This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TIGeR: A Unified Framework for Time, Images and Geo-location Retrieval

David G. Shatwell Sirnam Swetha Mubarak Shah
Institute of Artificial Intelligence, University of Central Florida
david.shatwell@ucf.edu swetha.sirnam@ucf.edu shah@crcv.ucf.edu

Abstract

Many real-world applications in digital forensics, urban monitoring, and environmental analysis require jointly reasoning about visual appearance, location, and time. Beyond standard geo-localization and time-of-capture prediction, these applications increasingly demand more complex capabilities, such as retrieving an image captured at the same location as a query image but at a specified target time. We formalize this problem as Geo-Time Aware Image Retrieval *and propose* TIGeR*, a unified framework for* ***T****ime,* ***I****mages and* ***Ge****o-location* ***R****etrieval.* TIGeR *supports flexible input configurations (single-modality and multi-modality queries) and uses the same representation to perform (i) geo-localization, (ii) time-of-capture prediction, and (iii) geo-time–aware retrieval. By preserving the underlying location identity despite large appearance changes,* TIGeR *enables retrieval based on where and when a scene was captured, rather than purely on visual similarity. To support this task, we design a multistage data curation pipeline and propose a new diverse dataset of 4.5M paired image–location–time triplets for training and 86k high-quality triplets for evaluation. Extensive experiments show that* TIGeR *consistently outperforms strong baselines and state-of-the-art methods by up to 16% on time-of-year, 8% time-of-day prediction, and 14% in geo-time aware retrieval recall, highlighting the benefits of unified geo-temporal modeling.*

1. Introduction

Understanding how a location visually appears across different points in time is essential for many real-world applications, including digital forensics, environmental and climate change analysis, cultural heritage preservation, and augmented reality. In these settings, the goal is not simply to retrieve images that look similar to a query, but to retrieve images that correspond to the *same physical location* at a *specified target time*. For example, a query image of a city square in winter should retrieve an image of the same square in summer, despite large differences in foliage, lighting, pedestrian patterns, or ongoing construction. This capability requires models that represent location identity robustly while explicitly accounting for temporal variation in how that location appears.

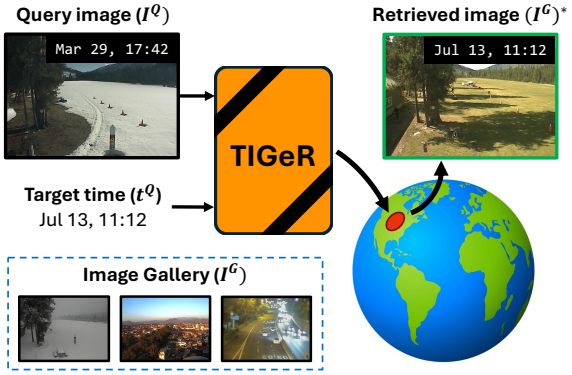


Figure 1. TIGeR unifies time, image, and location understanding, enabling geo-localization, time prediction, and geo-time aware image retrieval: given a query image and a target time, it retrieves an image captured at the same location at the specified time.

We formalize this setting as *Geo-Time Aware Image Retrieval*. Given a query image and a target time, the task is to retrieve an image captured at the same location at the specified time, rather than one chosen only based on visual similarity (Figure 1). The central challenge is to learn representations that factor out appearance changes driven by time while preserving the underlying geo-location semantics of a place. However, existing approaches do not address this challenge. Standard image retrieval methods primarily rank images based on appearance similarity [3, 10, 45], but are trained to be invariant to the time-of-capture. Composed retrieval methods can modify visual attributes (e.g., "with snow," "at night") [2, 26], but these transformations operate purely in the image domain and do not guarantee that the retrieved image comes from the same geographic location. Geo-localization models, on the other hand, aim to estimate where an image was taken, and prior work [29, 41] extends this to jointly infer where and when an image was captured. Yet these models typically encode each modality (image, location, time) independently and align only

the final embeddings through contrastive learning. Without an explicit cross-modal fusion mechanism, they struggle to capture rich geo-temporal relationships and to model how visual appearance at a fixed location evolves over time.

To address this challenge, we propose a novel approach using a multi-modal transformer that leverages self-attention to jointly process visual, geo-location, and temporal information. One or more modality-specific embeddings are passed through the multi-modal transformer, aggregating information across modalities and producing a unified representation that is then projected into a joint geo-temporal embedding space. By allowing modalities to attend to each other, the model learns geo-temporal correlations *directly*, rather than aligning separately learned embeddings post hoc, thereby enabling robust retrieval even when metadata is incomplete or noisy.

The learned representation serves as a unified geo-temporal embedding that supports: (i) geo-localization, (ii) time-of-capture prediction, and (iii) geo-time aware image retrieval. Given an image and a target time, `TIGeR` performs retrieval based on *where* and *when* the scene is, rather than purely on how it looks, making it more robust to substantial seasonal, structural, and lighting changes.

However, developing such models also requires large, high-quality image datasets paired with geo-locations and timestamps. Existing datasets are often geographically biased toward the Northern Hemisphere [18] or contain substantial degraded or low-quality imagery [8, 9], making them insufficient for rigorous evaluation at global scale. To address this gap, we introduce a new benchmark for geo-time aware image retrieval. Our benchmark is constructed from a curated subset of the AMOS webcam corpus [8, 9], which offers global geographic coverage and continuous image capture across seasons, weather conditions, and times of day. We develop a multi-stage curation pipeline to (i) remove noisy, low-quality, or corrupted frames, (ii) normalize and balance geographic coverage, and (iii) construct train/test splits that preserve both spatial and temporal diversity. The resulting dataset provides a challenging yet reliable platform for evaluating models that must reason jointly over images, locations, and timestamps.

In summary, our key contributions are as follows:

- We introduce geo-time aware image retrieval, a new task that requires retrieving images consistent with both the geographic location and target time-of-capture of a query.
- We propose `TIGeR`, a multimodal transformer-based framework that learns a unified geo-temporal embedding space through cross-modal attention over image, location, and time inputs.
- We curate both a large-scale training dataset and a benchmark dataset with global geographic coverage, broad temporal diversity, and extensive quality filtering, providing the first large-scale testbed tailored to this task.
- We show that `TIGeR` outperforms prior methods, achieving average improvements of 16% on time-of-year prediction, 8% on time-of-day prediction, and up to 14% over state-of-the-art baselines for time-of-capture prediction and geo-time aware retrieval.

2. Related Work

Time-aware understanding. Modeling how visual appearance evolves over time has recently attracted growing attention in computer vision. Early works explored predicting the *time-of-capture* from single images using either physical or data-driven cues. Physically inspired methods [14, 34] estimate time or verify timestamp consistency by analyzing shadows or the Sun's position, but they require strong priors such as known camera orientation, sky visibility, or external metadata. Data-centric approaches [22, 28, 41] instead train neural networks to predict temporal attributes such as time-of-day or month from labeled images, often in combination with geo-tags. However, these models typically treat time as a classification target and are limited to coarse, discretized predictions. GT-Loc [29] extends this direction by learning a shared embedding space for image, location, and time, but processes each modality independently and aligns them only via a contrastive loss at the embedding level, restricting its ability to capture rich cross-modal geo-temporal relationships. In contrast, `TIGeR` explicitly learns these relationships by enabling cross-modal attention between modalities, producing a unified representation that models how visual appearance varies jointly across geolocation and time.

Global geo-localization. Geo-localization aims to infer the geographic coordinates of an image or retrieve images taken at similar locations from a large gallery. Traditional methods either discretize the globe into geo-cells and frame localization as classification [12, 19, 23, 37, 38], or adopt a retrieval-based paradigm that matches query embeddings to stored geospatial features [25, 29, 30, 36, 43, 44]. Recent approaches such as GeoCLIP [36] leverage CLIP-style contrastive training to align image and GPS embeddings in a shared feature space, improving large-scale retrieval. PIGEON [6] combines classification and retrieval for fine-grained localization, while Img2Loc [42] and GAEA [4] use multimodal large language models to reason jointly over image content and metadata. However, these approaches still operate with largely independent encoders and do not explicitly model cross-modal relationships governing geo-temporal variation. `TIGeR` bridges this gap by using attention-based fusion to learn a shared representation space in which both *where* and *when* matter: the model can retrieve images captured at the *same* location but at a *different* time, enabling genuine geo-temporal understanding.

Geo-temporal and multi-modal representation learning. Beyond standard geo-localization, several works seek

to learn representations that combine multiple modalities such as images, satellite, and GPS. SatCLIP [11] and related approaches [16, 17, 40] extend CLIP-like dual encoders approach to align satellite, ground-level views, or GPS for cross-domain geospatial tasks. While these dual-encoder methods highlight the effectiveness of contrastive learning for spatial alignment, they still encode each modality independently and rely on fixed modality pairs. `TIGeR` departs from this paradigm by employing a shared multimodal transformer and joint contrastive learning that allows image, GPS, and time to *attend to each other*, inspired by [31, 32]. This design enables the model to learn fine-grained geo-temporal correlations (e.g., how a specific location changes across seasons or lighting conditions) and to perform retrieval across arbitrary modality combinations, rather than being constrained to pre-defined pairs.

3. Method

Our goal is to learn a unified representation that jointly embeds the location, visual, and time modalities, namely, the image I, GPS coordinates l, and time t into a shared geo-temporal feature space. The representation is designed to preserve location identity while capturing temporal variation, enabling flexible cross-modal retrieval. Our proposed model is capable of various tasks, such as, (i) retrieve an image of the same place at a different time, (ii) estimate the geo-location from image alone, and (iii) predict a plausible time-of-capture from single image or (iv) combined image and location cues.

3.1. Problem Statement

Given a dataset $S_{\text{train}} = \{(I_i, l_i, t_i)\}_{i=1}^{N}$, where I_i is an image, l_i denotes its GPS coordinates, and t_i its capture time, our goal is to learn a model $f : \mathcal{M} \to \mathbb{R}^d$ that maps an input $\mathcal{M}$ from the modality set $\{I, l, t, (I, l), (I, t), (l, t)\}$ into a shared feature space $\mathbb{R}^d$. The model is trained so that features from inputs describing similar geo-temporal conditions (*i.e.*, captured at nearby locations and times) have high cosine similarity, while unrelated inputs are pushed apart.

3.2. Architecture

Our model comprises three stages (see Figure 2). The first stage includes an image encoder $\mathcal{V}(I) = V \in \mathbb{R}^{N_V \times d}$, a location encoder $\mathcal{G}(l) = L \in \mathbb{R}^{N_L \times d}$, and a time encoder $\mathcal{T}(t) = T \in \mathbb{R}^{N_T \times d}$, where N_V, N_L, and N_T denote the number of tokens per modality and d the embedding dimension. The image encoder is a frozen CLIP ViT [24] that outputs one `[CLS]` token and $N_V - 1$ patch embeddings. The location and time encoders use random Fourier features (RFF) to project 2D inputs into N_L and N_T high-dimensional vectors, modulated by frequencies $\sigma_i \in \{2^{2i}\}_{i=0}^{N-1}$. Each modality embedding is then linearly projected and layer-normalized.

The second stage employs a multi-modal transformer $\mathcal{F}$ to refine the modality-specific embeddings, learning fine-grained relationships between them via self-attention. Since there are six possible input combinations, $\{V, L, T, [V; L], [V; T], [L; T]\}$, we perform six forward passes during training. When two modalities are provided, we concatenate their embeddings along the token dimension before fusion, i.e., $X \in \{V, L, T, [V; L], [V; T], [L; T]\}$ and obtain $X' = \mathcal{F}(X)$.

For each refined embedding X', we perform pooling to aggregate token representations and then project them into the shared embedding space via learned projections $h_x(\cdot)$:

$$\mathbf{x} = \tfrac{1}{N_x} \sum_{i=1}^{N_x} X'_{i,:}, \quad \mathbf{x}' = h_x(\mathbf{x}), \quad f(x) = \bar{\mathbf{x}} = \tfrac{\mathbf{x}'}{\|\mathbf{x}'\|},$$

Similarly for bimodal inputs, we first pool and project each constituent modality to obtain embeddings $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, following the same procedure used for unimodal inputs. The resulting normalized embeddings are then fused by averaging:

$$f(x_1, x_2) = \frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{\|\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\|}.$$

This yields six embeddings $\{\bar{\mathbf{v}}, \bar{\mathbf{l}}, \bar{\mathbf{t}}, \bar{\mathbf{vl}}, \bar{\mathbf{vt}}, \bar{\mathbf{lt}}\}$, all residing in a shared geo-temporal feature space.

3.3. Losses

`TIGeR` is optimized using a combination of contrastive and classification losses with soft metric targets. The contrastive objectives encourage embeddings from different input modalities to align within the shared feature space, forming the foundation for all geo-temporal retrieval tasks. In parallel, a classification loss is applied only to the final image embeddings $\bar{\mathbf{v}}$, enforcing a smooth, structured organization of the embedding space guided by the soft metric targets.

Contrastive losses. We employ five contrastive objectives to align all meaningful modality pairs, excluding the individual location and time embeddings $(\bar{\mathbf{l}}, \bar{\mathbf{t}})$, which remain uncorrelated. Specifically, we optimize the pairs $(\bar{\mathbf{v}}, \bar{\mathbf{l}})$, $(\bar{\mathbf{v}}, \bar{\mathbf{t}})$, $(\bar{\mathbf{v}}, \bar{\mathbf{lt}})$, $(\bar{\mathbf{l}}, \bar{\mathbf{vt}})$, and $(\bar{\mathbf{t}}, \bar{\mathbf{vl}})$. Each pair $(\mathbf{x}_i, \mathbf{y}_i)$ is trained using the InfoNCE loss [20]:

$$\mathcal{L}_C^{(i)}(\mathbf{x}_i, \mathbf{y}_i) = -\log \frac{\exp(\mathbf{x}_i^\top \mathbf{y}_i / \tau)}{\sum_j \exp(\mathbf{x}_i^\top \mathbf{y}_j / \tau)}, \tag{1}$$

where τ is a temperature parameter controlling the sharpness of the distribution. The overall contrastive objective is the sum of all pairwise components:

$$\begin{aligned} \mathcal{L}_C = \mathcal{L}_C(\bar{\mathbf{v}}, \bar{\mathbf{l}}) + \mathcal{L}_C(\bar{\mathbf{v}}, \bar{\mathbf{t}}) + \mathcal{L}_C(\bar{\mathbf{v}}, \bar{\mathbf{lt}}) \\ + \mathcal{L}_C(\bar{\mathbf{l}}, \bar{\mathbf{vt}}) + \mathcal{L}_C(\bar{\mathbf{t}}, \bar{\mathbf{vl}}). \end{aligned} \tag{2}$$

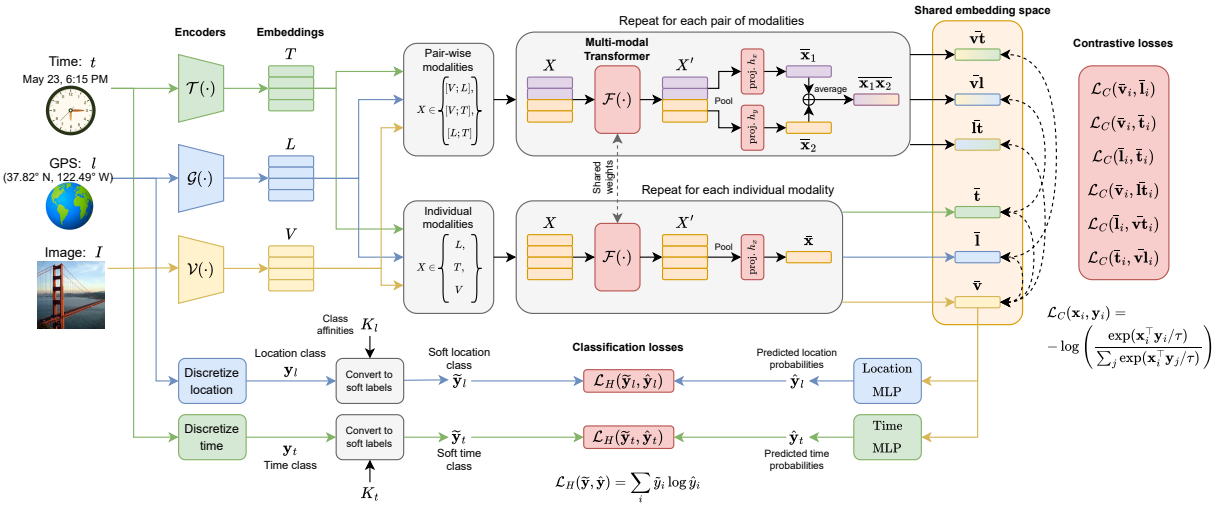


Figure 2. **Architecture of `TIGeR`.** Given an image I, GPS coordinates l, and timestamp t, modality-specific encoders produce visual (V), location (L), and temporal (T) embeddings. Individual and pair-wise modality tokens are passed through a *shared multi-modal transformer* $\mathcal{F}(\cdot)$, which employs self-attention to learn geo-temporal interactions. Pooled embeddings are projected into a *shared embedding space*, where a **contrastive loss** aligns unimodal and fused representations across all modality pairings (e.g., V–L, V–T, L–T). In parallel, the **classification losses with metric targets** on discretized location and time classes encourage the model to learn hierarchical cross-modal representations that vary smoothly across the embedding space.

Classification losses. To encourage the model to learn geo-temporal representations that vary smoothly across the embedding space, we employ two classification losses with soft targets on the image embedding $\bar{\mathbf{v}}$.

Geo-location classes. We define spatial classes using the Hierarchical Equal Area isoLatitude Pixelisation (HEALPix) [5] scheme, which divides the Earth's surface into equal-area regions whose centers lie on discretized latitude circles. The base resolution yields 12 cells, and each subdivision increases the count fourfold. We use the fourth hierarchy level ($12 \times 4^3 = 768$ regions). For each image–GPS pair (I, l), we compute the HEALPix label as $(\mathbf{y}_l, l_C) = \text{HEALPix}(l)$, where $\mathbf{y}_l$ is the one-hot class vector and l_C the coordinates of the corresponding cell center.

Time classes. Temporal classes are defined by binning the time-of-day (ToD) and time-of-year (ToY) into 1-hour and 1-month intervals, producing $24 \times 12 = 288$ bins. Following GT-Loc [29], each timestamp is represented as a point on the flat torus $t = (\theta, \phi) \in \mathbb{T}^2$, partitioned into 288 equal-area regions. For each (I, t) pair, we assign $(\mathbf{y}_t, t_C) = \text{TimeClass}(t)$, where $\mathbf{y}_t$ is the one-hot class vector and t_C the center of the corresponding region.

Soft targets. To obtain soft supervision, we measure the similarity between class centers via a metric κ. For geo-location classes, κ is the Haversine (great-circle) distance on the sphere. For time classes, we use the geodesic distance on the torus:

$$\kappa(t_i, t_j) = \sqrt{\sum_{\alpha \in \{\theta, \phi\}} \min\big(1 - |\Delta\alpha_{i,j}|, |\Delta\alpha_{i,j}|\big)^2}. \quad (3)$$

Given a temperature γ, we construct a normalized affinity matrix

$$K_{i,j} = \frac{\exp[-\kappa(C_i, C_j)/\gamma]}{\sum_j \exp[-\kappa(C_i, C_j)/\gamma]}, \quad (4)$$

where C_i and C_j are the centers of two classes of the same modality. The soft target is then $\tilde{\mathbf{y}} = K\mathbf{y}$, which spreads probability mass over neighboring classes. The final classification loss is the cross-entropy between the predicted probabilities $\hat{\mathbf{y}} = \sigma(\text{MLP}(\bar{\mathbf{v}}))$ and the soft target:

$$\mathcal{L}_H = -\sum_i \tilde{y}_i \log \hat{y}_i. \quad (5)$$

4. Benchmark Dataset Curation

Constructing a dataset suitable for time-aware geolocalization presents significant challenges. One potential dataset is AMOS [8, 9], which provides a vast collection of static outdoor imagery distributed across diverse regions, months, and times-of-day, however, its raw form is highly unsuitable for direct use. The dataset suffers from pervasive noise and corruption, stemming from both hardware limitations and environmental conditions. Many cameras frequently generate unusable frames due to broken hardware or missing camera feeds. Low-cost sensors introduce additive Gaussian-like noise and compression artifacts, while night time images often collapse into near-black frames with severely limited dynamic range. Other failure modes include misconfigured sensors producing banding or dead pixels, placeholders or stale frames with uniform content,

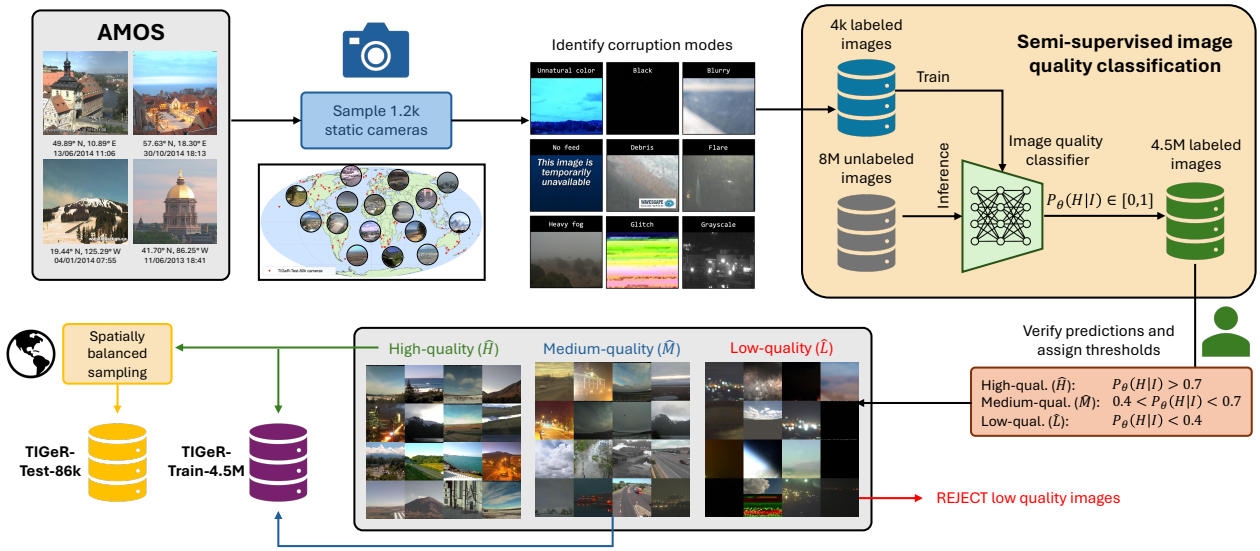


Figure 3. **Benchmark dataset curation pipeline.** Starting from the AMOS corpus, we randomly sample 1,255 static cameras with broad global coverage and identify diverse corruption modes. We manually label 4,000 images for quality, train a semi-supervised image quality classifier on frozen visual features, and apply it to 8M unlabeled frames to predict a quality score $P(H|I)$. Using thresholds T_H=0.7 and T_L=0.4, images are partitioned into high-, medium-, and low-quality subsets, and low-quality frames are discarded. High-quality images are then used to construct a geographically balanced test set by sampling cameras across $10^\circ \times 10^\circ$ latitude-longitude bins and retaining only cameras with at least 500 frames spanning a year, while the remaining high- and medium-quality images form the training set. The resulting benchmark contains 4.5M training images and 86k test images, with no camera overlap between splits.

and weather-induced distortions (rain, snow, fog) that obscure critical scene details. Without careful filtering, these corruptions overwhelm useful signal, making curation a central challenge. We list all the corruption modes and describe them in detail in the supplementary section A.

To overcome these challenges, we design a multi-stage curation pipeline (Figure 3), that systematically filters noise while preserving both geographic and temporal diversity. To ensure global coverage, we begin by randomly sampling 1,255 cameras from the AMOS dataset across both hemispheres and multiple continents. To build a supervised filtering model, we manually annotate 4,000 samples into high- and low-quality categories, capturing the diverse corruption modes described above. Formally, if an image I has any form of corruption, we assign it to the low-quality set L with a label $q = 0$; otherwise it is placed in the high-quality set H with $q = 1$.

Using these seed sets, we train a lightweight classifier on frozen CLIP [24] and DINOv2 [21] visual features, with a linear probe optimized for binary quality prediction. The output of the model is a confidence score $p(q = H|I) \in [0, 1]$, indicating the probability that an image is of high-quality. We then evaluate the model's performance on a hold-out set of 400 images resulting in a binary classification accuracy of 91%. Based on the cumulative score distribution to select two thresholds $T_H = 0.7$ and $T_L = 0.4$ which we use to group the predictions into high- ($\hat{H}$), medium- ($\hat{M}$), and low-quality ($\hat{L}$) sets as follows:

$$I \in \begin{cases} \hat{H}, & \text{if} \quad P(H|I) \geq T_H \\ \hat{M}, & \text{if} \quad T_L \leq P(H|I) < T_H \\ \hat{L}, & \text{if} \quad P(H|I) < T_L \end{cases}. \tag{6}$$

The images assigned to the low-quality set $\hat{L}$ are removed from the dataset, since excessive noise can obscure the geo-temporal signals needed to train and evaluate our model. In contrast, the high-quality images in $\hat{H}$ are well suited for building the test set, as they typically exhibit good illumination, mild weather conditions, and reliable camera feeds. To construct a representative benchmark, we seek a test set with a balanced geographic distribution across both hemispheres. This is particularly important for tasks involving time-of-year, since seasonal appearance is strongly correlated with calendar month, while seasons are inverted between the Northern and Southern Hemispheres.

Beyond balancing the number of test images across hemispheres, we also aim to avoid over-representing regions with denser camera coverage. To this end, we partition the globe into $10^\circ \times 10^\circ$ latitude-longitude bins and select cameras that are approximately evenly distributed across them. From these candidate cameras, we retain only those with at least 500 frames spanning a full year, ensuring coverage across multiple months and hours at each location. All remaining cameras from the high- and medium-quality

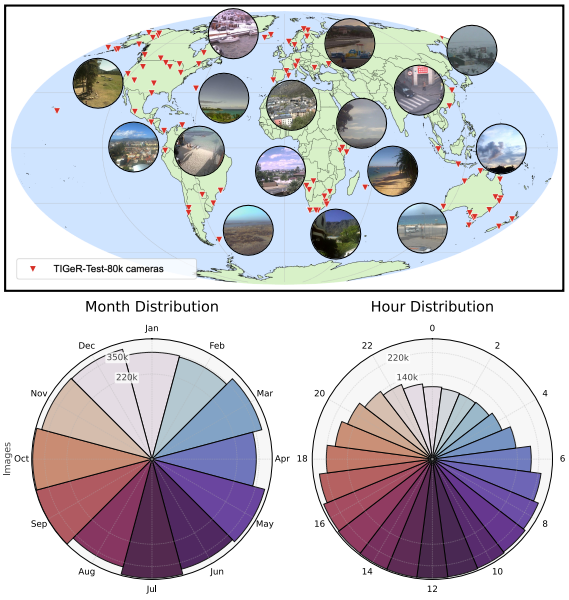


Figure 4. **Dataset statistics and distributions. Top**: Geographical distribution of camera locations across the world in the proposed dataset, showing wide coverage across multiple continents. **Left**: Time-of-year (month) distribution of the captured images. **Right**: Time-of-day (hour) distribution of the captured images.

subsets are assigned to the training set. Importantly, the training and test splits share no cameras, allowing our evaluation to better measure the model's ability to generalize to unseen locations.

After all the filtering stages, the curated dataset contains $4.5M$ training images and $86k$ test images. Figure 4 illustrates the resulting geographic and temporal distributions. This pipeline transforms the noisy AMOS corpus into a clean, balanced, and diverse benchmark, enabling robust training and reliable evaluation for downstream geo-localization and time prediction tasks.

5. Benchmarking

We evaluate TIGeR on three retrieval tasks to demonstrate its ability to learn strong and generalizable geo-temporal representations: geo-time aware image retrieval, time-of-capture prediction and geo-localization.

5.1. Geo-time Aware Image Retrieval

This task aims to retrieve, from a global gallery, the image that best matches a query in terms of both location and time. Prior work [29] studied an oracle variant where the ground-truth GPS location l^Q is given, and the query is formed by fusing the location and target time t^Q embeddings. Here we consider a more challenging and realistic setting: given a query image I^Q and target time t^Q, the goal is to retrieve an image $\{I_i^G\}_{i=1}^N$ captured at the same physical location as I^Q and at time t^Q. In this case, the model must infer and preserve location identity directly from the image, rather than relying on explicit GPS.

We address this problem by encoding the image and time with TIGeR and fusing their embeddings into a unified geo-temporal representation. We precompute gallery embeddings $\bar{\mathbf{v}}_i^G = f(I_i^G)$ and, at inference, obtain the fused query representation $\bar{\mathbf{vt}}^Q = f(I^Q, t^Q)$. Retrieval is then performed by maximizing cosine similarity:

$$(I^Q)^* = \arg\max_{I^G} \left(f(I^Q, t^Q)^\top f(I^G) \right). \tag{7}$$

We evaluate on two complementary benchmarks: our TIGER-TEST-86K split and a balanced partition of CVT [27] across hemispheres. TIGER-TEST-86K is built from webcams with multiple observations of the same camera across the year, enabling fine-grained geo-temporal retrieval at fixed locations but varying times. CVT instead consists of social media images at arbitrary locations and times, without repeated cameras; here, retrieval is considered correct if the result lies within 125 km of the query location and matches the specified time.

We compare TIGeR against two main baselines. The first is the triple-encoder classifier of Zhai et al. [41], originally proposed for joint geo-location and time-of-capture classification. In addition to the original InceptionV2 [33] backbone, we also evaluate a stronger variant with a CLIP ViT-L/14 [24] backbone for fair comparison. Following [29], we adapt this model to retrieval by first predicting the query geo-location class $\hat{l}^Q$, then computing location and time posteriors $P(l \mid I_i^G)$ and $P(t \mid I_i^G)$ for each gallery image and defining

$$\text{score}(I_i^G) = P(l = \hat{l}^Q \mid I_i^G)\, P(t = t^Q \mid I_i^G),$$

which is used to rank the gallery.

The second baseline is GT-Loc [29], which maps locations, images, and times into a shared feature space but lacks TIGeR's multimodal transformer. Here we encode image and time independently and perform late fusion by averaging their embeddings, which cannot explicitly model higher-order interactions between geo-location and time.

Table 1 reports quantitative results. TIGeR consistently outperforms all baselines, achieving higher recall at most ranks on both datasets. This indicates that the learned representations are well aligned across visual, temporal, and geo-location modalities, and that the fused embeddings successfully encode both location and time information rather than over-emphasizing purely semantic visual similarity. Qualitative examples comparing TIGeR and the baselines are shown in Figure 5. We show additional examples on the supplementary section B.

Table 1. **Geo-time Aware Image Retrieval on images from unseen locations**. Given a query image (I) and target time (t), the goal is to retrieve an image captured at the *same geographic location* as the query but at the *specified target time*. Our model significantly outperforms prior approaches, demonstrating strong geo-temporal alignment and robustness to temporal variation. *indicates methods we replicate, closely adhering to protocols outlined by prior work

Method	Retrieval	TIGeR-test-86k			CVT		
		R@1 (%)	R@5 (%)	R@10 (%)	R@1 (%)	R@5 (%)	R@10 (%)
Zhai et al. [41]*	$It \to I$	1.97	7.05	11.23	0.45	1.59	2.38
Zhai et al. w/ CLIP [41]*	$It \to I$	2.60	8.80	13.70	2.95	8.76	13.31
Time-Loc [29]*	$It \to I$	0.38	1.73	3.19	15.23	19.28	21.18
GT-Loc [29]*	$It \to I$	0.34	1.56	2.94	**16.45**	20.44	22.59
TIGeR (Ours)	$It \to I$	**3.51**	**23.30**	**37.51**	14.55	**24.46**	**29.98**

5.2. Time Prediction and Geo-localization

In this task, the objective is to estimate both the geo-location and time-of-capture of a query image by comparing its embedding against dedicated galleries of location and time embeddings, respectively, and selecting the most compatible entries according to cosine similarity.

Let I^Q be a query image and let $\{x_i^G\}_{i=1}^{N_G}$ denote a gallery of targets from a modality x (either GPS coordinates or timestamps). Our model produces an image embedding $\bar{\mathbf{v}}^Q \in \mathbb{R}^d$ and a gallery embedding $\mathbf{x}_i^G \in \mathbb{R}^d$ for each gallery item. In addition, a modality-specific classifier predicts a probability vector $\hat{\mathbf{y}}$ over B coarse classes (e.g., HEALPix bins for GPS or time bins for timestamps). We write $b(x_i^G) \in \{1, \dots, B\}$ for the class index associated with gallery item x_i^G.

We combine a continuous retrieval signal with a discrete prior. Treat the cosine similarity as a temperature-scaled log-likelihood of a match, and the classifier output as a prior over classes. The resulting posterior-like score for each gallery item x_i^G is

$$\text{score}(x_i^G) \;=\; \frac{(\bar{\mathbf{v}}^Q)^\top \mathbf{x}_i^G}{\psi} \;+\; \beta(I^Q)\, \log P\big(b(x_i^G) \mid I^Q\big)\,, \tag{8}$$

where $\psi > 0$ is a temperature parameter and $\beta(I^Q) \geq 0$ controls the influence of the classifier prior for the query.

A fixed β can help on easy (confident) queries and hurt on hard (uncertain) ones. We therefore make β *query-adaptive* using the classifier entropy:

$$\beta(I^Q) \;=\; \beta_{\max}\left(1 - \frac{H(\hat{\mathbf{y}})}{\log B}\right), \tag{9}$$

where $H(\hat{\mathbf{y}})$ is the Shannon entropy and $\log B$ the maximum entropy for B classes. Thus, $\beta(I^Q) \in [0, \beta_{\max}]$: it approaches $\beta_{\max}$ when the classifier is confident (low entropy) and decreases toward 0 when it is uncertain (high entropy).

Equation (8) allows the retrieval term to *pull* fine-grained neighbors while the prior *guides* the search toward plausible classes. The entropy-adaptive weighting in (9) ensures the prior has strong influence only when it is informative, mitigating boundary effects and domain shift.

In the time prediction task, rather than conditioning onlt on the image embedding $\bar{\mathbf{v}}^Q$, we can optionally incorporate the known location as auxiliary input. In this setting, we use the fused embedding $\bar{\mathbf{vl}}^Q = f(I^Q, G^Q)$ as the query representation and compute cosine similarity against gallery time embeddings. This variant often improves temporal prediction accuracy by constraining retrieval to locations consistent with the provided GPS context.

Tables 2 and 3 summarize the performance of our method on time prediction and geo-localization, respectively. Overall, `TIGeR` achieves state-of-the-art results on both tasks. For time-of-capture prediction, `TIGeR` consistently outperforms prior methods on both month and hour estimation. Moreover, when geo-location is provided as an auxiliary input and fused with the image representation, performance improves substantially. This behavior is expected, as the visual appearance of time is strongly coupled with latitude and season (e.g., higher latitudes exhibit more pronounced seasonal variation than regions near the Equator).

For geo-localization, `TIGeR` also attains competitive performance compared to strong baselines. However, in this setting, using time as an auxiliary signal does not always lead to further gains. Unlike location, time-of-capture is a comparatively weak and ambiguous cue for disambiguating places: many geographically distant locations can share similar illumination and seasonal appearance at a given time, and the time predictions themselves may be noisy. Incorporating this uncertain temporal signal can therefore introduce additional variance or bias the model toward spurious correlations, which sometimes offsets the benefits of joint time–location reasoning for geo-localization.

5.3. Ablation Analysis

We conduct an ablation study to quantify the contribution of each component of `TIGeR` on geo-localization and time-of-capture prediction. For both tasks, we report errors av-

Table 2. **Time prediction** on unseen cameras. *indicates methods we replicate, closely adhering to protocols outlined by prior work.

Method	Retrieval	TIGeR-test-86k ToY Error↓	TIGeR-test-86k ToD Error↓	CVT-test ToY Error↓	CVT-test ToD Error↓
Zhai et al. [41]*	$I \rightarrow t$	68.38	3.97	87.37	3.28
Zhai et al. w/ CLIP [41]*	$I \rightarrow t$	57.51	3.22	68.95	2.8
Time-Loc [29]*	$I \rightarrow t$	74.87	4.02	65.10	2.86
GT-Loc [29]*	$I \rightarrow t$	74.58	3.52	78.95	**2.68**
TIGeR (Ours)	$I \rightarrow t$	**51.49**	**3.13**	**62.88**	2.73
Zhai et al. [41]*	$Il \rightarrow t$	66.36	3.88	73.15	3.15
Salem et al. [28]*	$Il \rightarrow t$	74.29	3.57	75.48	3.1
Zhai et al. w/ CLIP [41]*	$Il \rightarrow t$	55.15	3.19	60.42	2.77
Salem et al. w/ CLIP [28]*	$Il \rightarrow t$	55.4	3.24	64.05	2.71
TIGeR (Ours)	$Il \rightarrow t$	**48.86**	**3.06**	**47.00**	**2.35**

Table 3. **Geo-localization accuracy**. *indicates methods we replicate, closely adhering to protocols outlined by prior work.

Method	Retrieval	TIGeR-test-86k 200 km	TIGeR-test-86k 750 km	CVT-test 200 km	CVT-test 750 km
Zhai et al. [41]*	$I \rightarrow l$	22.05	46.20	3.50	10.50
Zhai et al. w/ CLIP [41]*	$I \rightarrow l$	28.58	64.85	23.13	57.74
GeoCLIP [36]	$I \rightarrow l$	21.58	37.58	43.62	60.77
GT-Loc [29]*	$I \rightarrow l$	21.07	35.93	42.63	59.49
TIGeR (Ours)	$I \rightarrow l$	**48.63**	**65.61**	**51.90**	**66.53**
Zhai et al. [41]*	$It \rightarrow l$	21.99	46.62	4.46	13.41
Zhai et al. w/ CLIP [41]*	$It \rightarrow l$	28.83	65.16	23.82	58.97
TIGeR (Ours)	$It \rightarrow l$	**48.34**	**64.51**	**53.40**	**67.15**

Table 4. Cumulative ablation study on the components of TIGeR.

Setting	Month Error	Hour Error	Geo-location Error (km)
Base model	76.77	3.10	668.97
+ multi-modal transformer	57.58	2.98	238.58
+ Classification heads	57.20	2.95	234.53
+ Entropy adaptive reranking	57.20	2.93	203.93

eraged over TIGeR-test-86k and CVT-test. For geo-localization, we use mean geodetic error (in kilometers) rather than recall at distance thresholds, as this yields a single, directly comparable metric.

Table 4 summarizes the results. We start from a *base model* that includes image, location, and time encoders trained with only two contrastive losses, analogous to GT-Loc [29]. Adding the *multi-modal transformer* together with the five cross-modal contrastive losses from Eq. 2 yields the largest performance gain across both temporal and geo-location errors, highlighting the importance of joint geo-temporal modeling. Next, we introduce *classification heads* as auxiliary supervision during training, which provides a modest additional improvement. Finally, we leverage these heads at test time through our *entropy-adaptive reranking* scheme, which uses the classifier predictions as a confidence-weighted prior. This step produces a substantial reduction in geo-localization error, resulting in the full TIGeR model. Additional ablations and analyses are provided in the supplementary section C.

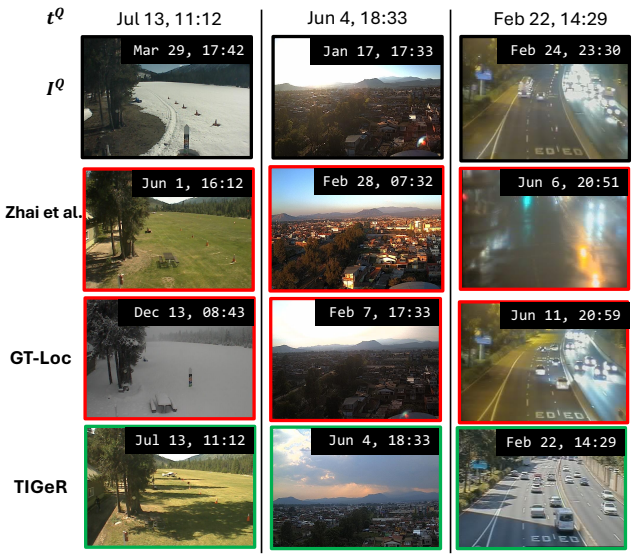


Figure 5. **Qualitative results on geo-time aware image retrieval with images from unseen locations**. Given a query image I^Q and a target time t^Q (shown above each column), the goal is to retrieve an image captured at the *same location* as the query but at the specified target time. For visualization, the actual time-of-capture for each image is overlaid on top of it. Each row shows top-1 retrieval results from different methods. While prior approaches (Zhai et al. [41], GT-Loc [29]) often retrieve images captured at incorrect times, our TIGeR model accurately retrieves images corresponding to the target time while maintaining location consistency. Best viewed in color.

6. Conclusion

We introduced TIGeR, a unified framework for jointly modeling images, geo-location, and time, and formalized the task of *geo-time aware image retrieval*, where the goal is to retrieve an image captured at the same physical location as a query but at a specified target time. To enable rigorous evaluation, we curated a large-scale benchmark with globally distributed, temporally diverse, and carefully filtered imagery. TIGeR combines modality-specific encoders with a multi-modal transformer to map all modalities into a shared geo-temporal embedding space, and is trained with contrastive objectives and entropy-adaptive reranking to capture fine-grained relationships between where and when a scene is observed. Extensive experiments show that TIGeR achieves state-of-the-art performance across geo-localization, time-of-capture prediction, and geo-time aware image retrieval.

Acknowledgements

Supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, et al. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21967–21977, 2024. 6

[2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15338–15347, 2023. 1

[3] Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2886–2892, 2025. 1

[4] Ron Campos, Ashmal Vayani, Parth Parag Kulkarni, Rohit Gupta, Aritra Dutta, and Mubarak Shah. Gaea: A geolocation aware conversational model. *arXiv e-prints*, pages arXiv–2503, 2025. 2

[5] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759–771, 2005. 4

[6] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12893–12902, 2024. 2

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[8] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Consistent temporal variations in many outdoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 2007. 2, 4

[9] Nathan Jacobs, Walker Burgin, Nick Fridrich, Austin Abrams, Kylia Miskell, Bobby H. Braswell, Andrew D. Richardson, and Robert Pless. The global network of outdoor webcams: Properties and applications. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, pages 111–120, 2009. 2, 4

[10] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9 (2):1286–1293, 2023. 1

[11] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023. 3

[12] Parth Parag Kulkarni, Gaurav Kumar Nayak, and Mubarak Shah. Cityguessr: City-level video geo-localization on a global scale. In *European Conference on Computer Vision*, pages 293–311. Springer, 2024. 2

[13] Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth J.F. Jones. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE MultiMedia*, 24(1):93–96, 2017. 6

[14] Xiaopeng Li, Wenyuan Xu, Song Wang, and Xianshan Qu. Are you lying: Validating the time-location of outdoor images. In *Applied Cryptography and Network Security: 15th International Conference, ACNS 2017, Kanazawa, Japan, July 10-12, 2017, Proceedings 15*, pages 103–123. Springer, 2017. 2

[15] Dongheng Lin, Han Hu, and Jianbo Jiao. What time tells us? an explorative study of time awareness learned from static images. *Transactions on Machine Learning Research*, 2025. 4, 5

[16] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[17] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. CSP: Self-supervised contrastive spatial pretraining for geospatial-visual representations. In *Proceedings of the 40th International Conference on Machine Learning*, pages 23498–23515. PMLR, 2023. 3

[18] Radu Paul Mihail, Scott Workman, Zach Bessinger, and Nathan Jacobs. Sky segmentation in the wild: An empirical study. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–6, 2016. 2

[19] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–579, 2018. 2

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[21] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou,

Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 5, 2

[22] Rafael Padilha, Tawfiq Salem, Scott Workman, Fernanda A Andaló, Anderson Rocha, and Nathan Jacobs. Content-aware detection of temporal metadata manipulation. *IEEE Transactions on Information Forensics and Security*, 17: 1316–1327, 2022. 2

[23] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. In *European Conference on Computer Vision*, pages 196–215. Springer, 2022. 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5, 6, 2

[25] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019. 2

[26] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19305–19314, 2023. 1

[27] Tawfiq Salem, Scott Workman, and Nathan Jacobs. Learning a Dynamic Map of Visual Appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[28] Tawfiq Salem, Jisoo Hwang, and Rafael Padilha. Timestamp estimation from outdoor scenes. In *Annual ADFSL Conference on Digital Forensics, Security and Law*, number 2. Embry-Riddle Aeronautical University Commons, 2022. 2, 8

[29] David G. Shatwell, Ishan Rajendrakumar Dave, Sirnam Swetha, and Mubarak Shah. Gt-loc: Unifying when and where in images through a joint embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2025. 1, 2, 4, 6, 7, 8, 3, 5

[30] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020. 2

[31] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 20020–20029, 2022. 3

[32] Sirnam Swetha, Mamshad Nayeem Rizve, Nina Shvetsova, Hilde Kuehne, and Mubarak Shah. Preserving modality structure improves multi-modal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21993–22003, 2023. 3

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6

[34] Tsung-Hung Tsai, Wei-Cih Jhou, Wen-Huang Cheng, Min-Chun Hu, I-Chao Shen, Tekoing Lim, Kai-Lung Hua, Ahmed Ghoneim, M Anwar Hossain, and Shintami C Hidayati. Photo sundial: estimating the time of capture in consumer photos. *Neurocomputing*, 177:529–542, 2016. 2

[35] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2

[36] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 8, 4

[37] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 2621–2630, 2017. 2

[38] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 37–55. Springer, 2016. 2

[39] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020. 6

[40] Angelos Zavras, Dimitrios Michail, Begüm Demir, and Ioannis Papoutsis. Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment. *arXiv preprint arXiv:2402.09816*, 2024. 3

[41] Menghua Zhai, Tawfiq Salem, Connor Greenwell, Scott Workman, Robert Pless, and Nathan Jacobs. Learning geo-temporal image features. In *British Machine Vision Conference*, 2019. 1, 2, 6, 7, 8, 5

[42] Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2749–2754, 2024. 2

[43] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 2

[44] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022. 2

[45] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19370–19380, 2023. 1