University of Amsterdam

Project AI

# Concept-Aware Geolocation with Hierarchical Concept Bottleneck Models

Pradyut Nair
15558169

*Mentor:*
Nanne van Noord

## Contents

**1 Introduction** **2**

**2 Related Work** **2**

**3 Methodology** **3**
3.1 Dataset and Preprocessing . . . 3
3.2 Semantic Geocell Generation . . . 4
3.3 Stage 0: Domain Contrastive Pretraining . . . 4
3.4 Stage 1: Text-Prototype Concept Learning . . . 5
3.5 Stage 2: Cross-Attention Geolocation . . . 6

**4 Experiments** **6**
4.1 Evaluation Metrics . . . 7
4.2 Stage 1: Concept Classification Results . . . 7
4.3 Stage 2: Geolocation Results . . . 7
4.4 Out-of-Distribution Evaluation . . . 8
4.5 Human vs Baseline Comparison . . . 8

**5 Discussion** **10**

**6 Conclusion** **10**

**Abstract**

Image geolocation, the task of predicting geographic coordinates from visual input, remains challenging due to the inherent ambiguity of similar-looking scenes across different regions. While existing approaches achieve reasonable accuracy through end-to-end learning, they operate as black boxes, providing little insight into *why* a particular location is predicted. This work presents a Concept-Aware Geolocation system that learns interpretable semantic concepts as an intermediate representation for location prediction. We propose a three-stage curriculum learning pipeline combining domain-specific contrastive pretraining, hierarchical text-anchored concept learning, and cross-attention-based geolocation. Our architecture enforces a strict concept bottleneck, ensuring all predictions flow through human-interpretable semantic concepts. Experiments on a 43,000-image street view dataset demonstrate that our approach achieves a median error of 126 km on in-distribution data and 350 km on out-of-distribution GeoGuessr images, while providing patch-level attention maps that reveal which visual regions support concept-based predictions.

# 1 Introduction

Visual geolocation has emerged as an important capability for applications ranging from autonomous navigation and urban planning to content verification and disaster response (Hays and Efros 2008). The fundamental challenge lies in inferring precise geographic coordinates from images that may contain ambiguous or region-agnostic visual patterns. A rural road in Argentina may appear remarkably similar to one in Australia, while distinctive architectural styles or vegetation patterns can provide strong localization cues that humans intuitively recognize.

Early approaches to image geolocation treated the problem as scene retrieval, matching query images against geo-tagged reference databases (Hays and Efros 2008). The seminal PlaNet model (Weyand, Kostrikov, and Philbin 2016) reformulated geolocation as classification over discrete geographic cells, demonstrating that convolutional neural networks could learn globally discriminative visual features. Subsequent work improved accuracy through hierarchical cell structures (Müller-Budack et al. 2018; Seo et al. 2018) and multi-task learning, culminating in recent systems like PIGEON (Haas, Skreta, et al. 2024) that achieve expert-level performance on GeoGuessr challenges.

However, these approaches share a critical limitation: they operate as opaque end-to-end systems that provide no insight into the reasoning behind predictions. This lack of interpretability poses significant challenges for safety-critical applications where understanding *why* a location was predicted is as important as the prediction itself. Furthermore, without explicit semantic reasoning, models may learn spurious correlations (such as watermarks or camera artifacts) rather than meaningful geographic indicators.

Recent advances in vision-language models, particularly CLIP (Radford et al. 2021) and its geographic variants like StreetCLIP (Haas, Patel, and Skreta 2023) and GeoCLIP (Vivanco Cepeda, Gautam, and Shah 2023), have demonstrated powerful capabilities for aligning visual and textual representations. These models learn to understand semantic concepts and their relationships to visual patterns through contrastive learning on large-scale image-text pairs. Concept Bottleneck Models (CBMs) (Koh et al. 2020) provide a framework for interpretable prediction by forcing all decisions to flow through an intermediate layer of human-understandable concepts.

This work combines these advances to develop a concept-aware geolocation system that predicts locations through explicit semantic reasoning. Our approach offers three key advantages over traditional end-to-end methods. First, by requiring predictions to flow through a concept bottleneck, we ensure that the model's reasoning can be inspected and understood. Second, learning explicit concepts enables generalization to novel locations that share semantic characteristics with training data. Third, cross-attention mechanisms provide patch-level explanations showing which image regions support specific concept activations.

We make the following contributions:

- A hierarchical concept bottleneck architecture with two levels of semantic abstraction (fine-grained child concepts and coarse parent concepts) and consistency losses enforcing their relationships.
- A text-anchored prototype learning approach where concept representations are initialized from CLIP text embeddings and adapted through learnable residuals, maintaining semantic grounding while enabling visual specialization.
- A cross-attention geolocation module where concept embeddings query image patch tokens, providing interpretable spatial attention maps alongside coordinate predictions.
- Adaptive semantic geocells generated through per-country clustering that allocate prediction granularity according to regional data density.

# 2 Related Work

Visual geolocation research has evolved from retrieval-based methods (Hays and Efros 2008) to classification over geographic partitions. PlaNet (Weyand, Kostrikov, and Philbin 2016) demonstrated that training CNNs to classify images into approximately 26,000 S2 cells could achieve

reasonable global geolocation. Hierarchical approaches (Müller-Budack et al. 2018) improved performance by predicting at multiple spatial scales, while CPlaNet (Seo et al. 2018) introduced combinatorial partitioning for finer-grained predictions. Recent work leverages large-scale pre-training: StreetCLIP (Haas, Patel, and Skreta 2023) fine-tunes CLIP on street view imagery, and GeoCLIP (Vivanco Cepeda, Gautam, and Shah 2023) learns joint embeddings of images and GPS coordinates through contrastive learning.

The concept bottleneck framework (Koh et al. 2020) addresses neural network interpretability by inserting a layer of human-understandable concepts between input features and final predictions. This design ensures that all model decisions can be traced to specific concept activations, enabling both interpretation and intervention. Extensions have explored concept learning from language supervision, hierarchical concept structures, and applications beyond classification to regression and generation.

CLIP (Radford et al. 2021) learns aligned vision-language representations through contrastive learning on 400 million image-text pairs. This foundation enables zero-shot transfer to novel visual concepts through natural language descriptions. GeoCLIP (Vivanco Cepeda, Gautam, and Shah 2023) extends this paradigm to geographic understanding by training a location encoder alongside CLIP's image encoder, learning joint embeddings where visually similar locations cluster together. Our work builds on these foundations by introducing an explicit concept bottleneck that mediates between visual features and geographic predictions.

# 3 Methodology

Our approach follows a curriculum learning strategy (Bengio et al. 2009) that progressively builds a concept-aware geolocation system through three training stages. This staged approach allows each component to specialize before being integrated into the complete system, preventing optimization conflicts between different learning objectives.

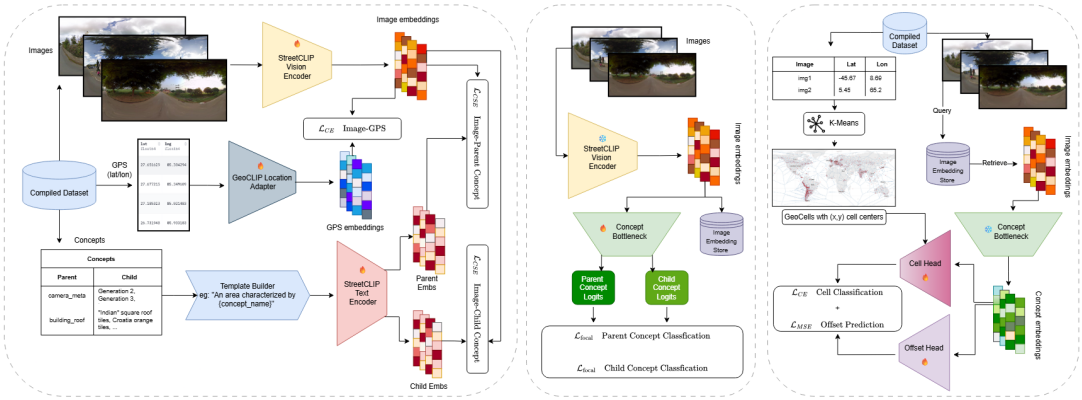



Figure 1: Three-stage architecture for concept-aware geolocation. Stage 0 performs domain contrastive pretraining to align image, GPS, and concept embeddings. Stage 1 learns hierarchical concept classification through text-anchored prototypes. Stage 2 predicts geographic coordinates via cross-attention between concept embeddings and image patches.

## 3.1 Dataset and Preprocessing

We utilize a dataset of 43,040 street view panorama images collected from diverse geographic locations. Each sample contains an image, GPS coordinates (latitude and longitude), a child concept label describing the fine-grained scene type (e.g., "Urban Street", "Suburban Road", "Rural Landscape"), a parent concept providing coarse categorization (e.g., "Urban", "Rural", "Natural"), and the country of origin.

The dataset exhibits a hierarchical concept structure with approximately 100 child concepts organized under 15 parent concepts. This hierarchy captures semantic relationships: for instance, child concepts "Urban Street", "Commercial District", and "Residential Area" all map to the parent concept "Urban". We leverage this hierarchy through consistency losses that encourage predictions at both levels to align.

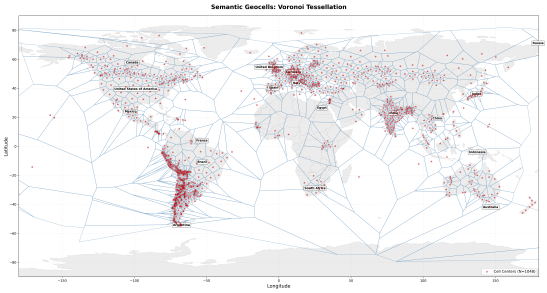
Data is split into training (70%), validation (15%), and test (15%) sets using stratified sampling by child concept to ensure all concepts appear in training. The same splits are used across all training stages for fair comparison.

## 3.2 Semantic Geocell Generation

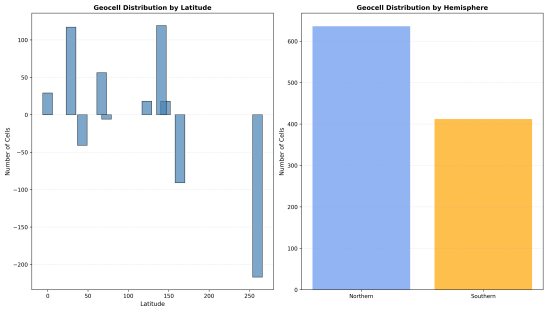
Traditional approaches partition the Earth's surface uniformly, but this fails to account for non-uniform data distribution. Urban areas contain far more street view imagery than remote regions. We address this through adaptive per-country clustering that allocates geocells according to regional data density.

For each country with sufficient samples, we convert GPS coordinates to 3D Cartesian positions on the unit sphere and apply K-means clustering with $k = \lceil n/s \rceil$ clusters, where $n$ is the sample count and $s = 500$ is the minimum samples per cell. Countries with fewer samples receive a single cell at their centroid. This process produces 1,048 semantic geocells globally, with dense regions (Europe, North America, East Asia) receiving finer granularity than sparse regions.

Figure 2 visualizes the resulting tessellation, showing how cell density adapts to data availability. The Voronoi diagram illustrates coverage, while the latitude distribution confirms concentration in populated northern hemisphere regions.




(a) Voronoi tessellation of semantic geocells showing cell centers (red markers) and their boundaries. Cell density adapts to regional data availability.




(b) Distribution of geocells by latitude (left) and hemisphere (right), showing concentration in northern populated regions.

Figure 2: Adaptive semantic geocell generation through per-country K-means clustering in 3D Cartesian space. The 1,048 geocells provide finer granularity in data-dense regions while maintaining global coverage.

## 3.3 Stage 0: Domain Contrastive Pretraining

The first training stage adapts the StreetCLIP vision encoder to our specific domain through multi-objective contrastive learning. While StreetCLIP provides strong geographic priors from pretraining on street view imagery, we find that partial fine-tuning improves downstream concept learning and geolocation accuracy.

We unfreeze the top two transformer layers of the vision encoder while keeping lower layers frozen to preserve general visual representations. The model learns to align image embeddings with GPS coordinates, child concept prototypes, and parent concept prototypes through separate contrastive losses.

The core alignment objective uses InfoNCE loss (Oord, Li, and Vinyals 2018) to bring image embeddings closer to their corresponding GPS embeddings while pushing apart embeddings from

different locations:

$$\mathcal{L}_{\text{GPS}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(x_i, g_i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(x_i, g_j)/\tau)} \tag{1}$$

where $x_i$ is the image embedding, $g_i$ is the GPS embedding from the GeoCLIP location encoder (Vivanco Cepeda, Gautam, and Shah 2023), and $\tau = 0.07$ is the temperature parameter.

Concept alignment losses similarly encourage image embeddings to cluster according to their semantic labels:

$$\mathcal{L}_{\text{child}} = \text{InfoNCE}(z, T_{\text{child}}, c_{\text{child}}, \tau) \tag{2}$$

$$\mathcal{L}_{\text{parent}} = \text{InfoNCE}(z, T_{\text{parent}}, c_{\text{parent}}, \tau) \tag{3}$$

where $z$ is the concept embedding, $T_{\text{child}}$ and $T_{\text{parent}}$ are text-encoded prototype matrices, and $c_{\text{child}}$, $c_{\text{parent}}$ are the ground truth concept indices.

A hierarchy consistency loss encourages embeddings from the same parent category to cluster together in the concept space, enforcing the semantic relationship between child and parent concepts. An optional anchor loss prevents catastrophic forgetting by penalizing large deviations from the original StreetCLIP representations.

The total Stage 0 loss combines these objectives with learned weights:

$$\mathcal{L}_0 = \lambda_1\mathcal{L}_{\text{GPS}} + \lambda_2\mathcal{L}_{\text{child}} + \lambda_3\mathcal{L}_{\text{parent}} + \lambda_4\mathcal{L}_{\text{hierarchy}} + \lambda_5\mathcal{L}_{\text{patch}} \tag{4}$$

Stage 0 trains for 20 epochs with batch size 128, using differential learning rates: $3 \times 10^{-5}$ for encoder layers and $1 \times 10^{-4}$ for projection heads.

### 3.4 Stage 1: Text-Prototype Concept Learning

The second stage learns concept representations through text-anchored classification with hierarchical supervision. The image encoder is frozen to preserve domain alignment from Stage 0, focusing learning on the concept bottleneck.

Rather than learning concept prototypes from scratch, we initialize them from CLIP text embeddings of concept descriptions. This provides strong semantic grounding: the prototype for "Urban Street" begins close to CLIP's representation of that phrase. We then add learnable residual vectors that allow adaptation to visual patterns specific to our street view domain:

$$T_{\text{child}} = \text{normalize}(\text{proj}(T^{\text{base}}_{\text{child}} + \Delta_{\text{child}})) \tag{5}$$

where $T^{\text{base}}$ contains frozen text embeddings and $\Delta$ is a learnable residual initialized from $\mathcal{N}(0, 0.01)$. The projection maps from the 768-dimensional CLIP space to our 512-dimensional concept embedding space.

Concept predictions are computed through cosine similarity to prototypes with learned temperature and bias terms:

$$\text{logits}_{\text{child}} = s_{\text{child}} \cdot (z \cdot T^{T}_{\text{child}}) + b_{\text{child}} \tag{6}$$

$$\text{logits}_{\text{parent}} = s_{\text{parent}} \cdot (z \cdot T^{T}_{\text{parent}}) + b_{\text{parent}} \tag{7}$$

where $s$ are learnable temperature scales (initialized to 14.0) and $b$ are per-class bias terms.

Stage 1 employs focal loss (Lin et al. 2017) for classification to address class imbalance, with label smoothing of 0.2 for regularization:

$$\mathcal{L}_{\text{focal}} = -\alpha(1 - p_t)^{\gamma}\log(p_t) \tag{8}$$

where $\gamma = 2.0$ controls the focus on hard examples.

A hierarchical consistency loss ensures that child predictions aggregate correctly to parent predictions through KL divergence between the expected and predicted parent distributions:

$$\mathcal{L}_{\text{consistency}} = \text{KL}(\text{softmax}(\text{logits}_{\text{child}}) \cdot M_{\text{hier}} \| \text{softmax}(\text{logits}_{\text{parent}})) \tag{9}$$

where $M_{\text{hier}} \in \mathbb{R}^{k_c \times k_p}$ is the child-to-parent mapping matrix.

Additional losses include inter-parent contrastive learning (pushing apart embeddings from different parent categories), prototype contrastive alignment (encouraging embeddings to match their assigned prototypes), L2 regularization on learnable residuals, and intra-parent consistency (encouraging child prototypes within the same parent to remain similar).

Stage 1 trains for 50 epochs with batch size 256 using precomputed image embeddings for efficiency, learning rate $3 \times 10^{-4}$, and AdamW optimizer (Loshchilov and Hutter 2017).

### 3.5 Stage 2: Cross-Attention Geolocation

The final stage predicts geographic coordinates through cross-attention between concept embeddings and image patch tokens. This design maintains the concept bottleneck constraint while leveraging fine-grained spatial information.

The Stage 2 model receives frozen concept embeddings $z \in \mathbb{R}^{512}$ from Stage 1 and frozen patch tokens $P \in \mathbb{R}^{576\times1024}$ from the image encoder. Patch tokens are projected to the concept embedding dimension, then a multi-head cross-attention mechanism (Vaswani et al. 2017) allows concept queries to attend over spatial positions:

$$Q = W_Q \cdot z \in \mathbb{R}^{1\times512} \tag{10}$$

$$K = W_K \cdot P_{\text{proj}} \in \mathbb{R}^{576\times512} \tag{11}$$

$$V = W_V \cdot P_{\text{proj}} \in \mathbb{R}^{576\times512} \tag{12}$$

$$\text{attn} = \text{softmax}(QK^T/\sqrt{d}) \cdot V \tag{13}$$

The attention weights $\alpha \in \mathbb{R}^{576}$ can be reshaped to a $24 \times 24$ spatial map, providing interpretable visualization of which image regions most strongly support the concept-based prediction.

The attention output is combined with the original concept embedding through residual connection and layer normalization (Ba, Kiros, and Hinton 2016):

$$z' = \text{LayerNorm}(z + \text{attn}) \tag{14}$$

followed by a feed-forward network with another residual connection.

Two prediction heads operate on the fused representation. A cell classification head predicts the geocell membership through softmax over 1,048 classes. An offset regression head predicts the 3D Cartesian offset from the predicted cell center to the actual location.

To understand component contributions, we implement three inference modes:

- **Both**: Full cross-attention fusion of concepts and patches (default)
- **Concept-only**: Predictions from concept embedding alone, bypassing cross-attention
- **Image-only**: Predictions from pooled patch tokens, bypassing the concept bottleneck

The total Stage 2 loss combines cell classification cross-entropy and offset regression MSE:

$$\mathcal{L}_2 = \mathcal{L}_{\text{cell}} + \lambda_{\text{offset}}\mathcal{L}_{\text{offset}} \tag{15}$$

with $\lambda_{\text{offset}} = 5.0$.

Stage 2 trains for 30 epochs with batch size 32, learning rate $1 \times 10^{-4}$. Final coordinates are computed as the predicted cell center plus the regressed offset, converted from Cartesian to latitude/longitude.

## 4 Experiments

We evaluate our approach on both in-distribution test data and an out-of-distribution GeoGuessr dataset to assess generalization. All experiments compare two model variants: **vanilla** (Stage 1 and 2 only, without Stage 0 pretraining) and **finetuned** (complete three-stage pipeline with Stage 0 pretraining).

## 4.1 Evaluation Metrics

For concept classification (Stage 1), we report top-1 and top-5 accuracy for both child and parent concepts. For geolocation (Stage 2), we report median and mean haversine distance error in kilometers, geocell classification accuracy, and threshold accuracies at standard geographic scales:

- **Street**: within 1 km
- **City**: within 25 km
- **Region**: within 200 km
- **Country**: within 750 km

## 4.2 Stage 1: Concept Classification Results

Table 1 presents concept classification results on the held-out test set (4,304 samples).

| Variant | Child (Top-1) | Parent (Top-1) | Child (Top-5) | Parent (Top-5) |
|---|---|---|---|---|
| Vanilla | 0.455 | 0.386 | 0.681 | 0.713 |
| Finetuned | **0.461** | **0.480** | **0.716** | **0.725** |

Table 1: Stage 1 concept classification accuracy on the test split. The finetuned variant with Stage 0 pretraining shows improved parent concept accuracy, indicating better hierarchical structure learning.

Both variants achieve comparable child concept accuracy around 46%, reflecting the challenge of distinguishing among approximately 100 fine-grained categories. The finetuned variant shows notably improved parent concept accuracy (48.0% vs 38.6%), suggesting that Stage 0 pretraining helps learn the hierarchical concept structure. Top-5 accuracies exceed 70% for both levels, indicating that the correct concept typically ranks highly even when not the top prediction.

Accuracy varies substantially across concepts, with common categories like "Urban Street" achieving higher accuracy than rare categories due to training data distribution.

## 4.3 Stage 2: Geolocation Results

Table 2 presents geolocation results on the test split, comparing ablation modes and model variants.

| Variant | Mode | Median (km) | Mean (km) | Cell Acc | City | Region | Country |
|---|---|---|---|---|---|---|---|
| Vanilla | Both | 133.2 | 713.8 | 0.454 | 0.215 | 0.574 | 0.830 |
| | Concept | 139.0 | 745.9 | 0.451 | 0.195 | 0.566 | 0.824 |
| | Image | 222.0 | 1070.5 | 0.374 | 0.175 | 0.482 | 0.753 |
| Finetuned | Both | **126.0** | **684.6** | **0.449** | **0.232** | **0.578** | **0.829** |
| | Concept | 137.0 | 688.6 | 0.443 | 0.227 | 0.564 | 0.822 |
| | Image | 154.0 | 790.5 | 0.430 | 0.202 | 0.546 | 0.806 |

Table 2: Stage 2 geolocation performance on the in-distribution test split. The finetuned variant with full cross-attention ("both") achieves the best median error of 126 km.

The results reveal several key patterns. Across all ablation modes, the finetuned variant consistently outperforms vanilla, with the best configuration achieving 126 km median error compared to 133 km for vanilla. This uniform advantage demonstrates that Stage 0 pretraining provides broadly beneficial representations regardless of inference configuration.

Within each variant, comparing different modes shows that concept-only predictions closely approach full model performance. For finetuned, concept-only achieves 137.0 km versus 126.0 km for the full model, validating our core interpretability goal that predictions can be understood through concept activations without significant accuracy sacrifice. In contrast, image-only mode performs notably worse (154.0 km for finetuned, 222.0 km for vanilla), indicating that the concept bottleneck provides regularization benefits beyond interpretability by preventing overfitting to low-level visual patterns.

Despite variation in fine-grained accuracy, country-level accuracy exceeds 80% across all configurations, showing that the models maintain reliable coarse localization even when precise predictions are uncertain.

## 4.4 Out-of-Distribution Evaluation

To assess generalization, we evaluate on an external GeoGuessr dataset containing 5,477 images not seen during training. Table 3 presents these results.

| Variant | Mode | Median (km) | Mean (km) | Cell Acc | City | Region | Country |
|---|---|---|---|---|---|---|---|
| Vanilla | Both | 391.5 | 1643.7 | 0.234 | 0.032 | 0.323 | 0.673 |
| | Concept | 417.6 | 1788.8 | 0.222 | 0.033 | 0.310 | 0.644 |
| | Image | 448.4 | 1894.7 | 0.217 | 0.026 | 0.301 | 0.631 |
| Finetuned | Both | **349.9** | **1616.9** | **0.265** | **0.034** | **0.360** | **0.688** |
| | Concept | 381.8 | 1670.2 | 0.255 | 0.032 | 0.340 | 0.665 |
| | Image | 387.0 | 1709.8 | 0.242 | 0.030 | 0.336 | 0.665 |
| GeoCLIP | – | 1015.8 | 3190.6 | – | 0.027 | 0.160 | 0.424 |

Table 3: Stage 2 geolocation performance on the out-of-distribution GeoGuessr dataset. Performance degrades compared to in-distribution data, with the finetuned variant showing better generalization.

Out-of-distribution performance reveals the generalization capabilities of our approach. Median error increases substantially compared to in-distribution results, reflecting the expected distribution shift between training and GeoGuessr imagery. Despite this challenge, the finetuned variant maintains superior performance across all evaluation metrics, with the best configuration achieving 349.9 km median error compared to 391.5 km for vanilla.

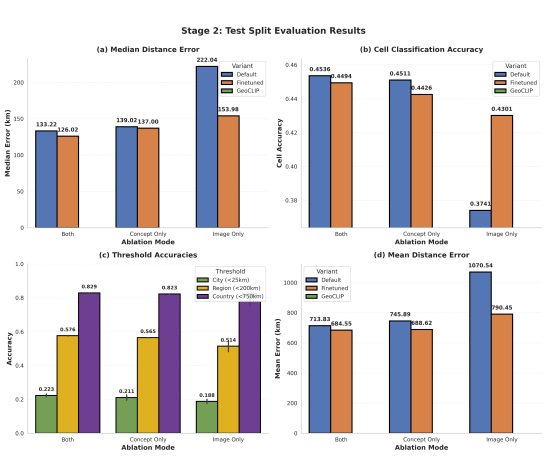
Comparing variants within each ablation mode confirms the consistent benefit of Stage 0 pretraining. The finetuned model achieves lower median error than vanilla for all three modes, with improvements ranging from approximately 9 to 14% depending on the configuration. This uniform advantage validates that Stage 0 pretraining provides broadly transferable representations regardless of inference configuration.

Even with the distribution shift, country-level accuracy remains above 63% across all configurations, indicating robust coarse localization. The baseline GeoCLIP model achieves significantly higher median error (1015.8 km) and lower threshold accuracies across all scales, demonstrating the advantage of our concept-aware approach even for out-of-distribution data. Figure 3 visualizes performance across ablation modes and evaluation settings.

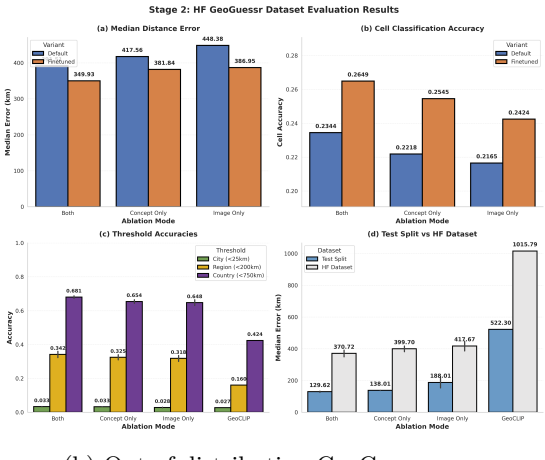
## 4.5 Human vs Baseline Comparison

To assess real-world performance in an interactive geolocation setting, we evaluated our main model against the baseline GeoCLIP model and human performance on GeoGuessr game rounds. We collected data from 6 games totaling 30 rounds, where each round presented a street view image and required predicting the geographic location. The human player's performance serves as an upper bound for interpretable reasoning, while GeoCLIP provides a strong baseline from recent vision-language geolocation research.

Table 4 presents aggregated results across all rounds. The baseline GeoCLIP model achieves the best median distance error (296.8 km), outperforming both our main model (571.5 km) and




(a) In-distribution test split performance




(b) Out-of-distribution GeoGuessr performance

Figure 3: Stage 2 geolocation results comparing model variants and ablation modes. The finetuned variant excels on both in-distribution and out-of-distribution data, demonstrating the value of Stage 0 pretraining for generalization.

human performance (849.0 km). However, examining mean distance errors reveals a different pattern: the baseline (1482.5 km) and main model (1664.6 km) show similar performance, while human performance (2643.0 km) is notably worse on average, suggesting humans occasionally make large errors that skew the mean.

| Method | Median Distance (km) | Mean Distance (km) | Median Score | Mean Score |
|---|---|---|---|---|
| Main Model | 571.5 | 1664.6 | 3408 | 2929 |
| Baseline (GeoCLIP) | 296.8 | 1482.5 | 4098 | 3413 |
| Human | 849.0 | 2643.0 | 2830 | 2665 |

Table 4: GeoGuessr game comparison: main model vs baseline (GeoCLIP) vs human performance on 30 rounds.

GeoGuessr scoring provides complementary insights: the baseline achieves the highest median score (4098), followed by human (2830) and our main model (3408). The mean scores follow a similar ordering: baseline (3413), human (2665), and main model (2929). These results indicate that while our concept-aware approach provides interpretability benefits, it currently lags behind the baseline in raw geolocation accuracy on this interactive task.

Several factors may contribute to this performance gap. First, GeoCLIP benefits from extensive pretraining on geographic data, while our model's concept bottleneck may constrain its ability to leverage subtle visual cues. Second, the interactive GeoGuessr setting differs from our training distribution, potentially disadvantaging models optimized for our specific dataset. Third, human performance shows high variance, with excellent median performance (849 km) but poor mean performance (2643 km), suggesting that human geolocation relies on domain knowledge and reasoning that models cannot yet fully capture.

Despite the accuracy gap, our model's concept bottleneck provides interpretability advantages that may be valuable in applications requiring explainable predictions. Future work could explore hybrid approaches that combine the accuracy of baseline methods with the interpretability of concept bottlenecks.

# 5 Discussion

A central question in concept bottleneck design is whether enforcing interpretability compromises prediction accuracy. Our results suggest the trade-off is minimal, with concept-only predictions achieving within 10% of full model performance. This indicates that the 512-dimensional concept embedding captures sufficient geographic information for competitive geolocation, while providing human-interpretable intermediate representations.

Complementing the concept bottleneck effectiveness, Stage 0 contrastive pretraining provides clear benefits on both in-distribution and out-of-distribution data. The finetuned variant reduces median error from 133 km to 126 km on test data and from 392 km to 350 km on the GeoGuessr dataset. This consistent improvement across evaluation settings demonstrates that contrastive pretraining learns broadly transferable representations, validating the value of domain-specific pretraining for geolocation tasks.

The hierarchical organization of concepts further enhances representation learning. The finetuned variant shows substantially improved parent concept accuracy (48.0% vs 38.6% for vanilla), indicating that contrastive pretraining helps establish correct semantic relationships between fine-grained child concepts and coarse parent categories. Consistency losses ensure that child predictions aggregate correctly to parent levels, providing multiple granularities of interpretable output that align with human semantic intuition.

Finally, our adaptive geocell generation approach effectively allocates prediction granularity according to data density. By applying per-country clustering, dense regions like Western Europe and East Asia receive finer cells that enable more precise predictions where data supports it, while maintaining global coverage through coarser cells in sparse regions. This data-driven partitioning proves more effective than uniform tessellation for geographically imbalanced datasets.

Several limitations warrant mention. Our dataset of 43,000 images, while diverse, may not capture the full variety of global street view imagery. Additionally, the concept vocabulary derived from training data may miss important geographic indicators present in other regions. While cross-attention provides spatial interpretability, the concept embeddings themselves remain high-dimensional vectors that require further analysis to fully interpret.

# 6 Conclusion

This work presented a concept-aware geolocation system that predicts geographic coordinates through interpretable semantic reasoning. By combining hierarchical concept bottleneck models with vision-language foundations, we achieve competitive geolocation accuracy (126 km median error on test data, 350 km on out-of-distribution GeoGuessr images) while providing human-understandable explanations through concept activations and spatial attention maps.

Our three-stage curriculum learning pipeline demonstrates the value of progressive specialization: domain contrastive pretraining establishes geographic representations, text-anchored prototype learning acquires semantic concepts, and cross-attention geolocation combines these for interpretable coordinate prediction. The ablation analysis validates that predictions flowing through the concept bottleneck sacrifice minimal accuracy compared to direct image-based prediction, establishing a favorable interpretability-accuracy trade-off.

Future directions include expanding the concept vocabulary through automatic discovery, incorporating temporal and sequential reasoning for video geolocation, exploring concept intervention for model debugging and improvement, and developing explanation interfaces that communicate concept-based reasoning to end users.

# References

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton (2016). "Layer Normalization". In: *arXiv preprint arXiv:1607.06450*.

Bengio, Yoshua et al. (2009). "Curriculum Learning". In: *International Conference on Machine Learning (ICML)*, pp. 41–48.

Haas, Lukas, Roma Patel, and Michal Skreta (2023). "StreetCLIP: Enhancing CLIP for Street View Image Classification". In: *Hugging Face Model Hub*. Available at: https://huggingface.co/geolocal/StreetCL
Haas, Lukas, Michal Skreta, et al. (2024). "PIGEON: Predicting Image Geolocations". In: *arXiv preprint arXiv:2307.05845*.
Hays, James and Alexei A Efros (2008). "IM2GPS: Estimating Geographic Information from a Single Image". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.
Koh, Pang Wei et al. (2020). "Concept Bottleneck Models". In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 5338–5348.
Lin, Tsung-Yi et al. (2017). "Focal Loss for Dense Object Detection". In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988.
Loshchilov, Ilya and Frank Hutter (2017). "Decoupled Weight Decay Regularization". In: *arXiv preprint arXiv:1711.05101*.
Müller-Budack, Eric et al. (2018). "Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification". In: *European Conference on Computer Vision (ECCV)*, pp. 575–592.
Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation Learning with Contrastive Predictive Coding". In: *arXiv preprint arXiv:1807.03748*.
Radford, Alec et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 8748–8763.
Seo, Paul Hongsuck et al. (2018). "CPlaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps". In: *arXiv preprint arXiv:1808.02130*.
Vaswani, Ashish et al. (2017). "Attention is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30.
Vivanco Cepeda, Vicente, Gaurav Gautam, and Mubarak Shah (2023). "GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 36.
Weyand, Tobias, Ilya Kostrikov, and James Philbin (2016). "PlaNet - Photo Geolocation with Convolutional Neural Networks". In: *European Conference on Computer Vision (ECCV)*. Springer, pp. 37–55.