Cross-View Geolocalization and Disaster Mapping with Street-View and VHR Satellite Imagery: A Case Study of Hurricane IAN

Hao Li[a,*], Fabian Deuser[a,b], Wenping Yin[a,c], Xuanshu Luo[a], Paul Walther[a], Gengchen Mai[d], Wei Huang[e] and Martin Werner[a]

[a]*Professorship of Big Geospatial Data Management, School of Engineering and Design, Technical University of Munich, Munich, 85521, Bavaria, Germany*

[b]*Institute of Distributed Intelligent Systems, University of the Bundeswehr Munich, Neubiberg, 85579, Bavaria, Germany*

[c]*School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou, 221116, China*

[d]*Spatially Explicit Artificial Intelligence Lab, Department of Geography and the Environment, University of Texas at Austin, Austin, 78712, Texas, USA*

[e]*College of Surveying and Geo-Informatics, Tongji University, Shanghai, 200092, China*

ABSTRACT

Nature disasters play a key role in shaping human-urban infrastructure interactions, Effective and efficient response to natural disasters is essential for building resilience and sustainable urban environment. Two types of information are usually the most necessary and difficult to gather in disaster response. The first information is about the disaster damage perception, which shows how badly people think that urban infrastructure has been damaged. The second information is geolocation awareness, which means how people's whereabouts are made available. In this paper, we proposed a novel disaster mapping framework, namely CVDisaster, aiming at simultaneously addressing geolocalization and damage perception estimation using cross-view Street-View Imagery (SVI) and Very High-Resolution satellite imagery. CVDisaster consists of two cross-view models, where CVDisaster-Geoloc refers to a cross-view geolocalization model based on a contrastive learning objective with a Siamese ConvNeXt image encoder and CVDisaster-Est is a cross-view classification model based on a Couple Global Context Vision Transformer (CGCViT). Taking Hurrican IAN as a case study, we evaluate the CVDisaster framework by creating a novel cross-view dataset (CVIAN) and conducting extensive experiments. As a result, We show that CVDisaster can achieve highly competitive performance (over 80% for geolocalization and 75% for damage perception estimation) with even limited fine-tuning efforts, which largely motivates future cross-view models and applications within a broader GeoAI research community. The data and code are publicly available at: https://github.com/tumbgd/CVDisaster.

1. Introduction

Given the fast development in Remote Sensing (RS) technology, the availability of large-scale and high-quality Earth observation (EO) data has significantly benefited timely humanitarian responses to natural disasters (Van Westen, 2000; Dong and Shan, 2013; Li et al., 2023a). Meanwhile, recently, Street View imagery (SVI) has gained significant momentum in urban studies and computer vision in the last few years (Zhang et al., 2018, 2019; Biljecki and Ito, 2021), and has shown great potential in complementing traditional satellite imagery analysis by providing a unique and informative cross-view perspective on the ground (Zhu et al., 2022).

In a disaster mapping scenario, two types of information are critical for timely and accurate disaster response and relief. The first type of information is the disaster damage perception, which refers to the ways in which individuals and groups evaluate, subjectivize, and perceive damages to the urban built environment due to the disaster. This information is usually estimated from RS data based on expert knowledge and intensive manual efforts. The second type of information is geolocation awareness, which is basically how accurately people can geographically locate themselves on the map. By combining both information, an ideal disaster mapping framework is able to simultaneously estimate human perception of the damage levels and provide accurate geolocations in the affected areas.

However, it is not a trivial task to build such a framework due to two major challenges: on the one hand, traditional RS data can become insufficient for fine-grained damage perceptions, especially for distinct and sophisticated urban contexts, where a potential solution is to combine satellite imagery with the emerging source of SVIs to ensure a more fine-grained and cross-view of urban disaster damage perception. On the other hand, existing geolocalization approaches are often not satisfying, because they predominantly depend on satellite navigation systems, such as GPS, Galileo, and BeiDou, which typically lack the appropriate accuracy required for disaster response. Meanwhile, urban context and weather conditions can bring another dimension of complexity where satellite signals are blocked. Fortunately, we have enough ingredients to address the latter challenge as cross-view geolocalization with satellite and
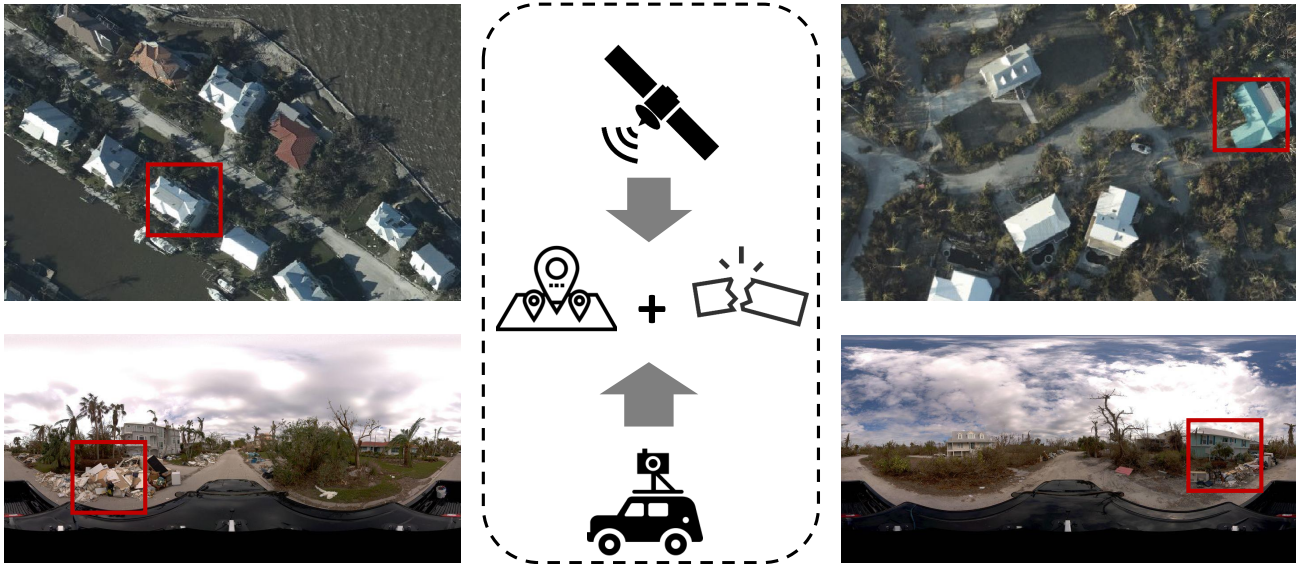
Figure 1: An overview of the proposed framework for **Cross-view** Geolocalization and **Disaster** mapping with street-view and satellite imagery, namely **CVDisaster**.

street-view imagery offers a sensible alternative. Herein, this technique can match real-time SVI obtained from carriers against a collection of satellite imagery with known geolocations so that the geographical coordinates of SVI can be decided. To the best of our knowledge, there is no such disaster mapping framework exists that can achieve damage perception and cross-view geolocalization at the same time.

In this paper, we fill the aforementioned research gap by developing a novel disaster mapping framework - **CVDisaster** - (see Figure 1). Specifically, our framework addresses the damage perception estimation and cross-view geolocalization at the same time by leveraging the head-view satellite and street-view imagery using state-of-the-art Geospatial Artificial Intelligence (GeoAI) models. To validate the proposed framework, we conducted a case study in Sanibel Island, Florida, which was hit by Hurricane Ian in 2022. Intensive experiments show the great potential of **CVDisaster** in providing timely damage perception and geolocation awareness with competitive accuracy, leading to substantial advantages for future disaster response applications. Moreover, we made the case study dataset (i.e., CVIAN) openly available to encourage related research in both computer vision and disaster mapping communities.

In Section 2, we give an overview of related works regarding state-of-the-art disaster mapping, street-view image-based urban analysis, and cross-view geolocalization, respectively. In Section 3, we elaborate on the detailed methodology design of the proposed framework, ranging from the problem statement to the training and inferencing of both geolocalization and damage perception estimation models. Next, Section 4 shows the experimental results from the case study of Hurricane IAN and summarizes the key findings, followed by Section 5 presenting a critical reflection of limitations and identifying future works. Last but not least,

Section 6 concludes the paper by highlighting the scientific contributions to a broader community.

2. Related Work

2.1. GeoAI for Disaster Mapping and Localization

Disaster mapping refers to the capability for even non-profession to assist in disaster response situations via mapping and other spatial analysis (Herfort et al., 2021; Li et al., 2022). The concept of disaster mapping has been successfully used to support disaster response and humanitarian aid activities, especially under a disaster scenario, where successful examples include the mapping tasks during the 2017 Hurrican Harvey (Feng et al., 2020), the 2019 Cyclone Idai and Kenneth in Mozambique (Li et al., 2020), and the 2023 Turkey Syria Earthquake (Wikipedia, 2023). However, considering the time-crucial nature of disaster responses and humanitarian aid, traditional disaster mapping workflows become less efficient and unsatisfactory in covering a large-scale area and providing timely damage assessment within a rather short time. In this context, the emergence of high-resolution satellite imagery allows for faster and better disaster mapping with GeoAI techniques (Salcedo-Sanz et al., 2020; Werner and Li, 2022), thus providing a promising solution to address this challenge that local stakeholders currently encounter. Early works in this direction (Herfort et al., 2019; Huck et al., 2021) report an interesting finding on improving the speed and accuracy of disaster mappings via a machine-assisted manner. In the meantime, there is a stream of GeoAI research focusing on extracting accurate location information during disasters, mainly from social media text data (e.g., Twitter) (Kumar and Singh, 2019; Hu and Wang, 2020; Mihunov et al., 2020; Hu et al., 2022, 2023b). One famous example is the news article published in the U.S. National Public Radio, titled "Facebook, Twitter

Replace 911 Calls For Stranded In Houston", which reported how affected people by Hurricane Harvey in 2017 used social media to share their location and asked for help, which significantly helps the rescue team to locate and reach those people in need. One can find a comprehensive survey on location reference recognition in Hu et al. (2023a).

However, a majority of existing disaster mapping and localization approaches either rely on post-disaster satellite imagery analysis for damage assessment or use geoparsing tools to georeference a social media text. Therefore, there is a pressing need for an intelligent disaster mapping and geolocalization solution, ideally within a single framework. To the best of our knowledge, **CVDisaster** is the first such integrated framework that can achieve large-scale damage perception and cross-view geolocalization at the same time.

2.2. Street-view Imagery for Urban Analytics

Due to its emerging availability, SVI has become a crucial data source for urban studies. Diakakis et al. (2017) conducted a comprehensive review of the applications of SVI in urban research, highlighting its growing significance in urban analysis. Their study indicates that most urban research utilizing SVI relies on Google SVI (GSVI). However, crowdsourced platforms like Mapillary and KartaView are also rapidly evolving and becoming key tools in urban research.

In urban analysis, SVI is extensively applied across various fields, such as the maintenance of spatial data infrastructure, studies of urban morphology and perception, and traffic flow analysis. For instance, Kim et al. (2020) and Li et al. (2023b) inferred urban features based on SVI to generate 3D urban models. Krylov et al. (2018) effectively detected utility poles and traffic signals using GSVI, demonstrating the unique efficacy of SVI in identifying streetlights and traffic signs. In urban morphology analysis, many scholars have estimated urban geometric indicators using SVI to study microclimates and light pollution. Hu et al. (2020) and Cicchino et al. (2020) extracted road variables from SVI to analyze the safety of walking and cycling in urban areas.

Researchers also used SVI to extract information about human health and well-being. By matching participants' movement trajectories with SVI, one can analyze the environmental features residents encounter in their daily activities, providing robust data support for public health policymaking. For example, Nguyen et al. (2018) investigated GSVI to extract derived indicators such as street greenery, crosswalks, and building types to describe the built environment at the postal code level in three US cities. The study found a correlation between community characteristics and the prevalence of obesity and diabetes. In addition to these key indicators, Keralis et al. (2020) demonstrated that factors such as overhead visible wires and whether roads are single-lane are associated with various health outcomes, including diabetes, psychological distress, and alcohol consumption. Further related studies include analyzing residents' air pollution exposure, stress levels, and infectious diseases based on street-view data (Apte et al., 2017; Han et al., 2022; Psyllidis et al., 2023).

More importantly, SVI plays an increasing role in disaster response, particularly in long-term recovery and reconstruction planning. It helps decision-makers understand changes in disaster-affected areas, providing crucial references for future disaster prevention and urban planning. Curtis and Mills (2012) and Curtis et al. (2013) explored recovery after tornadoes, hurricanes, and wildfires using GSVI. Mabon (2016) utilized GSVI from the evacuation zone around the Fukushima Daiichi Nuclear Power Plant to assess dynamic disaster recovery methods. Additionally, SVI has been used in disaster emergency response and risk assessment. Diakakis et al. (2017) used GSVI to identify the probability of buildings in Athens being flooded. Naik (2016) designed a crowdsourced sensing system for disaster response during catastrophic flooding in Chennai, India, helping residents in flood-affected areas and reducing casualties. SVI provides detailed ground-level information, such as the condition of damaged buildings, the extent of street flooding, and the state of infrastructure. This information is crucial for disaster assessment and emergency response. By combining SVI with RS data, we can obtain more accurate and comprehensive disaster information, thereby enhancing the precision and efficiency of disaster response and supporting post-disaster recovery and reconstruction. However, research that integrates SVI with RSdata in a disaster response scenario is still limited.

2.3. Cross-view Geolocalization

Unlike the single-image geolocalization task (Weyand et al., 2016; Cepeda et al., 2023; Zhou et al., 2024), cross-view geo-localisation enhances classic location-based services and navigation systems by matching ground-level imagery with overhead imagery. This enables accurate positioning in GNSS-denied environments, e.g., during a disaster. Workman et al. (2015) showed the superiority of CNN-based features for localizing a wide-ranging dataset with crawled Flickr images across the USA. In subsequent work, they introduced the first cross-view geo-localisation dataset, namely CVUSA Zhai et al. (2017). This dataset leverages street-view images from GSVI all across the US to match them against overhead imagery to locate the street-views. Since then multiple datasets have arisen with different focuses. CVACT Liu and Li (2019) aimed for a larger test set than CVUSA and included the region of Canberra, Australia, to test for cross-domain generalization. As an alternative to ground-level imagery, University-1652 Zheng et al. (2020) introduced drone views of buildings to match them against overhead imagery. Unlike CVUSA and CVACT, which rely on center-aligned street-view images for matching to satellite imagery, VIGOR Zhu et al. (2021) uses a novel approach. This method allows multiple street view images to be matched to a single satellite image at different positions, allowing precise regression of the exact offset. None of the previously released datasets have specifically addressed cross-view geo-localization in disaster scenarios,

which involve unique challenges such as destructed and altered environments.

By exploiting image similarities and differences, cross-view geolocation is characterized by contrastive learning. Vo and Hays (2016) pioneered soft-margin triplet loss and set a long-standing loss standard for this task. Further work introduced specialized aggregation methods like the NetVLAD layer Hu et al. (2018) or the SAFA-module Shi et al. (2019), enhancing the ability to capture and aggregate discriminative features from cross-view images. Zhu et al. (2022) are the first to introduce the Transformer architecture in this domain and following work by Zhu et al. (2023), they utilized the MLP-Mixer architecture with further performance gains. Deuser et al. (2023) introduced hard negative sampling based on the geographical distance as well as feature similarity and showed superior performance. Fervers et al. (2023) enhanced this previous work with a second stage for re-ranking the results and improved overall retrieval performance.

3. Methdology

3.1. Task statement

Given a set of street-view imagery $\{L_s\}$ and satellite imagery $\{L_a\}$ with G_a refers to the geographical locations (e.g., longitude and latitude) of satellite imagery, our objective to learn a cross-view embedding space $\mathbb{R}_{CV}$ (e.g., via a non-linear function $f(L_s, L_a) \rightarrow \mathbb{R}_{CV}$) in which two tasks are solved simultaneously: 1) each street-view imagery L_s is close to its corresponding satellite imagery L_a in the embedding space $\mathbb{R}_{CV}$ so that the correct geographical location can be retrieved based on their similarities in the embedding space; 2) each pair of street-view and satellite imagery $\{L_s, L_a\}$ is close to all other pairs where a similar level of damage perception are observed. Figure 2 shows how we achieve this objective by integrating two GeoAI models (i.e., CVDisaster-Geoloc and CVDisaster-Est), namely the disaster perception estimation model and the cross-view geolocalization model, into a single framework CVDisaster. In the rest of the section, we will elaborate on the detailed design specifics and model choice.

3.2. Cross-view Geolocalization via Contrastive Learning

In this paper, we formulate CVDisaster-Geoloc, the task of cross-view geolocalization, as a imagery retrieval problem, where an image encoder $f()$ is a nonlinear function $f(\mathbf{I}_i, \theta) : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^D$, which is parameterized by θ and maps the input image feature space (i.e., spatial dimension of $H \times W$ with three RGB bands) into a vector embedding representation of D dimension. Herein, cross-view means that $\mathbf{I}_i = \{L_s^i, L_a^i\}$ consisting of paired colocated SVI and satellite imagery, so that the corresponding geo-coordinates G_a from satellite imagery can be queried to use as the geographical coordinates of the input SVI. In this setting, two factors are of key importance for a good cross-view geolocalization model, which are the used image encoder

$f()$ and the vector embedding representation in the learned feature space $\mathbb{R}^D$.

3.2.1. Siamese Image Encoder with the modern ConvNeXt

To build a rock-solid image encoder for both SVI and satellite imagery, we follow the design in Deuser et al. (2023) by using a Siamese network that uses the modernized ConvNeXt as a backbone (Liu et al., 2022). Similar to the classic ResNet(He et al., 2016), ConvNeXt belongs to the Convolution Neural Network (CNN) family, which follows the classic sliding-window, fully convolutional paradigm, but brings in a list of modern neural architecture designs specificity for performance boosting, especially for high-resolution input, such as satellite imagery.

The key motivation for using ConvNeXt as the image encoder $f()$ is actually intuitive: first, it keeps the simplicity and effectiveness of classic CNN then modernizes the ResNet step by step towards the modern Swin Transformer (Liu et al., 2021) style to ensure performance gain. Figure 3 shows the architecture of the 4-stage ConvNeXt network and highlights a comparison between ConvNeXt and ResNet blocks. Herein, it is necessary to notice the following modification w.r.t a classic ResNet model.

Stage Compute Ratio: For classic ResNet, the computation distribution across different stages are decided empirically. For example, ResNet50 is featured with a number of blocks distributed into four stages with a ratio of (3,4,6,3), which makes the convolution operation heavy already in an early stage. One change in Swin Transformer is to reduce the stage compute ratio to 1:1:9:1, which has been introduced to ConvNeXt as well. As a result, the number of blocks in ConvNeXt50 becomes (3,3,9,3).

Patchify Layer: As natural images are inherently redundant, a common practice in the classic ResNet family is to use a stem cell for aggressively down-sampling. However, ViT's patch encoder makes this even more aggressive by adopting a large kernel size and non-overlapping convolution, namely the "patchify" layer. Similar designs are adopted in the new ConvNeXt with a 4×4 non-overlapped convolution layer to accommodate the network's multi-stage nature.

Inverted Boottleneck and Large Kerner: Following a similar idea in the Transformer block, the ConvNeXt block also uses an inverted bottleneck by keeping the dimension of the hidden layer four times of the input dimension. This idea has been proven to be beneficial in the popular MobileNetV2 (Sandler et al., 2018) and many more advanced CNN models (Koonce and Koonce, 2021). Moreover, the ConvNeXt benefits from its larger kernel-sized convolution design, which brings a significantly better performance based on the Liu et al. (2022). As a prerequisite for a larger kernel, the depthwise convolution layer is placed prior to the dense convolutional layers as shown in the comparison of Figure3.

Micro-scale Modification: The modification involves a list of micro-scale improvements, mostly related to the activation function and normalization layer. For instance, the
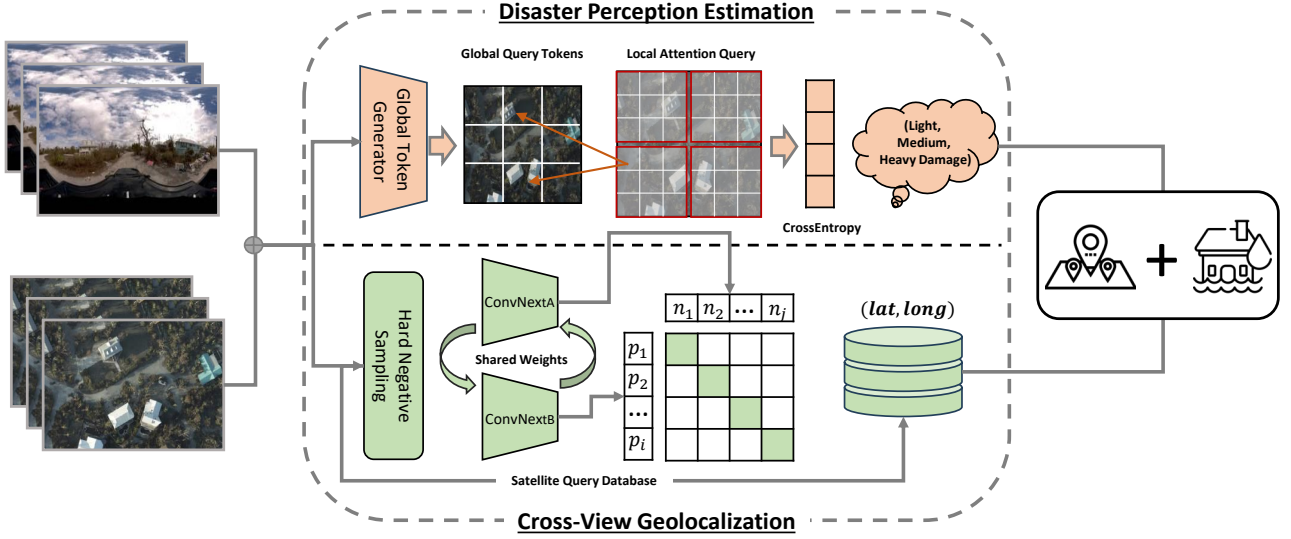
Figure 2: The proposed framework, namely CVDisaster, addresses two key tasks simultaneously, which are 1) CVDisaster-Geoloc: cross-view disaster perception estimation using coupled Global Context Vision Transformer; 2) CVDisaster-Est: cross-view geolocalization via contrastive learning.
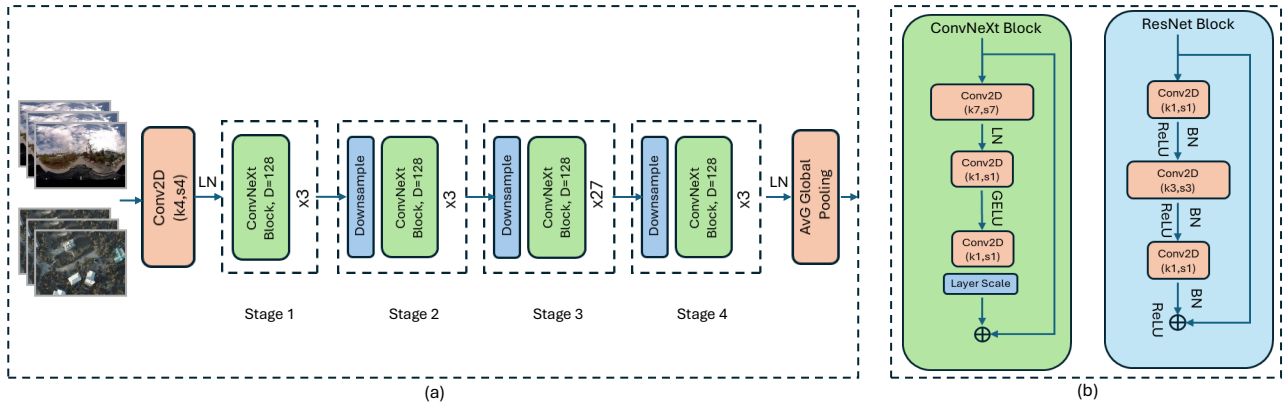

Figure 3: The siamese image encoder for cross-view geolocalization using (a) a four-stage ConvNeXt; (b) the comparison of ConvNeXt and ReseNet blocks.

ReLU used in ResNet is replaced by Gaussian Error Linear Unit (GELU) (Hendrycks and Gimpel, 2016), which is in fact a smoother variant of ReLU commonly used in modern transformer models. The Batch Normalization (BN) is replaced by the simpler Layer Normalization (Ba et al., 2016) (LN). Furthermore, the downsampling layers are added only between two different stages which are also inspired by the design of Swin Transformers.

Based on this modernized ConvNeXt backbone, we build a siamese network (Figure 3) as our image encoder $f()$ for both SVI and satellite imagery by adapting the network input to different spatial dimensions. Noticeably, although the Siamese network is trained on cross-view imagery, the inference can handle a single input of SVI as a query base.

3.2.2. Contrastive Learning with Hard Negative Sampling

The key to cross-view geolocalization is how to train a siamese ConvNeXt so that one can obtain the desired vector embedding representation of $\mathbf{I}_i = \{L_s^i, L_a^i\}$ in the learned feature space $\mathbb{R}^D$. Herein, we considered two factors to ensure efficient and effective representation learning in this cross-view setup: 1) contrastive pre-training on large-scale datasets and 2) fine-tuning with new cross-view imagery from the case study area.

Given the popularity of cross-view geolocalization, there are mainly three large-scale datasets, namely CVUSA (Workman et al., 2015), CVACT (Liu and Li, 2019), and VIGOR (Zhu et al., 2021), which have been made available to the research community. Different in their data sizes, landscape, and sample density, these three datasets form a good basis for pre-training a cross-view geolocalization model to gain nice general-sense vector representations. In this context,
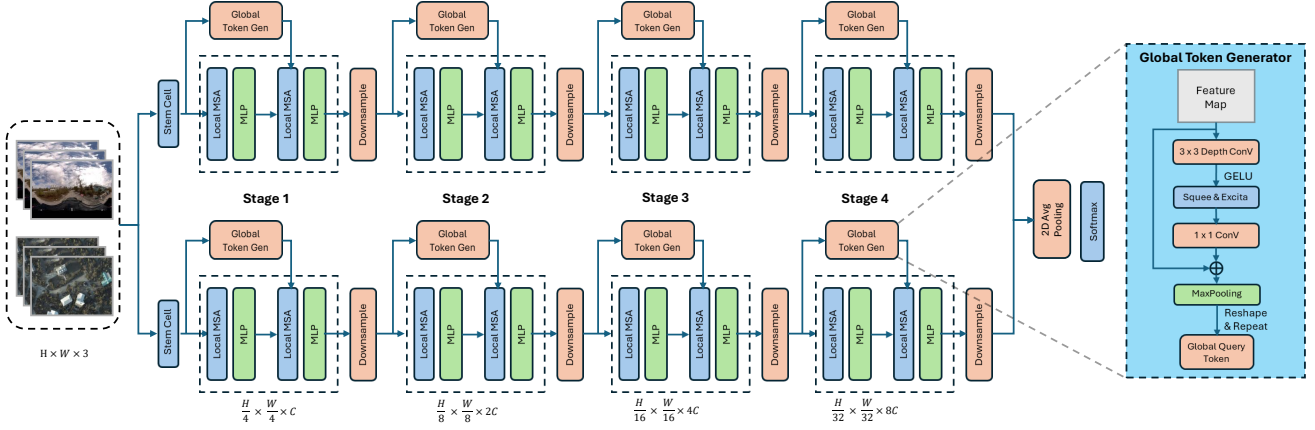
Figure 4: Coupled Global Context Vision Transformer for Cross-View Imagery Classification and Damage Perception Estimation. The global token generator is highlighed.

we pre-trained the siameses ConvNeXt network on all three datasets (i.e., CVUSA, CVACT, and VIGOR) by using the contrastive learning objective.

Following the "cluster" hypothesis that "closely associated documents tend to be relevant to the same requests" (Voorhees, 1985), the most common approach of contrastive learning is to simultaneously minimize the distance between the embeddings of the anchor t_a and the positive image t_p while maximizing the distance to the negative sample t_n. Therefore, a simple Triplet loss function looks like the following:

$$L_{triplet} = [||f(t_a) - f(t_p)||_2 - ||f(t_a) - f(t_n)||_2 + a]_+ \quad (1)$$

Here, $f()$ is the aforementioned image encoder (e.g., ConvNeXt) whose parameter θ will be learned. To prevent the encoder from pushing the negative image without limitation, a rectifier term with margin m is introduced to keep the maximum distance between the anchor and negative smaller than m.

Compared to the triplet loss, the InfoNCE (Oord et al., 2018; Radford et al., 2021) loss is often considered more robust as it is able to make use of all available negative samples with the batch. Specifically the InforNCE, in a supervised learning setting, computes categorical cross-entropy loss to identify the positive sample amongst a set of negative samples (Weng, 2021). Given a context vector c, the positive sample is drawn from a conditional distribution $p(x|c)$, where $(N-1)$ negative samples are drawn from the same distribution $p(x)$ but without condition. In this context, the probability of correctly selecting the positive samples can be formulated as follows:

$$p(C = \text{pos}|X, \mathbf{c}) = \frac{f(\mathbf{x}_{\text{pos}}, \mathbf{c})}{f(\mathbf{x}_{\text{pos}}, \mathbf{c}) + \sum_{j=1}^{N-1} f(\mathbf{x}_j, \mathbf{c})} \quad (2)$$

Here, N is the total number of samples in a batch, and $f(\mathbf{x}, \mathbf{c}) \propto \frac{p(\mathbf{x}|\mathbf{c})}{p(\mathbf{x})}$ is the similarity or scoring function between two samples.

Then, the InforNCE loss tries to optimize the negative log probability of correcting selecting the positive samples, thus can be calculated as follows:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log p(C = \text{pos}|X, \mathbf{c}) \right] \quad (3)$$

Although InforNCE has been intensively used in unsupervised and self-supervised representation learning (Mai et al., 2023; Vivanco Cepeda et al., 2024; Guo et al., 2024), it also offers a promising way for supervised representation learning in this cross-view setup. In this paper, we leverage the InforNCE as our contrastive learning loss in both the pretraining and fine-tuning stages for cross-view geolocalization. During the fine-tuning, we take the model weights pretrained on CVUSA data given its relatively large size and geographical closeness, then fine-tune the model on the new cross-view imagery collected from the study area in Sanibel Island (Florida, USA) after the Hurricane IAN. To this end, we also compare the geolocalization performance with and without the fine-tuning stage in Section 4 as an ablation study.

3.3. Damage Perception Estimation with Cross-view Imagery

Herein, CVDisaster-Est, specifically the task of damage perception estimation, is tackled as a multi-class image classification problem. Similarly to the geolocalization task, we define an image encoder $f()$ as a nonlinear function $f(\mathbf{I}_i, \theta) : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^B$ parameterized by θ and would map the input image feature space (again spatial dimension of $H \times W$ with RGB three bands) into a vector embedding representation of B dimension, but following by a softmax classification layer. In this manner, we can use exactly the same cross-view imagery pairs (e.g., $\mathbf{I}_i = \{L_s^i, L_a^i\}$) to

simultaneously estimate the damage perception level of the place when another model is trying to decide where SVI are collected.

Although this is a straightforward model design, we argue that this can bring key advantages for CVDisaster against existing disaster mapping approaches (Li et al., 2020; Herfort et al., 2019; Hu et al., 2023b). On the one hand, the data preprocessing is synchronized with zero overhead for preparing two datasets for distinct tasks (i.e., geolocalization and disaster mapping). On the other hand, the cross-view imagery can provide a unique combination of observation angles and opportunities to fasten and automate the traditional post-disaster survey with inherent geolocation metadata immediately available during the survey. This can be extremely helpful in such a time-crucial application scenario.

In the rest of this section, we will elaborate on how we tackle the damage perception estimation task in CVDisaster using the modern GeoAI-based imagery classification approach, specifically the couple GCViT model.

3.3.1. Coupled Global Context Vision Transformer

To tackle this cross-view image classification task, we develop a coupled GCViT model (CGCViT) as depicted in Figure 4) including two separate branches for SVI and satellite imagery, respectively. Unlike the siamese ConvNeXt, the design of CGCViT is driven by two special considerations: first, the appearance of disaster damages from two perspectives (head-view and street-view) differs significantly, therefore, requires highly-distinct image encoders $f()$ or sets of parameter θ; second, CGCViT can benefit from the complementary prediction capabilities learned from cross-view pairs at the same time. Moreover, the inference process also differs as the classification of damage perception level always relies on both views while the geolocalization inference actually uses only SVI imagery to query an existing satellite database. This is also why there is a single weight-shared image encoder designed for the cross-view geolocalization task.

As a backbone network, the core idea of GCVit is to advocate short- and long-range spatial dependencies with a multi-resolution architecture where self-attention is still computed in local windows but can reach long-range patch via global tokens (Hatamizadeh et al., 2023). Given a cross-view imagery pair $\mathbf{I}_i = \{L_s^i, L_a^i\}$ with the same dimension of $\mathbb{R}^{H \times W \times 3}$, the CGCViT consists of two branches of GCViT following by a 2D average pooling layer and a softmax classifier. Each branch will include four stages of local and global self-attention modules similar to ConvNeXt(Liu et al., 2022) and Swin Transformer (Liu et al., 2021), but with an increasing number of channels and decreasing spatial resolutions, both by a factor of 2. Herein, the difference between local and global self-attention modules lies in the access to global queried features from the global query generator.

Global Token Generator: As highlighted in Figure 4, the key advantage of GCViT comes from the fact that global attention is able to query long-range perception fields while keeping the local attention window unchanged. Herein, the global query token or so-called global self-attention can be pre-computed between each stage. Specifically, the global attention query $\mathbf{G}_q$ starts with a matrix of size $B \times C \times h \times w$, where B, C, $h \times w$ refers to batch size, channels, and spatial dimensional of the local window. In this way, the global query generator will repeat along batch dimension, and then be reshaped and added into multiple heads of local self-attention modules.

Global Self-Attention: Based on the global query token, the global self-attention can be formulated as follows:

$$\text{Global\_Attention}(\mathbf{g}_q, \mathbf{k}, \mathbf{v}) = \text{Softmax}(\frac{\mathbf{g}_q \mathbf{k}}{\sqrt{s}} + \mathbf{p})\mathbf{v} \quad (4)$$

where s, $\mathbf{p}$ refers to a scaling factor and a learnable relative position embedding vector. For instance, if the image patch position ranges from $[-b + 1, b - 1]$ then $\mathbf{p}$ will be generated based on spatial positions from a spatial grid of $\mathbb{R}^{(2b-1) \times (2b-1)}$. In this way, local self-attention has access to even long-range information from imagery regions outside of local windows, which provides an effective way of extending the reception field of self-attention without increasing the computation complexity.

In this paper, the CGCViT is able to extend this state-of-the-art ViT model into a dual branch setting and provide a rock-solid backbone for the cross-view damage classification task.

4. Experiment

4.1. Dataset overview

Hurricane IAN formed on September 23, 2022, causing severe storm surges and significant economic losses, making it one of the most devastating hurricanes in the history of Florida, USA. To this end, we have selected the renowned Sanibel Island and its surrounding area in southwest Florida, which was hit devastatingly by Hurricane IAN in 2022, as our case study area and created a novel cross-view dataset, namely CVIAN.

VHR Satellite Imagery: VHR satellite imagery provides extremely detailed overhead surface information, which is crucial for assessing disaster impacts, planning rescue operations, and formulating recovery strategies. For Hurricane IAN, the National Oceanic and Atmospheric Administration (NOAA) has collected relevant VHR satellite imagery. Each image is assembled into a mosaic distributed in tiles, with a ground sample distance of approximately 15 to 30 cm per pixel. In this study, we selected VHR imagery from September 30, 2022 from the NOAA open data portal[1], and divided it into five subareas to support the assessment of Hurricane IAN's damage extent. These images provide a fine-grained head-view of the study area right after the hurricane, thus enhance our understanding of the impact

---

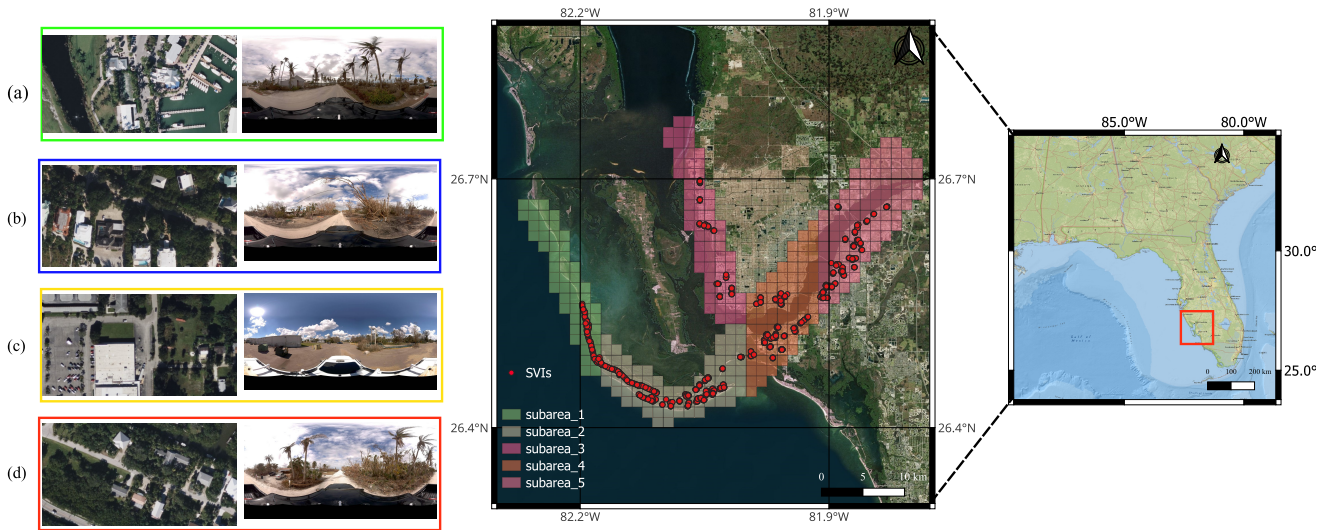[1]https://storms.ngs.noaa.gov/storms/ian

Figure 5: Overview of the study area together with the street-view and VHR satellite imagery. Five subareas are depicted in different colors in the middle image, where (a) to (d) are selected cross-view imagery pairs of CVDisaster.

Table 1
Overview of the CVIAN dataset, split into five subareas, respectively.

Subarea	VHR Reso (cm)	Image Pixel (Rows, Columns)	Num of raw SVIs	Num of selected SVIs	Aera (km2)
1	27	37,137 × 78,833	7,511	112	80.53
2	27	64,934 × 46,403	42,987	386	134.40
3	27	37,137 × 92,732	59,932	112	133.96
4	27	37,137 × 78,833	158,605	271	125.70
5	27	46,403 × 78,833	688,504	254	177.77

of the disaster and aid in developing effective response and recovery measures.

Street-view Imagery : The street-view images of Hurricane IAN used in our study were collected from the open-source Mapilliry platform, specifically from a mapping campaign conducted by Site Tour 360 in our study area. These images were captured by Site Tour 360 after access was restored post-disaster. Site Tour 360 utilized Mapillary as an mapping tool. The Mapillary platform can rapidly and openly disseminate these high-resolution images, which is crucial for disaster response. This enables rescue organizations and the public to promptly access the latest post-disaster images, aiding in identifying areas in urgent need of assistance and efficiently allocating resources.

For downloading and filtering these street-view images, we adopted the ZenSVI [2] tool. ZenSVI can efficiently download, process, and analyze large-scale street-view image data, providing valuable data support for planning post-disaster recovery efforts. In total, we have processed and filtered in total 957,539 SVI records from Mapiliary using geographic extents (i.e., five subareas) and their timestamps (i.e., only after 28th September, 2022), out of which we have selected 1,135 and manually labelled them for the damage perception level with a group of GIS and disaster experts.

[2]https://github.com/koito19960406/ZenSVI

The detailed split of SVI and extent of VHR satellite imagery is listed in Table 1.

Damage Perception Reference Data: Based on the aforementioned 1,135 SVI related to Hurricane IAN, we manually categorized them into three damage severity levels - light, medium, and heavy damages - based on a list of quantifiable and disaster-related indicators. Specifically, light damage images are characterized by a clean scene with no significant damage or only light damage, such as small areas of fallen trees or a few small road signs knocked down. Medium damage images are relatively cluttered and typically include larger or more extensive areas of fallen trees, as well as standing water around the trees. These images may also show more fallen road signs or road closure signs. Heavy damage images are very chaotic, featuring large or extensive areas of fallen trees, flooded roads, and housing trash. These indicators provide a extensible and subjective basis of disaster damage perception, which serves as the reference data for the subsequent cross-view imagery classification and validation.

As shown in Figure 6, we have elaborated some exemplary SVI in our CVIAN dataset with light, medium, and heavy damage based on different damage indicators. Among them, (a) and (b) are classified based on the amount fallen trees. (c) and (d) are classified according to the amount of housing trash. (e) and (f) are classified based on the damage

Figure 6: Stree-view imagery-based damage perception classification criteria with three level damages (light, medium, and heavy damage from low to high). (a)-(b): damage perception estimated based on fallen trees, (c)-(d): damage perception estimated based on housing trash, (e)-(f): damage perception estimated based on street signs or destroyed buildings, (g)-(h): damage perception estimated based on standing water in the street.

to road signs or destroyed buildings. (g) and (h) are classified according to the extent of standing water. This completes the damage perception reference data.

4.2. Experiment setup for Cross-View Geolocalization

In our experimental setup for CVDisaster-Geoloc, we employed a ConvNeXt-Base model, initialized using a pre-trained Sample4Geo model (Deuser et al., 2023). This pre-training on CVUSA allowed us to leverage a robust feature extraction as CVUSA features rural and urban environments. During pre-processing, we made sure that the street-view
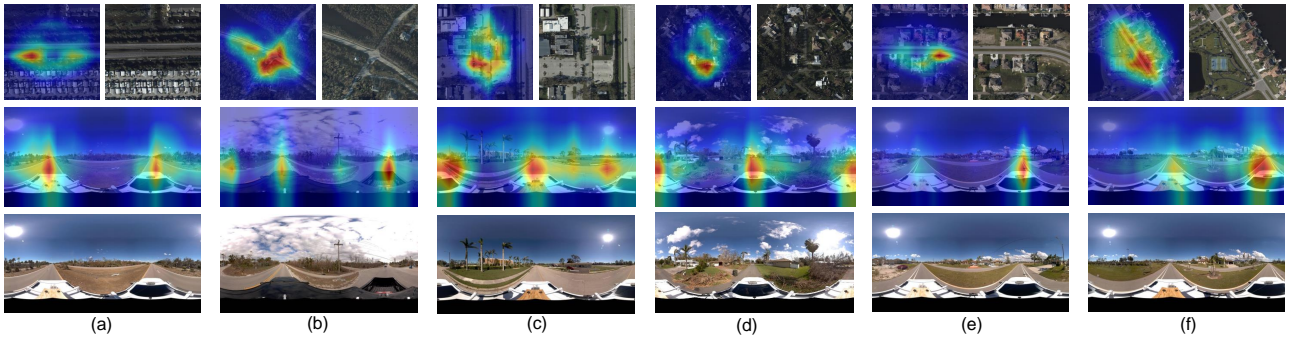
Figure 7: Heatmaps for correct geolocalized imagery pairs together with raw cross-view SVI and satellite imagery.

images were oriented north according to the CVUSA standard. We also cropped the top and bottom of the images to reduce their size and eliminate irrelevant information, thereby streamlining the input data for more efficient processing. As a result, the cross-view imagery pairs are of size 512×1024 pixels and 512×512 pixels for SVI and satellite imagery, respectively. During training, we used the InfoNCE loss function with label smoothing set to 0.1. This regularization technique helped to mitigate overconfidence in the predictions, thus promoting better generalization.

The model fine-tuning process was performed over 10 epochs using the AdamW (Kingma and Ba, 2014) optimizer with an initial learning rate of 0.0001. We use a learning rate scheduler with a warm-up of one epoch and a cosine decay for the remaining epochs. To address potential overfitting and improve generalization, we incorporate several data augmentation techniques. For both images, we use synchronous horizontal flipping and rotation to ensure consistent orientation with the corresponding street view images. In addition, we applied grid dropout and coarse dropout to prevent the model from focusing excessively on certain regions of the images. Color jittering is also used to improve the model's robustness to variations in lighting conditions.

As for CVDisaster-Est, we implemented the CGCViT model based on a backbone network of GCViT-Tiny with 20 million parameter pre-trained on ImageNet-1K dataset (Deng et al., 2009). Herein, we used exact the same cross-view imagery pairs as in CVDisaster-Geoloc with 512×1024 pixels for SVI and 512×512 pixels satellite imagery as a input feature size. Specifically, we fine-tuned the CGCViT with the AdamW (Kingma and Ba, 2014) optimizer for 100 epochs with an initial learning rate of 0.03, weight decay of 0.05 with acosine decay scheduler, and 10 warm-up epochs, respectively.

4.3. Geo-localization results

In our evaluation, we start with comparing state-of-the-art cross-view geolocalization models that were pre-trained on the CVUSA data, specifically TransGeo (Zhu et al., 2022), SAIG-D (Zhu et al., 2023), and Sample4Geo (Deuser et al., 2023). First, we directly apply three pre-trained models on the CVIAN dataset to serve as a baseline

of cross-view geolocalization performance. Next, we compare the fine-tuned model against pre-trained baseline models and conduct an ablation study w.r.t the ratio of train and test samples, ranging from 1:9 to 6:4. Since the CVDisaster-Geoloc is formulated as an imagery retrieval task, we consider four Recall@K evaluation metrics, namely, Recall@1, Recall@5, Recall@10, and Recall@1%, where K refers to the top K imagery that given by the query. A higher value (i.e., ranging from 0 to 100) simply means better accuracy.

Pre-trained Cross-view Geolocalization: Table 2 show a few interesting findings: 1) all three pre-trained models form a nice baseline of addressing cross-view geolocalization tasks in a completely unseen area with a R@10 around 80%. Noticeably, the pre-training dataset, namely the CVUSA dataset, differs with the CVIAN to a large extend in both SVI (one from GSVI and one from Mapillary) as well as Satellite imagery. These results confirm the promising value of cross-view geolocalization approaches as a pure vision-based alternative to classic positioning techniques (e.g., GPS, Wifi), especially in a disaster response scenario; 2) Given 30% of CVIAN imagery pairs for fine-tuning, CVDisaster-Geoloc achieves significant performance boosting at almost all Recall@K metrics (i.e., except Recall@1%) against three baseline models, with a relatively small improvement compared to Sample4Geo as they shared similar network architectures. This means one can easily adapt pre-trained cross-view geolocalization models to a new case study area with a limited cost of preparing "warming-up" contrastive learning samples for a much more affordable fine-tuning process than training an entirely new model from scratch, which will be an important feature desired for timely disaster response and geolocalization usage. In short, the preliminary results from implementing CVDisaster-Geoloc model on Hurricane IAN uncovers a promising avenue for leveraging pre-trained cross-view geolocalization techniques for low-cost and weather-resilient location awareness with such a time-critical task.

Moreover, by visualizing the correct geolocalized imagery pairs in Figure 7 (a) to (f), we see that the CVDisaster-Geoloc model was able to correlate SVI and satellite imagery based on landmarks, such as street, crossroad, building, which is similar to how human will spatially geolocalize

Table 2
Performance Metrics for Different Pre-trained Geolocalization Models

Method	Fine-tuned	R@1	R@5	R@10	R@1%
TransGeo Zhu et al. (2022)	✗	38.30	67.99	79.34	97.25
SAIG-D Zhu et al. (2023)	✗	43.62	72.43	83.33	98.05
Sample4Geo Deuser et al. (2023)	✗	74.56	91.22	95.48	**99.11**
CVDisaster-Geoloc (2:8)	✓	**81.84**	**96.68**	**98.45**	98.34

Table 3
Performance Metrics for Fine-tuned Geolocalization Models with Different Train/Test Ratios.

train:test	1:9		2:8		3:7		4:6		5:5		6:4	
Fine-tuned	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
Recall@1	75.69	79.72	75.97	81.84	78.86	87.47	79.03	87.74	83.33	92.20	82.30	91.37
Recall@5	92.22	95.87	92.47	96.68	93.80	97.97	94.39	98.23	96.99	99.47	96.02	98.89
Recall@10	96.06	98.13	96.23	98.45	97.47	99.11	97.49	99.26	98.94	99.82	98.67	99.56
Recall@top1	96.06	98.13	96.01	98.34	95.57	98.99	94.98	98.52	96.99	99.47	95.13	98.89

themselves in an unknown place. Unlike to existing benchmarks, the CVIANdataset is featured by massive and diverse damages (e.g., for streets, buildings, and vegetation as shown in Figure 6) caused by Hurricane IAN, which pose a unique challenge for our CVDisaster-Geoloc model. Unsurprisingly, the heatmap visualization confirms the robustness of our model in handling such sophisticated structure changes in the built environment at a much affordable cost and additional effort. From our perspective, these results can motivate future works to investigate even spatial-temporal changes in a cross-view setup.

Ablation study: As an ablation study, we examine the effect of different train and test ratios on the model performance of CVDisaster-Geoloc. Specifically, we consider two variables here: 1) pre-trained or fine-tuned models, 2) how many samples are used for fine-tuning. Table 3 shows the comparative performance changes w.r.t these two variables. Two key findings deserve extra attention: first, an obvious finding is that more fine-tuning samples lead to generally higher performance boosting with an exception from 5:5 to 6:4, where the fine-tuned model performance start to drop. A potential reason is related to the size of our CVIAN dataset and its relatively small geographical coverage, which can be a future work direction to consolidate the finding here; second, we see already a satisfying performance boosting

using limited fine-tuning with up to a few hundred imagery pairs (e.g, 1:9 and 2:8). This provides extra flexibility and reduced deploying time for the proposed framework during a real-world disaster response.

4.4. Disaster mapping results

To evaluate the CVDisaster-Est model on cross-view damage perception estimation, we first compare the CGViT model trained on cross-view imagery pairs with two single-view models (i.e., one trained on SVI and one trained on VHR satellite imagery). Next, we conduct a similar ablation study as in CVDisaster-Geoloc to investigate the effect of fine-tuning ratios on the model performance. In this context, we consider mainly four multi-classification metrics, namely Precision (P), Recall (R), Overall Accuracy (OA), and the F1 score.

Cross-view Damage Perception Estimation: Table 4 and 5 compare the performance of our CVDisaster-Est model with two single-view models based on either SVI or VHR satellite imagery. Specifically, Table 4 shows a detailed class-wise evaluation metrics w.r.t three level of damage labelled in the CVIAN dataset (i.e., Light, Medium, and Heavy Damages). A clear pattern is that medium damages are often more challenging (with low F1 scores in all three cases) than light and heavy damages. This can be attributed

Table 4
Class-wise Performance Comparison of Precision (P), Recall (R) and F1 score for SVI, VHR Satellite, and CVDisaster-Est.

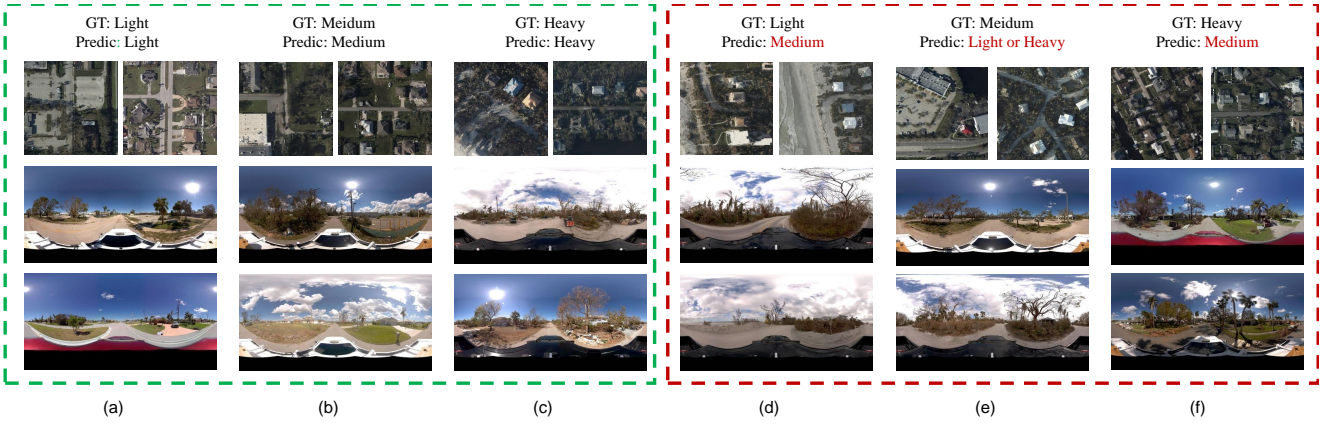|	SVI			VHR Satellite			CVDisaster-Est		
	P	R	F1	P	R	F1	P	R	F1
Light	78.61	**88.15**	0.83	72.52	82.37	0.77	**82.64**	87.64	**0.85**
Medium	62.04	56.30	0.59	55.69	39.08	0.45	**64.29**	**66.81**	**0.65**
Heavy	81.88	72.19	0.76	65.80	**75.15**	0.70	**89.15**	71.88	**0.79**

Figure 8: Visualizations of CVDisaster-Est classification results. (a) to (c) are correctly predicted imagery pairs and (d) to (f) are wrongly predicted imagery pairs. In each image, we show both Ground Truth (GT) and predicted (Predic) labels.

Table 5
Overall Performance Metrics for SVI, Satellite, and CVDisaster-Est.

	P	R	OA	F1
SVI	74.17	72.21	74.50	0.73
VHR Satellite	64.67	65.69	67.07	0.65
CVDisaster-Est(5:5)	**78.69**	**75.44**	**77.96**	**0.77**

Table 6
Performance Metrics for Cross-View Damage Perception Estimation with different Train Test ratios.

train:test	1:9	2:8	3:7	4:6	5:5	6:4
P	65.65	73.00	68.87	73.63	78.69	70.99
R	65.36	72.12	69.06	72.45	75.44	70.36
OA	66.84	73.80	70.05	73.89	77.96	71.31
F1	0.66	0.73	0.69	0.73	0.77	0.71

to how the damage perception levels are classified in the reference dataset (see Figure 5) as medium damages involves both qualitative and quantitative analysis of those damage indicator we considered (as introduced in Section 4.1), thus pose a general challenge to disaster mapping approaches (Dong and Shan, 2013). More importantly, we see a stimulating accuracy improvement when extending single-view to cross-view, for instance, CVDisaster-Est outperforms both SVI and VHR satellite in all three classes in P and F1 score. Nevertheless, Table 5 confirms the advantage of using cross-view imagery for disaster perception estimation rather than any single-view models.

Moreover, the aforementioned statement can be strengthened when we start to visualize the CVDisaster-Est classification results w.r.t where it works and where it fails. Figure 8 demonstrates both correct and incorrect classification cases using the cross-view CVDisaster-Est model. The correct cases (e.g., Figure 8(a) to (c)) basically echo the previous finding where SVI provides major hints based on

various damage indicators (e.g., fallen trees, housing trash, etc). As for incorrect examples (e.g., Figure 8 (d) to (f)), one interesting pattern is that though heavy damage might be misclassified as medium damage, there is not a single case that light damage cases are classified as heavy ones. Herein, medium damages remain challenging as they may be confused with both light and heavy damages.

Ablation Study: Similarly to CVDisaster-Geoloc, we conduct an ablation study to examine the influence of train and test ratios in the classification performance as listed in Table 6. Herein, one can see that the training process of cross-view damage perception estimation involves more uncertainties when exposed to an increasing number of training samples. Our assumption is that disaster damages are often spatial auto-correlated due to environmental and human factors, thus a random sampling is insufficient to ensure all possible damage types are well-covered. In this context, a possible solution is to explicitly consider the spatial auto-correlation of cross-view imagery early enough in the sampling process, such as using metric auto-correlation (Wang et al., 2024) or locality sensitive sampling (Luo and Shrivastava, 2019).
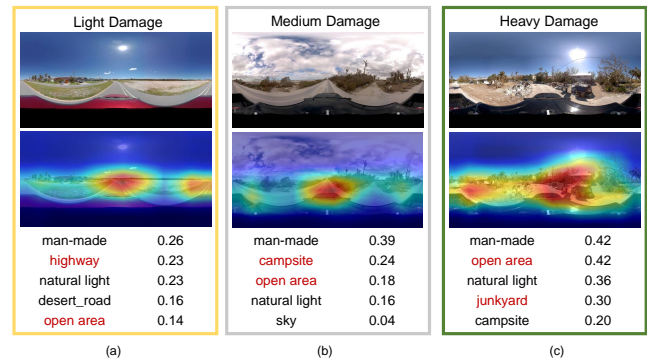

Figure 9: Selected example of SVI scene classification with Place365 classes for three-level damages in the CVIAN dataset (from left to right: light, medium, and heavy damage).

5. Discussions

In this paper, we propose CVDisaster, as a novel framework to simultaneously tackle two important tasks in a disaster response scenario, which are cross-view geolocalization and disaster perception estimation. As a case study, we constructed a first-of-this-kind dataset (i.e., the CVIAN dataset) based on SVI and VHR satellite imagery collected around Sanibel Island after Hurricane IAN, based on which we conducted extensive evaluation of the proposed framework and its major components. Despite the promising results, there are a few limitations that deserve future attention: 1) Though pre-trained geolocalization models offer a good baseline performance for CVDisaster-Geoloc, CVDisaster still requires training efforts mainly to learn and align the classification feature space (CVDisaster-Est) to various damage indicators spotted by human experts. Therefore, a future work direction is to automatically extract and quantify those damage related indicators from SVI, for instance by adopting the Place365 scene classification categories (Figure 9). Of course, it would be great if disaster-related scenes or targets, such as flooding or housing trash, can be added to those pre-training datasets. To this end, we see CVDisaster make a unique contribution towards inspiring a list of integrated cross-view applications with the GeoAI research community. 2) there are still many unsolved challenges in such a cross-view disaster mapping scenario. For example, Figure 10 shows an interesting case where from the VHR satellite imagery that NOAA collected on 30th September 2022 (two days after Hurricane IAN), one can observe massive standing water on the street and cares need to risk for access. However, from the Mapillary SVI taken on 2nd October (four days after Hurricane IAN), water is cleaned with only road signs ("Detour") left as a sign of potential damages. Together with the inherent ambiguity of the damage perception level, the rapid temporal change during the disaster poses another level of difficulty to this task. Future cross-view models should definitely take such temporal changes into considering during the model pre-training. Last but not the least, our future work will focus on extending the CVIAN dataset to cover multiple areas and disaster types around the world.

6. Conclusions

In this work, we present a novel framework, namely CVDisaster, to a time-crucial application scenario of disaster mapping, where two types of information are key, which are disaster damage perception and geolocation awareness. To the best of our knowledge, CVDisaster is the first of this kind framework that can simultaneously achieve cross-view geolocalization (CVDisaster-Geoloc) and disaster damage perception estimation (CVDisaster-Est). A case study on the CVIAN dataset collected from Hurricane IAN confirms the advantages of our CVDisaster framework over classic positioning techniques (e.g.., GPS, Wifi) as well as damage assessment approaches purely based on Very High Resolution (VHR) satellite imagery. We show that one can achieve highly competitive performance (over 80% for geolocalization and 75% for damage perception estimation) with limited fine-tuning efforts by benefiting from state-of-the-art pre-trained vision models, like ConvNeXt and CGCViT. We hope our findings can motivate future cross-view models and applications within a broader GeoAI research community.

Acknowledgements

Disclosure statement

No potential conflict of interest was reported by the author.

References

Apte, J.S., Messier, K.P., Gani, S., Brauer, M., Kirchstetter, T.W., Lunden, M.M., Marshall, J.D., Portier, C.J., Vermeulen, R.C., Hamburg, S.P., 2017. High-resolution air pollution mapping with google street view cars: exploiting big data. Environmental science & technology 51, 6999–7008.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450 .

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and gis: A review. Landscape and Urban Planning 215, 104217.

Cepeda, V.V., Nayak, G.K., Shah, M., 2023. Geoclip: clip-inspired alignment between locations and images for effective worldwide geo-localization, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 8690–8701.

Cicchino, J.B., McCarthy, M.L., Newgard, C.D., Wall, S.P., DiMaggio, C.J., Kulie, P.E., Arnold, B.N., Zuby, D.S., 2020. Not all protected bike lanes are the same: Infrastructure and risk of cyclist collisions and falls leading to emergency department visits in three us cities. Accident Analysis & Prevention 141, 105490.

Curtis, A., Mills, J.W., 2012. Spatial video data collection in a post-disaster landscape: The tuscaloosa tornado of april 27th 2011. Applied Geography 32, 393–400.

Curtis, J.W., Curtis, A., Mapes, J., Szell, A.B., Cinderich, A., 2013. Using google street view for systematic observation of the built environment: analysis of spatio-temporal instability of imagery dates. International journal of health geographics 12, 1–10.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Deuser, F., Habel, K., Oswald, N., 2023. Sample4geo: Hard negative sampling for cross-view geo-localisation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16847–16856.

Diakakis, M., Deligiannakis, G., Pallikarakis, A., Skordoulis, M., 2017. Identifying elements that affect the probability of buildings to suffer flooding in urban areas using google street view. a case study from athens metropolitan area in greece. International journal of disaster risk reduction 22, 1–9.

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS Journal of Photogrammetry and Remote Sensing 84, 85–99.

Feng, Y., Brenner, C., Sester, M., 2020. Flood severity mapping from volunteered geographic information by interpreting water level from images containing people: A case study of hurricane harvey. ISPRS Journal of Photogrammetry and Remote Sensing 169, 301–319.
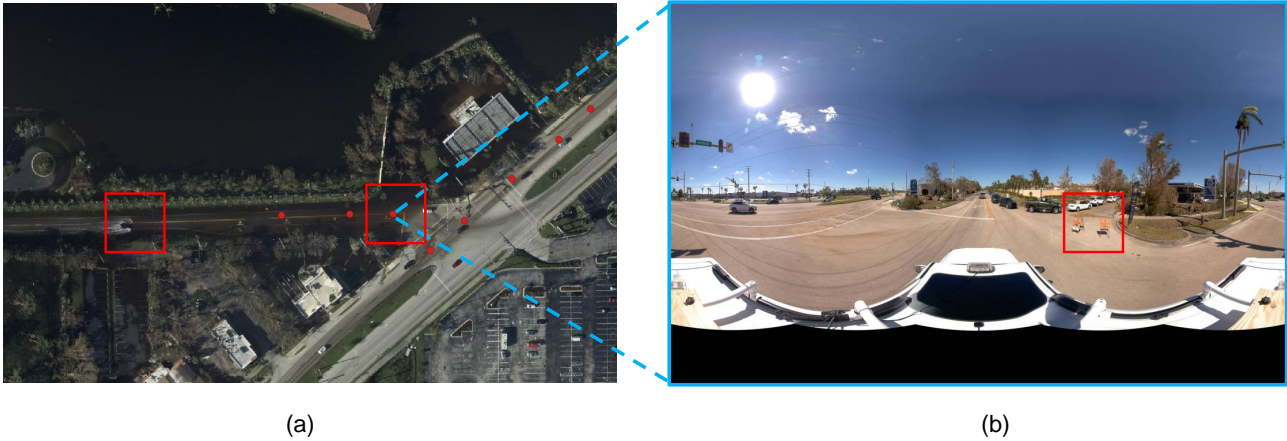
(a)   (b)

Figure 10: A Challenge case in the CVDisaster dataset. (a) VHR satellite imagery from 30th September, which is right after Hurricane IAN landed in the city; (b) SVI in Mapillary collected on 2nd October, where water on the street has already been cleaned but a road-sign of "Detour" was left as a proof of damage caused by Hurricane IAN.

Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R., 2023. C-bev: Contrastive bird's eye view training for cross-view image retrieval and 3-dof pose estimation. arXiv preprint arXiv:2312.08060 .

Guo, D., Yu, Y., Ge, S., Gao, S., Mai, G., Chen, H., 2024. Spatialscene2vec: A self-supervised contrastive representation learning method for spatial scene similarity evaluation. International Journal of Applied Earth Observation and Geoinformation 128, 103743.

Han, X., Wang, L., Seo, S.H., He, J., Jung, T., 2022. Measuring perceived psychological stress in urban built environments using google street view and deep learning. Frontiers in public health 10, 891736.

Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P., 2023. Global context vision transformers, in: International Conference on Machine Learning, PMLR. pp. 12633–12646.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hendrycks, D., Gimpel, K., 2016. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 .

Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., Zipf, A., 2021. The evolution of humanitarian mapping within the openstreetmap community. Scientific reports 11, 3037.

Herfort, B., Li, H., Fendrich, S., Lautenbach, S., Zipf, A., 2019. Mapping human settlements with higher accuracy and less volunteer efforts by combining crowdsourcing and deep learning. Remote Sensing 11.

Hu, L., Wu, X., Huang, J., Peng, Y., Liu, W., 2020. Investigation of clusters and injuries in pedestrian crashes using gis in changsha, china. Safety science 127, 104710.

Hu, S., Feng, M., Nguyen, R.M.H., Lee, G.H., 2018. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Hu, X., Al-Olimat, H.S., Kersten, J., Wiegmann, M., Klan, F., Sun, Y., Fan, H., 2022. Gazpne: annotation-free deep learning for place name extraction from microblogs leveraging gazetteer and synthetic data by rules. International Journal of Geographical Information Science 36, 310–337.

Hu, X., Zhou, Z., Li, H., Hu, Y., Gu, F., Kersten, J., Fan, H., Klan, F., 2023a. Location reference recognition from texts: A survey and comparison. ACM Comput. Surv. 56. URL: https://doi.org/10.1145/3625819, doi:10.1145/3625819.

Hu, Y., Mai, G., Cundy, C., Choi, K., Lao, N., Liu, W., Lakhanpal, G., Zhou, R.Z., Joseph, K., 2023b. Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages. International Journal of Geographical Information Science 37, 2289–2318.

Hu, Y., Wang, J., 2020. How do people describe locations during a natural disaster: an analysis of tweets from hurricane harvey. arXiv preprint arXiv:2009.12914 .

Huck, J.J., Perkins, C., Haworth, B.T., Moro, E.B., Nirmalan, M., 2021. Centaur VGI: A Hybrid Human–Machine Approach to Address Global Inequalities in Map Coverage. Annals of the American Association of Geographers 111, 231–251.

Keralis, J.M., Javanmardi, M., Khanna, S., Dwivedi, P., Huang, D., Tasdizen, T., Nguyen, Q.C., 2020. Health and the built environment in united states cities: Measuring associations using google street viewderived indicators of the built environment. BMC public health 20, 1–10.

Kim, S., Kim, D., Choi, S., 2020. Citycraft: 3d virtual city creation from a single image. The Visual Computer 36, 911–924.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Koonce, B., Koonce, B., 2021. Efficientnet. Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization , 109–123.

Krylov, V.A., Kenny, E., Dahyot, R., 2018. Automatic discovery and geotagging of objects from street view imagery. Remote Sensing 10, 661.

Kumar, A., Singh, J.P., 2019. Location reference identification from tweets during emergencies: A deep learning approach. International journal of disaster risk reduction 33, 365–375.

Li, H., Herfort, B., Huang, W., Zia, M., Zipf, A., 2020. Exploration of openstreetmap missing built-up areas using twitter hierarchical clustering and deep learning in mozambique. ISPRS Journal of Photogrammetry and Remote Sensing 166, 41–51. URL: https://doi.org/10.1016/j.isprsjprs.2020.05.007.

Li, H., Herfort, B., Lautenbach, S., Chen, J., Zipf, A., 2022. Improving openstreetmap missing building detection using few-shot transfer learning in sub-saharan africa. Transactions in GIS 26, 3125–3146.

Li, H., Wang, J., Zollner, J.M., Mai, G., Lao, N., Werner., M., 2023a. Rethink geographical generalizability with unsupervised self-attention model ensemble: A case study of openstreetmap missing building detection in africa, in: Proceedings of the 31st International Conference on Advances in Geographic Information Systems, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3589132.3625598, doi:10.1145/3589132.3625598.

Li, H., Yuan, Z., Dax, G., Kong, G., Fan, H., Zipf, A., Werner, M., 2023b. Semi-supervised learning from street-view images and openstreetmap for automatic building height estimation, in: 12th International Conference on Geographic Information Science (GIScience 2023), Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Liu, L., Li, H., 2019. Lending orientation to neural networks for cross-view geo-localization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5624–5633.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11976–11986.

Luo, C., Shrivastava, A., 2019. Scaling-up split-merge mcmc with locality sensitive sampling (lss), in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4464–4471.

Mabon, L., 2016. Charting disaster recovery via google street view: A social science perspective on challenges raised by the fukushima nuclear disaster. International journal of disaster risk science 7, 175–185.

Mai, G., Lao, N., He, Y., Song, J., Ermon, S., 2023. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations, in: International Conference on Machine Learning, PMLR. pp. 23498–23515.

Mihunov, V.V., Lam, N.S., Zou, L., Wang, Z., Wang, K., 2020. Use of twitter in disaster rescue: lessons learned from hurricane harvey. International Journal of Digital Earth 13, 1454–1466.

Naik, N., 2016. Flooded streets—a crowdsourced sensing system for disaster response: A case study, in: 2016 IEEE International Symposium on Systems Engineering (ISSE), IEEE. pp. 1–3.

Nguyen, Q.C., Sajjadi, M., McCullough, M., Pham, M., Nguyen, T.T., Yu, W., Meng, H.W., Wen, M., Li, F., Smith, K.R., et al., 2018. Neighbourhood looking glass: 360º automated characterisation of the built environment for neighbourhood effects research. J Epidemiol Community Health 72, 260–266.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Psyllidis, A., Duarte, F., Teeuwen, R., Salazar Miranda, A., Benson, T., Bozzon, A., 2023. Cities and infectious diseases: Assessing the exposure of pedestrians to virus transmission along city streets. Urban Studies 60, 1610–1628.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.

Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., Izquierdo-Verdiguier, E., Muñoz-Marí, J., Mosavi, A., Camps-Valls, G., 2020. Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources. Information Fusion 63, 256–272.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520.

Shi, Y., Liu, L., Yu, X., Li, H., 2019. Spatial-aware feature aggregation for image based cross-view geo-localization. Advances in Neural Information Processing Systems 32.

Van Westen, C., 2000. Remote sensing for natural disaster management. International archives of photogrammetry and remote sensing 33, 1609–1617.

Vivanco Cepeda, V., Nayak, G.K., Shah, M., 2024. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. Advances in Neural Information Processing Systems 36.

Vo, N.N., Hays, J., 2016. Localizing and orienting street views using overhead imagery, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer. pp. 494–509.

Voorhees, E.M., 1985. The cluster hypothesis revisited, in: Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA. p. 188–196. URL: https://doi.org/10.1145/253495.253524, doi:10.1145/253495.253524.

Wang, Z., Mai, G., Janowicz, K., Lao, N., 2024. Mc-gta: Metric-constrained model-based clustering using goodness-of-fit tests with autocorrelations. arXiv preprint arXiv:2405.18395 .

Weng, L., 2021. Contrastive representation learning. lilianweng.github.io URL: https://lilianweng.github.io/posts/2021-05-31-contrastive/.

Werner, M., Li, H., 2022. Atlashdf: an efficient big data framework for geoai, in: Proceedings of the 10th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, pp. 1–7.

Weyand, T., Kostrikov, I., Philbin, J., 2016. Planet-photo geolocation with convolutional neural networks, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, Springer. pp. 37–55.

Wikipedia, 2023. the 2023 turkey and syria earthquakes. URL: https://wiki.openstreetmap.org/wiki/2023_Turkey_Earthquakes.

Workman, S., Souvenir, R., Jacobs, N., 2015. Wide-area image geolocalization with aerial reference imagery, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3961–3969.

Zhai, M., Bessinger, Z., Workman, S., Jacobs, N., 2017. Predicting ground-level scene layout from aerial imagery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 867–875.

Zhang, F., Wu, L., Zhu, D., Liu, Y., 2019. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. ISPRS journal of photogrammetry and remote sensing 153, 48–58.

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H.H., Lin, H., Ratti, C., 2018. Measuring human perceptions of a large-scale urban region using machine learning. Landscape and Urban Planning 180, 148–160.

Zheng, Z., Wei, Y., Yang, Y., 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization, in: Proceedings of the 28th ACM international conference on Multimedia, pp. 1395–1403.

Zhou, Z., Zhang, J., Guan, Z., Hu, M., Lao, N., Mu, L., Li, S., Mai, G., 2024. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation, in: ACM SIGIR 2024.

Zhu, S., Shah, M., Chen, C., 2022. 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, pp. 1162–1171.

Zhu, S., Yang, T., Chen, C., 2021. Vigor: Cross-view image geo-localization beyond one-to-one retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3640–3649.

Zhu, Y., Yang, H., Lu, Y., Huang, Q., 2023. Simple, effective and general: A new backbone for cross-view image geo-localization. arXiv preprint arXiv:2302.01572 .