Temporal Resilience in Geo-Localization: Adapting to the Continuous Evolution of Urban and Rural Environments

Fabian Deuser
University of the Bundeswehr Munich
Neubiberg, Germany
fabian.deuser@unibw.de

Wejdene Mansour
Technical University of Munich
Munich, Germany
wejdene.mansour@tum.de

Hao Li
Technical University of Munich
Munich, Germany
hao_bgd.li@tum.de

Konrad Habel
University of the Bundeswehr Munich
Neubiberg, Germany
konrad.habel@unibw.de

Martin Werner
Technical University of Munich
Munich, Germany
martin.werner@tum.de

Norbert Oswald
University of the Bundeswehr Munich
Neubiberg, Germany
norbert.oswald@unibw.de

Abstract

Static cross-view geo-localization datasets fail to capture the dynamic nature of real-world environments, as they do not account for rapid urban development and seasonal changes. As a result, models trained on such datasets experience degraded performance when confronted with more recent data, as they struggle to adapt to temporal variations such as newly constructed buildings or changing landscapes. To accurately assess the performance gap and evaluate model robustness, it is essential to use temporally diverse data that allows us to measure how well models can handle temporal shifts and remain resilient to such changes. To address this need, we have enriched the CVUSA dataset with recent satellite and Street View imagery, creating the CVTemporal dataset. This enhanced dataset is critical for testing how well geo-localization models can adapt to temporal discrepancies and identify persistent, invariant features.

In this work, we also examine the impact of temporal changes on the performance of selected well-known cross-view geo-localization models.Furthermore, we present a re-ranking approach based on existing satellite imagery from both datasets, which leads to significant performance improvements. Despite temporal variations in the data we achieve with our models remarkably good results especially on R@1. Additionally, we investigate strategies for identifying which temporally changed data should be collected to update pre-trained models, minimizing the labor-intensive process of recollecting entire datasets. Experiments with the CVUSA datasets demonstrate that it is possible to improve temporal alignment and model performance with only a small fraction of newly collected data.

1. Introduction

Accurate geo-localization is the foundation of many technological applications. It is critical for autonomous driving, enabling vehicles to navigate seamlessly and manage route uncertainty [1, 5]. In agriculture, precise location data underpins targeted pesticide application, increasing efficiency and sustainability [3]. In tourism, customized self-guided tours use localization to enhance the visitor experience [7]. Urban environments are complex and include buildings as well as other structures that can interfere with satellite signals, a phenomenon known as the urban canyon effect. To address these challenges, cross-view geo-localization presents a viable alternative or an additional refinement. This technique matches real-time street-level imagery from vehicles against a database of satellite images with known coordinates. The geo-position of the vehicle is then determined based on the satellite image that best aligns with the Street View. Recent advancements in cross-view geo-localization are leveraging deep learning for image retrieval, utilizing deep learning models to embed both satellite and street-level images [4,16,20,22,29,31]. The process

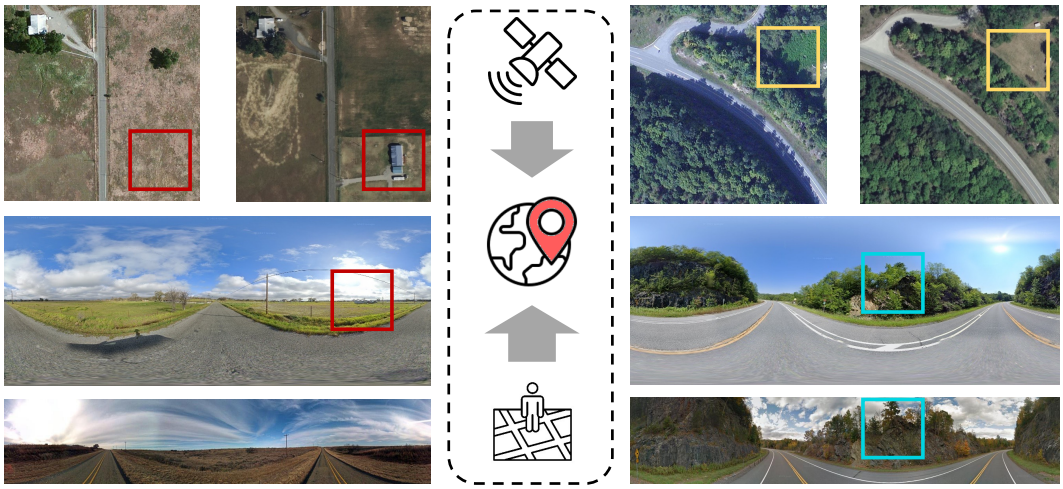

Figure 1. **Examples of the new CVTemporal dataset.** On the left, we observe structural changes (red) that appear in both the new satellite image and street image, whereas in version 1.0, these are not present. On the right side, we see different light influences (yellow) or changes in the flora (turquoise).

involves searching for the nearest neighbor based on similarity metrics, like cosine similarity, to find the best match.

Despite advancements in geo-localization datasets [10, 22, 24, 30], the geo-temporal dynamics reflecting continual environmental and structural changes remain underexplored. Seasonal shifts, weather impacts on road conditions, and urban developments dramatically alter landscapes, yet existing work [15, 25] has not fully captured these dynamics. Our contribution leverages updated Street View and satellite imagery to analyze temporal shifts, thus offering a deeper understanding of geo-temporal dynamics in cross-view geo-localization.

To address the challenge of outdated geo-localization data, we introduce the CVTemporal dataset, an enhanced version of the original CVUSA dataset [22], which includes newly collected Street View and satellite imagery from the same locations. This update better reflects current environmental conditions, as shown in Figure 1. Our primary goal is to provide a robust resource for evaluating state-of-the-art geo-localization models under real-world temporal variations. We begin by evaluating several baseline models, widely accepted in the community, that have been pre-trained on the original CVUSA 1.0 dataset. Based on this evaluation, we identify the best performing model as the basis for further optimization. Our focus is on improving its effectiveness against temporal changes.

Given the labor-intensive process of updating Street View data, we use the more readily available satellite imagery to estimate which specific Street View segments need to be updated. This targeted approach ensures that we can significantly improve the model's adaptability to temporal changes without the need for extensive data recollection. In addition, we are exploring a re-ranking strategy using older satellite data to further refine the model's performance, with the goal of balancing efficient data management with maintaining robust and accurate geo-localization capabilities.

To summarize, our key contributions in this work are:

- The introduction of CVTemporal, an enhanced geo-localization dataset incorporating the latest satellite and Street View imagery as a baseline to assess the impact of temporal variations.
- A comprehensive evaluation of state-of-the-art cross-view geo-localization models, pinpointing their deficiencies in temporal robustness.
- We investigate selection strategies to determine the minimum amount of newly collected data needed to effectively update our model, allowing it to adapt to temporal shifts using both current and historical satellite imagery.

2. Related Work

The first dataset for cross-view geo-localization, CVUSA, introduced by Workman et al. [22] featured over 1.5 million geo-tagged street and Flickr images corresponding to 880k aerial views. The images are distributed over the whole USA and feature a wide variance of different urban and rural areas. In their work, they used off-the-shelf CNN features for geo-localization. In a further iteration of the dataset, Zhai et al. [24] sampled a smaller subset of 44, 416 satellite and Street View pairs to train a deep learning pipeline for cross-view geo-localization.

The distinct flora and architectural elements of each location call for representative datasets of the area of interest.

Liu et al. [10] collected CVACT with an urban-focused setting around Canberra (Australia), whereas the University-1652 [27] dataset is set around university buildings.

More datasets have been compiled over the years to accommodate the emergence of various new subtasks and challenges in the field of geo-localization [9, 14, 17]. Unlike University-1652, SUES-200 [28] captures various scenes around a university in Shanghai at four different altitudes from 50 drone views, totaling 40k images. The scarcity of panoramic views in numerous regions of the world motivated the collection of sequences of limited Field-Of-View images in [26]. The spatial embeddings retrieved from the ground image sequence are consolidated with a Temporal Feature Aggregation Module. To enrich the non-panoramic geo-localization problem with valuable information, Vyas et al. [19] introduce the first cross-view video dataset. GAMa consists of $51,535$ ground videos, 40 seconds each, selected from the BDD100k dataset [23], paired with a matching large overhead image. For each one-second clip, GPS data is employed to extract a centered tile from the aerial view, which is input to the localization network with an aerial image encoder and a ground video encoder.

In their recent work, Ma et al. [12] take on the cross-time challenge and propose MTGL40-5, a multi-temporal geo-localization dataset solely for high-resolution satellite views of ports and airports. MTGL40-5 highlights the substantial changes in the satellite images over five years, which impair the cross-time image matching accuracy. Similarly, the Cross-View Time (CVT) dataset proposed by Salem et al. [15] comprises over 98k ground views sampled from 50 outdoor webcams, capturing various seasons and different times of the day.

The task of Temporal Domain Adaption in Earth Observation applications necessary due to temporal distribution shifts for the same geographical area is an under-explored topic, considering the short revisit time of satellites. To the best of our knowledge, Capliez et. al [2] is the latest of the few works to research this topic in depth, focusing on temporal transfer learning in land cover mapping between a source and a target year (domain).

3. CVTemporal

3.1. Problem Statement

Current datasets like CVUSA 1.0 [22, 24], CVACT [10] and VIGOR [30] for cross-view geo-localization are static and do not represent an important aspect of the world, namely the temporal dimension. Other dynamic datasets like CVT [15] or MTGL40-5 [12] choose not to use Street View images and thereby miss out on their detailed visual information. Additionally, both datasets are limited to a single view, either ground-level or satellite, which results in less precise and robust outcomes. As we show in Figure 1, our enhanced version of the CVUSA dataset features different aspects of temporal variations, e.g., changes in building structures, seasonal changes in vegetation, or improved quality of the captured images.

3.2. Data Collection

For our data collection, we use the provided locations of the CVUSA 1.0 dataset. We query the BingMap API for satellite imagery, and due to its tile-based nature, we sample a 5×5 grid around the given location. The original version of the dataset is center-aligned, meaning that the Street View is captured in the center of the satellite imagery. Using the GPS coordinates provided by the BingMap API, which originate from the upper left corner of the imagery, we calculate the necessary offset to accurately align the old and new images, and then crop them to a uniform size of 750×750 pixels. We use the same zoom level of 19, which corresponds to a scaling of 30 centimeters per pixel.

To acquire the 360° panorama views, we use the provided locations to sample the newest available images. For both satellite and Street Views, some of the imagery was taken before 2015, so these images are not very different from those in the original CVUSA dataset. Unlike CVUSA 1.0, which crops out the top and bottom of Street View images to omit unnecessary sky and street details, we retain this information in our dataset. This decision allows us to capture important features, such as high-rise buildings and street markings, that are essential for accurate geo-localization. As a result, our Street View images have a resolution of 1024×2048, compared to the 224×1232 resolution used in CVUSA 1.0. Depending on when the street images were taken, there are black borders in the lower part of the image which we crop, and then resize the images to 1024×2048 to have the same proportions as newer street images. Furthermore, we align the images similar to CVUSA 1.0 by the given orientations. As a result, the geographic north is in the middle of the street image and the upper middle of the satellite image. An example of this orientation can be found in the supplementary material.

3.3. Dataset Statistics

Our CVTemporal dataset consists of $35,313$ training and $8,828$ validation pairs. It is compatible with the version 1.0, allowing for evaluations between new satellite images and old Street View images, and vice versa. The slight difference in the number of images is due to occasional unavailability of images from the API. In Figure 2, we show the capture date distribution of the acquired images for both the satellite and the street images. This statistic also confirms our assumption that new satellite images are easier to obtain, as about 99.65% of all satellite images are acquired after 2015. The situation is different for panoramic images, where only about 56.81% are acquired after 2015. 2015 was

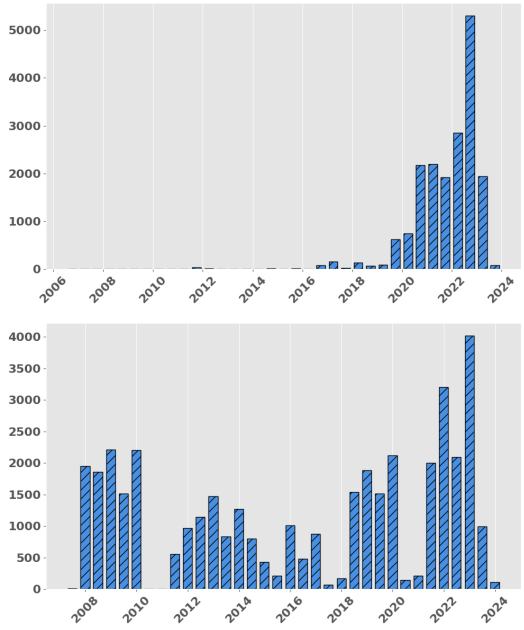

Figure 2. Distribution of the recording date of the satellite images (above) and the street images (below) in 6-month aggregation.

chosen as the criterion because CVUSA 1.0 was crawled in this year. In both cases, however, the APIs did not provide timestamps for every image, so we can only assume that the data distribution remains the same for the images without timestamps.

3.4. Temporal Data Distillation for Geo-localization

Our analysis shows that satellite imagery is updated more frequently than Street View imagery, making it a valuable indicator for identifying which Street Views need to be refreshed to reflect the latest environmental and structural changes. Similar to dataset distillation [21], which condenses a dataset to its most essential elements for model training, our approach focuses on pinpointing the most impactful parts of the dataset that require updating. This allows us to selectively refresh the data rather than indiscriminately updating the entire dataset, ensuring that the model remains current and effective with minimal effort.

We propose to use the frequent updates of satellite imagery as a metric to guide our selective refresh of Street View data. This method not only prioritizes recollection efforts toward areas of significant change, but also aligns with the broader goal of maintaining a lean, efficient dataset that adapts over time without the exhaustive resource requirements typically associated with full data recollection. In Section 4.2, we explore various strategies for using updated satellite imagery as a strategic filter to effectively select which data segments should be recollected. This ensures that our geo-localization models remain both accurate and adaptable to temporal changes without the unnecessary overhead of broader data collection.

4. Methodology

In our first experiments, we test the generalization of previously trained approaches. Therefore, we compare four well-known models TransGeo [29], SAIG-D [31], CDE [18] and Sample4Geo [4] on the new data. All models are pre-trained on CVUSA 1.0 and through the different architectures we feature a wide range of comparisons. Since the resolution and the aspect ratio changed between the two versions of CVUSA we adapt the inference resolution. To allow a fair comparison between the individual models, we crop the upper and lower parts of the Street View evenly, but not the width. This reflects the aspect ratio of CVUSA 1.0 where this crop was also used. Thus, we use the same aspect ratio on which they were trained for the comparison of all models. Figure 3 provides an overview of the initial phase of our experiments. All models utilize an image encoder pre-trained on CVUSA 1.0. During inference, we introduce the newly acquired satellite and Street View imagery, encode both, and use cosine similarity to determine a localization match. As illustrated in the Figure 3, some environmental changes are quite drastic, and while both the satellite and Street View images capture these changes, the models struggle to generalize to these new conditions. We evaluate both versions of the dataset to demonstrate the degradation in performance when images from the same locations, but taken at different times, are used for inference.

4.1. Model Overview

TransGeo is tested with 320×320 for the satellite images and 112×616 for the Street View images. For the MLP-Mixer approach SAIG-D, we stick to the resolution 256×256 for the satellite images and 128×512 for the Street View images. Similarly, we need to resize to 112×616 as the input resolution for CDE and 256×256 for the satellite images. Toker et al. adapt a GAN to translate the satellite image into a street image to make the subsequent matching task easier, thus the SAFA-retrieval model [16] receives the 112×616 size for both views. Sample4Geo [4] infers with 384×384 for the satellite images and 140×768 for the Street View images.

All four architectures are contrastively trained, with the difference that TransGeo, CDE, and SAIG-D use triplet loss, and Sample4Geo uses InfoNCE loss. Additionally, Sample4Geo uses hard negative sampling to leverage harder examples during training.

After selecting the best-performing model from our evaluations, we re-train it on all newly collected data to establish an upper performance benchmark achievable with the most recent dataset. We then select a small, strategically chosen portion of this newly acquired data to fine-tune another

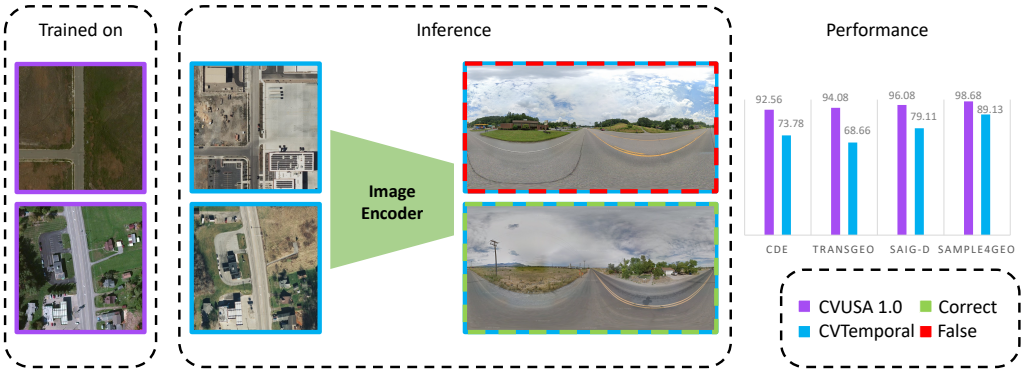

Figure 3. Examples of the CVUSA 1.0 satellite images (purple), CVTemporal (blue) and the predictions during inference with a pre-trained model [4]. While the model can correctly assign the lower satellite image, it fails with the upper one, although temporal changes can be seen on both. The performance bar chart indicates the performance drop for several approaches [4, 18, 29, 31] between the two versions of the CVUSA dataset and highlights the challenge of temporal alignment.

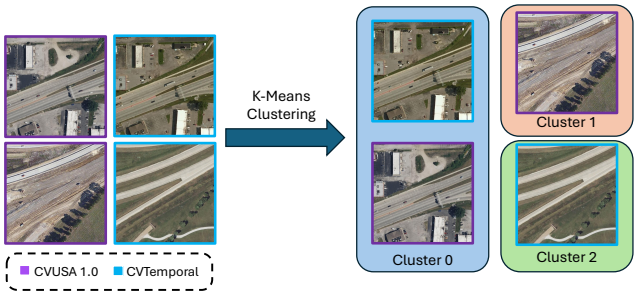

Figure 4. Our selection strategy based on satellite imagery. We apply k-Means Clustering to both historical and recent satellite images, creating pseudo-classifications without explicit labels. By evaluating the deviations between clusters of old and new images, we accurately select the samples needed for adaptation training.

model that has been pre-trained on the older dataset. The purpose of this step is to measure how closely we can match the performance of the fully updated model using only a minimal subset of the new data.

4.2. Data Selection Strategies

In our effort to refine model training with strategically selected data, we compare three different selection strategies.

The first strategy is straightforward: after collecting all the images, we evaluate the model's performance to determine which training examples are incorrectly predicted. This method allows us to directly assess the value of both correctly and incorrectly predicted data points for refining the model. However, this approach is only practical when the entire dataset has already been collected, and serves primarily as a first step in assessing the usefulness of the information.

The second strategy uses simple vector change analysis. This method quantifies the magnitude of differences by measuring changes in feature vectors between new and existing satellite images. While this technique provides a simple way to detect significant changes, it overlooks important issues such as spatial coherence and contextual relevance.

The third strategy, illustrated in Figure 4, uses k-Means clustering to categorize both historical and recent satellite images using feature vectors from a pre-trained model in the absence of labels. This pseudo-classification method groups paired old and new images, allowing for the detection of significant shifts in cluster membership. Images with minor changes remain clustered together, while those with significant deviations form new clusters. This not only emphasizes visual and spatial relationships, but also enhances contrastive training through implicit hard negative sampling. As suggested in [4], the use of hard negatives, which are visually similar but differently categorized images, is beneficial for training models using the InfoNCE loss function, thereby improving the effectiveness of the model by emphasizing more challenging comparisons.

4.3. Time-invariant Re-Ranking

We propose a simple but effective re-ranking strategy using historical satellite data. Typically, a similarity matrix (e.g., cosine similarity) is computed between satellite and Street View features to match images and determine geo-locations. In our approach, we extract features from satellite images over two time epochs, CVUSA 1.0 and the newly collected CVTemporal, and fuse them by computing their mean. This fusion creates features approaching time invariance that emphasizes consistent geographic elements and improves robustness to temporal changes. The advantage of this approach is its scalability - as satellite imagery from additional time steps becomes available.

4.4. Implementation Details

For our contrastive training, we apply the InfoNCE loss as in Sample4Geo. When adapting the pre-trained Sample4Geo, we train for 10 epochs, using a learning rate of $5e{-}4$ with a cosine decay schedule and a warm-up of 1 epoch. Conversely, retraining on the full dataset involves 40 epochs with a learning rate of $1e{-}3$ and initialization from an ImageNet pre-trained ConvNeXt-Base [11]. Both training strategies involve augmenting the images with random rotation, horizontal flipping, color jitter, blur, sharpening, grid dropout, and coarse dropout. Specifically, for random rotation, we rotate the satellite view in 90-degree increments and then roll the pseudo Street View accordingly to maintain alignment.

5. Evaluation

5.1. Pre-trained on CVUSA 1.0

In our experiments, we evaluate the four state-of-the-art models from Section 4.1 with various data. The initial tests with CVUSA 1.0 establish the baselines. Afterward, we evaluate the models in different settings using the CVTemporal dataset. In the first setting (CVTemporal New Sat), we replace the satellite images with the newly captured ones. In the second setting (CVTemporal New Pano), we swap only the Street Views and in the last setting, both image sources are replaced by all newly captured images. As mentioned before, for testing we crop the test images by keeping the same aspect ratio to look similar to the training images from CVUSA 1.0. Otherwise, depending on the approach, the performance drops even more than what is recorded in Table 1, as we show in our supplementary material.

Table 1 indicates that the Sample4Geo approach outperforms others in handling temporal variations in images, demonstrating superior generalization. Models with higher scores on CVUSA 1.0 generally maintain their performance on version 2.0, with TransGeo being an exception. Notably, Sample4Geo, a CNN-based model, exhibits significant resilience to image changes, particularly when only one type of input is updated.

It's worth noting that new satellite imagery results in a less sharp performance decline than introducing new Street View imagery. In particular, the Sample4Geo approach shows only minimal performance degradation under these conditions. This is likely because the dominant feature used for mapping to Street View data - the overall road structure - remains largely consistent in most satellite images, even when new data is introduced. As we can observe, there is a strong performance decrease as soon as both views are updated, underlining our proposal to strengthen temporal alignment.

Dataset	R@1	R@5	R@10	R@1%	Δ in % for R@1
CVUSA 1.0					
CDE [18]	92.56	97.55	98.33	99.57	-
TransGeo [29]	94.08	98.36	99.04	99.77	-
SAIG-D [31]	96.08	98.72	99.22	99.86	-
Sample4Geo [4]	**98.68**	**99.68**	**99.78**	**99.87**	-
CVTemporal (New Sat)					
CDE [18]	84.87	93.88	95.81	98.96	−9.06%
TransGeo [29]	81.48	92.78	95.04	98.81	−15.46%
SAIG-D [31]	88.19	96.12	97.51	99.42	−8.94%
Sample4Geo [4]	**95.21**	**98.51**	**98.96**	**99.61**	−3.64 %
CVTemporal (New Pano)					
CDE [18]	76.46	87.39	90.39	96.09	−21.05%
TransGeo [29]	75.04	89.97	93.04	98.19	−25.37%
SAIG-D [31]	81.64	92.31	94.59	98.34	−17.68%
Sample4Geo [4]	**91.30**	**96.04**	**97.02**	**98.56**	−8.08%
CVTemporal					
CDE [18]	73.78	85.48	88.66	94.99	−25.45%
TransGeo [29]	68.66	84.94	89.06	96.83	−37.02%
SAIG-D [31]	79.11	90.00	92.78	97.85	−21.45%
Sample4Geo [4]	**89.13**	**94.87**	**96.11**	**98.04**	−10.71%

Table 1. **Comparison between state-of-the-art approaches pre-trained on CVUSA 1.0 and evaluated on both versions of the CVUSA dataset.** The aspect ratio of the new Street View images are adjusted to match the appearance of the CVUSA 1.0 dataset. In our supplementary we provide results without this adjustment.

5.2. Training on CVTemporal

After evaluating several approaches, we selected Sample4Geo [4] as our base model due to its relatively low loss of generalization compared to other methods tested. Our goal is to equalize the performance of CVUSA 1.0 and CVTemporal datasets, using only a small portion of data for retraining. Therefore, we test different selection strategies in Table 2, where the subset indicates the amount of training data used. In addition to our selection strategy, we emphasize the impact of our proposed re-ranking method, which provides an additional performance boost. We conducted experiments using 1 %, 5 %, 10 %, 20 %, and 30 % (denoted as x) of the dataset to highlight the importance of proper selection strategies on smaller subsets. While 30 % gives the best results, the 10 % subset seems to be a good compromise between computational cost and achievable performance, since false predictions are only 10.87 %. We report the remaining subset portions in our supplementary material. To compare what the maximum achievable performance is, we also train the Sample4Geo model on the full dataset, denoted by 100 %.

However, the false prediction approach results in poor performance due to label noise and the presence of examples that are particularly difficult to distinguish. These hard examples, if overly difficult, can degrade model performance. The second selection method (Magnitude) uses the magnitude of the vector change between the old and new satellite images. We select x% of the samples with the largest change and achieve an improvement over the false-prediction selection.

Model Type	R@1	R@5	Trained on		Subset
			CVUSA 1.0	CVTemporal	
Baseline	89.13	94.87	X	-	100%
Baseline + ReRank	92.64	96.45	X	-	100%
False preds	91.40	97.97	-	X	10%
Magnitude	91.75	97.55	-	X	10%
Clustering	92.10	97.53	-	X	10%
Clustering + ReRank	**94.65**	**98.51**	-	X	10%
False preds	93.70	98.70	-	X	30%
Magnitude	93.71	98.44	-	X	30%
Clustering	93.95	98.44	-	X	30%
Clustering + ReRank	**95.60**	**99.00**	-	X	30%
Full	95.00	98.44	-	X	100%
Full	95.48	98.83	X	X	100%
Full + ReRank	**97.21**	**99.24**	X	X	100%

Table 2. **Comparison of Selection Strategies:** Training on the full dataset with re-ranking delivers the best overall performance. However, when training on a subset of the data, the k-Means clustering method with re-ranking outperforms other selection strategies, achieving the highest R@1. The metrics R@1 & R@5 are reported on the validation split of the CVTemporal dataset.

Our third selection method (Clustering) based on k-Means clustering performs best when only training on $x\%$ of the newly collected imagery. Once we introduce historical satellite data as a re-ranking in our inference, denoted by ReRank in Table 2, the performance improves further. Through this combination, we achieve competitive performance when compared to a full retraining (Full) on the CVTemporal dataset. Through our re-ranking, where we extract and average the features from the old and new satellite images, we created time-invariant features. Since the uneven data distribution shown in Figure 2 indicates that satellite imagery is updated more frequently, while some Street View imagery remains outdated, features from historical satellite imagery are beneficial in inference because they help align with older Street View data that may not reflect recent changes. To assess whether re-ranking benefits model training on new data, we re-ranked the Full model predictions using the old satellite images (Full+ReRank). Performance improved, likely due to the aforementioned distribution.

Our results confirm the effectiveness of the clustering method described in Section 4.2 and demonstrate its superior performance when training on a subset of the data without prior knowledge of which data to update. The re-ranking strategy we explore improves performance by providing time-invariant features that are applicable both when training on a subset and when retraining on the full dataset. This approach effectively narrows the performance gap and maintains competitiveness in the field.

6. Ablation and Visualization

In our ablation study, we explore the optimal selection of k-Means cluster size for effective data categorization. In addition, we provide visual insights into examples where temporal changes are evident, enhancing our understanding of the dynamics of the model's performance.

Cluster Size	R@1	R@5	R@10	R@1%
5	92.03	97.37	98.32	99.58
10	92.00	97.44	**98.53**	99.64
20	**92.10**	**97.53**	**98.53**	99.64
30	92.07	97.50	**98.53**	**99.69**

Table 3. **Comparison between different cluster sizes based on the 10 % clustering subset.**

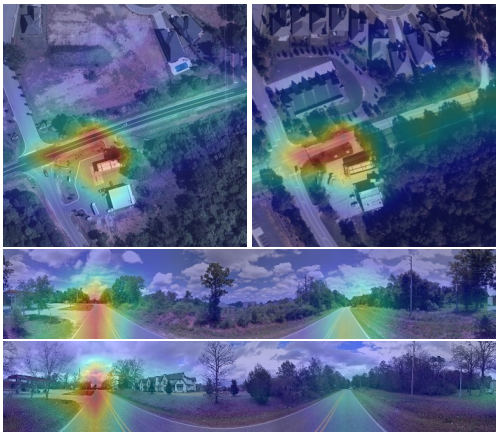

Figure 5. **Heatmaps for a correct prediction with our adaption model.** CVUSA 1.0 data (left satellite and upper Street View images) show clear temporal changes compared to the CVTemporal images (right satellite and bottom Street View). The model still focuses on the street layout, with new building information playing a minor role.

6.1. Cluster Size

In our analysis of different k-Means cluster sizes, see Table 3, we found that the number of clusters did not significantly affect the overall performance improvement. Despite varying cluster sizes, all configurations consistently outperformed other methods such as change vector analysis and using incorrectly predicted samples for selection. This suggests that while the choice of cluster size is flexible, the shift in cluster assignment is more critical to improving model performance.

6.2. Visualization

Human spatial reasoning employs landmarks such as streets, buildings, and vegetation. The investigation of our fully trained model's activations, as shown in Figure 5, reveals analogous patterns. Selected examples illustrate accurate predictions amidst significant structural changes, including new constructions and altered vegetation. The com-

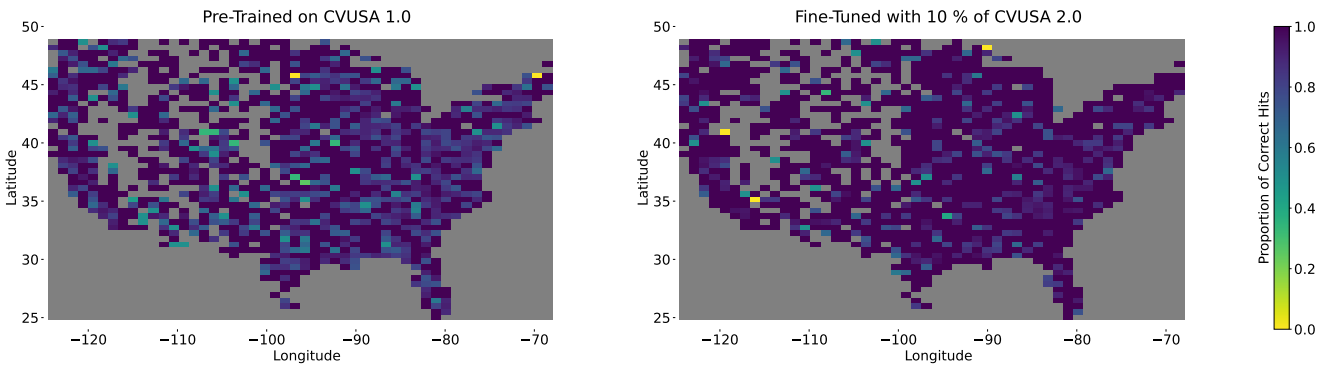

Figure 6. **Geographical Distribution of Model Predictions:** Predictions are aggregated in a 50×50 grid to examine regional performance discrepancies. The plot demonstrates areas where model performance varies, highlighting regions with frequent prediction errors and showing an overall improvement in model accuracy. Grey areas indicate that there are no samples from the dataset in this area.

parative analysis utilizes satellite and Street View images from the CVUSA 1.0 dataset against activations from the CVTemporal dataset.

The findings indicate minimal variation in model activations over time, with a predominant reliance on basic features like road layouts, and a lack of emphasis on new buildings. This underscores our assertion that temporal alignment remains unresolved, with the model adopting a simplistic approach to learning.

In our ablation study, we examined the geographic distribution of predictions by grouping them into a 50×50 grid, Figure 6. This allowed us to assess whether the model had difficulty with certain regions. The analysis revealed that regions with higher rates of incorrect predictions tended to be remote areas characterized by roads with few distinctive building features. These areas pose significant challenges to the model, primarily due to the lack of unique landmarks and the homogeneity of the surrounding vegetation typical of these geographic regions. Such environments often result in higher confusion rates, especially when the model relies on less diverse features to make predictions.

7. Conclusion

In this work, we present CVTemporal, an enhanced version of the established CVUSA dataset that allows for more accurate evaluation of cross-view geo-localization models under temporal variations. As the world continues to change, it is critical to have diverse datasets that reflect these temporal shifts and provide a robust basis for further evaluation of model performance. CVTemporal provides a rich variety of conditions to address the evolving challenges posed by temporal variations, setting a new benchmark for evaluating cross-view geo-localization models in dynamic environments.

Our approach improves model performance by combining targeted data selection with a powerful re-ranking strategy that leverages historical data. While identifying specific Street Views for updating improves performance after retraining, the re-ranking approach itself provides significant gains without additional training. This re-ranking technique consistently improves temporal alignment and outperforms baseline models, whether the models are retrained or not. Together, these strategies demonstrate that with careful data selection and the application of re-ranking, models can effectively adapt to ongoing change while minimizing the need for extensive new data collection.

8. Discussion

Our experiments on the CVTemporal dataset vividly illustrate the significant impact that data refreshing has on model performance. Consistent updating of the dataset is critical, as evidenced by the notable improvements in models trained on fresh data.

However, there is a caveat to the timestamp distribution in Figure 2: not all Street View images are updated, which may lead to an overestimation of performance improvements. Pre-trained models are likely to perform similarly on unchanged images, suggesting that our performance estimates may be overly optimistic. This underscores the importance of a more thorough investigation of the update process to ensure truly comprehensive updates.

Additionally, enhancing model adaptability and understanding for complex geo-localization scenarios requires the integration of diverse spatiotemporal factors into our datasets. Including temporal dependencies, spatial heterogeneity, and varying zoom levels will enable models to more effectively handle challenges such as cross-country or temporally variant geo-localization and variations in image resolution. This approach aligns with recommendations by [6, 8, 13] and is vital for advancing the field of geo-localization.

References

[1] Eli Brosh, Matan Friedmann, Ilan Kadar, Lev Yitzhak Lavy, Elad Levi, Shmuel Rippa, Yair Lempert, Bruno Fernandez-Ruiz, Roei Herzig, and Trevor Darrell. Accurate visual localization for automotive applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[2] Emmanuel Capliez, Dino Ienco, Raffaele Gaetano, Nicolas Baghdadi, and Adrien Hadj Salah. Temporal-domain adaptation for satellite image time-series land-cover mapping with adversarial learning and spatially aware self-training. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:3645–3675, 2023. 3

[3] Nived Chebrolu, Philipp Lottes, Thomas Läbe, and Cyrill Stachniss. Robot localization based on aerial images for precision agriculture tasks in crop fields. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1787–1793, 2019. 1

[4] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16847–16856, October 2023. 1, 4, 5, 6

[5] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21621–21631, 2023. 1

[6] Michael F Goodchild and Wenwen Li. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, 118(35):e2015759118, 2021. 8

[7] Chris D. Kounavis, Anna E. Kasimati, and Efpraxia D. Zamani. Enhancing the tourism experience through mobile augmented reality: Challenges and prospects. *International Journal of Engineering Business Management*, 4:10, 2012. 1

[8] Hao Li, Jiapan Wang, Johann Maximilian Zollner, Gengchen Mai, Ni Lao, and Martin Werner. Rethink geographical generalizability with unsupervised self-attention model ensemble: A case study of openstreetmap missing building detection in africa. In *Proceedings of the 31th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '23, New York, NY, USA, 2023. Association for Computing Machinery. 8

[9] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. *arXiv e-prints*, pages arXiv–2208, 2022. 3

[10] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5624–5633, 2019. 2, 3

[11] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6

[12] Jingjing Ma, Shiji Pei, Yuqun Yang, Xu Tang, and Xiangrong Zhang. Mtgl40-5: A multi-temporal dataset for remote sensing image geo-localization. *Remote Sensing*, 15:4229, 08 2023. 3

[13] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*. PMLR, 2023. 8

[14] 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. 3

[15] Tawfiq Salem, Scott Workman, and Nathan Jacobs. Learning a dynamic map of visual appearance. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12432–12441, 2020. 2, 3

[16] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1, 4

[17] Yuxi Sun, Yunming Ye, Jian Kang, Ruben Fernandez-Beltran, Shanshan Feng, Xutao Li, Chuyao Luo, Puzhao Zhang, and Antonio Plaza. Cross-view object geo-localization in a local region with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 3

[18] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 4, 5, 6

[19] Shruti Vyas, Chen Chen, and Mubarak Shah. Gama: Cross-view video geo-localization. In *European Conference on Computer Vision*. Springer, 2022. 3

[20] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, 2021. 1

[21] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 4

[22] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3961–3969, 2015. 1, 2, 3

[23] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. pages 2633–2642, 06 2020. 3

[24] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, pages 867–875, 2017. 2, 3

[25] Menghua Zhai, Tawfiq Salem, Connor Greenwell, Scott Workman, Robert Pless, and Nathan Jacobs. Learning geo-temporal image features. *arXiv preprint arXiv:1909.07499*, 2019. 2

[26] Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Cross-view image sequence geo-localization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2913–2922, 2023. 3

[27] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. pages 1395–1403, 10 2020. 3

[28] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4825–4839, 2023. 3

[29] 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. 1, 4, 5, 6

[30] 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, 3

[31] Yingying Zhu, Hongji Yang, Yuxin Lu, and Qiang Huang. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572*, 2023. 1, 4, 5, 6