HybridMark: A Scalable, High-Quality, and Cost-Efficient Framework for IP Geolocation Landmark Collection and Evaluation

Zhongshun Zhang, Ziling Wei, Fei Wang, Jiayao Wang and Shuhui Chen

***Abstract*—IP geolocation is a fundamental component of cyberspace mapping and an essential feature of various internet services. Accurate IP geolocation is critical for ensuring network security, optimizing content delivery, and maintaining compliance with regional legal regulations. A high-quality, large-scale landmark dataset serves as the cornerstone for achieving accurate IP geolocation. However, traditional methods for collecting landmark data face significant challenges due to the sensitivity of geographic information as well as the substantial time and economic costs involved. In this paper, we propose HybridMark, a novel framework for IP geolocation landmark collection that leverages multiple sources of open-source intelligence (OSINT) to gather and filter qualified landmarks. HybridMark is designed to provide high-coverage and high-precision data to support network measurement-based IP geolocation services. Furthermore, we introduce a comprehensive evaluation methodology for assessing the quality of landmark datasets from multiple dimensions and conduct experimental assessments of HybridMark. The results demonstrate that HybridMark significantly outperforms existing state-of-the-art research and open-source datasets in terms of quantity, coverage, density, stability, and accuracy.**

***Index Terms*—Geolocation, landmark, large language model**

I. Introduction

IP geolocation is widely applied in Internet services, such as online advertising, network attack attribution, and Content Delivery Network (CDN) services. By relying solely on a target IP address, geolocation methods can determine the country, city, and even street-level location. Many companies and organizations provide IP geolocation databases that link IP addresses to geographical locations. However, these databases still face challenges in terms of accuracy and coverage, with some not achieving the claimed 70% accuracy at the city level and poor performance when querying infrastructure-related data [1]–[3]. As a result, researchers continue to explore new algorithms, including machine learning techniques, to improve geolocation accuracy.

All IP geolocation algorithms based on network measurements require IP addresses with known geographical locations as reference points for the localization algorithm, referred to as landmarks, which are positions of reference hosts with well-known geographic locations and serve as possible location estimates for the target IP. Many IP geolocation algorithms can accurately estimate the geographical location of a target IP address by comparing network characteristics, such as the network latency between the probing host and the target IP address or landmarks [4]–[7]. As the Internet increasingly permeates socio-economic activities, the demand for precise IP geolocation has escalated. For example, in scenarios such as online advertising, cybersecurity monitoring, and emergency response, precise IP geolocation is regarded as a critical factor for ensuring efficient services and informed decision-making. At the same time, network measurement-based IP geolocation is increasingly dependent on high-quality landmarks. Through the use of more high-quality landmarks can more accurate and diverse network state information be obtained by network measurement-based IP geolocation algorithms. This allows a more comprehensive understanding of the target IP's network environment and geographical characteristics to be achieved, ultimately enabling higher accuracy in IP geolocation. Therefore, the significance of landmarks in enhancing IP geolocation accuracy is being increasingly recognized. However, there is a lack of widely distributed, highly stable, low-cost, and highly accurate landmark databases on the Internet [8]. While some open-source or commercial databases are available, their coverage and update frequency often fail to meet the demands of high-precision geolocation. Moreover, due to the absence of effective evaluation methods, it is challenging to objectively select and compare these datasets, making it difficult to determine which database is superior or inferior.

Previously, many studies have attempted to increase the number of landmarks through web service mining [7], [9]–[15], a classic source of landmarks in the field of IP geolocation. However, with the advancement of cloud services and related technologies, an increasing number of web services are no longer deployed locally but instead hosted on cloud servers. This shift means that the actual physical location of many web servers no longer corresponds to the geographic information embedded in their web content or to the physical locations of their respective owners. Furthermore, as web technologies continue to evolve, the HTML source code of web pages no longer strictly adheres to specific coding paradigms. This evolution has made traditional rule-based web-based landmark mining methods less effective in extracting large-scale, accurate geographic information. Leveraging large language model (LLM) for processing web page content presents a promising approach to addressing this challenge. Departing from traditional rule-based approaches, LLM employs sophisticated deep

This work was supported by the National Natural Science Foundation of China (No. 62202486 and U22B2005), the Key Research and Development Project of Jiangsu Province (No. BE2023004-4)

The authors are with the College of Computer Science and Technology, National University of Defense Technology, No. 137 Yanwachi Street, Changsha, Hunan, 410073, P. R. China(emails: zhangzhongshun@nudt.edu.cn; weiziling@nudt.edu.cn; wangfei09a@nudt.edu.cn; shchen@nudt.edu.cn)

Manuscript received April 19, 2021; revised August 16, 2021.

learning techniques to analyze the semantics and structure of web pages, leveraging contextual insights from source code to extract geographically relevant information embedded in natural language content. For instance, LLM can extract addresses, location details, contact information, and even deduce geographic context through nuanced textual analysis. A key advantage of LLM lies in their adaptability to diverse web page structures, enabling the extraction of high-quality geographic information even in the absence of standardized formatting. The rise of cloud services has enabled users to access cloud-based resources over the Internet without the need for local device deployment or maintenance. Due to their convenience and flexibility, cloud services have gained immesadnse popularity, prompting cloud service providers to invest heavily in building cloud data centers worldwide to attract more users. However, through OSINT, many cloud service providers expose the real geographic locations of their data centers, which can be used to generate landmarks. Internet of Things (IoT) devices, which are physical devices connected to the Internet and capable of collecting, sending, and receiving data, have seen rapid growth in both number and geographic distribution in recent years. According to statistical data, the global number of IoT devices has reached billions, spanning diverse fields such as smart homes, urban infrastructure, and industrial production. Unlike other network devices, IoT devices generally possess relatively fixed geographic locations and tend to demonstrate greater stability and reliability [8], [16], making them well-suited to meet the requirements for high-quality landmarks.

This study focuses on two core objectives. First, we develop a large-scale, low-cost, and highly accurate landmark generation method, referred to as HybridMark. Second, we establish a comprehensive evaluation framework for landmark datasets. We employ LLM and OSINT to collect web-based, cloud-based, and IoT-based geolocation landmarks. Additionally, we propose a comprehensive landmark dataset evaluation framework, enabling a multidimensional quantitative assessment of the dataset in terms of quantity, coverage, density, stability, and accuracy. The main contributions of this study are as follows:

- We propose and implement a method for generating a landmark dataset on a million-scale and successfully collect such a large volume of landmark data.
- HybridMark is the first to utilize LLM for web-based landmark generation, integrating multi-source data such as WHOIS information, HTTP header data, and DNS resolution to validate the collected landmarks.
- We are the first to leverage OSINT for generating cloud-based landmarks and conduct confidence-based accuracy assessments of geolocation databases across different regions.
- We introduce a novel multidimensional quantitative evaluation framework for landmark datasets and validate the collected data, demonstrating that HybridMark achieves state-of-the-art performance in terms of quantity, coverage, density, stability, and accuracy.

The rest of this paper is organized as follows. Section II reviews related work in the field of IP geolocation. Section III introduces the structure of HybridMark. Section IV presents several novel evaluation metrics and methods for assessing landmark databases and compares HybridMark landmark dataset with other databases. Section V discusses specific implementation details in the experiments and highlights some limitations of our study. Finally, Section VI concludes the paper.

II. RELATED WORK

In IP geolocation, methodologies are primarily categorized into three types based on their underlying algorithms: information inference, database query, and network measurement. Information inference-based techniques estimate the geographic location of an IP address by analyzing publicly available data. Database query-based methods rely on pre-constructed geographic databases that map IP addresses to specific locations. Network measurement-based techniques estimate the geographic position of an IP address by analyzing network characteristics, providing dynamic and adaptable solutions based on real-time network behavior. This paper integrates contributions from all three methodological categories, combining their strengths to address geolocation challenges. Additionally, it explores research on landmark collection, which serves as the foundation for IP geolocation algorithms. This section provides a comprehensive review of relevant prior works in these domains.

Information Inference: Methods based on information inference primarily involve analyzing publicly or semi-publicly available data sources on the internet, such as domain names, IP address registration details, routing data, and network topology. By extracting geographic information embedded in these data and associating it with IP addresses, these methods enable the determination of geographic locations. DroP [17] constructed a comprehensive dictionary linking geographic strings (e.g., airport codes) to specific locations. It then searched these strings in large datasets of router DNS names, associating IP addresses with geographic locations by matching the dictionary entries to the DNS names. Structon [10] extracted the geographic information of web server IP addresses from web pages, using potential indicators like city names, state or province names, postal codes, and telephone area codes to infer the possible geographic location of web servers. Additionally, Dan et al. [18] extracted location information from domain names registered to IP addresses. Other methods [19], [20] also leverage Whois database queries to establish the correspondence between IP addresses and geographic locations.

Database query: This method involves maintaining a database containing IP addresses and their corresponding geographic information (e.g., country, city, latitude, and longitude). When the geographic location of an IP address needs to be determined, the IP address is simply queried in the database to retrieve its geographic information. These databases are typically provided by major internet service providers, regional internet registries, and specialized data collection organizations, and they encompass rich geographic data, including multi-level information on countries, provinces, and cities. Numerous companies currently offer IP geolocation database

services, specializing in collecting, organizing, and providing data linking IP addresses to their geographic locations, such as MaxMind [21], IP2Location [22], IPinfo [23], DB-IP [24], IPIP [25], and IP138 [26], among others. However, several studies have highlighted that relying exclusively on commercial databases may introduce errors of up to several thousand kilometers [1]–[3], [18], [27]–[33].

Network Measurement: Unlike methods relying on pre-constructed databases or information inference, network measurement approaches leverage real-time network data to determine geolocation. These methods typically involve the use of probing tools such as ping or traceroute to collect network characteristics, including latency, path information, and hop count between the target IP and other network points. By analyzing the relationships between the target IP and landmarks, network measurement techniques infer the target's relative geographic location. Depending on the localization algorithm employed, network measurement methods can be categorized into latency vector similarity, latency distance models, probabilistic formulas, machine learning techniques, or landmark selection strategies. A prominent and foundational algorithm in this domain is the Constraint-Based Geolocation (CBG) algorithm proposed by Gueye et al. [4], which utilizes multi-point measurements and distance constraints to infer the geographic location of the target host. By establishing a continuous answer space rather than relying on discrete values, CBG improves the accuracy of converting latency measurements into geographic distance constraints.

Landmark Collection: Landmarks are fundamental in IP geolocation, acting as reference points for network measurements that infer the geographic location of target IP. By measuring network characteristics, such as latency and hop counts, between the target IP and landmarks, these points serve as the foundation for geolocation inference. Regardless of the specific geolocation methodology employed, the number and distribution of landmarks have a direct impact on the accuracy and coverage of localization [7], [34]. Consequently, landmarks are considered the cornerstone of IP geolocation research, making their selection and management a crucial step in enhancing geolocation precision.

In the early stages, numerous collaborative initiatives between research institutions and commercial entities led to the establishment of various platforms for internet measurement and monitoring. These include open-source landmark databases such as PlanetLab [34], RIPE Atlas [35], Measurement Lab [36], and NLNOG Ring Nodes [37]. However, these platforms are often constrained by limitations such as a limited number of landmarks, restricted geographic coverage, and high operational costs. Additionally, over time, many of these landmark databases have faced challenges related to inadequate maintenance, resulting in the obsolescence of numerous nodes [7]. To overcome these issues, researchers have proposed various methods for landmark collection, aimed at improving the robustness and functionality of geolocation systems.

Ciavarrini et al. [12] developed an IP geolocation method that leverages crowdsourcing platforms, utilizing users' smartphones as landmarks. They proposed that various internet-connected devices, including home PCs, DNS servers, web servers, routers, and smartphones, could serve as landmarks. VoteGeo [16] addressed challenges such as uneven distribution and low applicability by employing widely distributed and cost-effective IoT devices (e.g., cameras) as landmarks. IoT-based landmarks, with their extensive distribution, operational stability, and fixed geographical locations, were used to enhance geolocation accuracy. Guo et al. [10] linked web server IP addresses to geographic locations by extracting organizational information from web service providers, thereby improving the accuracy and coverage of IP geolocation databases. Jiang et al. [13] gathered datasets from three distinct sources: RIPE Atlas probes (primarily located in residential and academic networks), university websites (in academic networks), and city government websites (predominantly in commercial networks). These diverse data sources significantly outperformed previous studies in variety, thereby boosting geolocation accuracy. VLOC [14] used multiple arbitrary web servers as external landmarks, estimating distances via network latency measurements. The GeoGet [11] method focused on collecting a large number of web servers with accurate geographic locations to serve as landmarks. Wang et al. [8], [9] proposed GeoCAM, a system that uses online cameras monitoring physical environments as high-quality sources of landmarks. GeoCAM regularly monitors websites hosting real-time camera feeds, identifies pages with video streams using machine learning, and extracts IP addresses and geographic coordinates of cameras through natural language processing, facilitating large-scale landmark generation.

In recent years, researchers have conducted in-depth studies on the accuracy of landmark. For example, Li et al. [38] proposed a nearest-router association method that evaluates the reliability of landmarks and enhances geolocation accuracy by analyzing the delay-distance relationship between landmarks and their closest routers. Ma et al. [39] leveraged router hostname recognition techniques to extract implicit geographic information from network infrastructure, improving the accuracy of city-level landmarks. Furthermore, addressing the challenge of quantifying landmark errors in existing evaluation methods, Yang et al. [40] introduced an upper-bound error estimation model that defines error ranges through clustering analysis and probabilistic modeling. Evaluator [41] employs DNS queries and reverse verification techniques to comprehensively assess the reliability of web-based landmarks while integrating gradient descent to optimize model parameters, enhancing robustness and coverage. These studies have contributed to refining landmark evaluation methodologies, providing crucial support for improving the precision and applicability of IP geolocation. However, despite significant advancements in landmark accuracy assessment, certain limitations remain. First, most studies focus primarily on improving landmark accuracy, often lacking a comprehensive, multidimensional evaluation framework that considers factors such as density, coverage, and reliability. This gap makes it challenging to quantify the applicability of these methods across different scenarios. Second, existing research typically employs independent evaluation methodologies with varying standards, lacking a unified set of quantitative evaluation metrics. As

a result, comparing different approaches directly is difficult, limiting the objective assessment of landmark dataset quality.

III. Design

In this section, we provide a detailed description to the design and implementation of HybridMark. As illustrated in Figure 1, the overall framework for the landmark collection method is composed of three distinct modules: IP address acquisition, geographic information extraction, and IP landmark validation. Each type of landmark—web-based, cloud-based, and IoT-based—undergoes systematic processing within these modules. For web-based landmarks, IP addresses and webpage source codes are obtained through search engine queries and DNS resolution. Hidden geographical information within the webpage source code is extracted using LLM. These landmarks are subsequently validated by integrating multi-source data to ensure reliability. Cloud-based landmarks are derived from public information provided by cloud service providers, which includes IP addresses and corresponding geographic locations. IoT-based landmarks, in contrast, are acquired through cyberspace mapping engines, which collect IP addresses and geographic data from a wide range of IoT devices, spanning different manufacturers and device types. The validation of both cloud-based and IoT-based landmarks is conducted using third-party IP geolocation databases. To improve validation accuracy, we evaluate and differentiate the precision of these databases across different geographic regions. As a result, HybridMark not only achieves the largest number of landmarks compared to prior research but also significantly enhances the positioning accuracy of geolocation algorithms.

A. Web-based Landmark Generation

Web landmarks refer to web server nodes whose geographical information can be inferred through OSINT. These landmarks have been extensively utilized in prior research [7], [8], [11], [15], [18]. As landmarks in IP geolocation, web servers offer several distinct advantages.

1) **Broad geographical distribution:** Web servers are not confined by specific geopolitical constraints, enabling web service providers to establish operations across diverse global regions.
2) **Abundant open-source geographical information:** Many web servers register their geographic details during setup, which can be queried through databases such as WHOIS [19], [20], [42]. Moreover, some web pages explicitly include the server's location information.
3) **High availability and stability:** Web services are typically designed for high reliability, characterized by continuous operation and unrestricted public accessibility. As a result, the number of unreachable nodes is significantly minimized when employing web-based landmarks in IP geolocation.
4) **Fast response times and low latency:** Web service providers often prioritize optimizing their servers' network performance. This focus on connectivity results in lower network latency, reducing measurement errors in network-based geolocation processes.

For web-based landmarks, we primarily utilize a combination of web scraping and LLM for collection. Initially, we use web scraping techniques to capture the source code of target websites, obtaining their web content. To ensure the accuracy of the landmarks and the reliability of their geolocation, we perform multi-source information validation on the scraped webpages. This step is critical for identifying and filtering out non-locally deployed web servers, such as content delivery networks, cloud hosting services, and proxy servers, which do not reflect local geographic deployment. For webpages that pass this validation process, their source code is used as input data for LLM to perform in-depth analysis. Leveraging its advanced natural language processing capabilities and contextual understanding, LLM extract organizational information and associated geographic details from the unstructured data in the source code. This approach effectively reduces the cost of landmark collection while significantly enhancing accuracy by capitalizing on the model's strengths in processing complex and unstructured datasets.

Web-based Landmark IP Address Acquisition and Geolocation Extraction: Previous studies indicate that academic institutions are more likely to host their web servers locally [13]. Building on this insight, this research leverages the 2025 QS World University Rankings [43] as a dataset to demonstrate our landmark collection methodology. The names of 1,503 universities worldwide were extracted from the rankings, and search engines were employed to identify the official websites associated with each institution, yielding their corresponding domain names. Subsequently, we performed DNS resolution to extract the IP addresses associated with these domains. Using web scraping techniques, we crawled the source code of these websites and directly input the source code of the target websites into LLaMA 3.1. The model was tasked with extracting organizational information and geographic details from the source code. By analyzing textual content, structured data, and hidden metadata (e.g., meta tags and link structures) in the source code, the model automatically identifies and extracts critical information reflecting the website owner and its geographic attributes. This approach eliminates the need for manual preprocessing of source code, fully utilizing the semantic understanding and information extraction capabilities of the large model. By leveraging the model's contextual understanding of natural language and web content, the method adapts effectively to diverse web structures and non-standardized formats without requiring additional feature engineering. It significantly enhances the accuracy and automation of extracting landmark information from websites. Ultimately, the IP addresses of these websites are associated with specific latitude and longitude coordinates, represented as (IP, (Latitude, Longitude)).

Web-based Landmark Verification: Prior to confirming the dataset as valid landmarks, a multi-source verification process is conducted to ascertain whether the target web server is locally hosted by the corresponding university. If the university's website is hosted on a cloud server or utilizes CDN, the actual geographical location will have no correlation

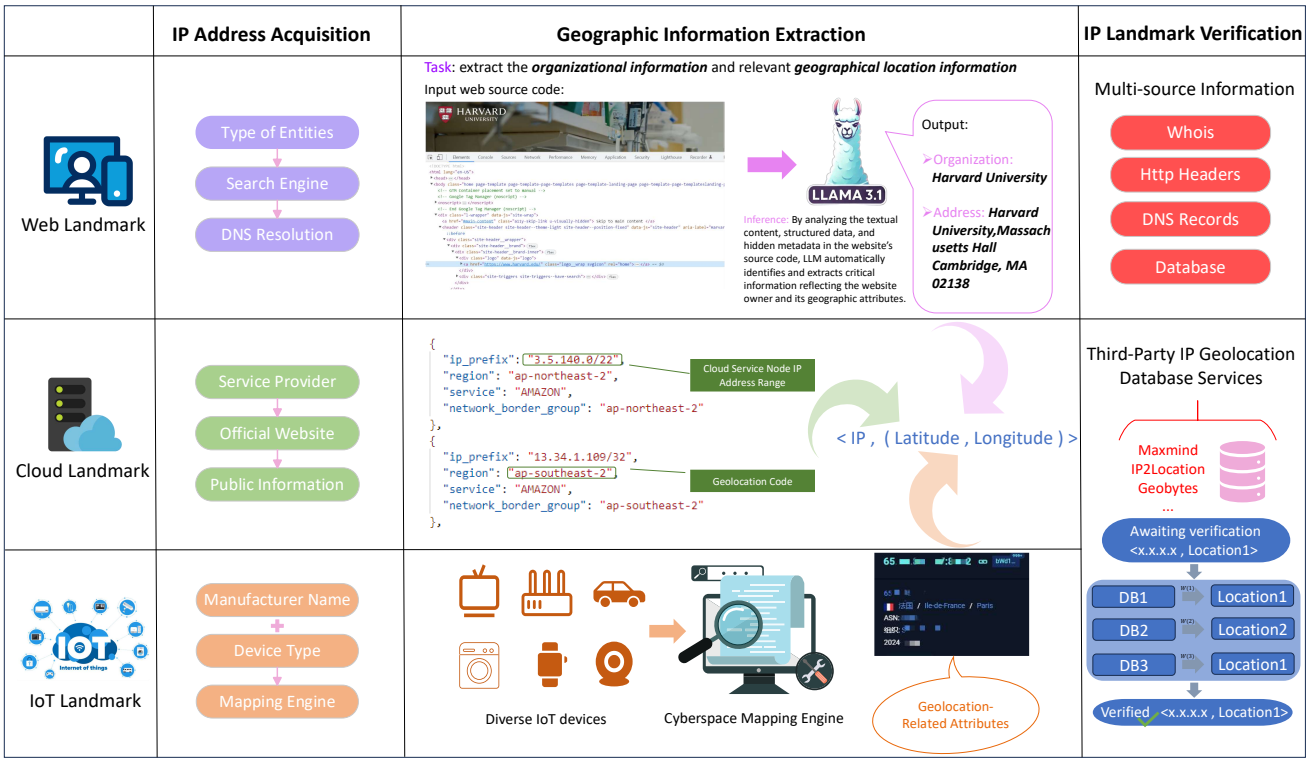

Fig. 1. The framework of HybridMark.

with the webpage content. This process can be broken down into four steps, as summarized below.

1) **Whois Information Query.** We begin by using the Whois protocol to query the registration details of the target domain, extracting and analyzing fields such as "domain_name," "registrar," "name_servers," "emails," "org," and "name." These fields often reveal the actual owner of the domain. If certain field values correspond to globally recognized internet companies (e.g., Google, Amazon, Cloudflare), it can be preliminarily inferred that the domain is likely utilizing third-party cloud or CDN services. To improve filtering efficiency, we have built a blacklist of famous internet companies, cloud service providers, and CDN service providers, with a subset shown in Table I. Query results are systematically cross-referenced against this blacklist to exclude domains that are unlikely to be locally hosted.
2) **HTTP Header Information Analysis.** Certain HTTP header fields can provide insights into the deployment environment of a web server. For example, the "X-Cache" field, often added by caching proxy servers such as CDNs or reverse proxies, indicates whether the request was served from the cache. The "Via" field reveals the intermediary proxy servers that processed the request, commonly used for debugging or diagnosing routing issues. Similarly, the "X-CDN" field typically includes information about the service provider. The presence of these fields strongly suggests the use of CDN or proxy server technologies, indicating that the website is not locally hosted. Such websites are excluded from the dataset. Additionally, the "Server" field, which provides information about the server software, may also disclose the service provider's name. This information is cross-referenced against the blacklist compiled in the first step to further filter out candidates that are unlikely to be locally deployed.
3) **DNS Record Analysis.** Next, we analyze various DNS records, including CNAME, NS, MX, and PTR, to refine the filtering process. If any of these records contain names of prominent service providers (e.g., Google, AWS), it indicates that the domain is likely hosted using third-party services. Additionally, discrepancies in A records for the same domain, observed across multiple geographically distributed DNS servers, may suggest the use of CDNs or load balancing mechanisms. To ensure thorough analysis, we resolve each domain using 30 globally distributed DNS servers. If all servers return the same IP address for a given domain, it is likely locally hosted. Conversely, variations in the resolved IP addresses imply reliance on CDNs or similar technologies.
4) **Third-Party IP Geolocation Database Validation.** Finally, third-party IP geolocation databases, such as MaxMind [21] and IP2Location [22], are employed to validate whether the actual IP address corresponds to the geographical location predicted by LLM. In practice, the

TABLE I
GLOBAL INTERNET SERVICE PROVIDERS.

Region	Internet Service Providers
United States	Google, Amazon, Microsoft Azure, IBM Cloud, Oracle Cloud, DigitalOcean, Linode, Vultr, Akamai, Cloudflare, Fastly, CDNetworks, Verizon Media, Lumen Technologies, StackPath, Rackspace, Facebook, Apple, Salesforce, Dropbox, Netlify, GitHub, Heroku, Squarespace
China	Alibaba Cloud, Tencent Cloud, Baidu Cloud, Huawei Cloud, ChinaCache, ChinaNetCenter, JD Cloud, Sina, Weibo, Xiaomi, Kingsoft Cloud, Qihoo 360, YunDun, UCloud, ByteDance, Meituan, Baidu, Tencent, Alibaba, Huawei
Europe	OVHcloud, Hetzner, Scaleway, Gandi, IONOS, Leaseweb, Swisscom, Orange Business Services, Telefónica, Vodafone, Deutsche Telekom, BT Group, Telenor, KPN, Proximus, Clouvider, A1 Telekom Austria, Zayo Group
Japan	NTT Communications, SoftBank, Rakuten, KDDI, IDC Frontier, Sakura Internet, GMO Internet, Fujitsu, NEC, Hitachi
South Korea	KT Corporation, SK Broadband, LG Uplus, Naver, Kakao, Samsung SDS
India	Tata Communications, Bharti Airtel, Reliance Jio, HCL Technologies, Wipro, Infosys, ZNetLive
Russia	Yandex, Mail.Ru, Rostelecom, Selectel, DataLine, Xelent, Croc
Australia	Telstra, Optus, NextDC, Macquarie Telecom, Aussie Broadband, Anchor
Canada	Rogers Communications, Bell Canada, Telus, Cogeco, Shaw Communications, Tucows
South America	Lumen Latin America, UOL Host, Locaweb, Globenet, Tigo Business, Movistar, Vivo
Middle East	Etisalat, Saudi Telecom Company, Zain, Ooredoo, Bezeq, Mobily
Africa	MTN, Vodacom, Orange Egypt, Liquid Telecom, Seacom, Telkom SA, Econet Wireless
Other Global Providers	ProxyMesh, Bright Data, ScraperAPI, GeoSurf, Oxylabs, Storm Proxies, Zyte

validation is conducted at the provincial or city level, and only those landmarks where the geolocation data from the database and the model results align at this level are retained, thereby enhancing the overall accuracy.

By applying the four-step process outlined above, landmark data from websites of 1,503 universities were systematically collected and rigorously validated, ultimately yielding 857 Web landmarks that meet the established criteria for validity.

B. Cloud-based Landmark Generation

The rapid expansion of cloud services in the modern internet landscape has driven major service providers to accelerate the development of cloud infrastructure, particularly through the construction of large-scale data centers. This trend creates opportunities to leverage cloud service nodes as landmarks in IP geolocation. Cloud service nodes are widely distributed across the globe, featuring high availability and low network jitter. Moreover, many cloud service providers make the IP address ranges and corresponding regions of their cloud nodes publicly available on their official websites [44]–[46]. By knowing the physical location of a data center within a specified region, it is possible to associate the IP address of a cloud service node with a precise geographic location. In practice, numerous Data Center Location Service Providers offer detailed data center location information [47]–[50], enabling the efficient collection of high-density landmark datasets. While some studies have attempted to validate the authenticity of geographic locations provided by cloud service providers, no existing research has yet utilized cloud service nodes for the construction of landmark datasets.

This section presents a method for collecting cloud-based landmarks, aimed at identifying and locating cloud service nodes by leveraging publicly available information from cloud service providers. Specifically, we gather public IP address ranges and their associated geographical location data from the official websites of major cloud providers. Many cloud providers share details about the IP ranges of their nodes and their approximate geographical locations, typically at the city level. This information helps users select service nodes closer to their physical locations, thereby reducing network latency. Once these public datasets are obtained, any additional information regarding the precise locations of cloud data centers within specific cities can further refine the mapping of public IP addresses to the exact latitude and longitude of the data centers. By analyzing these IP address ranges, we can extract reachable IP addresses that serve as valid IP geolocation landmarks. The process can be divided into three main steps.

IP Address Collection: We selected globally recognized cloud service providers, including Amazon, Microsoft, and Oracle, and accessed their official websites to retrieve publicly available IP address range information. Many providers publish the IP address ranges associated with their cloud service nodes. As shown in Figure 2, the data is sourced from publicly available information provided by Amazon Web Services (AWS). On the left, the entries list specific IP prefixes related to the AWS infrastructure. Two key data points of interest are:

- ”ip_prefix”: This specifies a range of IP addresses.
- ”region”: This indicates the geographic region associated with the IP prefix.

By publishing IP address ranges and their associated geographic locations, cloud service providers enable users to select data centers or nodes closer to their physical location. This helps reduce network latency, improves data transfer speeds, and enhances overall service quality. Furthermore, such transparency allows users to gain a clearer understanding of the deployment regions and infrastructure distribution of the services. This openness fosters greater user trust, particularly among those with stringent data privacy and compliance requirements. In certain industries and regions, specific geographic requirements for data storage and processing are enforced. For instance, the European Union’s General Data Protection Regulation (GDPR) [51] mandates that data must be stored and processed within designated regions. By disclosing geographic information, cloud providers help customers ensure regulatory compliance, making such disclosures a logical step for providers.

Data Center Location Identification: When publishing IP address ranges, providers also provide corresponding geographic region information. These regions are typically represented by codes, which, although varying across providers, generally correspond to city- or province-level areas. For

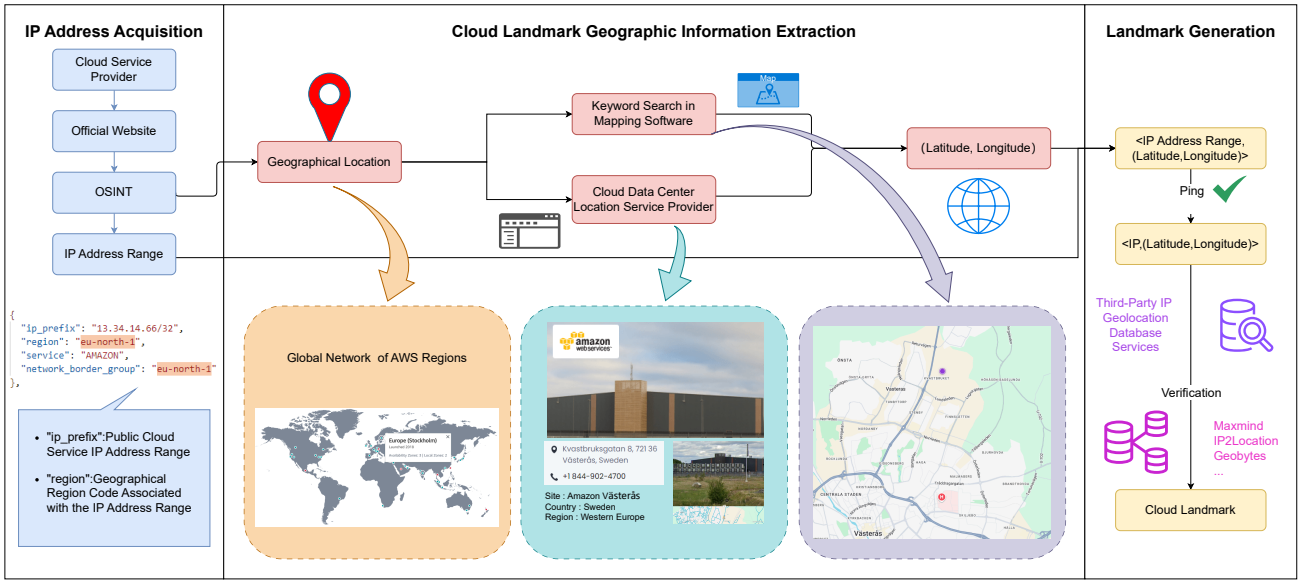

Fig. 2. The framework of cloud-based landmark generation.

example, the code "eu-north-1" represents AWS data centers in Northern Europe. By combining publicly available IP address ranges with geographic region codes, we can broadly identify the geographic distribution of cloud service nodes. For each IP range and its approximate region, we aim to identify the precise location of the cloud service provider's data center within that area. In some instances, specific nodes can be located by searching for key terms such as the "provider name" combined with "cloud data center" in mapping software. Additionally, data center location services can be utilized.

Data center location service providers typically aggregate information related to data center hosting, cloud services, connectivity, hardware, software, and managed services from global vendors [47]–[50]. These services enable users to search data center directories, compare provider offerings, and procure technological solutions. One critical piece of information they provide is the geographical location of cloud service providers' data centers, which directly influences service latency for users. These services compile data on data center locations, facilities, and providers through various methods, including the presentation of addresses, descriptions, images, and video tours. By leveraging these resources, we can efficiently identify the geographic locations of cloud service providers' data centers.

For example, as shown in Figure 2, we collected detailed information about Amazon's data centers, including geographic coordinates, contact numbers, site names, and regional data. In some instances, on-site photos were also available. Using third-party mapping software, we cross-referenced these locations and validated them with features such as real-world imagery. Ultimately, we associated specific IP address ranges with precise latitude and longitude coordinates.

Landmark Generation: In this phase, reachable and active IP addresses are filtered from the identified IP address ranges through a ping scan. Specifically, for each IP address within a given range, a network probe is performed using the ICMP protocol, sending a ping request to assess the network connectivity of the target IP address. By analyzing the ping response, we can determine which IP addresses respond successfully, indicating their availability and reachability at the network layer. This process ensures the robustness and stability of the selected IP addresses.

To optimize the process and minimize network impact, a maximum of four ping requests are sent to each target IP address. If no response is received, the IP address is excluded from further consideration. For an entire IP range scan, if two live landmarks are found, the range is skipped. By adjusting this strategy during actual collection, the number of landmarks gathered can be significantly increased. Finally, after verification through multiple database sources (see section C for details), the selected IP addresses are confirmed as valid and reliable cloud-based landmarks.

C. IoT-based Landmark Generation

IoT-based landmarks refer to IoT devices that are connected to the internet, have public IP addresses, and whose geographic location information can be retrieved. These devices are widely distributed across various regions and network environments, such as industrial networks and residential communities, where they serve as environmental monitoring sensors, smart home devices, and other types of equipment. The diversity of device types further contributes to the variability of their geographic location data. In this section, we propose a systematic method for collecting IoT-based landmarks, as detailed below.

IoT-based Landmark IP Address Acquisition and Geolocation Extraction: First, we gather a set of information about IoT devices, including basic data such as the manufacturers, device types, and their names. Next, this information is integrated into search keywords by combining attributes like "manufacturer" and "device type" (e.g., "Hikvision +

camera"). These keywords are then input into a cyberspace mapping engine [52]. Utilizing the engine's advanced search and analysis capabilities, we retrieve relevant data about the devices, including IP addresses, port numbers, domain names, operating systems, and associated countries or regions. From this, we extract key details to generate (IP, (Latitude, Longitude)) key-value pairs. The mapping engine employs refined query algorithms and device fingerprinting techniques to associate each matched IoT device with its corresponding IP address and geographic location, thus enabling the collection of a large amount of IoT device data.

IoT-based Landmark Verification: Building on this, we further validate and filter the data. First, to avoid issues arising from Network Address Translation (NAT), where multiple IoT devices share a single IP address, we eliminate duplicate IP address entries. It is well-known that most IP geolocation databases often underperform compared to their claimed accuracy, with many providing suboptimal results in practice [28], [53]. Additionally, geolocations provided by cyberspace mapping tools may rely on single-source database data [16], which can lead to inaccurate geolocation estimates for IoT devices. To address this issue and improve the accuracy of the landmarks, VoteGeo [16] employs multi-source database verification. For each IP address, queries are made to multiple distinct IP geolocation databases, and only IP addresses for which all results are consistent across the databases are considered valid landmarks. Assuming P_i is the accuracy rate claimed by each geolocation database provider, the overall accuracy of the landmarks validate using the multi-source database method can be calculated using (1).

$$P = \frac{P_1 P_2 \dots P_n}{P_1 P_2 \dots P_n + (1-P_1)(1-P_2)\dots(1-P_n)} \tag{1}$$

This approach inspire us to model conditional dependencies based on the characteristics of each database. In practice, certain databases exhibit stronger regional accuracy due to factors such as privacy policies or data localization. For example, IP138 [26] performs notably better in China compared to databases like MaxMind and IP2Location. By adjusting predictions according to the regional strengths and limitations of each database, we can enhance the overall accuracy.

To implement this, we define a region-specific characteristic function for each database. This function represents the relative confidence of database i for region A, serving as a weighting factor. For instance, if database i demonstrates higher accuracy in region A than its global average, the function value increases; conversely, it decreases when the database's accuracy is lower in that region. The overall accuracy of database i globally can be represented as a probability $P(\text{correct} \mid \text{DB}_i)$, and its accuracy in region A is computed as:

$$W(i) = P(\text{correct} \mid \text{DB}_i) \times R_i(A) \tag{2}$$

where $R_i(A)$ denotes the relative accuracy adjustment for region A.

For example:

- If database i is 20% more accurate in North America than globally, $R_i(\text{North America}) = 1.2$.
- If it is 10% less accurate in the Asia-Pacific region, $R_i(\text{Asia-Pacific}) = 0.9$.

When estimating the geographic location L of an IP address based on predictions from multiple databases, each database's prediction for region A is weighted by $W(i)$. The final result is obtained by summing the weighted predictions and selecting the one with the highest combined weight. If this result aligns with the region-level prediction from cyberspace mapping engine, the IP address is considered a valid geolocation landmark. We extract landmarks from multiple continents and use different databases to verify their accuracy. Table II lists the accuracy weights of the Maxmind and IP2Location databases across various continents, reflecting the differences in the geographic location performance of these databases.

TABLE II
ACCURACY WEIGHT VALUES OF GEOLOCATION DATABASES ACROSS DIFFERENT CONTINENTS

Database	Africa	Americas	Asia	Europe	Oceania
IP2Location	0.7532	0.5584	1.0649	0.9091	1.0519
Maxmind	1.1912	0.4706	1.0588	1.0441	1.3529

IV. EVALUATION

Currently, many IP geolocation studies evaluate landmark dataset, typically focusing on three aspects: quantity, distribution, and accuracy. Common evaluation methods include statistical analysis of landmark quantity, visual representation of distribution, and accuracy measurement through comparison with database results and error calculation [8], [13], [14], [16]. However, there are two main drawbacks in this evaluation approach. First, the evaluation dimensions are insufficiently comprehensive, as important factors such as density and stability—key indicators in many contexts—are not considered. Second, the lack of a unified, quantitative framework hinders intuitive comparisons and objective quality assessments across different datasets, reducing the practical utility and reference value of landmark databases.

In this section, we systematically evaluate the collected landmark dataset from five critical dimensions to assess its quality and practical applicability. The evaluation focuses on quantity, coverage, density, stability, and accuracy. To evaluate coverage and stability, we introduce innovative methodologies, including grid area analysis and time series analysis, for a quantitative evaluation. Additionally, we propose a new density metric to enhance the comprehensiveness of the evaluation. Moreover, we replicate established algorithms and perform comparative analyses using both our landmarks and existing open-source landmarks. This comparison aims to evaluate the accuracy of our landmarks and demonstrate their effectiveness in supporting geolocation tasks.

A. Landmark Quantity

Regarding the quantity of landmarks, we collect a total of 913,177 landmarks globally within a single month of continuous data gathering, covering 194 countries. The distribution among the three landmark categories is approximately balanced. To evaluate the effectiveness of the collected

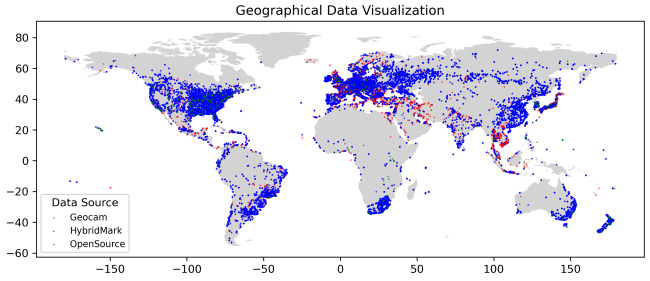

Fig. 3. Geospatial distribution of landmarks.

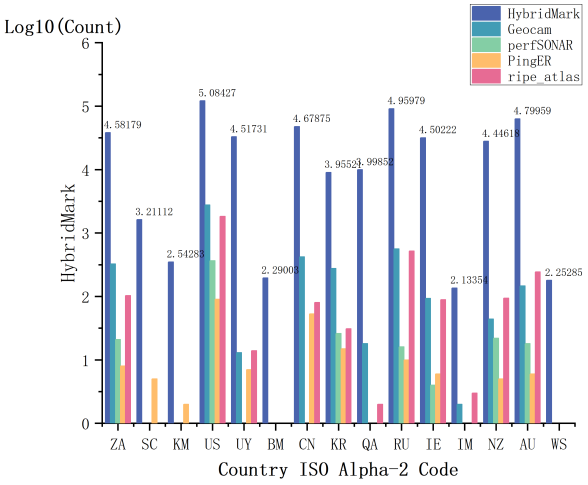

Fig. 4. Comparison of the number of landmarks.

landmarks, we compare our dataset with several well-known open-source landmark datasets, including perfSONAR [54], PingER [55], RIPE Atlas [35], and GeoCAM [8]. PerfSONAR, PingER, and RIPE Atlas are three widely used tools in the field of network measurement and monitoring, maintained by dedicated organizations. These tools utilize globally distributed probes which can be used as landmarks for IP geolocation to monitor network performance. GeoCAM, the current state-of-the-art research claims to include only 16,863 landmarks spanning 170 countries, making HybridMark approximately 54 times larger in terms of landmark quantity. To visually represent the global distribution of these landmarks, we create a geographical distribution map, with each point representing a cluster of nearby landmarks, as shown in Figure 3. Due to the large volume of data, each point aggregates multiple landmarks in proximity. In this map, blue nodes represent landmarks collected by HybridMark, red nodes correspond to GeoCAM landmarks, and green nodes denote open-source landmarks. In most regions, it is clear that the number of our nodes significantly exceeds those in the other datasets. However, in specific areas, such as the Middle East and Southeast Asia, the node count in HybridMark is lower than that of GeoCAM. Empirical validation focusing solely on the accessibility of landmarks reveals that, among the 2,971 landmarks reported by GeoCAM in these regions, only 1,283 are reachable, resulting in an accessibility rate below 50%. This finding highlights the challenges in ensuring the practical usability of GeoCAM's nodes in these areas, despite their numerical advantage.

Figure 4 illustrates a comparison of the number of landmarks in HybridMark versus four other datasets across different countries, presented in bar graph format. The horizontal axis represents ISO alpha-2 country codes, while the vertical axis represents the base-10 logarithmic transformation of the landmark counts, ensuring clarity in the visualization of the data distribution. Countries or regions from the Americas, Europe, Africa, Asia, and Oceania are selected based on geographic area, including the largest, smallest and median-sized country within each region. This selection strategy guarantees a diverse set of samples, covering various geographic scales and providing a solid foundation for performance evaluation. In total, three countries from each of the five regions—fifteen countries in total—are chosen for display and analysis. At the top of each bar group, the number of landmarks collected by HybridMark and the second-largest dataset are explicitly labeled. The results demonstrate that HybridMark ranks first in landmark quantity across all countries, outperforming the second-largest dataset by at least a factor of eight. Notably, for certain countries or regions with small geographical areas, HybridMark significantly exceeds others in landmark quantity, as detailed in Table III.

TABLE III
DISTRIBUTION OF LANDMARKS ACROSS COUNTRIES AND REGIONS BY DATABASE

ISO	HybridMark	GeoCAM	perfSONAR	PingER	ripe_atlas
WS	179	0	0	0	0
BN	32	0	0	5	1
GQ	47	0	0	0	1
FJ	81	0	1	1	2
BF	172	6	0	6	6
GM	227	0	0	2	0
KM	349	0	0	2	0
BQ	466	0	0	0	0
IR	883	238	0	9	102
IM	136	2	0	0	2
BM	195	0	0	0	0

B. Coverage

In the field of landmark data collection and evaluation, coverage analysis is typically conducted using visual mapping techniques, which provide an intuitive display of the spatial distribution of landmarks on geographical distribution maps. However, this method has significant limitations when dealing with large volumes of landmark data. The dense distribution of data can lead to visual clutter, making it difficult to distinguish individual landmarks effectively. Furthermore, relying solely on graphical representations lacks quantitative analysis tools, making it difficult to provide precise numerical metrics to evaluate the coverage quality of the landmark data, thus limiting objective comparisons between datasets. There is an urgent need to introduce a unified quantitative metric system to supplement the shortcomings of visual mapping, offering

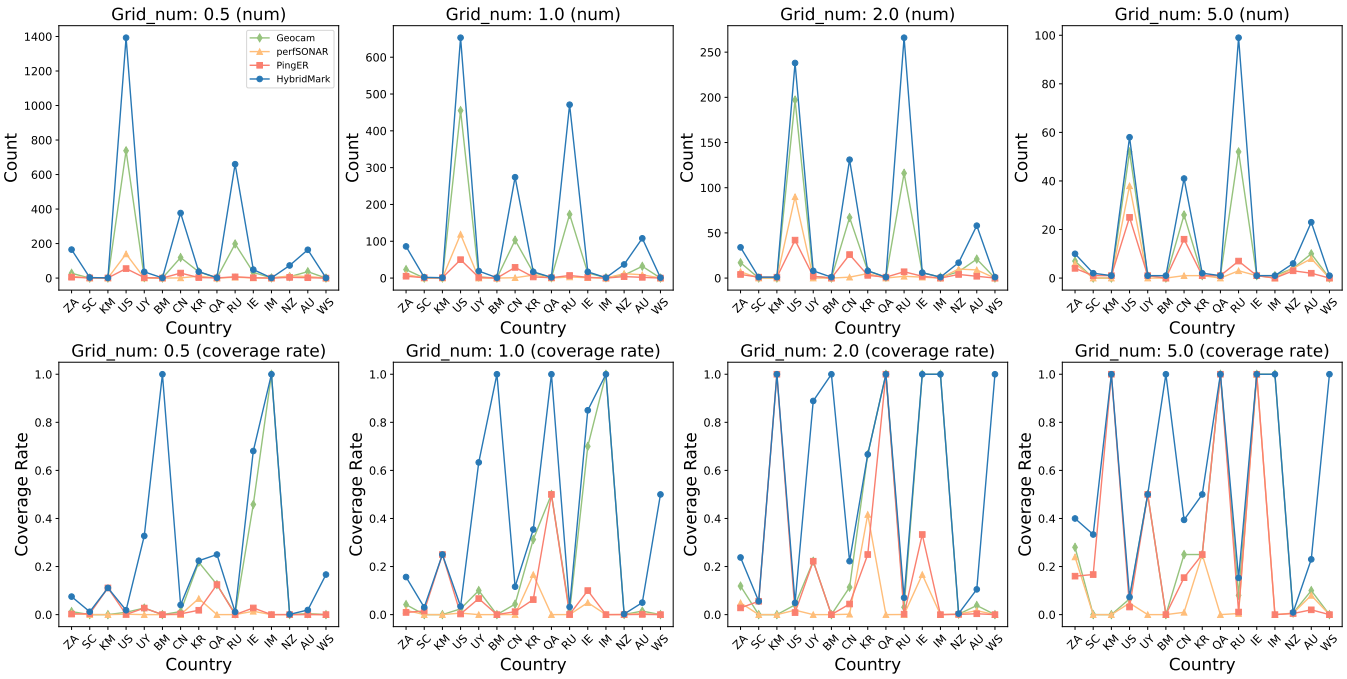

Fig. 5. Comparison of count and coverage metrics across datasets by grid number.

more comprehensive and scientifically grounded support for landmark data coverage analysis and optimization.

To scientifically assess the coverage of different landmark datasets in target areas, we propose a novel area coverage analysis method based on grid partitioning. This method divides the target area into fixed-size geographic grid cells to measure the spatial distribution and coverage of landmarks. Specifically, the target area is divided into regular grids according to fixed latitude and longitude intervals. Then, we count the number of data points from each dataset falling into these grids and calculate the proportion of the grid cells that are covered.

In this section, we analyze the coverage of four geographic datasets: GeoCAM, perfSONAR, PingER, and the landmark dataset collected by HybridMark. Due to the difficulties in obtaining global node data from RIPE Atlas, this dataset is excluded in this section. To comprehensively evaluate the coverage performance of these datasets, we selected 15 representative countries and set grid intervals for latitude and longitude at 0.5°, 1°, 2°, and 5°. We conduct the experiment using the grid area method, calculating two indicators: the number of covered grids (the total number of grid cells containing at least one landmark) and the coverage ratio (the proportion of covered grids to the total number of grids in the target area). Given the complexities of using national boundary data due to political factors, we selected individual countries and defined the target area as the rectangular region encompassing the country's maximum and minimum latitude and longitude. This method serves as an approximation for spatial boundaries and does not significantly affect the overall distribution characteristics of the data.

The result of the calculation and comparison of the number of covered grid cells and the coverage rates of different datasets under four grid sizes is shown in Figure 5. The top four line charts illustrate the comparisons of the number of covered grid cells across different grid scales, while the bottom four show the comparisons of coverage rates. The experimental results indicate that HybridMark outperforms other datasets in terms of both the number of covered grid cells and coverage rates across all grid sizes, with particularly notable performance at finer grids of 0.5° and 1°. In some smaller countries, our dataset's coverage rate is close to or even reaches 100%, demonstrating its stronger spatial adaptability across diverse geographic environments. While the GeoCAM exhibits a relatively high number of covered grid cells in larger countries (e.g., Russia and Brazil), its coverage rate is lower, suggesting that its points are concentrated in a few key regions and fail to span the entire geographic area. As the grid scale increases, both coverage rates and the number of covered grid cells increase across all datasets. This is due to the reduction in the total number of grids as grid cell size increases, which in turn obscures spatial details. Therefore, the grid resolution can be adjusted based on specific requirements and localization accuracy in practical evaluations. Additionally, the data from 15 countries highlight significant differences in coverage across continents and countries. In technologically advanced regions, such as the Americas and Europe, datasets exhibit higher coverage compared to Africa, which may be attributed to better network infrastructure and richer geographic information collection resources in those regions. For smaller countries (e.g., Monaco and Singapore), coverage rates are typically high, although the differences in the number of covered grid cells remain significant.

To further visually compare the coverage of different datasets, we select the United States as the target country for two main reasons: its vast geographic area and diverse

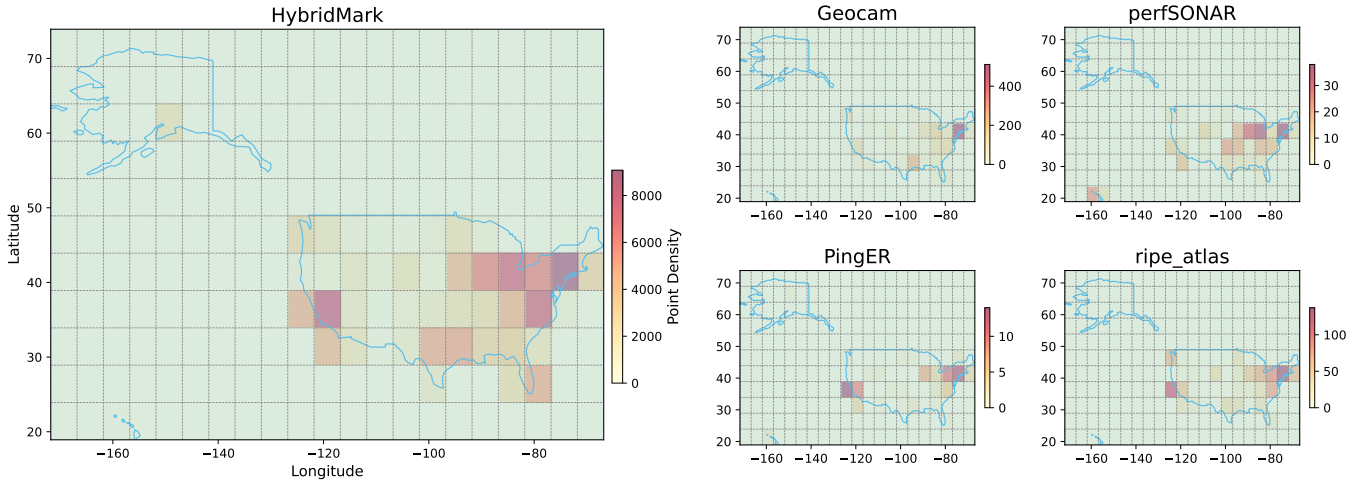

Fig. 6. Geospatial heatmap of landmark distribution in the United States.

terrain, which result in a broader and more varied landmark distribution, and its highly developed internet infrastructure with comprehensive network node coverage. Additionally, the U.S. exhibits a significant number of nodes in all existing landmark datasets. Using the grid area method, we generat heatmaps for each dataset to illustrate their coverage, as shown in Figure 6. In the heatmap, color intensity represents the grid point density (i.e., the number of data points within a grid cell). Darker regions correspond to higher data point densities, while lighter regions indicate sparse or uncovered areas, as indicated in the legend to the right of each map. Specifically, we display our dataset's extensive coverage in the U.S. through a large-scale heatmap, contrasting it with smaller heatmaps for the other four datasets, highlighting notable differences in coverage density and spatial range. The horizontal and vertical axes of the maps represent longitude and latitude, respectively, with grid intervals set to 5°. The results reveal that HybridMark significantly outperforms the others in terms of both coverage and point density. It demonstrates a wide and uniform coverage range, especially in remote and less internet-developed areas such as Hawaii and Alaska, where its performance surpasses that of the other datasets.

C. Density

In IP geolocation, achieving high-precision, fine-grained street-level localization for a specific region requires a sufficiently high density of landmarks within a small area. This necessitates not only a sufficient number of landmarks but also a high landmark density. In such cases, metrics such as coverage and quantity alone are insufficient. However, most existing research focuses primarily on the spatial distribution and quantity of landmarks, with relatively little attention given to density analysis. IP geolocation algorithms are highly sensitive to the density of landmark distribution. For instance, in high-density regions, network measurement and machine learning algorithms tend to yield greater accuracy, while in sparse regions, localization errors may increase significantly. Urban areas generally require higher landmark density to achieve fine-grained localization, whereas rural areas prioritize broader coverage. Therefore, evaluating density is essential for analyzing performance variations across regions and for designing hybrid algorithms tailored to different density distributions and application scenarios. In this paper, the average nearest neighbor algorithm is used to reflect the density characteristics of a target region by calculating the nearest neighbor distance among all landmarks in the area.

The Mean Nearest Neighbor Distance (MNN) is a classical method for quantitatively evaluating the spatial distribution density of a point set, widely applied in fields such as geographic information science, ecology, and cyberspace analysis. In the context of landmark dataset density analysis, MNN characterizes spatial distribution by calculating the nearest neighbor distance between points. Let the landmark dataset be represented as a point set $S = \{p_1, p_2, \ldots, p_n\}$, where each point p_i has coordinates (x_i, y_i). The nearest neighbor distance $d_{\text{NN}}(p_i)$ is defined as the minimum Euclidean distance from point p_i to all other points:

$$d_{\text{NN}}(p_i) = \min_{j \neq i} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3}$$

The MNN is then calculated as the average of the nearest neighbor distances of all points:

$$\text{MNN} = \frac{1}{n} \sum_{i=1}^{n} d_{\text{NN}}(p_i) \tag{4}$$

A smaller MNN value indicates a denser point set distribution and a more compact landmark data coverage, while larger values may suggest the presence of localization blind spots. Figure 7 illustrates the comparison of the average nearest neighbor distances across landmark datasets in different countries. The horizontal axis represents the 15 target countries selected in section A, while the vertical axis denotes the MNN of all landmarks within each country. For datasets with no landmarks in certain regions, we set the MNN distance to 500. From the figure, it is evident that the landmark dataset constructed by HybridMark exhibits lower MNN distances in most countries and regions. For geographically smaller

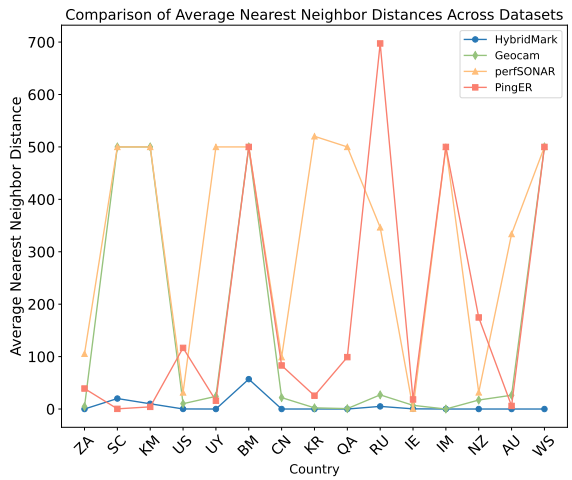

Fig. 7. Comparison of average nearest neighbor distances across datasets.

countries, such as Seychelles, the calculation of MNN distance for PingER appears significantly lower due to the extremely limited number of landmarks available in these regions. This is because the computation of MNN distance is highly correlated with the density of landmark distribution. When the number of landmarks is insufficient and their distribution is highly concentrated, the nearest-neighbor distances tend to be shorter, thereby lowering the overall average value. Overall, compared to existing datasets, HybridMark exhibits a distinct advantage in terms of density.

D. Stability

Landmark stability refers to the performance of a landmark node to maintain consistent network delay-related latency data across different time scales. Stable landmarks contribute to reducing localization inaccuracies caused by network fluctuations, equipment failures, or malicious attacks, thereby enhancing the overall accuracy and robustness of the localization system. Additionally, collecting highly stable landmarks can decrease the frequency of landmark data collection, thus lowering the cost of IP geolocation. However, existing studies lack a quantitative analysis of this stability metric. In this work, we employ the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model, a method from time series analysis, to quantify the stability of landmarks based on the characteristics of network latency data.

Time series analysis uses statistical methods to extract patterns or trends from data organized in temporal order, enabling the modeling and analysis of dynamic changes in behavior. In the evaluation of network stability, time series analysis can uncover the temporal characteristics of factors such as latency and bandwidth fluctuations. The GARCH model is a key method for handling the phenomenon of volatility clustering in time series data. Its core idea is that current volatility is influenced not only by past volatilities but also by error terms from previous time points. This model focuses on both mean fluctuations and dynamic changes in variance. Network latency data often exhibit volatility and non-stationarity, with fluctuations driven by factors such as network congestion, routing changes, or system failures. Consequently, latency data are typically non-stationary and exhibit strong temporal dependencies and volatility clustering. Widely used in finance, the GARCH model is a powerful statistical tool for capturing these characteristics, making it well-suited to modeling the fluctuations inherent in latency data.

Network latency data is often represented as a time series. Let $\{y_t\}_{t=1}^{T}$ denote a sequence of latency values, where each timestamp t corresponds to a latency value y_t, representing the transmission delay from the source node to the target node. Latency data typically exhibits the following characteristics: **Seasonality**: Variations in network load during different periods can cause periodic fluctuations in latency. **Burstiness**: Factors such as network congestion, failures, or routing changes may lead to significant spikes in latency. **Non-stationarity**: The variance of latency data often change over time, demonstrating volatility clustering, where some periods show high variance while others exhibit low variance. Thus, the stability of latency data is not only related to the mean level but also to its volatility. Compared to traditional autoregressive models, the GARCH model focuses on modeling the volatility patterns in latency data, particularly the dynamic changes in variance.

Suppose we have a set of latency data sequences $\{y_t\}_{t=1}^{T}$, where the data at each time t can be expressed as:

$$y_t = \mu + \epsilon_t, \tag{5}$$

where μ is the mean value of the latency sequence, and ϵ_t is the residual term, representing the deviation of the actual latency from the mean latency. The key to the GARCH model lies in the conditional volatility σ_t^2, which indicates the level of variability in latency data at time t. The mathematical form of the GARCH(1, 1) model is as follows:

$$\epsilon_t = \sigma_t z_t, \quad z_t \sim N(0,1), \tag{6}$$

where z_t is a standard normal random variable representing the error term. According to the GARCH(1, 1) model, the conditional variance σ_t^2 of the latency data can be expressed as:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \tag{7}$$

where:

- σ_t^2 is the conditional variance of latency data at time t, indicating its volatility.
- α_0 is a constant term, representing the long-term average volatility.
- α_1 is the coefficient of the autoregressive term, measuring the influence of past error terms (ϵ_{t-1}^2) on current volatility.
- β_1 is the coefficient of the moving average term, representing the impact of past volatility (σ_{t-1}^2) on current volatility.

By applying the GARCH model to latency data, we can quantify the volatility of latency at each time step and use this

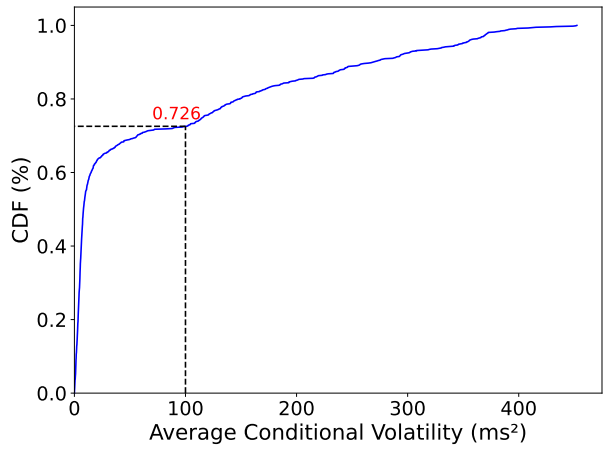

Fig. 8. CDF of average conditional volatility for IP addresses.

information to evaluate the stability of landmarks. For each IP address or network node with latency sequence $\{y_t\}_{t=1}^{T}$, the conditional volatility σ_t^2 can be calculated. The average volatility of the entire latency sequence can then be determined as:

$$\text{Average Volatility} = \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2. \tag{8}$$

We randomly select 10,000 landmarks and conducted latency tests at different times of the day (08:00, 16:00, and 00:00 CST, UTC+8) over a month. For each landmark, we calculate the average volatility. The resulting CDF chart is presented as Figure 8, which illustrates the distribution of the average volatility across all landmarks in the dataset. The x-axis represents the average volatility. From the graph, it is evident that when the average volatility is low (near the left end of the x-axis), the y-axis increases sharply, suggesting that most landmarks exhibit low volatility and thus high stability. As the average volatility rises, the curve flattens, indicating a decreasing proportion of landmarks with higher volatility. About 72.6% of the landmarks have an average volatility below 100, which means the standard deviation of latency fluctuations is approximately 10 milliseconds. This implies that the majority of latency data fluctuates within a ±10ms range, indicating that the landmarks we collected exhibit good stability.

By applying the GARCH model to analyze latency data, we obtain the conditional volatility at each time step, which is then used to calculate the average volatility, reflecting the extent of latency variation and its instability. In practical applications, higher conditional volatility typically correlates with more pronounced latency fluctuations and increased network instability, whereas lower conditional volatility indicates relatively stable latency and more reliable network performance. Furthermore, the GARCH model effectively addresses the non-stationarity in latency data, where volatility changes over time rather than remaining constant. Over extended observation periods, network latency volatility may fluctuate due to factors such as network load and topology changes. The GARCH model dynamically adjusts its volatility estimates, making it particularly well-suited for handling such non-stationary behavior. By modeling the latency data of different IP landmarks with GARCH, we can quantify the stability of each landmark and analyze stability differences among them, thus providing a solid foundation for landmark selection in geolocation algorithms.

E. Accuracy

We use two mainstream commercial geolocation databases, MaxMind GeoLite2 and IP2Location to evaluate the accuracy of the IP geolocation landmarks we collect. The geographic positions of the landmarks are verified by calculating the distance deviation between the latitude and longitude provided by these databases and the geographic coordinates of the HybridMark landmarks. Specifically, the latitude and longitude of each landmark served as the baseline. We query the two databases for each IP address to obtain their corresponding geographic positions, then calculate the geographic distance between these results and the baseline values. Figure 9 presents the CDF curves for the deviation distances between the commercial databases and the HybridMark landmarks. The horizontal axis represents the magnitude of deviation distances (in kilometers). From the overall characteristics of the CDF curves, it is evident that the GeoLite2 and IP2Location curves rise sharply for deviation distances under 30 kilometers, indicating that most landmarks have deviation concentrated within this range. MaxMind GeoLite2 achieves approximately 67.76% coverage of landmarks within a 50-kilometer deviation, while IP2Location reaches 72.66%. Considering that the typical diameter of most cities is around 50 kilometers, these results demonstrate high accuracy for city-level location representation. This aligns with the claims of commercial geolocation databases regarding their city-level precision and validates the reliability of HybridMark landmarks in city-level scenarios.

In addition, we reproduce the classic CBG algorithm in IP geolocation and compare its performance using the HybridMark landmark set with that of publicly available open-source landmark sets to evaluate the accuracy of HybridMark landmarks. Specifically, we apply the CBG algorithm to estimate the geographic location of target nodes using two different landmark sets and compare the results with the true geographic location, calculating the positioning error for each test node. Figure 10 presents the CDF curves of the positioning errors for both landmark sets, visually illustrating the performance differences. The horizontal axis represents the error distance (in kilometers), while the vertical axis indicates the cumulative probability. The results show that the positioning accuracy using the HybridMark landmark set is significantly better than that achieved with the open-source set, as evidenced by the faster upward trend of the CDF curve. Specifically, 54% of the targets have an error of less than 500km, and 99% have an error of less than 1000km when using HybridMark. In contrast, only 29% of the targets achieve an error of less than 500 km, and 72% fall below 1,000 km with the open-source set. On average, the positioning

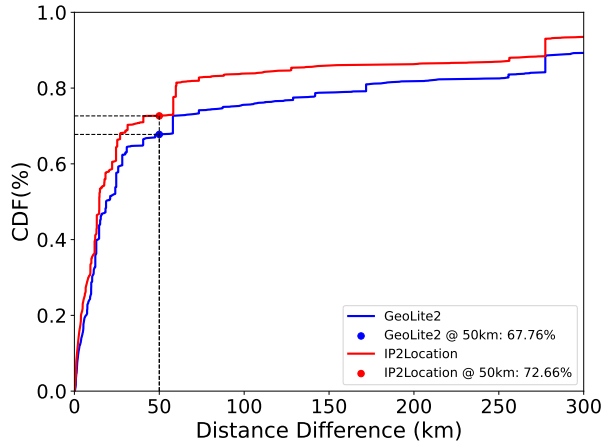

Fig. 9. CDF of distance differences across databases.

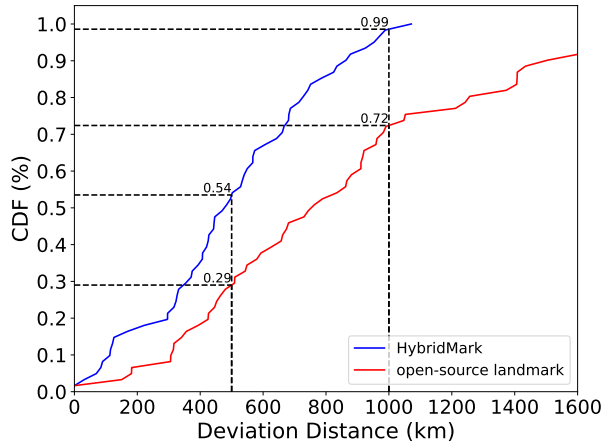

Fig. 10. CBG with HybridMark/open-source landmarks.

accuracy of the CBG algorithm improves by approximately 50% when using the HybridMark landmark set. These findings highlight the advantages of HybridMark in terms of geographic coverage, density, and stability, which provide higher-quality constraint information for the CBG algorithm and contribute to its improved performance.

V. Discussion

Our study has collected a large-scale, high-quality, and low-cost set of landmarks, significantly surpassing previous landmark datasets in terms of quantity, coverage, density, and accuracy. However, there remain areas where the current design could be further refined. Below is a discussion of the limitations and potential issues of the method presented in this paper:

Non-local Deployment of Web Sites: A challenging issue prior to collecting web-based landmarks is distinguishing whether a web page is locally deployed. The increasing reliance on cloud services and content delivery networks (CDNs) by modern organizations often decouples the geographical location of the server from the content's origin. Moreover, technologies such as proxy servers, load balancers, and other intermediaries further obscure the server's actual IP address. To address this issue, we adopt multi-source verification methods, including Whois queries, DNS resolution, HTTP header analysis, and consultation of third-party IP geolocation databases. However, these methods are not foolproof, as the heterogeneous and sometimes outdated nature of these information sources can still lead to inaccuracy. Future research will attempt to extract more accurate deployment features from network behavior patterns and design more efficient verification mechanisms to mitigate information asymmetry and adapt to dynamic changes, ultimately enhancing the accuracy and robustness of local deployment identification.

Cloud-based Landmark Accuracy Limitations: During our collection of cloud-based landmarks, we observe that cloud service providers may establish multiple data centers within the same city as part of disaster recovery strategies to mitigate the impact of force majeure events. Official public information typically provides only the city-level geographical location of the corresponding IP, making it challenging to pinpoint the exact data center location. To address this, we implement a method combining third-party data center location services with geometric center estimation, which partially reduces errors. However, for disaster recovery designs involving more dispersed regions, this method can lead to increased estimation errors. Specifically, the geometric midpoint approach may introduce a geographical error of approximately 10 kilometers. While this is acceptable for city-level analysis, it may pose limitations for tasks requiring higher geolocation precision. In the future, we aim to incorporate additional intra-city information, such as operator network topology, known data center distribution patterns, or user feedback, to refine geographical location estimations.

IoT-based Landmark Location Patterns: Our method for collecting IoT-based landmarks relies on utilizing cyberspace mapping engines to identify IP addresses and their corresponding geographic locations. However, the geographical location of an IP address does not always reflect the actual physical location of the device. For example, in scenarios involving NAT, multiple devices may share a single IP address, and certain service providers may obscure device locations using relays or proxies. To mitigate these issues, we filter out duplicate IP addresses from the dataset. Nonetheless, this approach has limitations. In future work, we plan to incorporate advanced proxy detection and relay path identification techniques to distinguish between direct device connections and those routed through proxies, thereby improving the reliability and accuracy of the landmark nodes.

Privacy Issues: Since the landmark collection process in this paper involves extensive probing and data collection, data privacy and security are important considerations. Our methodology is strictly based on publicly available and open-source data obtained from reputable online platforms, ensuring that all information used is accessible within the public domain. Importantly, no private or unauthorized personal data is accessed or processed at any stage of the collection. To further mitigate concerns, our probing operations are conducted at minimal frequencies and scales to reduce potential interference with target network resources. This approach ensures that the

data collection process adheres to ethical standards, avoiding excessive data retrieval or risks of violating data privacy.

VI. CONCLUSION

IP geolocation algorithms based on network measurements heavily rely on high-quality landmarks. In this paper, we propose HybridMark, a large-scale, high-quality, and low-cost IP geolocation landmark collection method, coupled with an innovative landmark quality evaluation framework. HybridMark fully automates the collection of high-quality IP geolocation landmarks on a large scale using open-source information. Experimental results validate the effectiveness of HybridMark, demonstrating its ability to efficiently and accurately collect millions of high-quality landmarks. The dataset collected is two orders of magnitude larger than the landmarks used by existing geolocation services. Consequently, HybridMark supports highly accurate and broadly covered geolocation services.

REFERENCES

[1] O. Dan, V. Parikh, and B. D. Davison, "Improving ip geolocation using query logs," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 2016, pp. 347–356.

[2] W. Xu, Y. Tao, and X. Guan, "Experimental comparison of free ip geolocation services, vol. 895," 2020.

[3] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos, "A look at router geolocation in public and commercial databases," in *Proceedings of the 2017 Internet Measurement Conference*, 2017, pp. 463–469.

[4] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 2004, pp. 288–293.

[5] B. Wong, I. Stoyanov, and E. G. Sirer, "Octant: A comprehensive framework for the geolocalization of internet hosts." in *NSDI*, vol. 7, 2007, pp. 23–23.

[6] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards ip geolocation using delay and topology measurements," in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, 2006, pp. 71–84.

[7] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards {Street-Level}{Client-Independent}{IP} geolocation," in *8th USENIX Symposium on Networked Systems Design and Implementation (NSDI 11)*, 2011.

[8] Q. Li, Z. Wang, D. Tan, J. Song, H. Wang, L. Sun, and J. Liu, "Geocam: An ip-based geolocation service through fine-grained and stable webcam landmarks," *IEEE/ACM Transactions on Networking*, vol. 29, no. 4, pp. 1798–1812, 2021.

[9] Z. Wang, Q. Li, J. Song, H. Wang, and L. Sun, "Towards ip-based geolocation via fine-grained and stable webcam landmarks," in *Proceedings of The Web Conference 2020*, 2020, pp. 1422–1432.

[10] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang, "Mining the web and the internet for accurate ip address geolocations," in *IEEE INFOCOM 2009*. IEEE, 2009, pp. 2841–2845.

[11] D. Li, J. Chen, C. Guo, Y. Liu, J. Zhang, Z. Zhang, and Y. Zhang, "Ip-geolocation mapping for moderately connected internet regions," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 2, pp. 381–391, 2012.

[12] G. Ciavarrini, V. Luconi, and A. Vecchio, "Smartphone-based geolocation of internet hosts," *Computer Networks*, vol. 116, pp. 22–32, 2017.

[13] H. Jiang, Y. Liu, and J. N. Matthews, "Ip geolocation estimation using neural networks with stable landmarks," in *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2016, pp. 170–175.

[14] M. Eskandari, A. S. De Oliveira, and B. Crispo, "Vloc: An approach to verify the physical location of a virtual machine in cloud," in *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*. IEEE, 2014, pp. 86–94.

[15] Y. Wang, H. Zhu, J. Wang, J. Liu, Y. Wang, and L. Sun, "Xlboost-geo: An ip geolocation system based on extreme landmark boosting," *arXiv preprint arXiv:2010.13396*, 2020.

[16] D. Jia, L. Liu, S. Jia, and J. Lin, "Votegeo: An iot-based voting approach to verify the geographic location of cloud hosts," in *2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC)*. IEEE, 2019, pp. 1–9.

[17] B. Huffaker, M. Fomenkov, and K. Claffy, "Drop: Dns-based router positioning," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 3, pp. 5–13, 2014.

[18] O. Dan, V. Parikh, and B. D. Davison, "Ip geolocation through reverse dns," *ACM Transactions on Internet Technology (TOIT)*, vol. 22, no. 1, pp. 1–29, 2021.

[19] D. Moore, R. Periakaruppan, J. Donohoe, and K. Claffy, "Where in the world is netgeo. caida. org," in *Proc. of the INET*, vol. 2000, 2000.

[20] P. T. Endo and D. F. H. Sadok, "Whois based geolocation: A strategy to geolocate internet hosts," in *2010 24th IEEE International Conference on Advanced Information Networking and Applications*. IEEE, 2010, pp. 408–413.

[21] I. MaxMind. (2024) Maxmind geoip2 and ip intelligence solutions. https://www.maxmind.com/en/home. Accessed: 2024-12-09.

[22] IP2Location. (2024) Ip2location: Ip address to location database. https://www.ip2location.com/. Accessed: 2024-12-09.

[23] IPinfo.io, "Ipinfo," 2024, accessed: 2024-12-10. [Online]. Available: https://ipinfo.io/

[24] DB-IP, "Db-ip," 2024, accessed: 2024-12-10. [Online]. Available: https://db-ip.com/

[25] IPIP, "Ipip.net," 2024, accessed: 2024-12-10. [Online]. Available: https://www.ipip.net/?origin=EN

[26] IP138, "Ip138 - ip address query and geolocation service," 2024, accessed: 2024-12-10. [Online]. Available: https://www.ip138.com/

[27] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for internet hosts," in *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, 2001, pp. 173–185.

[28] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "Ip geolocation databases: Unreliable?" *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 53–56, 2011.

[29] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.

[30] S. Laki, P. Mátray, P. Hága, I. Csabai, and G. Vattay, "A model based approach for improving router geolocation," *Computer Networks*, vol. 54, no. 9, pp. 1490–1501, 2010.

[31] S. Laki, P. Mátray, P. Hága, T. Sebők, I. Csabai, and G. Vattay, "Spotter: A model based active geolocation service," in *2011 Proceedings IEEE INFOCOM*. IEEE, 2011, pp. 3173–3181.

[32] Q. Zhao, F. Wang, C. Huang, and C. Yu, "Improving ip geolocation databases based on multi-method classification," in *2020 IEEE 14th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*. IEEE, 2020, pp. 44–48.

[33] Y. Luo, X. Jia, S. Fu, and M. Xu, "pride: Privacy-preserving ride matching over road networks for online ride-hailing service," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1791–1802, 2018.

[34] P. Consortium, "Planetlab: An open platform for developing, deploying, and accessing planetary-scale services," https://planetlab.cs.princeton.edu/, accessed: 2024-12-10.

[35] R. NCC, "Ripe atlas: A global internet measurement platform," https://atlas.ripe.net/, accessed: 2024-12-10.

[36] M. Lab, "Measurement lab: Open tools for internet performance measurement," https://www.measurementlab.net/, accessed: 2024-12-10.

[37] N. Ring, "Nlnog ring: Network measurement participants," https://ring.nlnog.net/participants/, accessed: 2024-12-10.

[38] R. Li, Y. Sun, J. Hu, T. Ma, and X. Luo, "Street-level landmark evaluation based on nearest routers," *Security and Communication Networks*, vol. 2018, no. 1, p. 2507293, 2018.

[39] T. Ma, F. Liu, F. Zhang, and X. Luo, "An landmark evaluation algorithm based on router identification and delay measurement," in *Artificial Intelligence and Security: 5th International Conference, ICAIS 2019, New York, NY, USA, July 26–28, 2019, Proceedings, Part III 5*. Springer, 2019, pp. 163–177.

[40] W. Yang, X. Liu, and M. Yin, "Street-level landmark evaluation with upper error bound," *IEEE Access*, vol. 7, pp. 112 037–112 043, 2019.

[41] M. Yin, W. Yang, X. Liu, and X. Luo, "Evaluator: A multilevel decision approach for web-based landmark evaluation," *Security and Communication Networks*, vol. 2020, no. 1, p. 8843188, 2020.

[42] Internet Assigned Numbers Authority. (2024) Whois - internet assigned numbers authority (iana). Accessed: 2024-12-05. [Online]. Available: https://www.iana.org/whois

[43] QS Quacquarelli Symonds. (2024) QS World University Rankings 2025. Accessed: 2024-12-05. [Online]. Available: https://www.topuniversities.com/world-university-rankings
[44] Amazon Web Services. (2024) Amazon web services (aws) - cloud computing services. Accessed: 2024-12-05. [Online]. Available: https://aws.amazon.com
[45] Oracle Corporation. (2024) Oracle cloud infrastructure. Accessed: 2024-12-05. [Online]. Available: https://www.oracle.com/cloud/
[46] Microsoft Corporation. (2024) Microsoft azure: Cloud computing services. Accessed: 2024-12-05. [Online]. Available: https://azure.microsoft.com/en-us/
[47] Data Center Map. (2024) Data center map: Data center locations worldwide. Accessed: 2024-12-05. [Online]. Available: https://www.datacentermap.com/
[48] Baxtel. (2024) Baxtel: Data center map and directory. Accessed: 2024-12-05. [Online]. Available: https://baxtel.com/
[49] Digital Realty. (2024) Digital realty: Data center solutions. Accessed: 2024-12-05. [Online]. Available: https://www.digitalrealty.com/
[50] DataCenters.com. (2024) Datacenters.com: Find data centers worldwide. Accessed: 2024-12-05. [Online]. Available: https://www.datacenters.com/
[51] E. Union, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation)," Official Journal of the European Union, 2016, accessed: 2024-12-05. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj
[52] FOFA, "Fofa: Cyberspace mapping engine," https://fofa.info/, accessed: 2024-12-10.
[53] D. Komosny, M. Voznak, and S. U. Rehman, "Location accuracy of commercial ip address geolocation databases," *Information technology and control*, vol. 46, no. 3, pp. 333–344, 2017.
[54] perfSONAR Development Team, "perfsonar: A network measurement toolkit," https://www.perfsonar.net/, 2024, accessed: 2024-12-11.
[55] P. P. Team, "Pinger: Internet end-to-end performance monitoring," http://www-iepm.slac.stanford.edu/pinger/, 2024, accessed: 2024-12-11.