GeoSight: Enhancing Object Geolocalization with Visual Similarity and Coordinate Referencing

Jooho Kim*
Institute for a Disaster Resilient Texas (IDRT)
Texas A&M University
College Station, TX, USA

Junho Kim
Department of Computer Science
Rice University
Houston, TX, USA

Sayok Bose
Department of Computer Science
Texas A&M University
College Station, TX, USA

Sarah Brasseaux
Department of Computer Science
Texas A&M University
College Station, TX, USA

Abstract

Imagery collected from community-driven efforts and field surveys after disasters provides critical data for assessing damage severity, location, and extent. These images enhance disaster preparedness and response by updating outdated damage maps. However, relying on embedded coordinate metadata for disaster assessments, such as flood mapping or tornado path analysis, often introduces spatial inaccuracies. These errors arise due to discrepancies between the camera's capture position and the actual location of objects within the image, leading to geolocation inconsistencies that undermine the reliability of disaster analytics and AI-based decision-support systems. This study presents **GeoSight**, a novel framework for improving object geolocation accuracy in disaster imagery by integrating image-based spatial referencing with coordinate data. Using a case study based on the NOAA Damage Assessment Toolkit, we evaluate GeoSight's performance in reducing geolocation errors within the 2023 tornado damage dataset from Norman, Oklahoma. The framework assesses four image similarity models (DINO, DreamSim, CLIP, and ViT) for retrieving the most relevant reference images to enhance spatial alignment. Among these models, DreamSim achieved the highest performance, with 86% accuracy in the top-3 rankings and 93% in the top-5 rankings, followed by CLIP, DINO, and ViT. GeoSight successfully corrects location errors by an average of 26.12 meters, and with a maximum distance offset of 46.52 meters, significantly improving disaster mapping accuracy. Our findings highlight the importance of refining geospatial data for disaster response and recovery. By reducing geolocation errors, GeoSight enables more reliable damage assessments, optimized resource allocation, and improved emergency decision-making.

github.com/jk-junhokim/geosight

Keywords: similarity, geolocation, georeference, disaster, DreamSim

1 Introduction

Geodatabases are fundamental to a wide range of applications from disaster mitigation to recovery. Their accuracy and reliability are critical for supporting informed decision-making, particularly in high-stakes scenarios such as emergency management and disaster recovery. With the increasing integration of machine learning (ML) and AI-driven applications, these databases serve as foundational inputs for analytical models and predictive systems. However, uncertainties in geospatial

*Corresponding author: jooho.kim@tamu.edu

data—stemming from errors in data collection, processing, and integration—can significantly hinder their utility (Tomaszewski, 2020; Kim et al., 2025). These inaccuracies compromise spatial analyses, reduce the effectiveness of AI-based disaster assessment tools, and lead to misinformed decisions with far-reaching consequences.

In disaster management, the availability of precise and timely geospatial information directly impacts response effectiveness. Many damage assessments rely on imagery collected from field surveys and community-driven sources, which play a crucial role in mapping disaster impacts (Stephens et al., 2024). However, a major challenge arises when relying on embedded GPS metadata in images, as the recorded coordinates often reflect the camera's position rather than the actual location of the objects being captured. This discrepancy introduces spatial inaccuracies, affecting disaster assessments, flood mapping, and tornado path analyses (Zuo et al., 2021; Haque et al., 2022). Furthermore, growing privacy concerns have led to restrictions or deletions of public geodata, complicating efforts to maintain accurate datasets for disaster response.

To address these challenges, this study introduces GeoSight, a hybrid framework for object localization that integrates image-based spatial referencing with coordinate data. Rather than relying solely on embedded geotags, GeoSight leverages image similarity methods and reference datasets to refine object geolocation accuracy. This approach enhances disaster-related geospatial data by systematically reducing locational uncertainty, allowing for improved decision-making in emergency response and recovery efforts.

To validate GeoSight, we conduct a case study on the 2023 tornado in Norman, Oklahoma, using damage assessment imagery from NOAA's Damage Assessment Toolkit (DAT). We evaluate the effectiveness of four similarity-based models—DINO, DreamSim, CLIP, and ViT—in retrieving the most relevant images for geolocation correction. Model performance is assessed using top-5 ranking results, with findings demonstrating GeoSight's ability to enhance georeferencing accuracy.

By improving the precision of disaster-related geospatial data, GeoSight enhances the accuracy of damage assessments, optimizes resource allocation, and strengthens AI-driven disaster management systems. This research contributes to ongoing efforts to ensure that geospatial technologies can reliably support critical disaster response operations.

The next section presents a detailed review of existing literature on data uncertainty and its implications for geodatabases. In Section 3, we describe the GeoSight framework, including methods and data collection for a case study. Section 4 discusses results from the case study of the 2023 tornado in Norman, Oklahoma, followed by a conclusion summarizing key findings, implications for future research, and potential applications of this work.

2 Literature Review

2.1 Image Georeferencing

Georeferencing street-level images is essential in disaster management for accurately mapping damage, improving response efforts, and integrating images into geodatabases for long-term recovery planning. Unlike satellite imagery, street-level images provide critical ground perspectives that capture localized damage, infrastructure failure, and obstructions that may not be visible from overhead sources. These images are often collected by emergency responders, unmanned aerial vehicles (UAVs), social media, and community-driven initiatives (Xing et al., 2023; Stephens et al., 2024; Li et al., 2025; Gu et al., 2025). However, integrating them into geodatabases requires precise georeferencing methods, particularly when embedded GPS metadata is missing or inaccurate.

Most disaster-related images contain embedded GPS coordinates within their metadata (i.e., EXIF data), allowing for direct georeferencing. However, issues such as missing metadata due to privacy settings, image compression, or social media platforms often require alternative georeferencing techniques (Chu et al., 2022; Comalada et al., 2025). In such cases, computer vision-based image matching techniques to align images with pre-existing georeferenced datasets have been applied—such as Google Street View (GSV), OpenStreetMap, and UAV-collected imagery—using features like Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), and Oriented FAST and Rotated BRIEF (ORB).

For images without embedded GPS coordinates, deep learning techniques have also been developed to estimate geographic locations based on visual features (Kustu & Taskin, 2023; Xin et al., 2023; Kim et al., 2025). Chu et al. (2022) applied Convolutional Neural Networks (CNNs) trained on geotagged street-level datasets to predict image coordinates by analyzing urban features, road signs, and architectural patterns. Additionally, cross-view image matching techniques compare ground-level images with aerial or satellite imagery to infer spatial positioning, improving georeferencing accuracy (Tian et al., 2017; Rao et al., 2023). These methods are particularly useful in disaster scenarios where satellite images may not provide immediate post-event data.

To improve georeferencing accuracy, multi-source data fusion integrates image-based feature matching with auxiliary data such as road network information, building footprints, and textual information extracted from images (e.g., street names, business signs, house numbers). Text recognition models can extract and cross-reference addresses or phone numbers from images with GIS databases to infer locations (Sathianarayanan et al., 2024). Additionally, 3D point cloud alignment, generated through Structure-from-Motion and LiDAR, can refine image placement by matching images to existing geospatial models of urban environments (Diaz et al., 2022; Stilla & Xu, 2023).

Crowdsourced and social media imagery contribute significantly to disaster response efforts, but their geospatial accuracy varies. To improve the locational accuracy of these data, Pereira et al. (2019) developed automated quality assessment pipelines to evaluate the reliability of crowdsourced images by analyzing GPS accuracy, timestamp consistency, and visual similarity to known locations. Chu et al. (2022) also proposed blockchain-based geodatabases to verify and authenticate georeferenced disaster images, ensuring integrity in disaster mapping applications.

Despite advances in automated georeferencing, several challenges remain. Post-disaster environments often introduce occlusions (e.g., debris blocking landmarks, damaged buildings not matching original building images), perspective distortions, and GPS errors that complicate image localization. Sensor fusion techniques—such as integrating GNSS, Inertial Measurement Units, and visual odometry—can mitigate these issues by combining multiple positioning signals for enhanced accuracy (Vo et al., 2023). Furthermore, updating geodatabases with real-time georeferenced imagery remains a key research focus to enhance situational awareness and response coordination in disaster management.

2.2 Object Detection

Object detection is a key task in computer vision that involves identifying and localizing objects within an image. Deep learning-based object detection models, particularly those using CNNs, have demonstrated high performance in this domain. These models are commonly trained on large and structured datasets (e.g., COCO (Lin et al., 2014), Pascal VOC (Everingham et al., 2015), ImageNet (Deng et al., 2009), and Open Images (Kuznetsova et al., 2020)). However, publicly available datasets, such as those mentioned above, are constrained by issues such as small scale and limited variation in image content. These limitations significantly impede the development and refinement of deep learning-based object detection methods tailored for specific applications.

Detection models trained on natural images emphasize smaller foreground objects with detailed profile features. Due to the difference in scale and positioning of objects between natural images and street view images, applying pretrained detection models to street view images limits the model's ability to capture the visual characteristics of large-scale structures like buildings (Singh & Davis, 2018). To address these challenges, we reviewed recent advancements in deep learning-based object detection, analyzing publicly available models and their characteristics. Table 1 summarizes architectural designs, datasets, and limitations for each model. Among the reviewed methods, Faster R-CNN has been widely adopted for building detection in street-view imagery (Cha et al., 2018; Lenjani et al., 2020; Wang & Zhang, 2018; Zhao et al., 2021). When fine-tuned on annotated building datasets, Faster R-CNN demonstrates strong performance, benefiting from its two-stage design that enhances accuracy and robustness.

2.3 Image Similarity and Retrieval Applications

Image similarity refers to the degree of visual resemblance between two images, a fundamental concept in computer vision and pattern recognition. This process involves identifying shared visual

Table 1: Comprehensive review of Object Detection models.

Model	Architecture	Data	Limitations
Faster R-CNN	Two-stage: Region Proposal Network + CNN	PASCAL VOC, COCO, ImageNet	Slower inference due to its two-stage architecture. (Ren et al., 2015)
YOLO	Single-stage: Grid-based CNN	PASCAL VOC, COCO, ImageNet	Lower accuracy for small objects compared to two-stage methods. (Redmon, 2016)
SSD	Single-stage: Multi-scale CNN	PASCAL VOC, COCO	Difficulty detecting small objects; less accurate than Faster R-CNN. (Liu et al., 2016)
R-FCN	Two-stage: Region-based Fully Convolutional Network	PASCAL VOC, COCO	Fixed-size region proposals may reduce accuracy for irregular shapes. (Dai et al., 2016)

features while accounting for variations in lighting, object positioning, and background settings. Traditional similarity metrics, such as PSNR (Setiadi, 2021), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018), focus primarily on pixel-level fidelity but often fail to capture high-level semantic relationships critical for complex image retrieval and recognition tasks. Recent advances in deep learning have led to the development of more advanced models for similarity analysis, including ViT (Dosovitskiy, 2020), CLIP (Radford et al., 2021), DINO (Caron et al., 2021), and DreamSim (Fu et al., 2023). These models leverage self-supervised learning and large-scale pretraining to capture both perceptual and semantic similarities, significantly improving image retrieval, classification, and object recognition. Unlike traditional methods, these deep-learning models can generalize across domains and adapt to complex imagery, making them more suitable in high-variance environments such as disaster scenarios.

In disaster management, few studies have explored the use of similarity models such as ViT and CLIP for disaster-related image retrieval and damage assessment. Ding et al. (2022) applied a Siamese Transformer Network (STN) combined with ViT-based feature extraction, enabling adaptive similarity learning to track damage progression in disaster-affected areas. Their model was designed to compare pre- and post-disaster images, ranking them based on structural damage patterns. By leveraging self-attention mechanisms, ViT captures long-range dependencies in disaster imagery, improving response prioritization. Liu et al. (2024a) used a Text-Guided Knowledge Transfer (TGKT) module to refine CLIP's pretrained embeddings for disaster-specific queries, improving retrieval accuracy in emergency scenarios. However, the model relied on structured metadata and was not fine-tuned on disaster datasets. In both studies, the authors highlighted that future research should fine-tune models on disaster datasets, optimize them for real-time use, and improve interpretability.

While only a few studies have so far explored the use of similarity models like CLIP and ViT, these early works demonstrate significant potential for improving image retrieval and damage assessment in emergency response. The ability to accurately match textual descriptions with disaster imagery, as well as compare pre- and post-disaster scenes, is critical for enhancing situational awareness and response efficiency. Given the promising results and growing interest in deep learning-based similarity analysis, it is expected that more research will emerge in the near future, further advancing the field and offering robust tools for disaster-related studies.

3 METHOD

3.1 MODEL FRAMEWORK

The model framework of GeoSight is designed to identify an object location using embedded coordinate, building detection model and similarity analysis (Figure 1). The process starts with a query dataset of images and their coordinates, which are compared to a target database of reference images. A search radius filters relevant images based on proximity to the query coordinates. If an image contains multiple objects (e.g. multiple buildings), a building detection step isolates a structure. Otherwise, similarity analysis compares the query image to the filtered target images. Finally, the model retrieves images based on similarity scores, updating the coordinate if a match image is found. Otherwise, the original coordinate is retained.

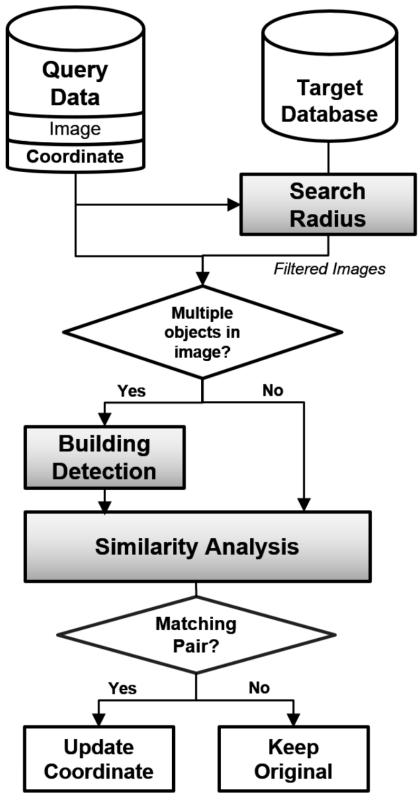

Figure 1: Framework for building detection and image matching through similarity models.

Search radius is designed to enhance computational efficiency and improve the performance of similarity analysis. As most disaster related imagery is focused on residential buildings, there are very similar building structures and external texture. So, the data in target database is filtered using coordinate metadata and the haversine formula ((Robusto, 1957)), which calculates the great-circle distance between two points on the Earth's surface based on their latitude and longitude as seen in Equation 1.

$$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{1}$$

where ϕ_1, ϕ_2 and $\Delta\lambda(= \lambda_2 - \lambda_1)$ are the latitudes and longitudes of point 1 and point 2, respectively, and r being Earth's mean radius. The haversine formula enables efficient spatial filtering by calculating distances based on geographic coordinates. This approach is widely employed in navigation and GIS applications such as disaster evacuation planning ((Swara, 2017)) and communication optimization ((Hafil et al., 2017)).

3.2 BUILDING DETECTION

The architectural design of Faster R-CNN is based on an end-to-end CNN and comprises two main components; Region Proposal Network (RPN) and Detection Network. The RPN generates candidate regions, also known as region proposals, that are likely to contain objects of interest. The Detection Network further processes these proposals, performing classification and bounding box regression to refine the locations and scales of the detected objects. This makes Faster R-CNN particularly well-suited for tasks requiring both object localization and classification, such as building detection.

We deployed Faster R-CNN for building detection based on the model from Ren (2015). We trained the Faster R-CNN model using the BEAUTY dataset ((Zhao et al., 2021)), which consists of 19,070 street-view images with 38,857 individually annotated buildings. While the dataset provides extensive annotations for buildings in urban environments, to the best of our knowledge, there are no large-scale datasets that specifically include annotated bounding boxes of damaged buildings. The currently available public BEAUTY dataset includes 4,769 images and 9,747 annotated buildings. We split the BEAUTY dataset into training, validation, and testing sets with a ratio of 75:15:10, resulting in 3,576 training images. To address the limited size of the training data and overfitting, we applied data augmentation techniques, including horizontal flipping ($p = 0.3$), color jittering (brightness = 0.2, contrast = 0.2, saturation = 0.2, $p = 0.3$), Gaussian blur (kernel size = 3–5, $p = 0.1$), Gaussian noise (variance = 10.0–50.0, $p = 0.1$), random brightness and contrast adjustments (± 0.1, $p = 0.1$), and hue-saturation-value shifts (hue = ± 10, saturation = ± 15, value = ± 10, $p = 0.1$). All images were normalized using a mean of (0.485, 0.456, 0.406) and a standard deviation of (0.229, 0.224, 0.225), followed by conversion to tensors. For optimization, we used stochastic gradient descent with a batch size of 4 images and a momentum coefficient of 0.9. The initial learning rate was set to 5.0×10^{-3}. During training, validation, and testing, all images were resized to 256×256 pixels.

3.3 SIMILARITY ANALYSIS

We applied four generalizable feature extraction models including DINO ((Caron et al., 2021)), DreamSim ((Fu et al., 2023)), CLIP ((Radford et al., 2021)), and ViT ((Dosovitskiy, 2020)). Each model is used to identify the five most similar images from the target database.

DINO (Distillation with No Labels) is a self-supervised learning model that derives meaningful representations from unlabeled data. It employs a self-distillation strategy, where a student network learns to replicate the output of a teacher network. DINO is well known for producing high-quality embeddings that capture both local and global image features ((Wanyan et al., 2024)). Leveraging the ViT as its backbone, DINO excels at encoding semantic relationships, even in datasets with complex and unstructured content.

DreamSim builds upon an ensemble of embedding extractors by fine-tuning them with synthetic datasets like NIGHTS ((Fu et al., 2023)). It incorporates a perceptual metric optimized using a triplet loss function, which ensures that the similarity between a query and a positive example (same class) surpasses the similarity between the query and a negative example (different class).

CLIP is a multimodal model trained on approximately 400 million image-text pairs. It aligns images and text within a shared embedding space. For image processing, CLIP divides images into fixed size patches, linearly embeds them, and passes them through a Transformer encoder. In image similarity tasks, CLIP computes embeddings for each image and determines relevance by measuring their cosine similarity.

ViT applies Transformer architecture to image analysis by segmenting an input image into non-overlapping patches (e.g., 16×16 pixels), flattening them, and projecting them into a sequence of embeddings. A Transformer encoder then processes these embeddings to capture the image's semantic relationships, which can be used for image similarity comparisons.

The four models above share key similarities, leveraging transformer-based architectures for feature extraction. DINO and ViT process image patches for encoding, while CLIP and DreamSim focus on similarity learning (i.e., CLIP in a multimodal space and DreamSim through perceptual constraints). Their key differences lie in training approaches: DINO is self-supervised, CLIP and DreamSim use supervised contrastive/triplet loss, and ViT serves as a general-purpose image encoder. Finally,

DINO refines ViT for self-supervised learning, while CLIP aligns images with text, and DreamSim fine-tunes embeddings for perceptual similarity.

3.4 Case Study: 2023 Tornado in Norman, Oklahoma

To evaluate the proposed model, we acquired building damage data from a tornado that struck Norman, Oklahoma during the February 2023 tornado outbreak, that produced 13 tornadoes across Oklahoma ((NOAA, 2023)). We selected the Norman tornado due to its extensive building damage in the urban area, providing more geolocation data points. Rated EF2, it caused 12 injuries and covered a 27-mile (43.45 km) path ((NOAA, 2023)). Our dataset includes 81 images from the NOAA DAT and 6,478 images from GSV.

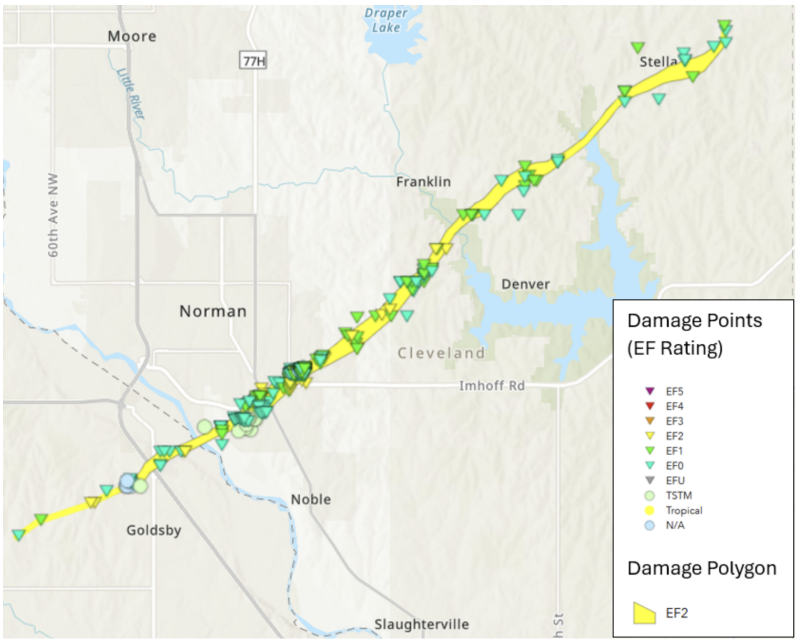

Figure 2: Norman, OK Tornado Path and Damaged Building Points.

Figure 2 shows the tornado's damage path and the NOAA DAT image locations ((NOAA, 2025)). The NOAA DAT is filtered to February 2023, and zoomed in to show the area around Norman, OK. The yellow polygon line on the map shows the path of the tornado, and the intensity associated with it (EF2). The triangles and circles on the map show each data point where damage was reported. The shade of the point indicates the rating of the tornado damage at that point. This tornado caused damage from EF0–EF2. The count of buildings reported with each EF number is as follows: 120 (EF0), 93 (EF1), 15 (EF2). This is the count of images in each category before the filtration based on quality of building images.

To evaluate buildings, and not include tree damage or other damage, we narrowed the data to 2–21 damage indicators ((NOAA, 2024)). Then, we excluded images that showed only parts of or sides of buildings. We started with 270 data points but then narrowed down to 81 points due to these constraints. This scope allows for a high-quality dataset of damaged buildings. After applying the scope on the dataset, our dataset only had a damage indicator 2. In Figure 3, the three buildings display damages. The buildings on the left and center show roofing damage, with tarps covering the roof. The image on the right has debris from the tornado in front of the building. The coordinates associated with each building are at the location on the street where the image was taken, not the building itself.

NOAA images, which are fixed in number, were directly scraped, while GSV images were generated by customizing parameters such as field of view (FOV), angle, and street position (Google, 2025). In Figure 3(b), the three buildings are from GSV. The images are taken from the street, but the coordinate for each is directly on the building geocoordinate. We used Microsoft Building Footprints

35.1936, -97.3988

35.1733, -97.4333

35.1937, -97.3987

(a) NOAA Dataset Sample Images

35.32178, -97.2123

35.24687, -97.32171

35.212072, -97.392560

(b) GSV Sample Images

Figure 3: Sample images used for evaluation.

(Microsoft, 2023) to generate a list of building coordinates in the NOAA dataset zone. This ensured that the images queried from GSV were solely buildings. These coordinates were then used in a Python script to query the street views of those buildings from GSV. We experimented with the data collection from different perspectives (e.g., angles, positions). The images show variability and occlusion. Some of the varying levels of exclusion are caused by viewing angles, natural elements (e.g., trees, cars, or mailboxes), and street-level obstructions. We chose FOV 75 because it captured the most while maintaining variability that came from natural settings.

We designated the NOAA dataset as the query dataset and the GSV dataset as the target dataset for the image retrieval process. We developed a ground truth dataset which is GSV images corresponding to each of the query NOAA images. These ground truth GSV images were added to the original GSV dataset as a metric for evaluation of image retrieval accuracy.

Our study also developed an additional building-centric dataset derived from the original dataset. This dataset was created using Faster R-CNN to detect and draw bounding boxes around buildings, effectively cropping out the background. The extracted building images were then saved as a building-centric dataset. The details of this process are discussed further in Section 4.1. Both the original and building-centric datasets were used to evaluate the performance of the image similarity model in retrieval tasks.

4 Results

4.1 Building Detection per Dataset

We evaluated the performance of our building detection system across three distinct datasets. The first dataset, BEAUTY, is used to assess the model's performance during validation and testing. The other two datasets are unseen datasets which are from NOAA and GSV images.

Precision-recall (PR) curves (Padilla et al., 2020), shown in Figure 4, and average precision (AP) metrics across different Intersections over Union (IoU) thresholds, presented in Table 2, demonstrate the higher performance of our fine-tuned model on the BEAUTY dataset. General-purpose models trained on large datasets as we mentioned in Section 2.2 often classify large buildings as part of the background, leading to suboptimal detection outcomes. In contrast, our fine-tuned model achieved significant improvements in the building detection task. Consequently, this model is deployed in subsequent experiments involving the unseen NOAA and GSV datasets.

Table 2: Comparison of Average Precision (AP) across varying Intersection over Unions (IoU).

IoU Threshold	Model	AP
0.5	**Fine-tuned Faster R-CNN**	**0.731**
	Pre-trained Faster R-CNN	0.017
0.8	**Fine-tuned Faster R-CNN**	**0.541**
	Pre-trained Faster R-CNN	0.001

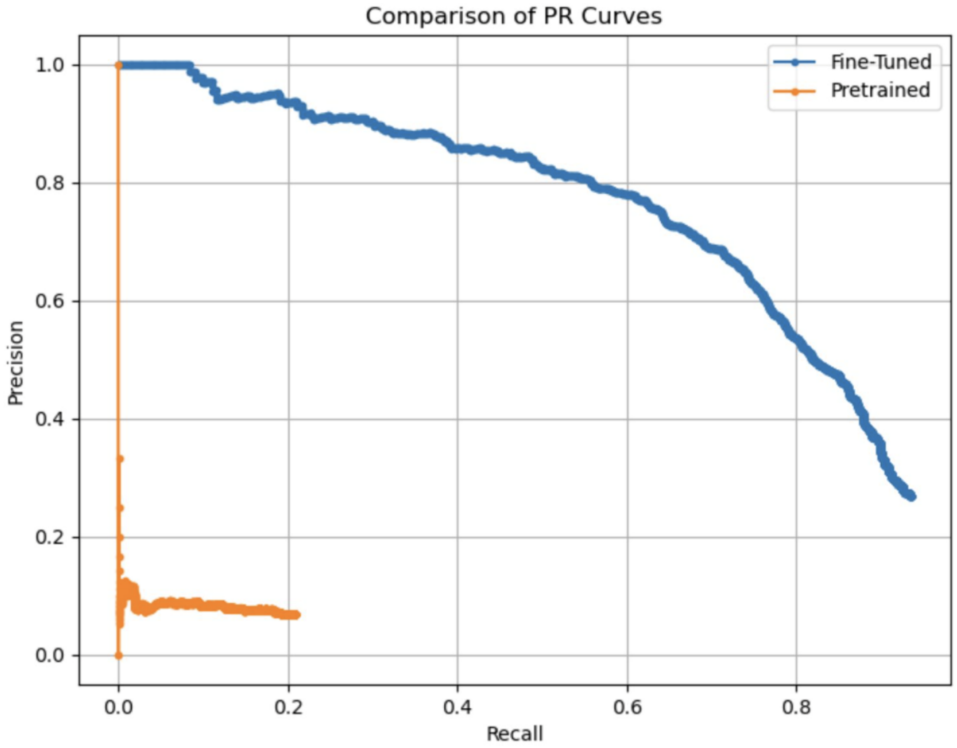

Figure 4: Precision-Recall Curve for Building Detection.

Table 3: Building detection count per dataset

	NOAA	GSV
Building Detected	81	6,178
Building Undetected	0	377
Total	81	6,555

We utilized the fine-tuned Faster R-CNN model to perform inference on the unseen NOAA query dataset. For the 81 images collected from the NOAA DAT database, all contain identifiable buildings when viewed by the human eye. Our model successfully detected buildings in all 81 images as shown in Table 3. Examples of original and cropped buildings for NOAA are presented in Figure 5. As shown in NOAA (b) of Figure 5, for images with multiple detected buildings, the bounding box with the highest confidence score is selected. The second dataset, GSV, introduces greater variation in building scale, occlusions, and missing buildings, posing additional challenges for accurate detection and bounding box generation. The detection performance indicates that the model achieves a 94% detection rate for GSV images. For the 377 images where our model failed to detect buildings, as shown in GSV (d) of Figure 5, we retained the original images for cases without detected buildings to ensure that all images from the GSV database were included.

(a) GSV Original

(b) GSV Bounding Box Detected

(c) NOAA Original

(d) NOAA Bounding Box Detected

Figure 5: Building detection results for GSV and NOAA datasets. Each row group displays original images and their corresponding bounding box detections. Rows (a) and (b) show GSV samples before and after detection, while (c) and (d) present NOAA samples. Red bounding boxes indicate the detected building with the highest confidence score.

4.2 Similarity Analysis

Our experiments evaluated the image retrieval performance of the selected similarity models using both original query images and a cropped dataset focused on buildings. We assessed the models' accuracy and interpretability to improve georeferencing by comparing the similarity between query images and target database. Both the query and target images are embedded as feature vectors. For each query, the model retrieved the top 5 similar images from the target database based on the embeddings, ensuring they also fall within a specified geolocation radius. The selected images' geolocation data is then used to update the query image's geolocation.

Search radius is critical for reducing the search dataset, improving the overall accuracy, and optimizing computational cost. By limiting the search space, a well-defined radius helps eliminate irrelevant candidates, leading to more precise retrieval results. To determine the optimal search radius for this case study, we conducted a series of tests evaluating retrieval accuracy at different distances using CLIP. Figure 6 shows the relationship between search radius (distance in meters) and CLIP's top 2 retrieval accuracy. The results show that accuracy increases sharply from 10 to 40 meters, reaching a peak around 40–50 meters where it stabilizes at approximately 77–80%. Beyond 60 meters, accuracy begins to decline, indicating a decrease in retrieval precision due to the inclusion of non-target structures. Based on these results, we selected a 50-meter search radius for this case study.

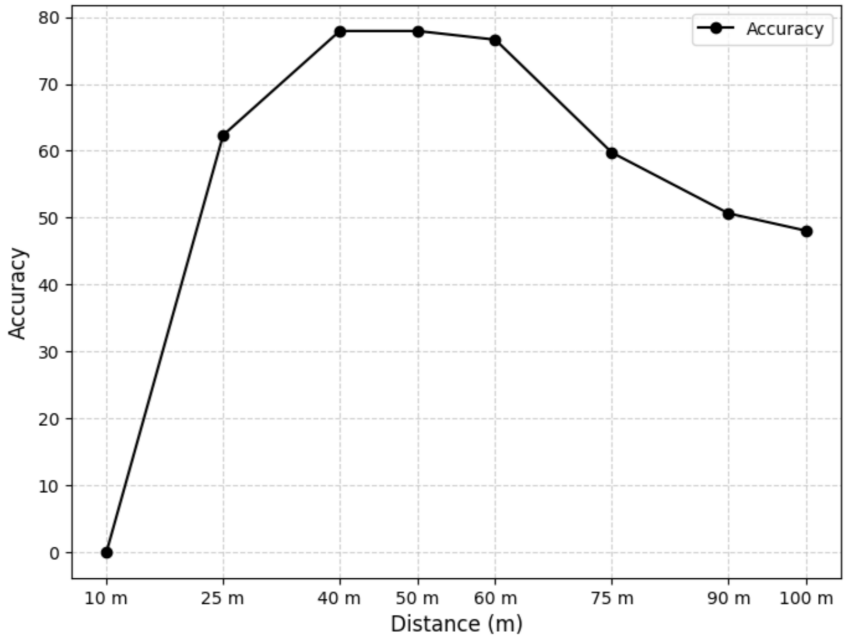

Figure 6: Impact of Search Radius on Image Retrieval Accuracy

The image retrieval accuracy was based on successful retrievals which are defined as finding at least one similar item to the query in top-N results (He & Hu, 2018; Zakizadeh et al., 2018). For each of our image similarity models including DINO, DreamSim, CLIP, and ViT, the accuracy for top-5 is shown in Table 4.

Table 4: Top-N Accuracy for Image Retrieval

Image Type	**Model**	**Accuracy (%)**				
		Top-1	**Top-2**	**Top-3**	**Top-4**	**Top-5**
Original Images	DINO	34	59	62	66	70
	DreamSim	**62**	**79**	**86**	**89**	**93**
	CLIP	47	62	78	84	90
	ViT	32	46	56	67	71
Building Centric	DINO	42	62	74	77	77
	DreamSim	**57**	**74**	**82**	**89**	**89**
	CLIP	43	60	75	81	91
	ViT	33	47	59	63	69

DreamSim achieved the highest performance across all metrics, with 86% accuracy for original images and 82% for building-centric images at the top-3 rank. CLIP followed closely with 78% and 75%, respectively, and performed well in higher-rank retrievals, though DreamSim outperformed it

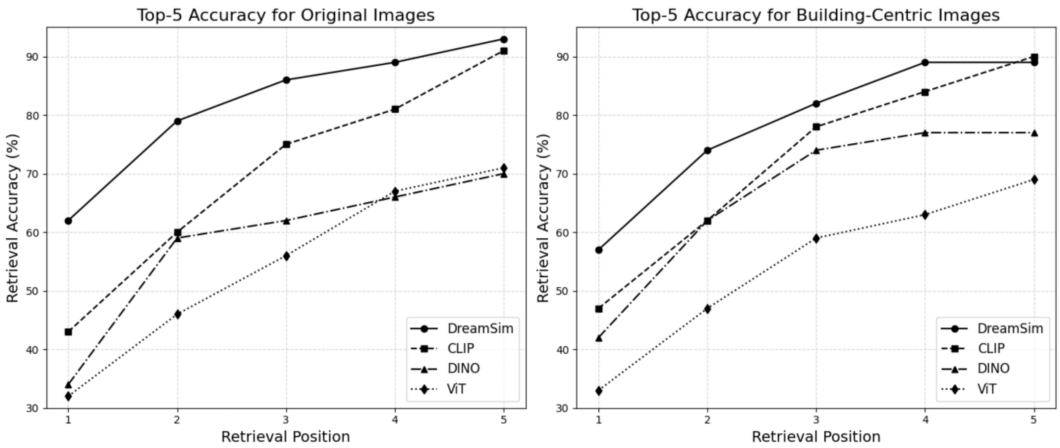

Figure 7: Top-5 retrieval accuracy results for retrieving the ground truth GSV image.

at top-1. DINO and ViT showed lower accuracy, with ViT performing the worst across all ranks. These results suggested that DreamSim and CLIP were the most effective models for image retrieval in our dataset. Performance was consistently lower for building-centric images across all models and ranks. Even DreamSim, the top-performing model, achieved below 90% accuracy at the top-5 rank for building-centric images.

Figure 8 presents examples of top-5 image retrieval results. (a)–(f) show top-1 retrieval examples, while (g)–(i) illustrate lower-ranked results. (f) demonstrates a case where only two images remained after applying the Haversine formula, resulting in retrieval from a limited pool of two images.

For each correct match, the geolocation data of the input image gets readjusted to the matched street view image to correct the geolocation data. We refer to the difference between the original coordinates and fixed coordinates as the distance offset. Figure 9 is a horizontal box plot for each image in different ranges of the distance offset. The mean distance offset was 26.12 meters with a standard deviation of 10.75 meters. The minimum and maximum distance offsets were 4.44 meters and 46.52 meters, respectively.

5 Discussion

GSV is a valuable source of street view imagery, widely used in various disaster-related applications, including damage assessment, urban resilience planning, and emergency response efforts (Zhai & Peng, 2020). With over 170 billion images across 87 countries, GSV continues to expand its coverage by mapping untraveled areas and frequently updating existing locations (Lookingbill & Russell, 2019; Google, 2025). However, some buildings in the NOAA DAT dataset lack corresponding GSV images. This limitation resulted from restricted street access, property owners choosing to blur their buildings, or temporary obstructions such as remodeling, construction, or environmental factors (Fan et al., 2025). Additionally, as privacy concerns grow, more building images may become unavailable over time, further limiting data accessibility for disaster-related applications.

To supplement missing GSV data, other street view image datasets can be explored as potential alternatives. One such dataset is Mapillary, a crowdsourced platform that provides over 2 million street view images collected from user-contributed photos taken with smartphones and action cameras. Mapillary leverages computer vision to generate street-level views and enhance mapping applications. Prior research has demonstrated its utility; Zarbakhsh & McArdle (2023) used Mapillary for urban data collection in New York City, and Çelik & Sümer (2020) assessed its potential for automated building image selection. In addition to Mapillary, other platforms such as KartaView (formerly OpenStreetCam) (Vo et al., 2023; Liu & Sevtsuk, 2024) and OpenStreetMap (Herfort et al., 2021) also offer alternative sources of street-level imagery, albeit with varying coverage and data quality. Furthermore, field surveys using specialized equipment from The Natural Hazards and Disaster Reconnaissance Facility (RAPID), such as the Applied Streetview 8K 360-degree panorama camera can capture high-resolution imagery while driving or walking, with the ability to extract

Figure 8: The image on the left is the query and the five on the right are retrieved images from GSV using DreamSim in the order of high to low similarity scores. The red box denotes the ground truth image.

individual images as needed. This approach provides a valuable option for obtaining site-specific data when publicly available datasets are insufficient. Considering the growing availability of such datasets and reconnaissance tools, future research could integrate multiple sources to improve coverage where GSV imagery is missing.

In our experiment, we observed the optimal search radius for image retrieval falls within the 40–60 m range (see Figure 6 in Section 4). This search radius is particularly effective in residential areas, where buildings often share similar architectural features. By setting the search radius within this range, we typically include 7–15 candidate buildings, which helps improve the overall accuracy of the image similarity model (Figure 10). The presence of multiple similar residential buildings within a community enhances the likelihood of retrieving the correct match while minimizing the inclusion

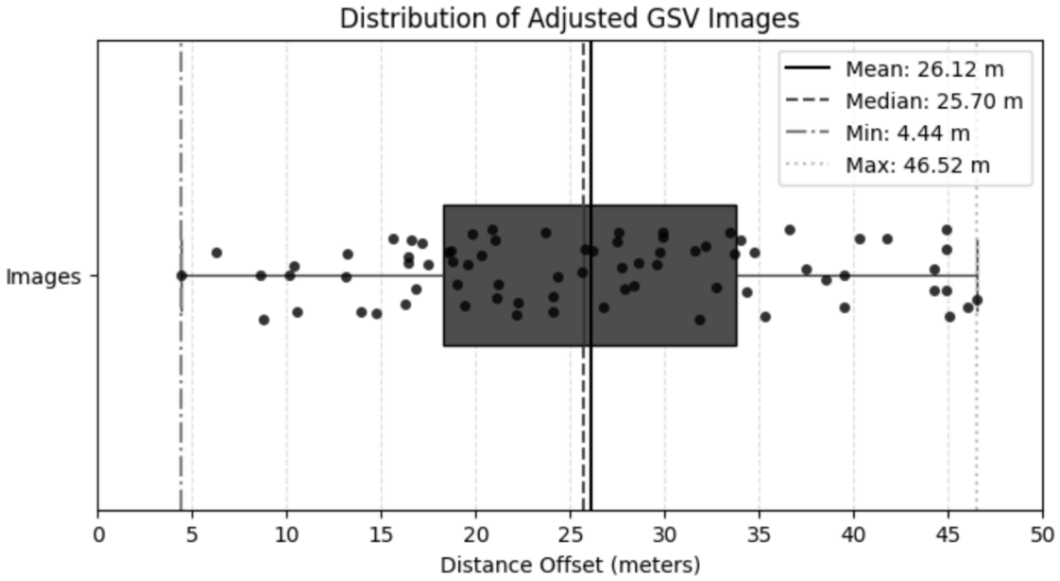

Figure 9: Distribution of corrections to the original geolocation data using DreamSim.

of irrelevant structures. This demonstrates the critical role of the search radius function in selecting the most suitable candidate buildings for comparison.

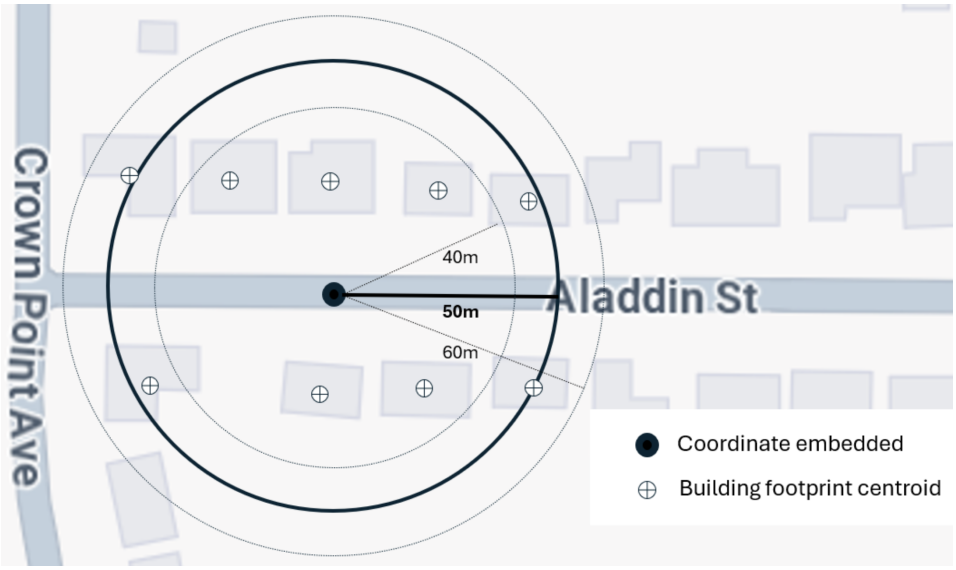

Figure 10: Search radius and the selected candidate buildings

Moreover, based on community layout characteristics, modelers might need to adjust the search radius parameters accordingly. For instance, residential buildings in rural or suburban areas are often more dispersed, requiring a larger search radius to capture an adequate number of candidate structures. In contrast, dense urban environments might benefit from a narrower radius to avoid retrieving too many irrelevant matches. This highlights the importance of adaptive search radius tuning based on the spatial context of the target area.

The quality and composition of the dataset directly impact the performance of similarity models, particularly when non-building objects and environmental variations introduce noise into the analysis. As shown in Figure 8 (see Section 5), both the NOAA and GSV datasets contain non-building objects such as trees, bushes, and cars that partially occlude buildings. Additionally, variations in weather

and time of capture result in differing color tones. To mitigate these factors, building detection was applied to crop out surrounding scenery, enabling similarity comparisons between building-centric images. However, as shown in Table 4 (see Section 4), top-1 accuracy for building-centric images remains lower. Depending on the image, our building detection model produced cropped regions that were too small, omitting building features such as edges and geometric structures. More accurate detection algorithms can enhance the performance of similarity models by preserving essential contextual information.

For our dataset, DreamSim achieves the best results for both original and building-centric images. By emphasizing human-aligned representations, DreamSim effectively captures a variety of global and dense features. While it outperforms the other three similarity models in our study, the findings indicate that houses with subtle, intricate structures present greater challenges for top-1 retrieval accuracy. For example, in Figure 8(h), the top-1 and top-3 retrieved images show only a minor difference in roof size while sharing the same triangular design. Similarly, in Figure 8(i), all retrieved images show high similarity with only subtle variations in window shapes. Beyond the top-3 retrieval results, we see cases with query images showing low retrieval accuracy correspond to structures with significant damage, such as missing windows or damaged rooftops. Additionally, the retrieved images share similar geometric attributes, such as the presence of two triangular roofs, one window on the left or right, or a centrally located entrance. These characteristics of both query and retrieved images contribute to low retrieval accuracy and result in lower overall similarity scores across all retrieved images. Furthermore, Figure 8(j) shows a case in which the ground truth image could not be retrieved in top-5 due to high density of houses within the given search radius or strong similarities in architectural features. For example, there are more than 15 images retrieved, and many houses have a single garage and window, but with variations in their positions or design. Considering these challenges in real-world applications, we need to develop a more reliable framework to enhance the performance of similarity analysis. This can be achieved by incorporating additional tasks, such as reducing candidate images by predicting the direction in which the photo was taken and utilizing text recognition to identify building numbers. Also, Sundaram et al. (2024) noted that DreamSim's performance declines on natural image datasets requiring fine-grained detail recognition. Thus, further fine-tuning on high-detail architectural datasets can enhance the performance of similarity models in post-disaster scenarios.

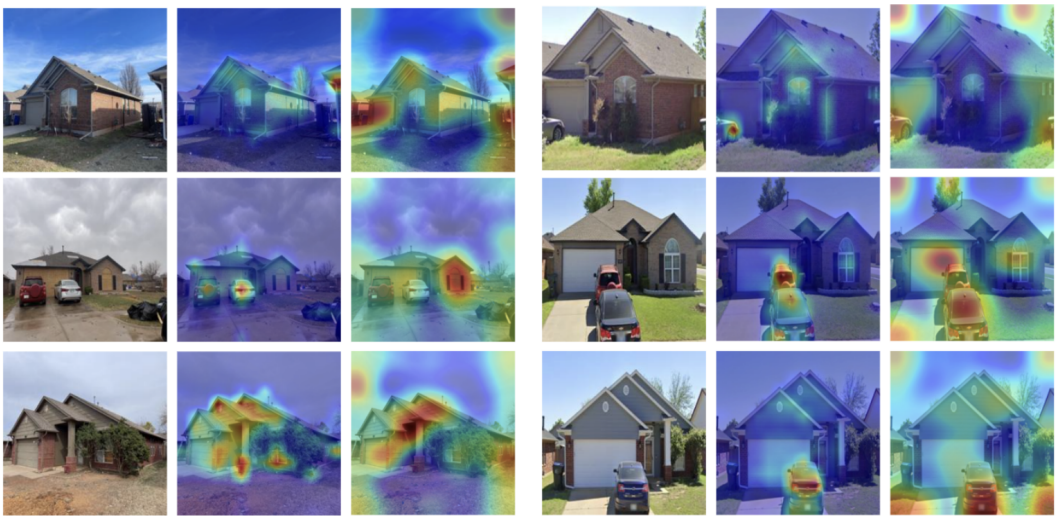

Figure 11: The right panel shows NOAA images used as query data, while the left panel displays the corresponding ground truth images from GSV. In (a) and (b), both DreamSim and CLIP achieve top-1 retrieval accuracy. In (c), CLIP maintains top-1 accuracy, whereas DreamSim is top-3.

Figure 11 presents transformer attention maps for the top-performing models, DreamSim and CLIP, offering insight into their similarity decisions for both NOAA and GSV images (Chefer et al., 2021). For DreamSim, we utilize attention from the DINO backbone (Fu et al., 2023), whereas for CLIP, we employ the ViT encoder backbone. In Figure 11(a), DreamSim exhibits an even distribution of attention, with a stronger focus on the large window and the neighboring building's roof, reflecting its

emphasis on standout features and foreground objects. In contrast, CLIP highlights multiple areas, including regions outside the house, suggesting a broader focus that incorporates both background and foreground elements. Notably, both models achieve top-1 retrieval accuracy for NOAA (a).

Our experiments reveal distinct focus patterns between CLIP and DreamSim. CLIP captures broader contextual information, attending to multiple areas within an image, while DreamSim emphasizes object-specific details and fine-grained features. This contrast is evident in their attention maps—DreamSim consistently focuses on structural elements such as windows and edges, with occasional noise from secondary objects like cars. CLIP, however, extends its focus to larger regions beyond the house, maintaining consistency across both query and target images, which contributes to accurate top-1 retrieval. These findings suggest that CLIP is more context-aware, whereas DreamSim aligns more closely with human-perceived similarity by prioritizing key objects and architectural details.

Figure 11(b) further illustrates top-1 retrieval accuracy for both models. Here, DreamSim relies more on a foreground object than on the house itself to retrieve the correct image, whereas CLIP distributes attention to the image corners, suggesting a less consistent filtering of background noise. Our dataset includes images with varying proportions of driveway, sky, and house. When images are captured from different angles, focusing solely on the house can lead to inaccurate retrieval. For example, in Figure 11(c), the NOAA image shows that while both models emphasize the distinctive cascading roof, CLIP also captures surrounding elements like the driveway, sky, and grass. This additional background context increases CLIP's similarity score, resulting in top-1 retrieval accuracy, while DreamSim, which consistently prioritizes foreground objects like cars, ranks the correct image at top-3.

These findings are supported by existing literature. Fu et al. (2023) demonstrated that DreamSim exhibits high sensitivity to foreground objects, likely due to its training on human-created datasets emphasizing semantic similarity. Moayeri et al. (2022) found that contrastive training, as used in models like CLIP, reduces foreground sensitivity in ViTs, with variations depending on image class. These differences underscore the importance of interpreting model behavior, which is influenced by factors such as image class (e.g., damaged buildings), size and angle, and the presence of foreground objects. Effective post-disaster image collection and dataset understanding are critical for selecting appropriate similarity models and ensuring accurate damage assessment.

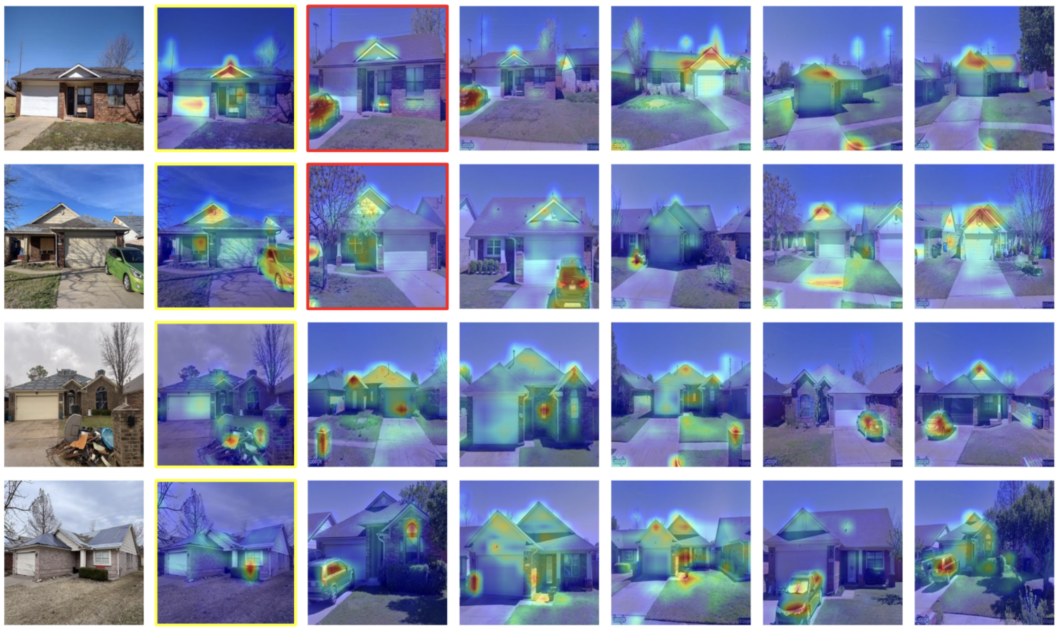

Figure 12: Attention maps from DreamSim. The top two rows (a, b) illustrate cases with top-1 retrieval accuracy, while the bottom two rows (c, d) show cases without accurate retrieval in the top-5.

Figure 12 shows examples of DreamSim's performance, comparing cases with successful top-1 retrieval accuracy to instances where it fails to retrieve the ground truth within the top 5. In (a) and (b), DreamSim emphasizes distinctive architectural features such as the triangular-shaped roof at the entrance, walkways, or nearby windows, consistently aligning these elements between the query and retrieved images. Although some attention is also directed toward foreground objects like cars, the shared focus on unique house features enhances similarity detection. In contrast, (c) and (d) illustrate retrieval failures where DreamSim primarily focuses on foreground objects that differ between the query and retrieved images, leading to incorrect matches. In (d), DreamSim attends to windows and roof patterns in the query but struggles with retrieval accuracy due to distractions from foreground elements or significant differences in viewing angles.

Our experiment further reveals that DreamSim performs better than CLIP in capturing building edges, particularly shallow-depth edges and key architectural details such as windows and pillars. However, when the query and target images exhibit significant differences in angle or pose, as shown in Figure 11(c), CLIP's broader attention to surrounding regions improves similarity detection. Nonetheless, since most of our street view images share common background elements such as sky, driveways, and grass, relying on these background features proves less effective than focusing on distinctive architectural details for achieving top-1 retrieval accuracy.

6 Conclusion

This study proposed GeoSight, a framework designed to enhance georeferencing accuracy in disaster-related imagery by leveraging coordinate-based alignment and image similarity techniques. Using the NOAA DAT as a case study, we demonstrated how refining object geolocation can improve the precision of disaster impact assessments and geospatial databases. Our approach was tested on a building damage dataset from the 2023 tornado in Norman, Oklahoma, where we evaluated the effectiveness of four similarity models in retrieving the most relevant images based on top-5 ranking results. The findings illustrate how integrating geospatial alignment with similarity-based retrieval methods enhances damage mapping and assessment accuracy. Our results highlight the critical role of precise georeferencing in disaster response and recovery. Reducing location errors in disaster imagery improves the reliability of damage assessments, optimizes resource allocation, and strengthens decision-making in emergency management. Furthermore, GeoSight contributes to the robustness of AI-driven applications in disaster management, enabling more precise damage classification, predictive modeling, and automated decision-support systems.

Future work will focus on further refining GeoSight by integrating additional contextual factors, including temporal changes in imagery (Li, 2021; Kim & Jang, 2023), shadow-based time and direction estimation (Wehrwein et al., 2015), and multi-view fusion techniques for improved object matching (Deng et al., 2021; Huang et al., 2024; Sun et al., 2024). Additionally, we aim to explore the potential of diffusion models to reconstruct damaged structures before feature extraction (Wang et al., 2023; Anciukevičius et al., 2023; Liu et al., 2024b). By addressing geolocation challenges, this research advances the development of more accurate disaster assessment methods.

References

T. Anciukevičius, Z. Xu, M. Fisher, P. Henderson, H. Bilen, N. J. Mitra, and P. Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12608–12618, 2023.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Y. J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyükoztürk. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*, 33(9):731–747, 2018.

H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

T. Chu, Y. Chen, H. Su, Z. Xu, G. Chen, and A. Zhou. A news picture geo-localization pipeline based on deep learning and street view images. *International Journal of Digital Earth*, 15(1): 1485–1505, 2022. doi: 10.1080/17538947.2022.2121437.

F. Comalada, O. Llorente, V. Acuña, J. Saló, and X. Garcia. Using georeferenced text from social media to map the cultural ecosystem services of freshwater ecosystems. *Ecosystem Services*, 72: 101702, 2025.

J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Y. Deng, H. Chen, and Y. Li. Mvf-net: A multi-view fusion network for event-based object classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8275–8284, 2021.

N. D. Diaz, W. E. Highfield, S. D. Brody, and B. R. Fortenberry. Deriving first floor elevations within residential communities located in galveston using uas based data. *Drones*, 6(4):81, 2022.

C. Ding, M. Wang, Z. Zhou, T. Huang, X. Wang, and J. Li. Siamese transformer network-based similarity metric learning for cross-source remote sensing image retrieval. *Neural Computing and Applications*, 35:8125–8142, 2022. doi: 10.1007/s00521-022-08092-6.

A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111: 98–136, 2015.

Z. Fan, C. C. Feng, and F. Biljecki. Coverage and bias of street view imagery in mapping the urban environment. *Computers, Environment and Urban Systems*, 117:102253, 2025.

S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023.

Google. Streetview request and response. https://developers.google.com/maps/documentation/streetview/request-streetview, 2025. Accessed: 2025-03-01.

D. L. Gu, Q. W. Shuai, N. Zhang, N. Jin, Z. X. Zheng, Z. Xu, and Y. J. Xu. Multi-view street view image fusion for city-scale assessment of wind damage to building clusters. *Computer-Aided Civil and Infrastructure Engineering*, 40(2):198–214, 2025.

C. Hafil, C. P. Harish, K. Mansoor, T. P. Saifudheen, and J. J. Francis. First response—collaborative disaster information gathering platform. In *2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 2, pp. 390–393. IEEE, 2017.

S. Haque, Z. Eberhart, A. Bansal, and C. McMillan. Semantic similarity metrics for evaluating source code summarization. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*, pp. 36–47, 2022.

T. He and Y. Hu. Fashionnet: Personalized outfit recommendation with deep neural network. *arXiv preprint arXiv:1810.02443*, 2018.

B. Herfort, S. Lautenbach, J. Porto de Albuquerque, J. Anderson, and A. Zipf. The evolution of humanitarian mapping within the openstreetmap community. *Scientific Reports*, 11(1):3037, 2021.

D. Huang, C. Tang, and H. Zhang. Efficient object rearrangement via multi-view fusion. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 18193–18199. IEEE, 2024.

J. Kim and K. M. Jang. An examination of the spatial coverage and temporal variability of google street view (gsv) images in small- and medium-sized cities: A people-based approach. *Computers, Environment and Urban Systems*, 102:101956, 2023.

Junho Kim, Sayok Bose, Sarah Brasseaux, and Jooho Kim. Enhancing object geolocations in imagery to improve disaster damage mapping and assessment. In *2025 AMS Annual Meeting*, New Orleans, Louisiana, 2025.

T. Kustu and A. Taskin. Deep learning and stereo vision based detection of post-earthquake fire geolocation for smart cities within the scope of disaster management: İstanbul case. *International Journal of Disaster Risk Reduction*, 96:103906, 2023.

A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

A. Lenjani, C. M. Yeum, S. Dyke, and I. Bilionis. Automated building image extraction from 360 panoramas for postdisaster evaluation. *Computer-Aided Civil and Infrastructure Engineering*, 35 (3):241–257, 2020.

H. Li, F. Deuser, W. Yin, X. Luo, P. Walther, G. Mai, and M. Werner. Cross-view geolocalization and disaster mapping with street-view and vhr satellite imagery: A case study of hurricane ian. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220:841–854, 2025.

X. Li. Examining the spatial distribution and temporal change of the green view index in new york city using google street view images and deep learning. *Environment and Planning B: Urban Analytics and City Science*, 48(7):2039–2054, 2021.

T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014.

A. Liu, B. Yang, W. Li, D. Song, Z. Sun, T. Ren, and Z. Wei. Text-guided knowledge transfer for remote sensing image-text retrieval. *IEEE Geoscience and Remote Sensing Letters*, 21:3504005, 2024a. doi: 10.1109/LGRS.2024.3374381.

F. Liu, W. Sun, H. Wang, Y. Wang, H. Sun, J. Ye, and Y. Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024b.

L. Liu and A. Sevtsuk. Clarity or confusion: A review of computer vision street attributes in urban studies and planning. *Cities*, 150:105022, 2024.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference*, volume 9905 of *Lecture Notes in Computer Science*, pp. 21–37. Springer, 2016.

A. Lookingbill and E. Russell. Beyond the map: How we build the maps that power your apps and business. https://mapsplatform.google.com/resources/blog/beyond-the-map-how-we-build-the-maps-that-power-your-apps-and-business, 2019. Accessed: 2025-05-20.

Microsoft. Microsoft building footprints. https://planetarycomputer.microsoft.com/dataset/ms-buildings, 2023. Accessed: 2025-05-20.

M. Moayeri, P. Pope, Y. Balaji, and S. Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19087–19097, 2022.

NOAA. The severe weather and tornado outbreak of february 26, 2023. https://www.weather.gov/oun/events-20230226, 2023. Accessed: 2025-05-20.

NOAA. The enhanced fujita scale (ef scale). https://www.weather.gov/oun/efscale, 2024. Accessed: 2025-05-20.

NOAA. Noaa storm damage viewer. https://apps.dat.noaa.gov/StormDamage/DamageViewer/, 2025. Accessed: 2025-05-20.

R. Padilla, S. L. Netto, and E. A. Da Silva. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 237–242. IEEE, 2020.

J. Pereira, J. Monteiro, J. Estima, and B. Martins. Assessing flood severity from georeferenced photos. In *13th Workshop on Geographic Information Retrieval (GIR'19)*, 2019. doi: 10.1145/3371140.3371145.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Z. Rao, J. Lu, C. Li, and H. Guo. A cross-view image matching method with feature enhancement. *Remote Sensing*, 15(8):2083, 2023. doi: 10.3390/rs15082083.

J. Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

C. C. Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

M. Sathianarayanan, P.-H. Hsu, and C.-C. Chang. Extracting disaster location identification from social media images using deep learning. *International Journal of Disaster Risk Reduction*, 104:104352, 2024. doi: 10.1016/j.ijdrr.2024.104352.

D. R. I. M. Setiadi. Psnr vs ssim: imperceptibility quality assessment for image steganography. *Multimedia Tools and Applications*, 80(6):8423–8444, 2021.

B. Singh and L. S. Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3578–3587, 2018.

K. K. Stephens, S. Varela Castro, Y. Xu, A. Juan, N. Diaz, R. Blessing, and S. D. Brody. Rectifying a flood data desert one step at a time: A co-created, engaged scholarship approach. *Journal of Applied Communication Research*, 52(3):421–434, 2024. doi: 10.1080/00909882.2024.2357131.

U. Stilla and Y. Xu. Change detection of urban objects using 3d point clouds: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:228–255, 2023. doi: 10.1016/j.isprsjprs.2023.04.008.

Z. Sun, Y. Fang, T. Wu, P. Zhang, Y. Zang, S. Kong, and J. Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13019–13029, 2024.

G. Y. Swara. Implementation of haversine formula and best first search method in searching of tsunami evacuation route. *IOP Conference Series: Earth and Environmental Science*, 97(1):012004, 2017.

Y. Tian, C. Chen, and M. Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3608–3616, 2017. doi: 10.1109/CVPR.2017.384.

Brian Tomaszewski. *Geographic Information Systems (GIS) for Disaster Management*. Routledge, 2020.

A. V. Vo, M. Bertolotto, U. Ofterdinger, and D. F. Laefer. In search of basement indicators from street view imagery data: An investigation of data sources and analysis strategies. *KI - Künstliche Intelligenz*, 37:41–53, 2023. doi: 10.1007/s13218-022-00792-4.

X. Wang and Q. Zhang. The building area recognition in image based on faster-rcnn. In *2018 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)*, pp. 676–680. IEEE, 2018.

Y. Wang, L. Li, T. Xue, and J. Gu. Reconstruct-and-generate diffusion model for detail-preserving image denoising. *arXiv preprint arXiv:2309.10714*, 2023.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

X. Wanyan, S. Seneviratne, S. Shen, and M. Kirley. Extending global-local view alignment for self-supervised learning with remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2443–2453, 2024.

S. Wehrwein, K. Bala, and N. Snavely. Shadow detection and sun direction in photo collections. In *2015 International Conference on 3D Vision (3DV)*, pp. 460–468, 2015. doi: 10.1109/3DV.2015.58.

W. Xin, C. Pu, W. Liu, and K. Liu. Landslide surface horizontal displacement monitoring based on image recognition technology and computer vision. *Geomorphology*, 431:108691, 2023.

Z. Xing, S. Yang, X. Zan, X. Dong, Y. Yao, Z. Liu, and X. Zhang. Flood vulnerability assessment of urban buildings based on integrating high-resolution remote sensing and street view images. *Sustainable Cities and Society*, 92:104467, 2023.

R. Zakizadeh, Y. Qian, M. Sasdelli, and E. Vazquez. Instance retrieval at fine-grained level using multi-attribute recognition. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 442–448. IEEE, 2018.

N. Zarbakhsh and G. McArdle. Points-of-interest from mapillary street-level imagery: A dataset for neighborhood analytics. In *2023 IEEE 39th International Conference on Data Engineering Workshops (ICDEW)*, pp. 154–161, 2023. doi: 10.1109/ICDEW58674.2023.00030.

W. Zhai and Z.-R. Peng. Damage assessment using google street view: Evidence from hurricane michael in mexico beach, florida. *Applied Geography*, 123:102252, 2020. doi: 10.1016/j.apgeog.2020.102252.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

K. Zhao, Y. Liu, S. Hao, S. Lu, H. Liu, and L. Zhou. Bounding boxes are all we need: street view image classification via context encoding of detected buildings. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2021.

R. Zuo, O. P. Kreuzer, J. Wang, Y. Xiong, Z. Zhang, and Z. Wang. Uncertainties in gis-based mineral prospectivity mapping: Key types, potential impacts and possible solutions. *Natural Resources Research*, 30(5):3059–3079, 2021. doi: 10.1007/s11053-021-09871-z.

N. Çelik and E. Sümer. Geo-tagged image retrieval from mapillary street images for a target building. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIV-4/W3-2020:151–158, 2020. doi: 10.5194/isprs-archives-XLIV-4-W3-2020-151-2020.