University of Amsterdam

Project AI

Concept-Aware Geolocation with Hierarchical Concept Bottleneck Models

Pradyut Nair
15558169

Mentor:
Nanne van Noord

Contents

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 (Astruc, Guerin, et al. 2024; Yiqi Li et al. 2024). 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, a learned gating mechanism combines concept and spatial information while providing interpretable insights into which geographic features rely on semantic concepts versus visual patterns.

We make the following contributions:

1. 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.
2. 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.
3. A learned gated fusion mechanism where concept embeddings and spatial image information are combined through a learnable gate that controls the contribution of each source per dimension, providing interpretable insights into which geographic features rely on semantic concepts versus spatial patterns.
4. 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 (Inc. 2021) 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 pretraining: 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, post-hoc transformation of pre-trained networks (Yuksekgonul et al. 2022), stochastic concept dependencies (Vandenhirtz and Bizer 2024), and interactive human-in-the-loop systems (Chauhan et al. 2023).

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. Cross-view approaches (Toker, Kira, and Lepetit 2021) have explored matching street-level imagery with overhead satellite views for localization. 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.

3.1 Dataset and Preprocessing

We utilize a dataset of 43,040 street view panorama images, each corresponding to a uniquely challenging GeoGuessr "meta" collected from the `learnablemeta.com` website. GeoGuessr is an online geographic discovery game (AB 2013), and `learnablemeta.com` curates collections of difficult or distinctive locations, referred to as "metas" (LearnableMeta 2023). For every sample in our dataset, we have an image, GPS coordinates (latitude and longitude), a fine-grained child concept label capturing the scene type (e.g., "Urban Street", "Suburban Road", "Rural Landscape"), a coarser parent concept label that reflects the broader category (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.

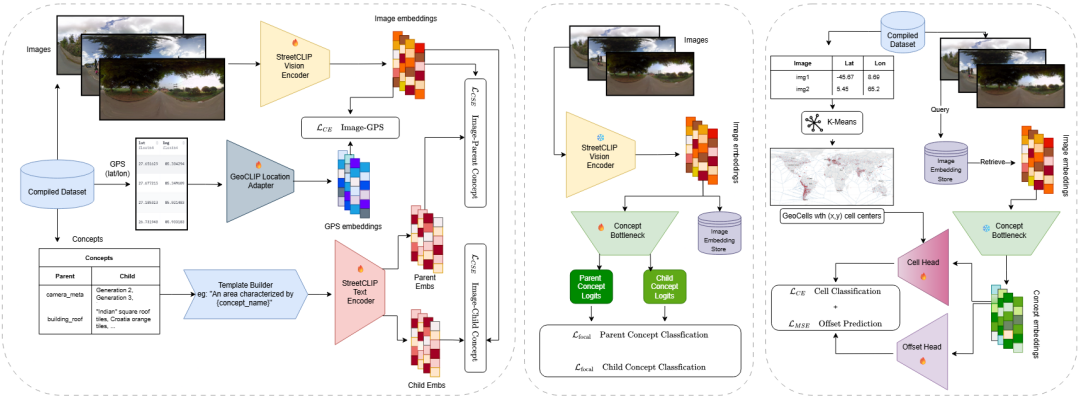


Figure 1: Three-stage architecture for concept-aware geolocation. Stage 0 performs domain contrastive pretraining to align image, GPS, and concept embeddings, with a GPS adapter (512d→768d) to align GeoCLIP location features with StreetCLIP image features. Stage 1 learns hierarchical concept classification through text-anchored prototypes with cosine similarity. Stage 2 predicts geographic coordinates via learned gated fusion: concept embeddings query image patches through cross-attention to extract spatial context, then a learned gate balances concept versus spatial information for each dimension.

3.2 Geocell Generation

Traditional approaches partition the Earth's surface uniformly using fixed-resolution grids like S2 cells (Inc. 2021), but this fails to account for non-uniform data distribution. Urban areas contain far more street view imagery than remote regions (Biljecki 2021; Wang, Zhang, and Liu 2025). 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 geocells globally, with dense regions (Europe, South 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.

3.3 Stage 0: Domain Contrastive Pretraining

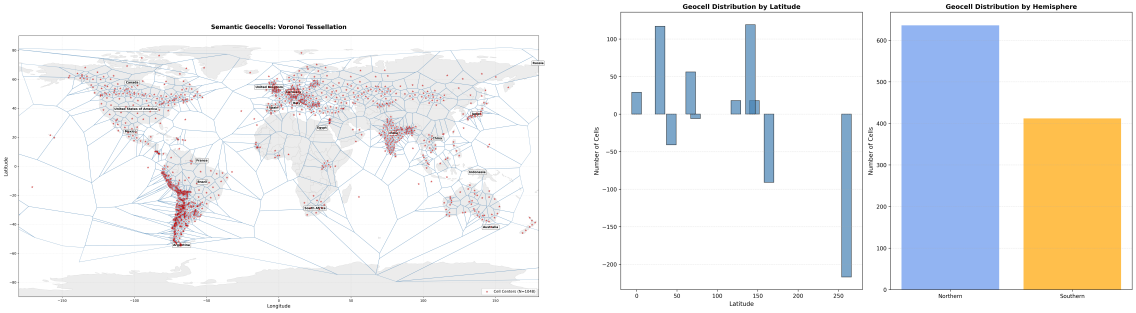
The first training stage adapts the StreetCLIP vision encoder to our specific domain through multi-objective contrastive learning. While StreetCLIP provides strong geographic priors from pretraining on street view imagery, we find that partial fine-tuning improves downstream concept learning and geolocation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

We unfreeze the top two transformer layers of the vision encoder while keeping lower layers frozen to preserve general visual representations. The model architecture includes several key components for alignment:

GPS Adapter. The GeoCLIP location encoder produces 512-dimensional GPS embeddings, but these must be aligned with the 768-dimensional StreetCLIP image feature space for contrastive learning. We introduce a GPS adapter network that projects location embeddings up to the image dimension:

$$g_{768} = \text{GPSAdapter}(g_{512}) = \text{LayerNorm}(\text{MLP}(\text{LayerNorm}(\text{MLP}(g_{512})))) \quad (1)$$

where each MLP layer uses GELU activation and dropout (0.1).



(a) Voronoi tessellation of 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 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.

Concept Bottleneck Projection. A concept bottleneck projects image features to a 512-dimensional embedding space that will be used for concept learning:

$$z = \text{ConceptBottleneck}(x) = \text{LayerNorm}(\text{MLP}(\text{LayerNorm}(\text{MLP}(x)))) \tag{2}$$

where $x \in \mathbb{R}^{768}$ is the image feature and $z \in \mathbb{R}^{512}$ is the concept embedding.

Prototype Projection. Text prototypes encoded by the frozen StreetCLIP text encoder (768-dimensional) are projected to the concept embedding space:

$$T_{\text{projected}} = \text{normalize}(W_T \cdot T_{\text{text}}) \tag{3}$$

where $W_T \in \mathbb{R}^{512\times768}$ is a learnable projection matrix.

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

$$\mathcal{L}_{\text{GPS}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(x_i, g_{768,i})/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(x_i, g_{768,j})/\tau)} \tag{4}$$

where x_i is the L2-normalized image embedding, $g_{768,i}$ is the GPS embedding projected to 768 dimensions, and $\tau = 0.07$ is the temperature parameter.

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

$$\mathcal{L}_{\text{child}} = \text{InfoNCE}(z, T_{\text{child}}, c_{\text{child}}, \tau) \tag{5}$$

$$\mathcal{L}_{\text{parent}} = \text{InfoNCE}(z, T_{\text{parent}}, c_{\text{parent}}, \tau) \tag{6}$$

where z is the L2-normalized concept embedding, T_{child} and T_{parent} are text-encoded prototype matrices projected to 512 dimensions, and c_{child}, c_{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. Additionally, a patch-level GPS alignment loss encourages patch tokens to align with GPS embeddings, providing spatial regularization.

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

$$\mathcal{L}_0 = \lambda_{\text{GPS}}\mathcal{L}_{\text{GPS}} + \lambda_{\text{child}}\mathcal{L}_{\text{child}} + \lambda_{\text{parent}}\mathcal{L}_{\text{parent}} + \lambda_{\text{hierarchy}}\mathcal{L}_{\text{hierarchy}} + \lambda_{\text{patch}}\mathcal{L}_{\text{patch-GPS}} \tag{7}$$

Stage 0 trains for 20 epochs with batch size 128, using differential learning rates: 3×10^{-5} for unfrozen encoder layers and 1×10^{-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.

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

$$T = \text{normalize}(W_T \cdot (T^{\text{base}} + \Delta)) \tag{8}$$

where $T^{\text{base}} \in \mathbb{R}^{k \times 768}$ contains frozen text embeddings, $\Delta \in \mathbb{R}^{k \times 768}$ is a learnable residual initialized from $\mathcal{N}(0, 0.01)$, and $W_T \in \mathbb{R}^{512 \times 768}$ is a learnable projection matrix that maps from the 768-dimensional StreetCLIP space to our 512-dimensional concept embedding space. The final prototypes are L2-normalized.

Cosine Similarity Classification. Concept predictions are computed through cosine similarity to prototypes, with learnable temperature scales and per-class bias terms:

$$\text{logits}_{\text{child}} = s_{\text{child}} \cdot (\hat{z} \cdot T_{\text{child}}^T) + b_{\text{child}} \tag{9}$$

$$\text{logits}_{\text{parent}} = s_{\text{parent}} \cdot (\hat{z} \cdot T_{\text{parent}}^T) + b_{\text{parent}} \tag{10}$$

where $\hat{z} = \text{normalize}(z)$ is the L2-normalized concept embedding, s_{child} and s_{parent} are learnable logit scales (initialized to 14.0, clamped to maximum 20), and $b_{\text{child}}, b_{\text{parent}} \in \mathbb{R}^k$ are per-class bias terms. The logit scale controls the "temperature" of the softmax distribution, with higher values producing sharper predictions.

Concept Bottleneck Options. The concept bottleneck that projects image features ($x \in \mathbb{R}^{768}$) to concept embeddings ($z \in \mathbb{R}^{512}$) can be implemented as either: (1) an MLP with LayerNorm and dropout (0.4), or (2) a TransformerBottleneck with self-attention and attention pooling for more expressive feature transformation. The TransformerBottleneck uses a learnable [CLS] token, positional encoding, and stochastic depth (0.2) for regularization.

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

$$\mathcal{L}_{\text{focal}} = -\alpha(1 - p_t)^\gamma \log(p_t) \tag{11}$$

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:

$$\mathcal{L}_{\text{consistency}} = \text{KL}(\text{softmax}(\text{logits}_{\text{child}}) \cdot M_{\text{hier}} \| \text{softmax}(\text{logits}_{\text{parent}})) \tag{12}$$

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 3×10^{-4}, and AdamW optimizer (Loshchilov and Hutter 2017).

3.5 Stage 2: Gated Fusion Geolocation

The final stage predicts geographic coordinates through learned fusion of concept embeddings and spatial image information. Unlike typical cross-attention architectures that simply concatenate features, our design uses a learned gating mechanism that explicitly controls how much concept versus spatial information contributes to each prediction.

Architecture. The Stage 2 model receives frozen concept embeddings $z \in \mathbb{R}^{512}$ from Stage 1 (via the frozen concept bottleneck) and frozen patch tokens $P \in \mathbb{R}^{576 \times 1024}$ from the image encoder. The architecture supports three ablation modes to understand component contributions:

1. Both: Learned gated fusion of concepts and spatial information (default)
2. Concept-only: Predictions from concept embedding alone
3. Image-only: Predictions from pooled patch tokens alone

Cross-Attention for Spatial Context. In the default "both" mode, we first use multi-head cross-attention (Vaswani et al. 2017) where the concept embedding serves as the query and patch tokens serve as keys/values:

$$Q = z \in \mathbb{R}^{512} \quad \text{(concept embedding as query)} \tag{13}$$

$$K = V = W_P \cdot P \in \mathbb{R}^{576\times512} \quad \text{(projected patch tokens)} \tag{14}$$

$$\text{attn} = \text{MultiHeadAttn}(Q, K, V) \in \mathbb{R}^{512} \tag{15}$$

where $W_P \in \mathbb{R}^{512\times1024}$ projects patch tokens to the concept embedding dimension. The attention weights can be reshaped to a 24×24 spatial map, providing interpretable visualization of which image regions most strongly support the prediction.

Learned Gating Mechanism. The key innovation is how we combine the concept embedding with the cross-attention output. Rather than simple concatenation or addition, we use a learned gate that controls the contribution of each information source:

$$z_{\text{spatial}} = \text{FFN}(\text{LayerNorm}(z + \text{attn})) \in \mathbb{R}^{512} \tag{16}$$

$$z_{\text{combined}} = \text{concat}(z, z_{\text{spatial}}) \in \mathbb{R}^{1024} \tag{17}$$

$$g = \sigma(W_g \cdot z_{\text{combined}}) \in \mathbb{R}^{512}, \quad g \in [0, 1] \tag{18}$$

$$z_{\text{gated}} = g \odot z + (1 - g) \odot z_{\text{spatial}} \tag{19}$$

$$z_{\text{final}} = \text{FusionMLP}(\text{concat}(z, z_{\text{gated}})) \tag{20}$$

where σ is the sigmoid function, $\odot$ is element-wise multiplication, and g is a 512-dimensional gate vector. Each dimension independently learns to balance concept versus spatial information.

Gate Interpretation. The gate values provide interpretability:

1. When $g_d \approx 1$: The d-th dimension relies primarily on concept information
2. When $g_d \approx 0$: The d-th dimension relies primarily on spatial (cross-attention) information
3. When $g_d \approx 0.5$: The d-th dimension balances both sources equally

Figure 3 shows the distribution of gate values during validation. The 90th percentile gate value of approximately 0.82 indicates that most dimensions tend to favor concept information, while the 10th percentile of 0.29 shows some dimensions rely more heavily on spatial context. This adaptive balancing allows the model to use concept information when it is discriminative while falling back to spatial features when concepts are ambiguous.

Prediction Heads. Two prediction heads operate on the final fused embedding z_{final}:

1. Cell Head: Softmax classification over 1,048 geocells
2. Offset Head: MSE regression predicting 3D Cartesian offset from cell center

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

$$\mathcal{L}_2 = \mathcal{L}_{\text{cell}} + \lambda_{\text{offset}} \mathcal{L}_{\text{offset}} \tag{21}$$

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

Stage 2 trains for 30 epochs with batch size 32, learning rate 1×10^{-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).

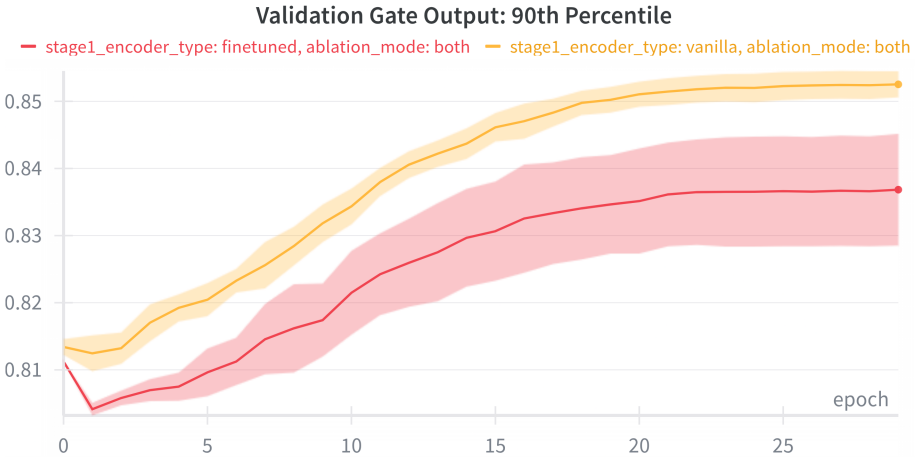


Figure 3: Distribution of learned gate values during validation. The gate controls the contribution of concept versus spatial (cross-attention) information for each of the 512 dimensions. Higher values (near 1) indicate greater reliance on concept information, while lower values (near 0) indicate greater reliance on spatial image features. The model learns to use a balanced combination, with median gate value around 0.58, though dimensions vary widely from concept-dominated (90th percentile: 0.82) to spatial-dominated (10th percentile: 0.29).

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:

1. Street: within 1 km
2. City: within 25 km
3. Region: within 200 km
4. 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	45.5%	38.6%	68.1%	71.3%
Finetuned	**46.1%**	**48.0%**	**71.6%**	**72.5%**

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 gated fusion ("both") achieves the best median error of 126 km by adaptively combining concept and spatial information.

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.

The gated fusion in the "both" mode achieves the best performance (126.0 km) by adaptively combining concept and spatial information. The learned gate values (Figure 3) show that the model uses a balanced combination, with median gate value around 0.58 indicating slight preference for concept information overall. However, the wide distribution (10th percentile: 0.29, 90th percentile: 0.82) reveals that different dimensions specialize: some rely primarily on concept information while others leverage spatial context from cross-attention.

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.

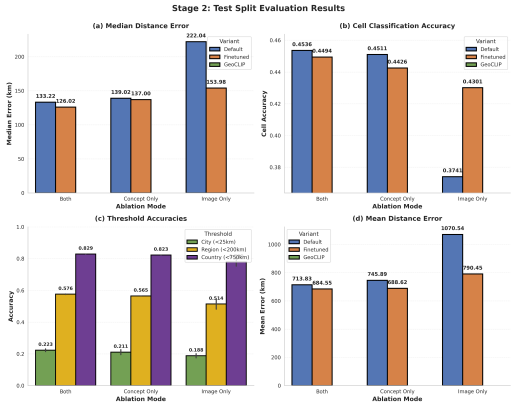
Out-of-distribution performance reveals the generalization capabilities of our approach. Median error increases substantially compared to in-distribution results, reflecting the expected distribution shift between training and GeoGuessr imagery. Despite this challenge, the finetuned variant maintains 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.

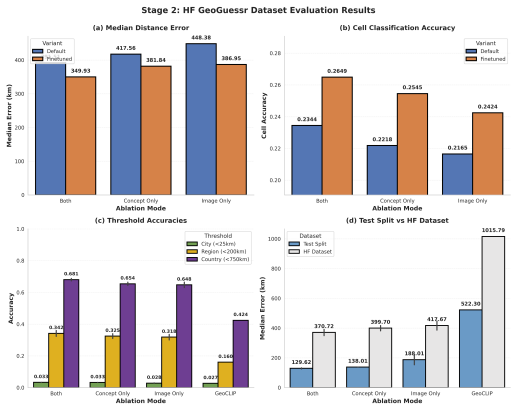
Even with the distribution shift, country-level accuracy remains above 63% across all configurations, indicating robust coarse localization. The baseline GeoCLIP model achieves significantly higher median error (1015.8 km) and lower threshold accuracies across all scales, demon-

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.



(a) In-distribution test split performance



(b) Out-of-distribution GeoGuessr performance

Figure 4: 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.

strating the advantage of our concept-aware approach even for out-of-distribution data. Figure 4 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 5 games totaling 25 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 (418.7 km) and mean distance error (1313.9 km), outperforming our main model (549.8 km median, 1527.0 km mean) and human performance (849.0 km median, 2643.0 km mean). However, examining GeoGuessr scores reveals a different pattern: our main model achieves the highest median score (3459) and mean score (3326), closely followed by GeoCLIP (3776 median, 3297 mean), while human performance lags behind (2830 median, 2665 mean).

GeoGuessr scoring provides complementary insights: GeoCLIP achieves the highest median score (3776), followed by our main model (3459) and human (2830). The mean scores show a similar competitive pattern: main model (3326), GeoCLIP (3297), and human (2665). These re-

Method	Median Distance (km)	Mean Distance (km)	Median Score	Mean Score
Main Model	549.8	1527.0	3459	3326
Baseline (GeoCLIP)	418.7	1313.9	3776	3297
Human	849.0	2643.0	2830	2665

Table 4: GeoGuessr game comparison: main model vs baseline (GeoCLIP) vs human performance on 25 rounds.

sults indicate that our concept-aware approach achieves competitive game performance with the baseline GeoCLIP model, with both models significantly outperforming human-level scoring.

Several factors may explain these results. GeoCLIP maintains an advantage in distance accuracy (418.7 km vs 549.8 km median), likely due to extensive pretraining on geographic data. However, our concept-aware model achieves competitive game scores (3459 vs 3776 median) and even slightly higher mean scores (3326 vs 3297), demonstrating that the concept bottleneck does not substantially compromise practical performance. The interactive GeoGuessr setting differs from our training distribution, and human performance shows high variance (849 km median but 2643 km mean distance), suggesting that human geolocation relies on domain knowledge and reasoning that remains challenging for models to fully capture.

Our model's concept bottleneck provides interpretability advantages while maintaining competitive performance with strong baselines like GeoCLIP. The small accuracy trade-off (549.8 km vs 418.7 km median error) is balanced by human-interpretable concept representations and spatial attention maps that enable explainable predictions. Future work could explore hybrid approaches that further narrow the accuracy gap while preserving interpretability benefits.

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.

Gated Fusion Interpretability. The learned gating mechanism in Stage 2 provides additional interpretability beyond concept activations. By examining gate values across dimensions (Figure 3), we can understand which aspects of geolocation rely on semantic concepts versus spatial visual patterns. The median gate value of 0.58 indicates overall slight preference for concept information, but the wide distribution reveals dimensional specialization: some dimensions (90th percentile: 0.82) rely primarily on concepts while others (10th percentile: 0.29) leverage spatial context. This suggests that certain geographic distinctions (e.g., architectural styles, road markings) are well-captured by semantic concepts, while others (e.g., vegetation patterns, lighting conditions) require spatial reasoning.

Stage 0 Contrastive Pretraining. 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 GPS adapter (512d→768d) is critical for aligning GeoCLIP location embeddings with StreetCLIP image features in the shared embedding space.

Hierarchical Concept Organization. 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, pro-

viding multiple granularities of interpretable output that align with human semantic intuition.

Adaptive Geocell Generation. 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 the gated fusion provides interpretability through gate values, 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 learned gate values.

Our three-stage curriculum learning pipeline demonstrates the value of progressive specialization: domain contrastive pretraining with GPS adapter establishes geographic representations, text-anchored prototype learning acquires semantic concepts through cosine similarity classification, and learned gated fusion combines concept and spatial information 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.

The learned gating mechanism provides additional interpretability by revealing which dimensions rely on semantic concepts versus spatial patterns. Gate values near 1 indicate concept-dependent features (e.g., architectural styles), while values near 0 indicate spatial-dependent features (e.g., vegetation patterns). This dimensional specialization allows the model to adaptively use the most appropriate information source for each aspect of geolocation.

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 both concept-based reasoning and gate-based dimensional specialization to end users.

References

AB, GeoGuessr (2013). *GeoGuessr - A Geography Discovery Game*. `https://www.geoguessr.com`. Accessed: 2025-01-07.

Astruc, Louis, Jean-Baptiste Guerin, et al. (2024). "OmniLoc: Towards Leveraging Multiple Perspectives for Image Geolocalization". In: *arXiv preprint arXiv:2508.00000*.

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

Biljecki, Filip (2021). "Street View Imagery in Urban Analytics and GIS: A Review". In: *ISPRS International Journal of Geo-Information* 10.10, p. 681.

Chauhan, Divi et al. (2023). "Interactive Concept Bottleneck Models". In: *International Conference on Machine Learning (ICML)*.

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.

Inc., Google (2021). "S2 Geometry Library". In: Open Source Library.

Koh, Pang Wei et al. (2020). "Concept Bottleneck Models". In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 5338–5348.

LearnableMeta (2023). *LearnableMeta - Challenging GeoGuessr Locations*. `https://learnablemeta.com`. Accessed: 2025-01-07.

Li, Yiqi et al. (2024). "Visual Geo-Localization from Images". In: *arXiv preprint arXiv:2407.14910*.

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*.

Toker, Asli, Zsolt Kira, and Vincent Lepetit (2021). "Coming Down to Earth: Satellite-to-Street View Synthesis for Geo-Localization". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12517–12527.

Vandenhirtz, Julian and Christian Bizer (2024). "Stochastic Concept Bottleneck Models". In: *arXiv preprint arXiv:2406.19272*.

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.

Wang, Xiaohui, Yuting Zhang, and Yang Liu (2025). "Assessing Google Street View Data Quality and Availability". In: *International Conference on Geoinformatics*.

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.

Yuksekgonul, Cem et al. (2022). "Post-hoc Concept Bottleneck Models". In: *arXiv preprint arXiv:2205.15480*.