

- Extract features from *frozen* StreetCLIP encoder
- Train simple linear classifier (logistic regression)
- Evaluate on three levels: Country, Region, City

If accuracy is high $\longrightarrow$ encoder already learned geography
If accuracy is low $\longrightarrow$ flow matching learns from scratch

Linear Probing: Dataset & Setup

Dataset: OSV-5M Test Set

- 50,000 images sampled from test set
- Geographic labels: Country, Region, City
- 217 countries, ∼2,050 regions, 7,292 cities

Method:

1. Extract 1024-dim features from frozen StreetCLIP encoder
2. Filter classes: Keep only labels with ≥ 2 samples
3. Train/test split: 80/20 (40k train, 10k test)
4. Train: Logistic Regression
5. Evaluate: Accuracy on held-out test set

Goal: Measure what encoder learned *before* flow matching

Linear Probing Results

Level	Linear Probe	PLONK (10000)	PLONK Paper	Classes
Country	**85.0%**	78%	76.2%	217
Region	**62%**	39%	44.2%	2,050
City	**8.2%**	6%	5.4%	7,292

Observations:

- Linear probe achieves strong accuracy on different levels
- Encoder captures substantial geographic knowledge from pre-training
- Different tasks: classification vs. coordinate regression

Understanding the Comparison

What We're Measuring:

- **Linear Probe:** Direct classification on frozen features
  - Shows: What encoder learned during pre-training
  - Baseline: How much knowledge before flow matching
- **PLONK Full Model:** GPS prediction $\rightarrow$ label lookup
  - Shows: End-to-end system performance
  - Reference: Published results from paper

Comparison helps quantify: How much geographic knowledge comes from the encoder vs. the flow matching process?

Attention analysis

Attention Analysis: Motivation

Question: What parts of images does StreetCLIP look at for geolocation?

Why This Matters:

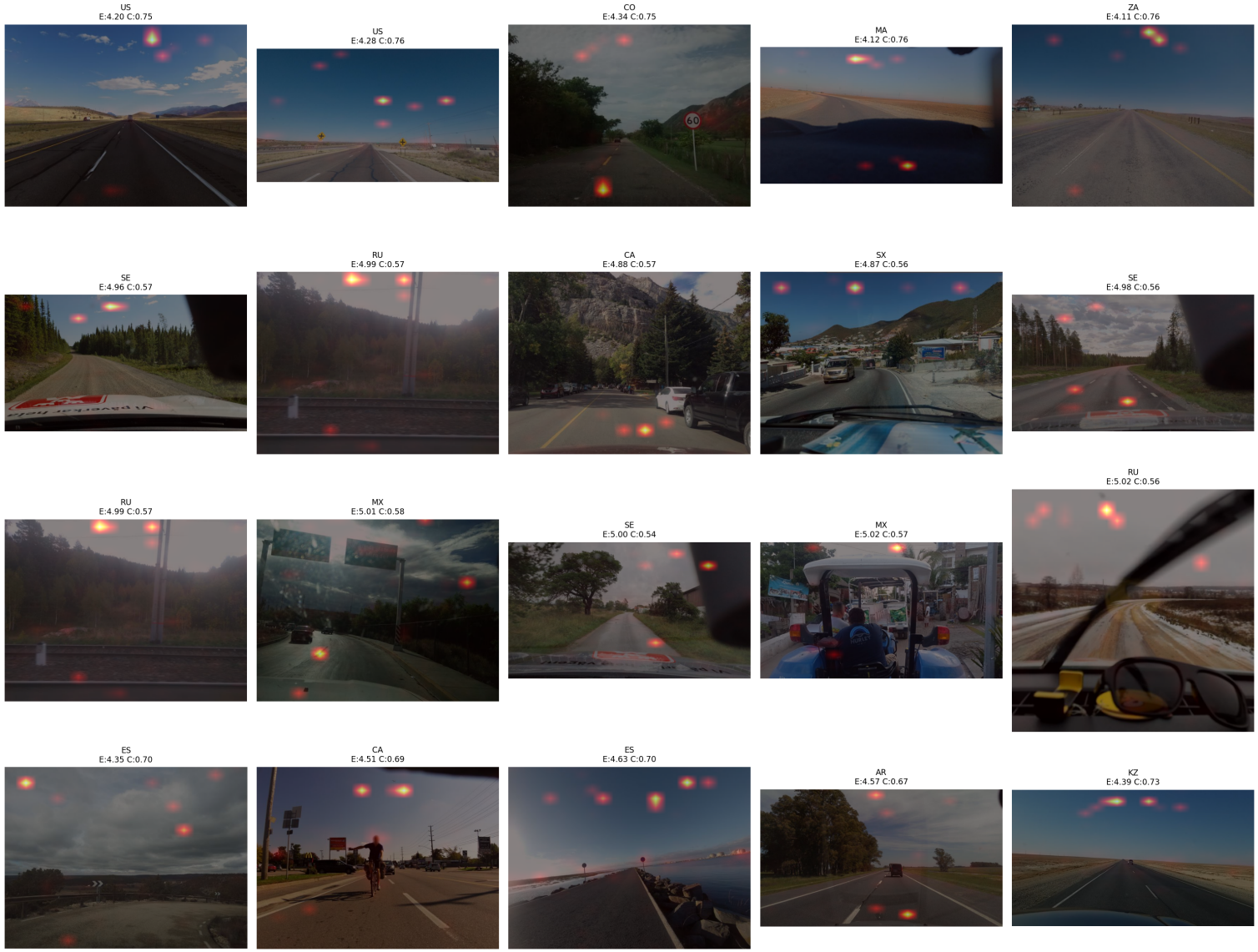
- Does it only focus on obvious landmarks?
- Does it use contextual cues (sky, vegetation)?
- Is attention consistent across images?

Approach:

1. Extract attention weights from last layer (50,000 images)
2. Visualize where model "looks" on 16×16 image patches
3. Compute statistics: spread, focus, patterns

Understanding attention helps explain how the encoder learns geography

Diverse examples



Attention Analysis Results

Metric	Value
Images Analyzed	**50,000**
Mean Entropy	**4.53**
Top-10% Concentration	**66.6%**
High Attention Patches	**95.2** (avg)

Interpretation:

- **Moderate entropy:** Model doesn't hyper-focus on single features
- **Selective attention:** 67% weight on top 10% of patches
- **Contextual coverage:** Uses 95 patches (37% of image)
- **Finding:** Model combines landmarks AND environmental cues (sky, vegetation, architectural patterns)

Discussion

PLONK's Two-Stage Architecture

Design Overview:

Stage 1: Encoder

- Pre-trained StreetCLIP
- Extracts visual features
- Captures geographic patterns
- Frozen during training

Stage 2: Flow Matching

- Maps features to coordinates
- Riemannian geometry (sphere)
- Generates probability distributions
- Trained end-to-end

Combining strong pre-trained encoder with geometric flow matching enables both high accuracy and uncertainty estimation

What Our Analysis Reveals

- Encoder already captures substantial geographic knowledge
- Strong performance
- Model attends to diverse visual features (not just landmarks)
- Uses environmental context: sky patterns, vegetation, architecture
- Attention distributed across 37% of image patches

Implication: The pre-trained encoder provides a strong foundation that flow matching can build upon for coordinate prediction

Conclusions

Summary

What We Did:

- Linear probing on frozen StreetCLIP features (country, region, city)
- Attention analysis on 50,000 images (1/4 of test set)

Key Findings:

- Pre-trained encoder contains substantial geographic knowledge
- Attention distributed across contextual features (sky, vegetation, environment)
- Model uses diverse visual cues, not just landmarks
- Attention patterns consistent across different countries
- Linear probing validates PLONK's frozen encoder design choice

Main Insight: Strong pre-trained encoder is critical for PLONK's performance

Future Directions

Potential Extensions:

- Replace frozen encoder with fine-tuned version and retrain flow matching
- Analyze attention patterns at different encoder layers
- Study encoder behavior on failure cases

Thank you for your attention!

Questions?