Crowd Detection Using Very-Fine-Resolution Satellite Imagery

Tong Xiao [a, b], Qunming Wang [b, *], Ping Lu [b], Tenghai Huang [b], Xiaohua Tong [b], Peter M. Atkinson [c, d]

[a] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing 100081, China

[b] College of Surveying and Geo-Informatics, Tongji University, 1239 Siping Road, Shanghai 200092, China

[c] Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YR, UK

[d] Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK

*Corresponding author. Email: wqm11111@126.com

Abstract: Accurate crowd detection (CD) is critical for public safety and historical pattern analysis, yet existing methods relying on ground and aerial imagery suffer from limited spatio-temporal coverage. The development of very-fine-resolution (VFR) satellite sensor imagery (e.g., ~0.3 m spatial resolution) provides unprecedented opportunities for large-scale crowd activity analysis, but it has never been considered for this task. To address this gap, we proposed CrowdSat-Net, a novel point-based convolutional neural network, which features two innovative components: Dual-Context Progressive Attention Network (DCPAN) to improve feature representation of individuals by aggregating scene context and local individual characteristics, and High-Frequency Guided Deformable Upsampler (HFGDU) that recovers high-frequency information during upsampling through frequency-domain guided deformable convolutions. To validate the effectiveness of CrowdSat-Net, we developed CrowdSat, the first VFR satellite imagery dataset designed specifically for CD tasks, comprising over 120k manually labeled individuals from multi-source satellite platforms (Beijing-3N, Jilin-1 Gaofen-04A and Google Earth) across China. In the experiments, CrowdSat-Net was compared with eight state-of-the-art point-based CD methods (originally designed for ground or aerial imagery and satellite-based animal detection) using CrowdSat and achieved the largest F1-score of 66.12% and Precision of 73.23%, surpassing the second-best method by 1.71% and 2.42%, respectively. Moreover,

extensive ablation experiments validated the importance of the DCPAN and HFGDU modules. Furthermore, cross-regional evaluation further demonstrated the spatial generalizability of CrowdSat-Net. This research advances CD capability by providing both a newly developed network architecture for CD and a pioneering benchmark dataset to facilitate future CD development. The source code is available at https://github.com/Tong-777777/CrowdSat-Net.

Keywords: Crowd detection; Very-fine-resolution (VFR) satellite imagery; Deep learning; Feature enhancement; Feature fusion.

1. Introduction

In recent years, population growth, continued urbanization and rapid economic development in many regions have led to an increase in the frequency of crowd activities in public areas. Such crowds can pose a series of public safety risks, including traffic congestion (Gazzawe and Albahar, 2024), crowd crushes (Al-Nami, 2023), security incidents (Feliciani et al., 2022) and public health risks (Pokhrel and Chhetri, 2021; Joiner et al., 2024). Alleviating these risks, for example through crowd control, has become an essential aspect of ensuring public safety in densely populated gatherings. Crowd detection (CD), which involves estimating the location and count of the individuals in a crowd in a specific place (Sam et al., 2020; Wan et al., 2021; Han et al., 2023), plays a crucial role in mitigating these risks. By managing and controlling crowd movements (Weng et al., 2023; Rezaee et al., 2024) effectively, CD can help reduce the likelihood of accidents and create a safer environment. In addition to crowd management, long-term CD can analyze future crowd distribution patterns by mining historical data trends, providing valuable insights for urban planning and infrastructure optimization (Luo et al., 2025). The definition of a crowd varies across research domains. In this research, we

adopt a broad definition: A gathering of individuals, whether isolated, dispersed or clustered, or any combination of these, is regarded as a crowd for detection.

To achieve CD effectively, early research relied primarily on analyzing imagery captured by ground surveillance or fixed cameras (Lempitsky and Zisserman, 2010; Zhang et al., 2021; Mo et al., 2024). These cameras provide diverse perspectives, varying scales and different illumination conditions with which to detect crowds. However, their coverage was inherently limited to specific areas within the camera's field of view. With the development of airborne and unmanned aerial vehicle platforms, aerial datasets, such as the DLR Aerial Crowd Dataset (Bahmanyar et al., 2019), Meynberg's dataset (Meynberg et al., 2016) and Mliki's dataset (Mliki et al., 2019), have been introduced for CD, enhancing the development of models to monitor and analyze crowd dynamics over large areas. Although aerial imagery provides many advantages for monitoring, especially its potentially fine spatial resolution, its effectiveness for long-term analysis is constrained by infrequent data updates. For example, in Hebei Province, China, the standard update cycle for aerial photography typically exceeds three years (Hebei Provincial People's Government, 2011), which limits its ability to capture continuous short-term variations in crowd distributions.

With advances in remote sensing satellite technology, fine spatial-temporal resolution imagery has increasingly been applied to detect small-scale (i.e., small area) objects, such as ships, airplanes and tiny buildings (Gao et al., 2020a,b; Yi et al., 2023b). Compared to aerial imagery, satellite sensor imagery (hereafter, satellite imagery) offers a broader spatial coverage and shorter revisit intervals (e.g., the Worldview-4 satellite has a revisit period of approximately 4.5 days (Rumora et al., 2020)), increasing the potential for consistent, large-scale monitoring and historical analysis. Nowadays, very-fine-resolution (VFR) satellite imagery from sources like the Beijing-3N (BJ3N) and Jilin-1 Gaofen-04A (JL4A) satellites, as well as the Google Earth platform, has enabled the discrimination of individuals with the naked eye (as shown in Fig. 1). This advance potentially allows more detailed and extensive monitoring, providing a valuable tool for applications requiring accurate object detection and tracking over wide areas. In light of this, as a primary contribution of this research, we presented a novel CD dataset, namely, CrowdSat, derived from the BJ3N and

JL4A satellites, as well as the Google Earth platform, covering 32 provincial-level divisions in China (except Guizhou Province and Macao). CrowdSat contains over 120k manually labeled individuals and consists of diverse regions with strong heterogeneity (e.g., built-up areas, snowy regions, beaches, desert regions, etc.). Compared to ground imagery, CrowdSat provides a much broader coverage, making it a more appropriate source for large-scale monitoring. Moreover, with its high revisit frequency, CrowdSat enables more continuous analysis of human activity patterns compared to aerial imagery, which is typically updated less frequently, especially at a large scale. To the best of our knowledge, CrowdSat is the first-ever CD dataset based on VFR satellite imagery, and it is intended to facilitate research on large-scale crowd analysis while uncovering historical patterns of human movement.

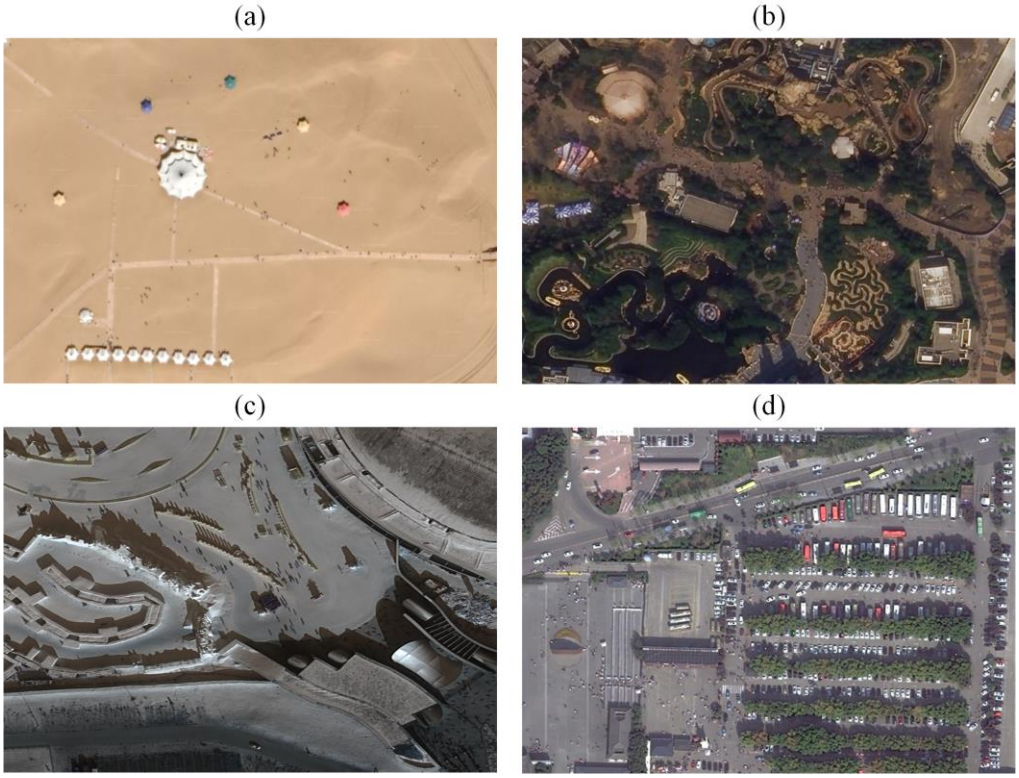

Fig. 1. Examples of VFR satellite imagery for CD. (a) Yuesha Island, Inner Mongolia, China, acquired by the Google Earth platform on Feb. 22, 2023 (© Google Earth 2025) with a spatial resolution of 0.30 m. (b) Shanghai Disney, China, acquired by the JL4A satellite on Feb. 16, 2023 with a spatial resolution of 0.31 m. (c) Harbin Ice and Snow World on Feb. 2, 2024 and (d) the Xi'an Emperor Qinshihuang's Mausoleum Site Museum in China on Oct. 29, 2023, both acquired by the BJ3N satellite with a spatial resolution of 0.30 m.

Compared to ground and aerial imagery, VFR satellite imagery is more stable and provided with greater consistency through time and can, thus, capture several sudden or unexpected crowd gathering cases. VFR satellite imagery, thus, has a greater range of potential applications. For example, VFR imagery can support post-disaster assessment and emergency management, long-term monitoring of urban mobility and public space usage, large-event planning and evaluation, and humanitarian and public-health interventions in refugee or quarantine areas. It can also assist with the sustainable management of tourism and cultural heritage sites by revealing how crowd densities evolve through time at popular destinations. These diverse applications highlight the need for advanced CD techniques based on VFR satellite imagery, particularly for tasks requiring precise individual-level information.

In the realm of CD, previous methods focused primarily on estimating crowd counts (Shu et al., 2022; Khan et al., 2023; Yi et al., 2023a). Counting, however, provides only coarse information and is insufficient for high-level tasks, such as individual tracking, activity recognition, anomaly detection and crowd flow or behavior prediction (Laradji et al., 2018; Song et al., 2021). Recent research, therefore, pays more attention to fine-grained localization of individuals and formulates CD as segmentation- (Meynberg et al., 2016), density map- (von Borstel et al., 2016), detection- (Ren et al., 2015) or point-based prediction (Zhou et al., 2019, 2020). In this research, we adopt a point-based formulation, which represents each person by a single point while retaining accurate localization and requiring relatively simple manual annotation.

Notably, the aforementioned approaches are employed primarily in images captured by ground surveillance or fixed cameras, or aerial vehicles. As illustrated in Fig. 2, individuals in such imagery exhibit typically clear and distinguishable features. When these clear signals become tiny and ambiguous in VFR satellite imagery, it is unclear whether the aforementioned methods can still work effectively. In the remote sensing field, related progress has already been made on dense small-object detection or counting using satellite imagery. Yuan et al. (2025) conducted comparison between YOLOv11, Single Shot MultiBox Detector (SSD), Faster Region-based Convolutional Neural Network (R-CNN) and Cascade R-CNN, for car detection using xView

images and bee-box detection using SkySat images, which shows that multi-scale feature extraction and anchor-free design are important for small objects. Guo et al. (2022) proposed a density map-based vehicle counting method for remote sensing imagery with limited spatial resolution, demonstrating that regressing vehicle density maps can yield accurate counts even when individual vehicles occupy only a few pixels. Bashir and Wang (2021) combined residual feature aggregation-based super-resolution with an object detector to enhance small-object detection in satellite and aerial images. Delplanque et al. (2023) proposed a single-stage CNN, which localizes each animal as a point. Wu et al. (2023) developed an ensemble of U-Net-based models to detect wildebeests from VFR satellite imagery over heterogeneous savanna landscapes, converting annotations to segmentation masks and using context-aware post-processing to obtain individual counts. These studies show that fine spatial resolution feature maps, multi-scale feature fusion and contextual priors can recover partially tiny and ambiguous signals of small objects in satellite imagery. Dense crowd detection using VFR satellite imagery, however, presents additional challenges. First, the signal of each individual in VFR satellite imagery is weaker than that of the small objects considered in previous remote sensing studies. As shown in Fig. 1, the signal size of each individual is approximately 3×3 pixels. During the convolution process, particularly in the pooling stages, this small-sized signal can lead to attenuation or even loss of the signal (Liu et al., 2021; Tong and Wu, 2022; Wei et al., 2024). Second, multi-scale feature fusion, which integrates the upsampled coarse and fine spatial resolution features, is potentially a useful method to alleviate this issue. However, traditional upsampling methods, such as nearest neighbor and bilinear interpolation, often fail to recover fine spatial details, leading to over-smooth boundaries (Li et al., 2024) and misalignment between the high-frequency details in the upsampled features and fine spatial resolution features (Dai et al., 2017), making it challenging to detect small-sized objects effectively.

Fig. 2. Examples of ground and aerial imagery. (a) Image from the ShanghaiTech Dataset[1], and (b) image from the DLR Aerial Crowd Dataset[2].

To overcome the aforementioned limitations, this paper proposed a novel point-based convolutional neural network (CNN) method, CrowdSat-Net, which was specifically designed for large-scale and long-term CD. CrowdSat-Net introduces two key contributions: 1) a Dual-Context Progressive Attention Network (DCPAN) to improve the individual instance feature presentation and 2) a High-Frequency Guided Deformable Upsampler (HFGDU) to replace traditional upsampling methods, which aims to recover the fine spatial information of individuals during the upsampling process.

In summary, the contributions of this research are three-fold:

1) To the best of our knowledge, this is the first research to utilize VFR satellite imagery for CD, which aims to facilitate studies on the characterization of human spatial distributions and temporal activities at

[1] https://github.com/desenzhou/ShanghaiTechDataset

[2] https://www.dlr.de/en/eoc/about-us/remote-sensing-technology-institute/photogrammetry-and-image-analysis/public-datasets/dlr-acd

a large-scale (both spatially and temporally).

2) To achieve this task, a novel CD dataset, CrowdSat, was collected by multi-source satellite platforms, which comprises over 120k labeled individuals and consists of diverse regions with strong heterogeneity, facilitating the development of CD methods. Additionally, during labeling, a point-like background removal strategy was introduced that uses multi-temporal VFR satellite imagery as auxiliary data to reduce mislabeling rates.

3) A novel point-based CD method using the CNN, termed CrowdSat-Net, was proposed to detect individuals in satellite imagery efficiently. Additionally, two innovational modules, DCPAN and HFGDU, were introduced to enhance individual feature presentation and recover the lost high-frequency information of individual features during upsampling.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive overview of CrowdSat, detailing its data collection and preprocessing, labeling process and data analysis. Section 3 presents the architecture of CrowdSat-Net, explaining its key components and design. Section 4 presents extensive experimental evaluations, while Section 5 discusses the broader applicability, inherent limitations of CrowSat and CrowSat-Net, and future directions. Finally, Section 6 summarizes the key findings of this research.

2. CrowdSat Dataset

In this paper, CrowdSat was presented for large-scale CD and the analysis of historical human movement patterns. To better understand the details of CrowdSat, this section provides a comprehensive overview, covering three key aspects: data collection and preprocessing, data labeling and data analysis.

2.1. Data Collection and Preprocessing

To ensure representative data for large-scale CD, three complementary remote sensing data sources, including the Google Earth platform and the BJ3N and JL04A satellites, were considered. The Google Earth imagery used in CrowdSat has an approximate spatial resolution of 0.30 m and originates primarily from Maxar (DigitalGlobe) WorldView-series satellites (e.g., WorldView-3/4). These data are pan-sharpened products produced by fusion of 0.30 m panchromatic and 1.20 m multispectral bands. It is noted that all Google Earth imagery used in this research was accessed solely for academic, non-commercial research purposes and in compliance with the Google Maps/Google Earth Additional Terms of Service. The released CrowdSat dataset does not include any Google Earth images. However, Google Earth imagery for some regions suffers from cloud or haze contamination, mosaic seams, compression artifacts or relatively sparse visible crowds, which limits its usefulness for dense crowd annotation. To complement these cases, we additionally used imagery from the BJ3N and JL04A satellites. The BJ3N satellite operates in a Sun-synchronous orbit at an altitude of 610 km, with a 5-day revisit period and an LTAN of 11:00. It captures RGB and panchromatic bands with a spatial resolution of 1.20 m and 0.30 m, respectively. The JL04A satellite follows a Sun-synchronous orbit at 535 km, with a shorter 3-day revisit period and an LTAN at 10:30, allowing for more frequent observations. The JL04A imagery used in CrowdSat has a spatial resolution of 0.31 m for the panchromatic band and 1.24 m for the multispectral bands.

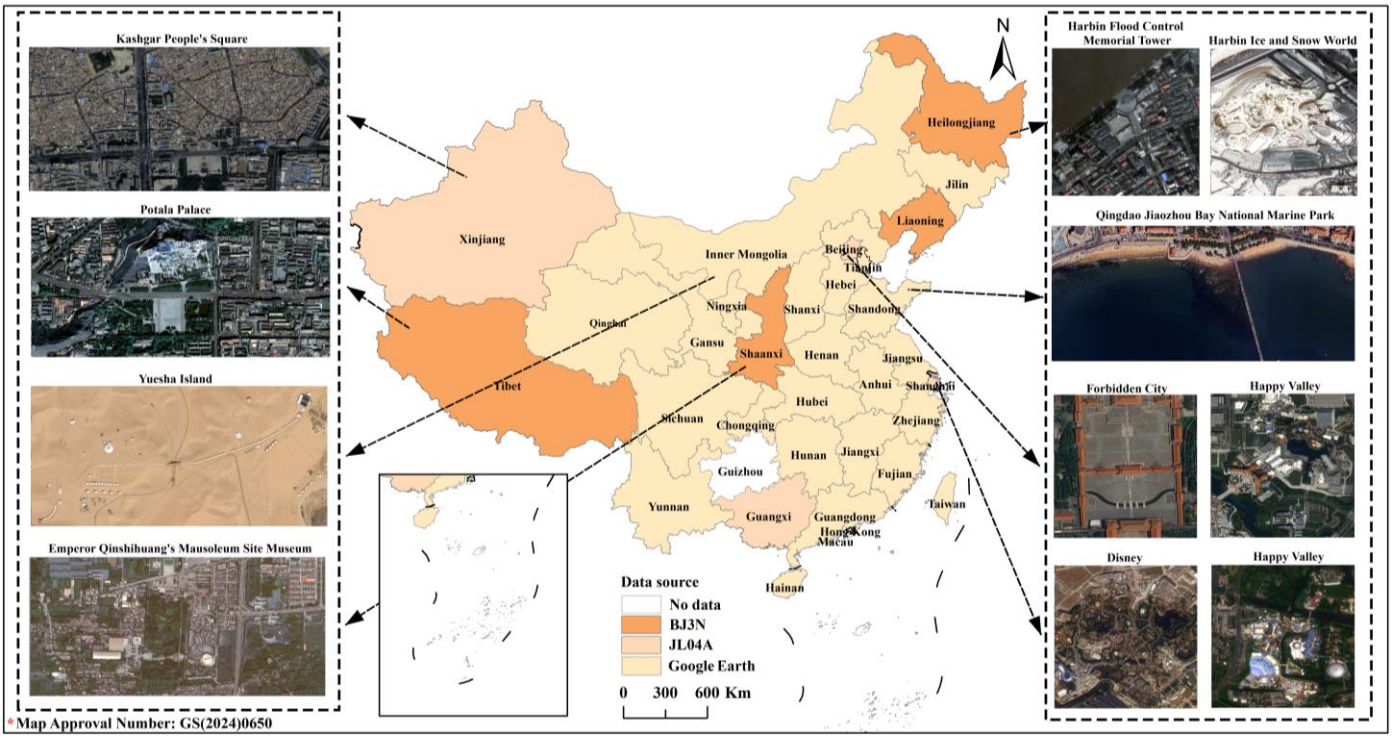

Fig. 3. Spatial distribution map of collected VFR satellite imagery for some locations where crowds typically gather. Part of the imagery displayed in this figure was obtained from the Google Earth platform (© Google Earth 2025).

To ensure broad spatial coverage and diverse conditions, we selected imagery spanning 32 provincial-level divisions in China (except Guizhou Province and Macao, due to fewer satellite images in these areas). Our dataset collection follows two main principles. First, the collected imagery must be visually reliable for annotation, i.e., cloud- or haze-free, with minimal shadow contamination and high radiometric quality. Second, the selected regions need to contain a sufficient number of clearly visible individuals. Within each provincial-level division, we first identified typical public gathering places (e.g., city squares, commercial streets, transport hubs and scenic sites), then reviewed imagery from BJ3N, JL04A satellites and the Google Earth platform manually, and selected the scenes that best satisfy these two principles. These regions exhibit substantial heterogeneity, encompassing various landscapes and urban settings, as illustrated in Fig. 3. These scenes include open public areas (e.g., the Forbidden City), built-up areas (e.g., Kashgar People's Square), snowy regions (e.g., Harbin Ice and Snow World), areas with lush vegetation (e.g., the Emperor Qinshihuang's Mausoleum Site Museum), beaches (e.g., Qingdao Jiaozhou Bay National Marine Park), desert regions (e.g.,

Yuesha Island), etc.. All the imagery was acquired between Feb. 20, 2023, and Jan. 2, 2025. In addition, before using the collected imagery, preprocessing steps such as geometric correction, clipping and radiometric normalization were performed to enhance the consistency between images from different sources.

Using the original 1.20 m BJ3N or 1.24 m JL4A multi-spectral imagery directly makes it challenging to distinguish individuals. To address this limitation, a state-of-the-art pan-sharpening technique known as area-to-point regression kriging (ATPRK) (Wang et al., 2015, 2016), which attained smaller relative global-dimensional synthesis error (ERGAS) and larger spectral angle mapper (SAM) values than classical fusion methods, was employed to increase the spatial resolution of the BJ3N and JL4A imagery to 0.30 m and 0.31 m, respectively. The fused imagery with finer spatial resolution enables a more accurate interpretation for subsequent analysis.

2.2. Data Labeling

Labeling individuals in VFR satellite imagery presents unique challenges compared to ground and aerial imagery. Due to the relatively blurred edges of individuals in VFR satellite imagery, it is difficult to delineate their contours. This increases the likelihood of mistaking other ground objects as individuals. For example, as shown in Fig. 4(a), objects within the red and cyan circles exhibit visual characteristics similar to individuals. However, the objects in the red circles correspond to road asphalt, while those in the cyan circles are street lamps. Similarly, other stationary objects, such as stone pillars, may also be labeled incorrectly, which can further mislead network training.

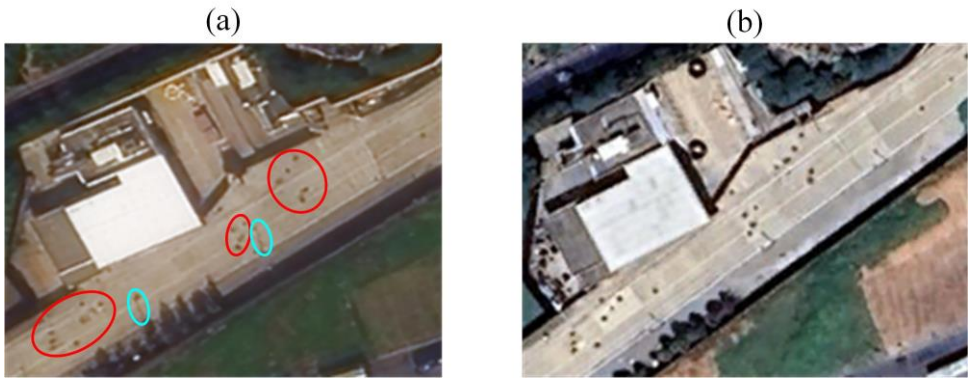

Fig. 4. Examples of mislabeling. (a) and (b) were captured in the same region on Feb. 16, 2023 from the BJ3N satellite and Aug. 7, 2024 from the Google Earth platform (© Google Earth 2025), respectively. The objects in the red and cyan circles correspond to road asphalt and street lamps, respectively.

To alleviate the above issue, multi-temporal VFR imagery from the Google Earth platform was used as auxiliary data, due to the extensive spatial and temporal coverage of Google Earth imagery (Li et al., 2023; Shi et al., 2023; Ni et al., 2024). For example, Fig. 4(b) was captured in the same region at a different time, making it easier to distinguish individuals from other small objects. By cross-referencing imagery captured at different time points, fixed objects were identified as non-individuals, reducing labeling errors. Moreover, to further ensure accurate annotations, a center-point labeling strategy was employed. Specifically, for each individual, a single point was placed at the center of the 3×3 pixel region to present the approximate position in the imagery. This approach prevents boundary ambiguity and minimizes the risk of mislabeling non-human objects.

To facilitate the use of the data in a deep learning architecture, following the collection of classical remote sensing datasets, such as LEVIR-CD (Chen and Shi, 2020) and WHU-CD (Ji et al., 2018), the collected imagery was cropped into 256×256 pixel patches without overlapping. Patches without individuals were removed, resulting in a total of 3,447 labeled patches.

2.3. Data Analysis

Compared to ground and aerial imagery, CrowdSat sets new benchmarks with its vast scale, diverse density distribution and comprehensive multi-environment coverage. It presents four key characteristics.

1)Multi-Density Representation. As illustrated in Fig. 5, the CrowdSat dataset consists of a total of 3,447 patches, covering 120,141 individuals. The number of individuals in each patch varies significantly, with a maximum of 808 individuals and a minimum of 1 individual. On average, each patch contains 34.9 individuals, while the median number per patch is 14.0. This wide range of crowd densities across different patches offers a comprehensive representation of various crowd scenarios, making the dataset valuable for evaluating crowd detection methods under diverse conditions.

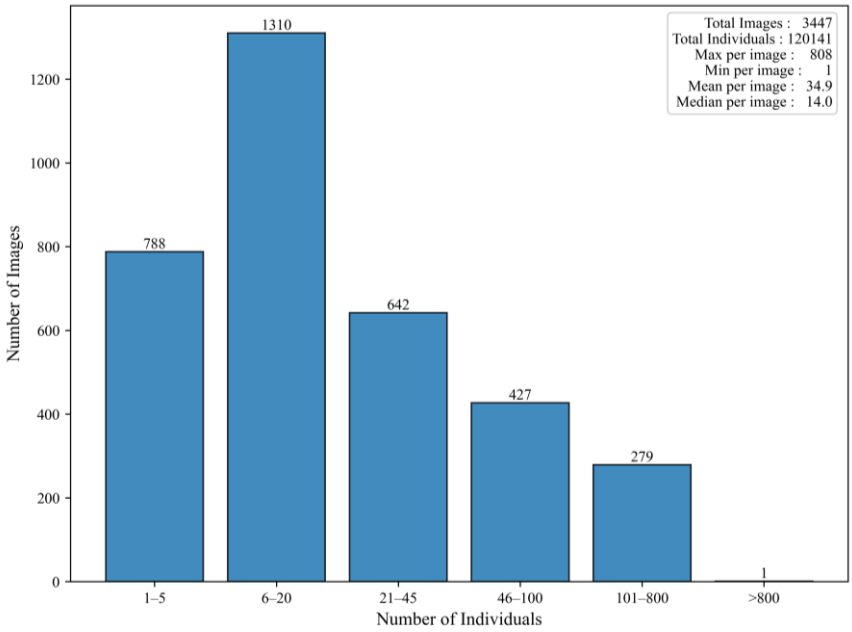

Fig. 5. Crowd count distribution in CrowdSat.

2) *Extensive National Coverage*. CrowdSat spans China (except Guizhou Province and Macao) and includes samples from 32 Chinese provincial-level divisions. As illustrated in Fig. 6, the number of patches per province ranges from 0 to 393, and the corresponding number of individuals ranges from 0 to 10,716. Provinces with frequent large public gatherings, such as Shandong (393 patches; 10,716 individuals) and Shanghai (330 patches; 6,374 individuals), contribute more samples, whereas sparsely populated or remote

regions, such as Tibet and Qinghai, contain fewer but still representative scenes. Guizhou and Macao currently have no labelled patches that satisfy our selection principles (cloud- and haze-free, minimal shadow contamination and a sufficient number of visible individuals). The vast coverage of CrowdSat provides a wide range of crowd types, promoting the generalizability of the dataset to various scenarios.

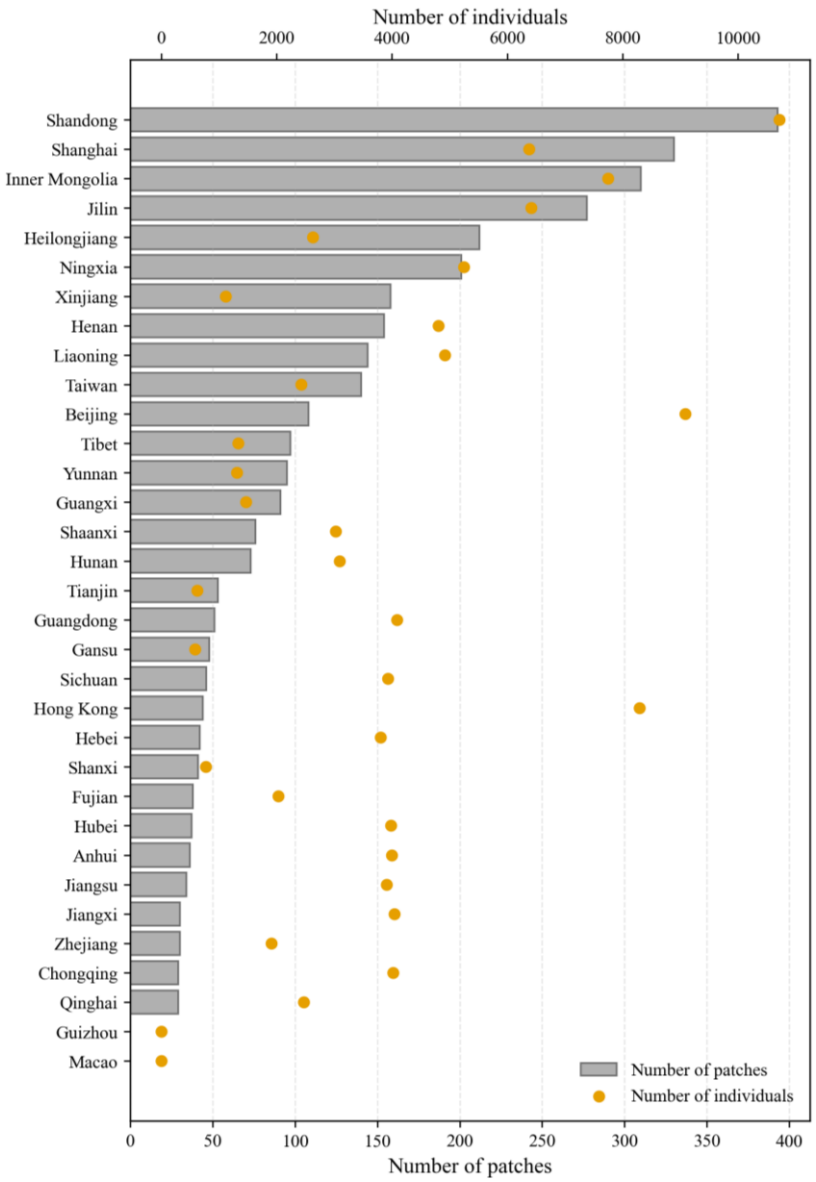

Fig. 6. Distribution of patches and individuals per province in CrowdSat.

3) *Multi-Season Adaptation*. CrowdSat was collected between Feb. 20, 2023 and Jan. 2, 2025, covering all four seasons. This wide temporal span allows models to adapt inherently to seasonal variations in terms of lighting conditions (e.g., summer glare and winter haze) and crowd movement patterns.

4) *Diverse Environmental Representations*. CrowdSat contains samples from distinct environmental categories across China, such as open areas, snowy regions, beaches, desert regions, etc.. As shown in Fig. 7, CrowdSat is dominated by open public areas, where mass gatherings occur commonly. Built-up areas and vegetated areas together account for most of the remaining samples, consisting of dense commercial streets, residential blocks and urban parks where crowd activities take place everyday. Beaches form a medium-sized component of the dataset, whereas snowy regions and desert regions are less frequent but still represented by dozens of patches and several thousand individuals. Traffic junctions contain a noticeable number of patches but fewer individuals per patch. By systematically encompassing scenarios from historical landmarks to modern transportation hubs, this cross-environment integration can alleviate location-specific bias.

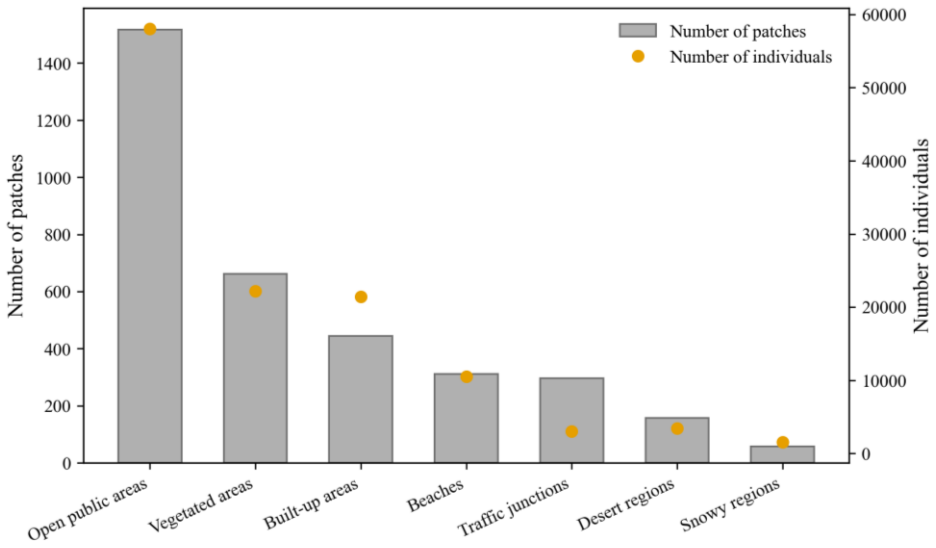

Fig. 7. Distribution of patches and individuals across different scene types in CrowdSat.

3. Methods

3.1. Overview of CrowdSat-Net

To detect crowds in large-scale VFR satellite imagery effectively, a novel point-based CNN, CrowdSat-Net, was proposed. It was specifically designed to address two key challenges: 1) gradual blurring and loss of small object signals during feature extraction and 2) high-frequency information loss during upsampling operations.

An overview of the CrowdSat-Net architecture is illustrated in Fig. 8. At the beginning, the label image is transformed into a Focal Inverse Distance Transform (FIDT) map (Liang et al., 2022a) using the FIDT method. CrowdSat-Net adopts a two-stage stacked Hourglass Network (Newell et al., 2016) (a typical network used for point feature extraction), with key modifications to increase the accuracy of CD. First, to improve small-object feature representation, a novel module, that is, DCPAN, is embedded in the image preprocessing stage. Second, during the upsampling operation, the traditional method (e.g., bilinear interpolation) is replaced with a novel module, HFGDU, which aims to restore lost high-frequency information. Within each Hourglass Network, a location map is generated, and the loss is computed using the FIDT map and the location map through the Focal Loss function (Lin et al., 2017b), a strategy known as Intermediate Supervision. The total loss L is computed as the sum of Focal Losses from each Hourglass Network. Finally, the location map from the last Hourglass Network is processed using a Local-Maxima-Detection-Strategy (LMDS) to obtain the final individual localizations.

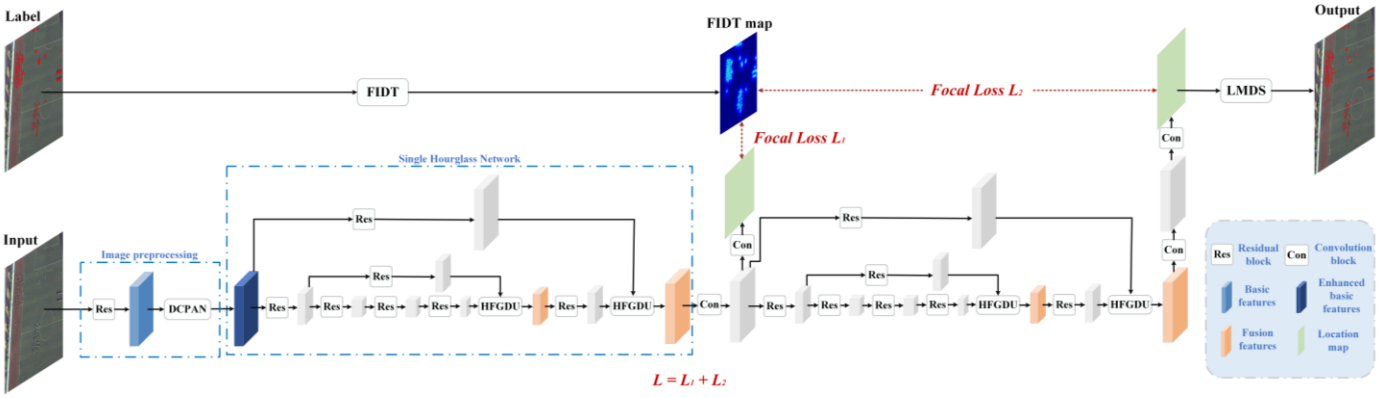

Fig. 8. Overview of the proposed CrowdSat-Net. First, the labeled image is transformed into the FIDT map. During each training iteration, CrowdSat-Net enhances the basic features in the image preprocessing using the DCPAN module. Then, these enhanced features pass through the two-stacked Hourglass Network. In each Hourglass Network, the traditional upsampling method is replaced with the HFGDU module to recover the lost fine details. Each Hourglass Network generates a location map, which is compared with the FIDT map to calculate the Focal Loss. The total loss L is the sum of Focal Losses from each Hourglass Network. After training, the location map conducted by the last Hourglass Network is transformed into the actual localization result using the LMDS method.

The remainder of this section introduces important components in detail, including the FIDT map, DCPAN, HFGDU, LMDS and model evaluation.

3.2. Focal Inverse Distance Transform (FIDT) Map

It is critical to identify individual localization in CrowdSat-Net. Traditional methods, such as binary-like maps (Liu et al., 2019), segmentation-like maps (Xu et al., 2022), topological maps (Abousamra et al., 2021) and independent instance maps (Gao et al., 2020c), struggle to distinguish overlapping objects in dense crowds due to their reliance on fixed thresholds or semantic boundaries. To overcome this problem, the FIDT (Liang et al., 2022a) method was employed, which enables overlap-free head localization by assigning higher pixel responses closer to head centers, thereby ensuring accurate localization in dense crowds. The FIDT is defined as follows:

$$I = \frac{1}{d(x,y)^{(\alpha \times P(x,y)+\beta)+C}} \tag{1}$$

where I represents the FIDT map, α and β are weight factors, set to the default values of 0.02 and 0.75, respectively (Liang et al., 2022a). C indicates a constant, which aims to prevent division by zero. $d(x, y)$

denotes the distance between the pixel at location (x, y) and its nearest annotated pixel at location (x', y'), in the set of all annotation **B**:

$$d(x,y) = \min_{(x',y')\in \mathbf{B}} \sqrt{(x-x')^2 + (y-y')^2} \quad (2)$$

3.3. Dual-Context Progressive Attention Network (DCPAN)

The image preprocessing stage aims to extract fundamental feature representations from the original inputs, serving as the foundation for downstream network architectures. Conventional implementations in the Hourglass Network employ typically a standard convolutional layer followed by a ResNet layer (Zhou et al., 2019), which ignores the gradual signal attenuation of small objects. This limitation can compromise model sensitivity to fine-grained features and increase false negative rates in individual detection.

To address the above issue, we first excluded the standard convolutional layer and then proposed the DCPAN module embedded after the ResNet layer, shown in Fig. 9. The DCPAN module combines synergistically a base spatial attention (SA) Encoding and two parallel branches (multi-scale feature extraction (MSFE) and local contrast enhancement (LCE) branches), aiming to improve the small-object signal presentation. Specifically, given a feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ generated by the ResNet layer, where C, H and W denote the channel, height and width of the feature, respectively, the DCPAN module generates small-object feature enhancement weights through three key operations detailed as follows.

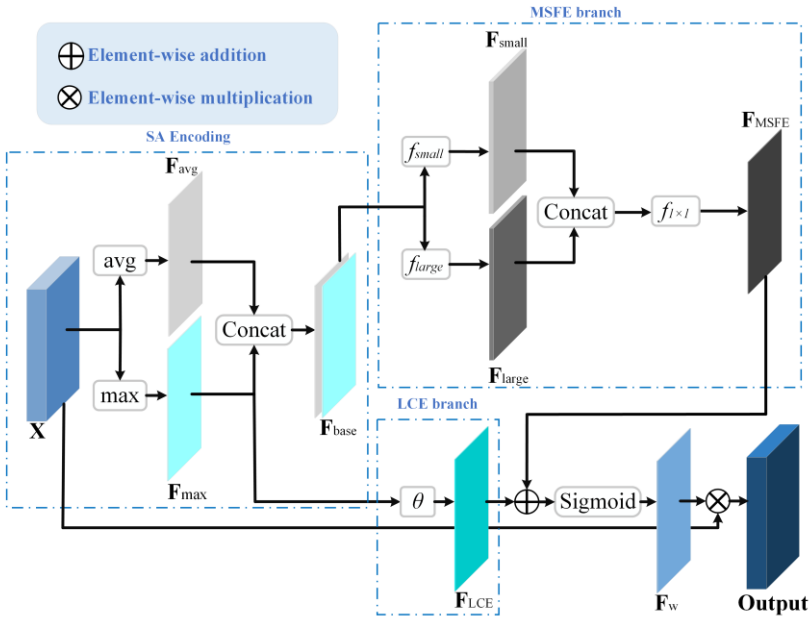

Fig. 9. Flowchart of the DCPAN module, which consists of three main components: SA Encoding, MSFE and LCE branches. The SA Encoding component extracts the base SA feature from the original features **X**. The MSFE branch captures contextual information from this base feature, while the LCE branch focuses on extracting its local contrast. Finally, the outputs of the MSFE and LCE branches are fused through element-wise addition, and the result is passed through a sigmoid activation function to obtain the enhancement weight map $\mathbf{F}_w$. This map is subsequently applied to the original feature **X** via element-wise multiplication.

1) SA Encoding. Small objects lack distinct semantic features, but exhibit unique spatial patterns. We first, thus, extracted the base spatial attention cues $\mathbf{F}_{base}$, achieved as follows:

$$\mathbf{F}_{\max} = \max(\mathbf{X}), \mathbf{F}_{\text{avg}} = \frac{1}{C}\sum_{c=1}^{C}\mathbf{X}_c \tag{3}$$

where $\mathbf{F}_{\max} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{F}_{\text{avg}} \in \mathbb{R}^{1\times H\times W}$ represent the max-pooled and average-pooled features, respectively. These two features are then concatenated to form the base spatial feature $\mathbf{F}_{\text{base}} \in \mathbb{R}^{2\times H\times W}$.

2) MSFE Branch. Although the base spatial feature can improve the presentation of prominent objects by highlighting global significant regions, the global pooling operations (e.g., max/avg) in SA Encoding fail to model positional dependencies between objects, leading to suboptimal results in cluttered scenes (Fu et al.,

2019). To alleviate this issue, a simple MSFE branch, which employs parallel dilated convolutional layers with complementary dilation rates, was proposed:

$$\mathbf{F}_{\text{small}} = f_{\text{small}}(\mathbf{F}_{\text{base}}), \mathbf{F}_{\text{large}} = f_{\text{large}}(\mathbf{F}_{\text{base}}) \tag{4}$$

where f_{small} and f_{large} are the dilated convolutional operations with dilation rates of 2 and 4, respectively. Then, the $\mathbf{F}_{\text{MSFE}} \in \mathbb{R}^{1\times H\times W}$ of these layers are concatenated and fused using a 1×1 convolutional layer:

$$\mathbf{F}_{\text{MSFE}} = f_{1\times1}\big(\text{Concat}(\mathbf{F}_{\text{small}}, \mathbf{F}_{\text{large}})\big) \tag{5}$$

3) LCE Branch. Standard pooling operations always blur high-frequency details, leading to degraded accuracy on fine-grained detection tasks (Fu et al., 2019). To address this problem, the LCE branch highlights locations with large heterogeneity by a local contrast generator θ:

$$\mathbf{F}_{\text{LCE}} = \theta(\mathbf{F}_{\text{max}}) = \partial(|\mathbf{F}_{\text{max}} - \text{AvgPool}_{3\times3}(\mathbf{F}_{\text{max}})|) \tag{6}$$

where $\mathbf{F}_{\text{LCE}} \in \mathbb{R}^{1\times H\times W}$ represents the contrast-sensitive weight map and $|\cdot|$ denotes absolute value computation. ∂ is the learnable contrast enhancement operator composed of a 3×3 convolutional layer, batch normalization, ReLU activation and a 3×3 convolutional layer.

Finally, the enhancement weight map $\mathbf{F}_w$ is generated by applying a sigmoid activation to the element-wise addition of $\mathbf{F}_{\text{MSFE}}$ and $\mathbf{F}_{\text{LCE}}$. This weight map is then multiplied pixel-wise with the original feature $\mathbf{X}$ to obtain the final feature.

3.4. High-Frequency Guided Deformable Upsampler (HFGDU)

Within each Hourglass network, the coarse spatial resolution features are upsampled and then fused with fine spatial resolution features to increase the ability to detect small objects. However, as mentioned in the Introduction, traditional upsampling methods, such as nearest neighbor and bilinear interpolation, often fail to recover fine spatial details, leading to over-smooth boundaries (Li et al., 2024) and misalignment of high-frequency features (Dai et al., 2017), making it challenging to detect small objects effectively.

To alleviate the above challenge, the HFGDU module was introduced. As shown in Fig. 10, HFGDU alleviates the dual challenges of high-frequency detail recovery and geometric-aware feature alignment through a three-stage architecture: Initial Upsampling, High-Frequency Detail Compensation (HFDC) and Deformable Alignment Fusion (DAF).

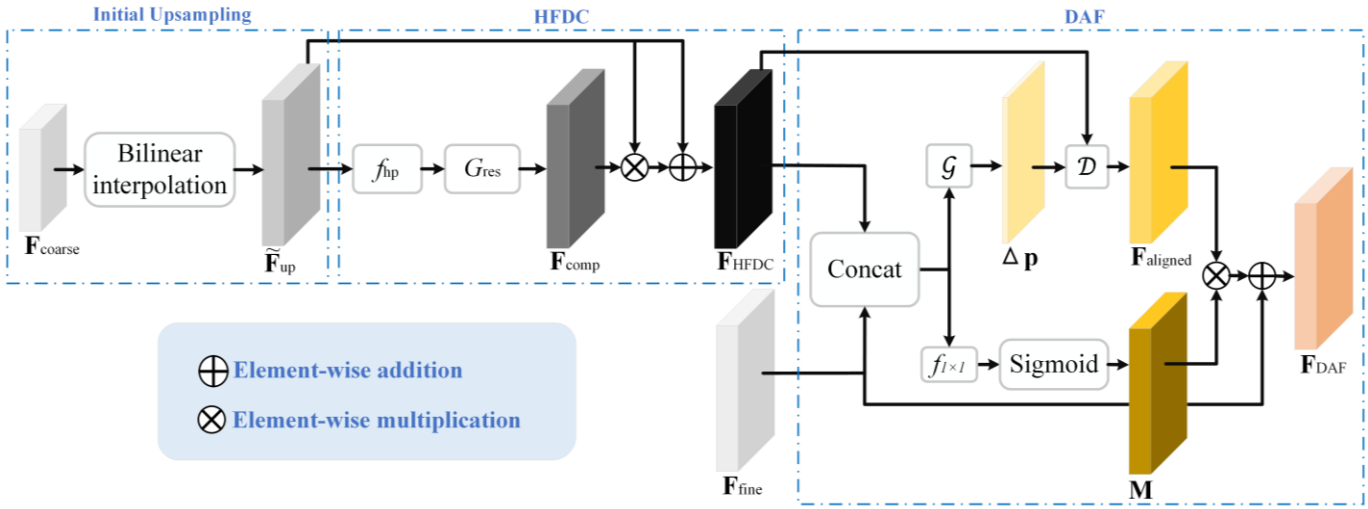

Fig. 10. The structure of the HFGDU module. It contains three main components: Initial Upsampling, HFDC and DAF. Initial Upsampling uses the bilinear interpolation to match the spatial resolution of coarse and fine features. HFDC recovers the high-frequency details of the upsampled features, while DAF refines the spatial alignment of the high-frequency features.

1) Initial Upsampling. Given an input coarse spatial resolution feature map $\mathbf{F}_{\text{coarse}} \in \mathbb{R}^{C' \times H' \times W'}$ and a fine spatial resolution feature map $\mathbf{F}_{\text{fine}} \in \mathbb{R}^{C' \times 2H' \times 2W'}$, to match the spatial size, we first applied the simple bilinear interpolation to $\mathbf{F}_{\text{coarse}}$ to obtain an upsampled feature map $\tilde{\mathbf{F}}_{\text{up}} \in \mathbb{R}^{C' \times 2H' \times 2W'}$. However, bilinear interpolation always smooths feature boundaries, failing to recover the lost high-frequency details that are

critical for small object detection.

2) HFDC. To recover the lost fine details during upsampling, we introduced a learnable high-pass filter f_{hp} to extract high-frequency components from the upsampled feature map:

$$\mathbf{F}_{\mathrm{hf}} = f_{\mathrm{hp}}(\tilde{\mathbf{F}}_{\mathrm{up}}) \tag{7}$$

where the filter f_{hp} is designed based on a Laplacian-like kernel, which is initialized with a center weight of 1 and surrounding weights of -1/8, to enhance edge structures. The extracted high-frequency feature map $\mathbf{F}_{\mathrm{hf}}$ is then refined using a residual compensation generator G_{res}:

$$\mathbf{F}_{\mathrm{comp}} = G_{\mathrm{res}}(\mathbf{F}_{\mathrm{hf}}) \tag{8}$$

where G_{res} is composed of a 3×3 convolutional layer followed by two additional 3×3 convolutional layers with ReLU activations, and a final 3×3 convolutional layer with a Sigmoid activation to constrain the compensation magnitude. The map $\mathbf{F}_{\mathrm{comp}}$ is then integrated with the upsampled feature map to generate the compensated high-frequency details map $\mathbf{F}_{\mathrm{HFDC}}$:

$$\mathbf{F}_{\mathrm{HFDC}} = \tilde{\mathbf{F}}_{\mathrm{up}} \cdot (\mathbf{1} + \mathbf{F}_{\mathrm{comp}}) \tag{9}$$

where $\cdot$ means element-wise multiplication. This step helps in recovering the fine high-frequency components, which can increase potentially the clarity and sharpness of the upsampled feature representation.

3) DAF. After compensating for high-frequency details, we employed DAF to refine the spatial alignment between the upsampled feature $\mathbf{F}_{\mathrm{HFDC}}$ and the fine spatial resolution feature $\mathbf{F}_{\mathrm{fine}}$. Traditional upsampling methods rely on fixed-grid sampling, which may cause spatial misalignment. To alleviate this, we introduced a deformable convolution layer $\mathcal{D}$ that predicts spatial offsets $\Delta\mathbf{p}$:

$$\mathbf{F}_{\text{aligned}} = \mathcal{D}(\mathbf{F}_{\text{HFDC}}, \Delta\mathbf{p}) \tag{10}$$

where $\mathbf{F}_{\text{aligned}}$ denotes the deformably sampled feature map used to correct misalignment, and the offsets $\Delta\mathbf{p}$ are adaptively learned from the feature map:

$$\Delta\mathbf{p} = \mathcal{G}(\text{Concat}(\mathbf{F}_{\text{HFDC}}, \mathbf{F}_{\text{fine}})) \tag{11}$$

In Eq. (11), $\mathcal{G}$ is a lightweight offset prediction network composed of two stacked 3×3 convolutional layers that estimates local displacements. This helps the model to align features dynamically based on the underlying structure.

Furthermore, to integrate selectively aligned features while suppressing redundant information, we introduced a feature modulation gate M that acts as an adaptive attention mechanism:

$$\mathbf{F}_{\text{DAF}} = \text{M} \odot \mathbf{F}_{\text{aligned}} + \mathbf{F}_{\text{fine}} \tag{12}$$

where $\mathbf{F}_{\text{DAF}}$ is the final fusion map and $\odot$ denotes element-wise multiplication, and the modulation gate $\mathbf{M}$ is computed as:

$$\mathbf{M} = \sigma(f_{1\times1}(\text{Concat}(\mathbf{F}_{\text{HFDC}}, \mathbf{F}_{\text{fine}}))) \tag{13}$$

where σ is the sigmoid activation. By integrating these three components, HFGDU aims to preserve high-frequency details and dynamically align features effectively, which can contribute to improving the detection of small objects in the Hourglass Network.

3.5. Local Maxima Detection Strategy

The FIDT map generated by the last Hourglass Network can identify the center point of each individual effectively, but it struggles with filtering false positives. To alleviate this issue, the LMDS was introduced to determine the correct position of each individual (Liang et al., 2022a). Specifically, a 3×3 max-pooling operation is employed to generate all candidate points, and an adaptive threshold, denoted as δ, is used to filter out false positives. This threshold is defined empirically as 100 / 255.0 times the maximum value of the FIDT map, and only points with values greater than or equal to δ are selected. Additionally, if the maximum value of the FIDT map is smaller than a tiny fixed threshold (set to 0.10) (Liang et al., 2022a), it indicates that there are no detected individuals in the input image.

3.6. Model Evaluation

To analyze the performance of the proposed CrowdSat-Net model comprehensively, both localization and counting metrics are introduced in this research. For localization metrics, the first step in assessing localization performance is to match the predicted individual point $\hat{P}$ with the ground reference individual point P. If the point-marching distance between $\hat{P}$ and P is less than a threshold γ, they are considered matched. In this study, γ was set to 2 pixels, and the nearest neighbors method (Kramer, 2013) was employed for matching, with each point being assigned to its single closest neighbor. After matching, the match matrix and counts of the number of True Positive (TP), False Positive (FP) and False Negative (FN) are obtained. Then, Precision, Recall, F1-score (Wang et al., 2020a) and Mean Localization Error (MLE) (Sam et al., 2021) were adopted based on TP, FP, FN and point-marching distance to evaluate the performance of the proposed model. The formulae are as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{14}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{15}$$

$$\text{F1-score} = \frac{2\times\text{Recall}\times\text{Precison}}{\text{Recall+Precision}} \tag{16}$$

$$\text{MLE} = \frac{1}{N_{\text{TP}}}\sum_{j=1}^{N_{\text{TP}}} ||\widehat{\mathbf{P}}_j - \mathbf{P}_j||_2 \tag{17}$$

where $\{\widehat{\mathbf{P}}_j\}_{j=1,\cdots,N_{\text{TP}}}$ and $\{\mathbf{P}_j\}_{j=1,\cdots,N_{\text{TP}}}$ denote the coordinates of the matched predicted and ground-reference points, respectively, and N_{TP} represent the number of TP. $||\widehat{\mathbf{P}}_j - \mathbf{P}_j||_2$ corresponds to the Euclidean distance between the predicted and ground-reference points. For counting metrics, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were included, which are defined as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |C_i^{pred} - C_i^{re}| \tag{18}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i^{pred} - C_i^{re})^2} \tag{19}$$

where C_i^{pred} and C_i^{re} are the predicted and ground-reference counts in the i-th image patch, respectively.

4. Experiments

In this section, comprehensive experiments on the Crowd-Sat dataset were conducted to examine the effectiveness of the proposed CrowdSat-Net method. Specifically, Section 4.1 details the experimental setup. Section 4.2 presents a series of ablation studies to validate the contributions of the DCPAN and HFGDU modules. Subsequently, Section 4.3 compares the CrowdSat-Net method with eight recent state-of-the-art CD benchmark methods (originally designed for ground or aerial imagery and satellite-based animal detection) to assess its performance. Section 4.4 examines the sensitivity of localization performance to the point-matching

519 threshold. Section 4.5 presents the computational efficiency and memory usage of different CD models.
520 Section 4.6 illustrates the performance of various backbones in CrowdSat-Net. Section 4.7 analyzes the
521 performance of CrowdSat-Net across different crowd density levels. Finally, Section 4.8 examines the
522 cross-regional generalization of the CrowdSat-Net method by applying it to unseen foreign regions.
523
524 *4.1. Experimental Setup*
525
526 In this research, the labeled patches in CrowdSat were divided randomly into training and validation sets in
527 a 4:1 ratio. To increase the robustness and generalization of models, we applied augmentation techniques,
528 including horizontal flipping, vertical flipping and CutMix (Yun et al., 2019). After augmentation, patches
529 without individuals were removed, resulting in 10,357 training patches and 704 validation patches.
530 After preparing the dataset, we trained the CrowdSat-Net model, all baseline models and other CD models
531 for 150 epochs using the Adam optimizer (Adam et al., 2014) with a base learning rate of 0.0003 and a weight
532 decay of 0.001, on the same augmented dataset described above to ensure fair comparison. For all models, the
533 batch size was set to 8, and they were implemented in PyTorch 1.21. All experiments were conducted on a
534 Windows 10 workstation equipped with a 13th Gen Core i7-13700 CPU and a 24 GB NVIDIA GeForce RTX
535 4090 GPU.
536
537 *4.2. Ablation Experiments*
538
539 *4.2.1. Ablation of DCPAN and HFGDU.*
540
541 To examine the contributions of the DCPAN and HFGDU modules in CrowdSat-Net, a basic ablation
542 experiment was conducted. In this experiment, the baseline refers to CrowdSat-Net with the DCPAN module
543 removed and the HFGDU module replaced by bilinear interpolation.

As presented in Table 1, introducing either DCPAN or HFGDU outperforms the baseline in terms of localization and counting metrics. For example, with DCPAN alone, the F1-score and Precision increase by 0.90% and 4.36%, respectively, while the MLE decreases by 0.02. The MAE and RMSE decrease by 0.83 and 2.74, respectively. A similar trend is observed when only HFGDU is enabled, that is, the F1-score rises to 65.52%, and MLE, MAE and RMSE are reduced to 0.40, 12.69 and 35.55, respectively. Notably, combining both modules achieves the best overall performance, with F1-score and Precision increased, and MLE, MAE and RMSE decreased compared with the baseline.

Table 1 Basic ablation study for DCPAN and HFGDU.

Methods	DCPAN	HFGDU	F1-score (%)	Recall (%)	Precision (%)	MLE	MAE	RMSE
Baseline	×	×	64.42	60.77	68.54	0.42	13.04	37.94
CrowdSat-Net	√	×	65.32	59.16	72.90	0.40	12.21	35.20
	×	√	65.52	**61.10**	70.63	0.40	12.69	35.55
	√	√	**66.12**	60.27	**73.23**	**0.39**	12.07	34.37

4.2.2. Ablation Study of Different Feature Enhancement Modules in CrowdSat-Net.

To further examine the effectiveness of DCPAN, a comparison with two well-known feature enhancement modules (spatial attention (SA) (Wang et al., 2017) and efficient channel attention (ECA) (Wang et al., 2020b)) was performed. In this experiment, all components were the same as the baseline method except for the feature enhancement module. The result is displayed in Table 2. It is seen that DCPAN achieves greater performance than SA and ECA, achieving the largest F1-score of 65.32% (+0.46% over SA, +0.58% over ECA) and Recall of 59.16% (+1.32% over SA, +1.02% over ECA). It also produces the smallest MAE (12.21) and RMSE (35.20), decreasing MAE by 0.93 and 1.23 and RMSE by 2.67 and 3.49 compared to SA and ECA, respectively.

Architecturally, DCPAN extends SA by integrating two novel branches: the MSFE and LCE. Specifically, MSFE utilizes dilated convolutions for multi-scale contextual information, mitigating false negatives effectively caused by the fixed receptive field of SA. Furthermore, in contrast to SA, the LCE enhances feature representation by amplifying local feature responses through contrast-based spatial attention. These innovations increase the localization accuracy of DCPAN collectively. Compared to ECA, which focuses on channel-wise attention exclusively, DCPAN alleviates spatial signal blurring through its dual-branch design.

Table 2 Comparison of DCPAN with SA and ECA in CrowdSat-Net.

Methods	F1-score (%)	Recall (%)	Precision (%)	MLE	MAE	RMSE
SA (Wang et al., 2017)	64.86	57.84	**73.82**	**0.38**	13.14	37.87
ECA (Wang et al., 2020b)	64.74	58.14	73.03	0.41	13.44	38.69
DCPAN	**65.32**	**59.16**	72.90	0.40	**12.21**	**35.20**

4.2.3. Ablation of Different Upsamplers in CrowdSat-Net.

To reveal comprehensively the effectiveness of HFGDU, a comparative analysis with four established benchmark upsamplers, including bilinear interpolation, Content-Aware ReAssembly of FEatures (CARAFE) (Wang et al., 2019), Deconvolution (Long et al., 2015), Pixel-shuffle (Shi et al., 2016), Dense Upsampling Convolution (DUC) (Zhao et al., 2017) and Dysample (Liu et al., 2023b), was conducted.

The result in Table 3 demonstrates that HFGDU exhibits greater localization performance over existing upsamplers, achieving the largest F1-score and Precision. This is attributed to its dual-domain enhancement mechanism: 1) The HFDC module employs a learnable Laplacian operator to amplify boundary gradients, mitigating the inherent edge blurring in bilinear interpolation effectively and yielding a 0.33% Precision gain, a 1.11% Recall increase and a 0.80% F1-score increase over bilinear interpolation. 2) The DAF mechanism based on dynamic migration prediction reduces the feature misalignment error by jointly optimizing the spatial correspondence between coarse and fine spatial resolution features. While Dysample achieves the largest

Recall, its Precision lags obviously behind HFGDU (-2.79%). While Deconvolution achieves a larger Recall, its fixed transposed convolution kernels lead to checkerboard artifacts, decreasing Precision by 2.44%. Although CARAFE generates upsampled cores via content-sensing, it struggles with small objects in dense scenes, leading to over-smooth results and a lower F1-score (65.17%) than bilinear interpolation (65.32%). Pixel-shuffle performs upsampling by first expanding the channel dimension and then rearranging the channels into a finer spatial grid. However, in this small-object detection task, its accuracy gain is limited. DUC yields slightly larger F1-score and Recall than bilinear interpolation, but the overall accuracy gain remains limited and still smaller than for HFGDU.

Table 3 Comparison between different upsamplers in CrowdSat-Net.

Methods	F1-score (%)	Recall (%)	Precision (%)	MLE	MAE	RMSE
Bilinear interpolation	65.32	59.16	72.90	0.40	12.21	35.20
CARAFE (Wang et al., 2019)	65.17	59.47	72.08	**0.38**	13.07	36.92
Deconvolution (Long et al., 2015)	65.38	60.74	70.79	0.41	12.76	35.84
Dysample (Liu et al., 2023b)	65.62	**61.42**	70.44	0.42	13.68	38.86
Pixel-shuffle (Shi et al., 2016)	64.98	59.44	71.66	0.42	13.35	37.14
DUC (Zhao et al., 2017)	65.46	60.12	71.84	0.41	12.88	36.14
HFGDU	**66.12**	60.27	**73.23**	0.39	**12.07**	**34.37**

4.3. Comparison Between CrowdSat-Net and Benchmark Methods

This section aims to compare the performance of the proposed CrowdSat-Net with six state-of-the-art CD methods (originally designed based on ground or aerial imagery): Point to Point Network (P2PNet) (Song et al., 2021), SCALNet (Wang et al., 2021), Crowd Localization TRansformer (CLTR) (Liang et al., 2022b), Point quEry Transformer (PET) (Liu et al., 2023a), Focal Inverse Distance Transform Map for Crowd Localization (FIDTMCL) (Liang et al., 2022a) and Auxiliary Point Guidance Crowd Counting (APGCC)

(Chen et al., 2024). Also, two satellite-based animal detection approaches, HerdNet (Delplanque et al., 2023) and U-Net-based ensemble model (Wu et al., 2023) (hereafter, UNE), were included.

As illustrated in Figs. 11 and A.1, five representative scenes, including traffic junctions, snowfields, dense urban regions, desert regions and other common impervious regions, were selected from 704 validation patches to demonstrate the detection performance across various scenarios. It is seen that SCALNet exhibits noticeable missed detections, particularly in cluttered junctions and dense urban areas where individual signals are easily confused with crosswalk markings, shadows and building edges within the yellow circles. FIDTMCL and APGCC alleviate this issue to some extent and detect more individuals in complex scenes. However, some of their predictions remain blurred around object boundaries (such as buildings, trees and tents), which leads to missed individuals at the periphery of gatherings. HerdNet captures several prominent clusters of individuals, yet the predictions are biased to high-density crowds and ignore frequently isolated individuals along roads or fences. P2PNet produces sparse predictions and often detects only a few scattered individuals in regions that contain clear gathering. The transformer-based methods, i.e., CLTR and PET, achieve limited visualization performance and, in densely populated areas, almost fail to detect crowds while producing anomalous outputs. UNE and CrowdSat-Net achieve greater detection performance with fewer missed or false detections, and, in the high- and low-density areas, their predictions are more accurate than those of the other benchmark methods.

Quantitative analysis on all 704 validation patches is shown in Table 4. It is clear that CrowdSat-Net achieves state-of-the-art localization performance with an F1-score of 66.12%, a Precision of 73.23% and an MLE of 0.39. Compared to the second most accurate method (UNE), CrowdSat-Net increases the F1-score by 0.80% and Precision by 6.38% and decreases the MLE by 0.36. Notably, CrowdSat-Net outperforms SCALNet by 5.71% in terms of F1-score, 8.41% in terms of Recall and 0.88 in terms of MLE, highlighting the effectiveness in small-object detection. Although APGCC achieves a larger Recall and the smallest MAE and RMSE, its Precision is 7.53% smaller than that of CrowdSat-NET, and its MLE is 0.75 larger, revealing inherent limitations in point-based auxiliary supervision for the precise localization of small objects. HerdNet

obtains an MLE of 0.50, but its F1-score is only 60.18%. P2PNet and CLTR present relatively greater counting performance, yet their localization performance is limited.

Table 4 Accuracy of nine CD methods on the CrowdSat dataset.

Methods	F1-score (%)	Recall (%)	Precision (%)	MLE	MAE	RMSE
SCALNET (Wang et al., 2021)	60.38	51.86	72.26	1.27	13.31	34.18
P2PNet (Song et al., 2021)	49.60	49.17	50.04	1.14	10.17	38.24
CLTR (Liang et al., 2022b)	13.50	13.27	13.73	1.24	9.26	22.06
PET (Liu et al., 2023a)	11.36	12.54	10.41	1.57	15.64	41.48
FIDTMCL (Liang et al., 2022a)	64.41	60.78	70.80	0.98	12.67	34.76
APGCC (Chen et al., 2024)	64.34	63.05	65.70	1.14	**8.31**	**17.32**
HerdNet (Delplanque et al., 2023)	60.18	57.31	63.35	0.50	14.49	38.93
UNE (Wu et al., 2023)	65.32	**63.87**	66.85	0.75	9.12	23.44
CrowdSat-Net	**66.12**	60.27	**73.23**	**0.39**	12.07	34.37

634

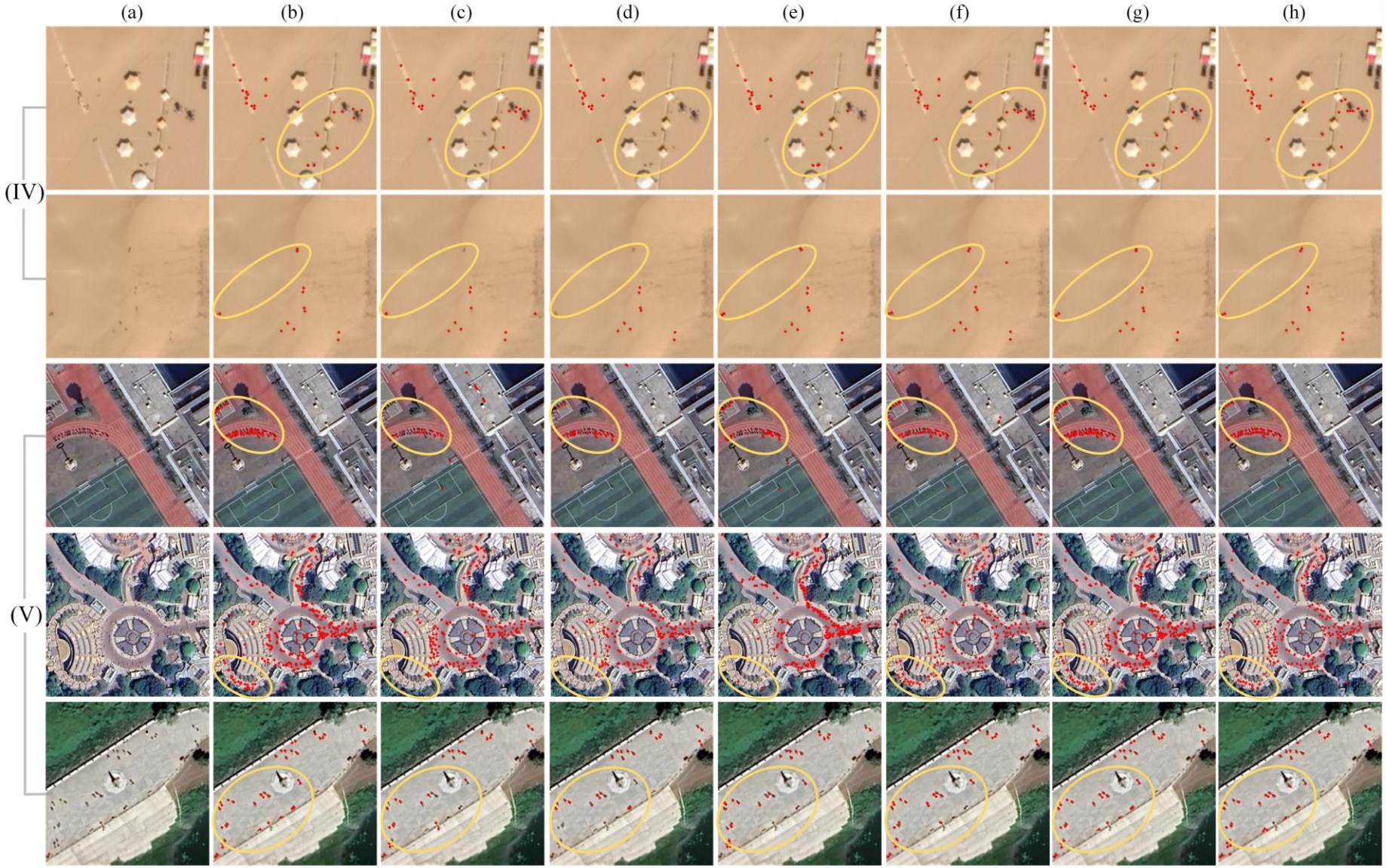

Fig. 11. Visual comparisons between different CD methods for the CrowdSat dataset. (a) Original image. (b) Reference. (c) SCALNet. (d) FIDTMCL. (e) APGCC. (f) HerdNet. (g) UNE. (h) CrowdSat-Net. (I) Traffic junctions. (II) Snowfields. (III) Dense urban regions. (IV) Desert regions. (V) Other common impervious regions. Some of the images in panels (I)-(V) were obtained from the Google Earth platform (© Google Earth 2025).

4.4. Sensitivity Analysis of Localization Performance to the Point-Matching Threshold

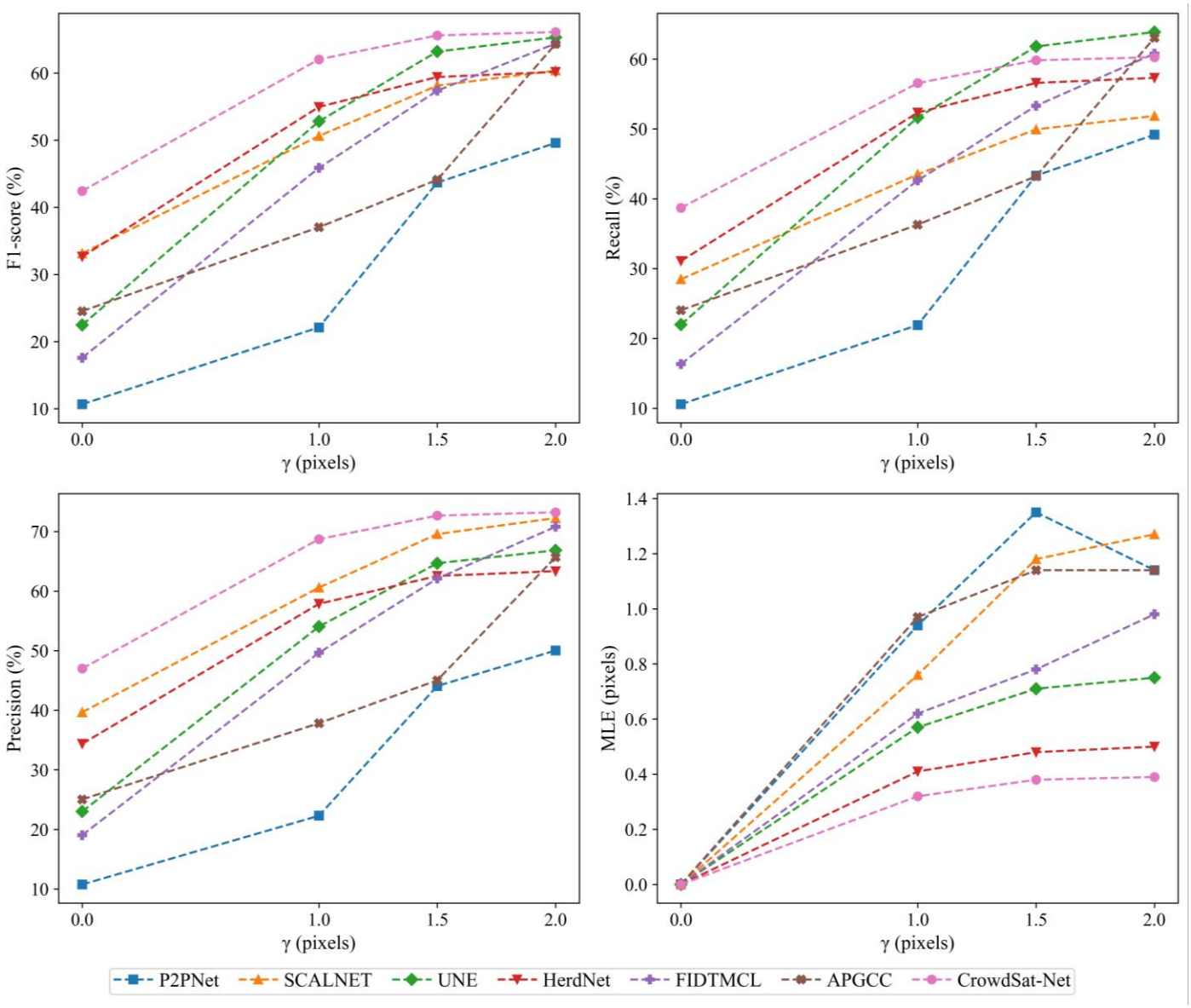

Fig. 12. Sensitivity of localization performance to the point-matching threshold across different CD methods.

To further examine the influence of the point-matching threshold, we conducted a sensitivity analysis with respect to the threshold γ. All methods (except for CLTR and PET) were evaluated under the selected values, i.e., 0.0, 1.0, 1.5 and 2.0. As illustrated in Fig. 12, CrowdSat-Net exhibits greater localization performance over all benchmark methods in two aspects. First, CrowdSat-Net generally achieves the greatest accuracy

under all values of γ. Notably, when γ is set to 0, which means that a prediction is counted as correct only if its pixel coordinates exactly coincide with those of a ground-reference point, CrowdSat-Net still obtains the largest F1-score of 42.44%, Recall of 38.69% and Precision of 47.00%. These values exceed the second most accurate results by 9.30%, 7.60% and 7.34%, respectively. Second, CrowdSat-Net is less sensitive to the selection of γ than the other methods. For example, when γ changes from 2.0 to 0.0 pixels, the F1-score and Precision of FIDTMCL decrease from 64.41% to 17.59% and from 70.80% to 19.04%, respectively. P2PNet and APGCC also show a similar trend. These steep drops illustrate that many of their predicted points lie near the matching point and are only counted as true positives when a relatively large γ is used.

4.5. Computational Efficiency and Memory Usage

This section assesses the practical applicability of different CD methods by comparing the computational cost per 256×256 pixel patch in terms of FLOPs, the number of parameters, runtime latency and peak GPU memory. As presented in Table 5, CrowdSat-Net is the most computationally demanding model. It has the largest FLOPs and parameter count, and its latency and peak memory are higher than those of the other methods. This overhead mainly stems from the Hourglass structure itself and the additional DCPAN and HFGDU modules. Despite this larger computational cost, CrowdSat-Net yields notable localization performance on CrowdSat, as discussed in Section 4.4. Moreover, in terms of practical deployment, for scenarios where computational resources are very limited and slightly lower localization accuracy is acceptable, methods such as SCALNET or APGCC may be preferred. For applications that require accurate localization of extremely small objects under strict matching criteria, CrowdSat-Net offers a more appropriate solution.

Table 5 Computational efficiency and memory usage for different CD methods (per 256×256 pixel patch).

Methods	FLOPs	Params	Runtime Latency	Peak GPU memory

SCALNET (Wang et al., 2021)	7.32G	18.64M	1.96ms	94.45M
P2PNet (Song et al., 2021)	26.18G	19.22M	2.21ms	122.99MB
CLTR (Liang et al., 2022b)	16.03G	59.89M	14.51ms	273.23MB
PET (Liu et al., 2023a)	221.06G	52.98M	12.24ms	300.85MB
FIDTMCL (Liang et al., 2022a)	35.60G	66.58M	15.02ms	317.24MB
APGCC (Chen et al., 2024)	23.49G	18.56M	2.039ms	113.88MB
HerdNet (Delplanque et al., 2023)	7.55G	18.16M	3.07ms	106.67MB
UNE (Wu et al., 2023)	40.19G	17.26M	2.87ms	182.42MB
CrowdSat-Net	297.29G	94.12M	20.25ms	701.05MB

4.6. Performance of Different Backbones in CrowdSat-Net

To investigate the performance of different backbones in CrowdSat-Net, we integrated DCPAN and HFGDU into three representative backbones, namely, U-Net (Ronneberger et al., 2015), Deep Layer Aggregation (DLA) (Yu et al., 2018) and HRNet (Sun et al., 2019), all of which include coarse and fine spatial resolution fusion strategies. As shown in Table 6, when γ=2, the F1-scores range for the three backbones are 59.80–62.30%, which are below those of the benchmark methods, such as UNE, FIDTMCL and APGCC (64.34–65.32%). However, as γ decreases, the accuracy of the three backbones becomes greater than that of the benchmark methods gradually. For example, when γ=0, the three backbones still retain F1-scores between 35.43% and 37.36%, while UNE, FIDTMCL and APGCC fall to 22.48%, 17.59% and 24.53%, respectively. With γ changes, the accuracies of the three backbones are more stable than those of the benchmark methods. Furthermore, compared to the Hourglass backbone, the three backbones produce clearly smaller localization accuracy at all point-matching thresholds, but they have fewer parameters than the Hourglass backbone, which makes them suitable replacements for Hourglass in applications that operate with more tolerant point-matching thresholds.

Table 6 Comparison between different backbones in CrowdSat-Net.

Methods	γ	Params	F1-score (%)	Recall (%)	Precision (%)	MLE	MAE	RMSE
U-Net (Ronneberger et al., 2015)	0	17.79M	35.43	38.09	33.13	0.00	16.56	33.77
	1		55.29	59.43	51.69	0.36		
	1.5		60.24	64.75	56.32	0.46		
	2		60.39	64.92	56.46	0.47		
DLA (Yu et al., 2018)	0	17.15M	35.51	37.63	33.61	0.00	15.59	33.54
	1		54.96	58.25	52.02	0.35		
	1.5		59.72	63.30	56.53	0.50		
	2		59.80	63.38	56.61	0.46		
HRNet (Sun et al., 2019)	0	10.16M	37.36	36.88	37.85	0.00	13.91	34.45
	1		57.50	56.77	58.26	0.35		
	1.5		60.95	60.17	61.75	0.41		
	2		62.30	61.50	63.11	0.40		

4.7. Performance of CrowdSat-Net with Different Crowd Densities

This section examines the performance of CrowdSat-Net systematically across different crowd densities to highlight its adaptive advantages in complex environments. The crowd count ranges were derived from the reference crowd count distribution using quantiles (0.1, 0.3, 0.6, 0.8, 0.95). The crowd density levels were then categorized as follows based on these quantile-derived ranges: 1–5 individuals within an image were classified as extremely sparse, 6–20 as sparse, 21–45 as moderate, 46–100 as relatively high, 101–800 as high and 800+ as extremely dense.

As illustrated in Fig. 13, CrowdSat-Net demonstrates consistent robustness in moderate-to-high crowd density scenarios, achieving peak F1-scores of 73.03% in the 46–100 group and 69.60% in the 21–45 group. However, performance degradation was observed in both extremely sparse and extremely dense scenarios. For extremely sparse groups (1–5 individuals), Precision decreases to 35.81%, and in extremely dense crowds (800+), Recall drops to 39.23%. To explain these situations, a visual diagram of these situations is shown in

Fig. 14. In extremely sparse scenes, the obvious decrease in Precision is attributed to an increased susceptibility to false positives caused by background clutter, such as image noise and other small objects (e.g., road asphalt, stone pillars and street lamps). Additionally, in extremely dense crowds, the decrease in Recall is due to severe occlusion among individuals, making true positive detection challenging.

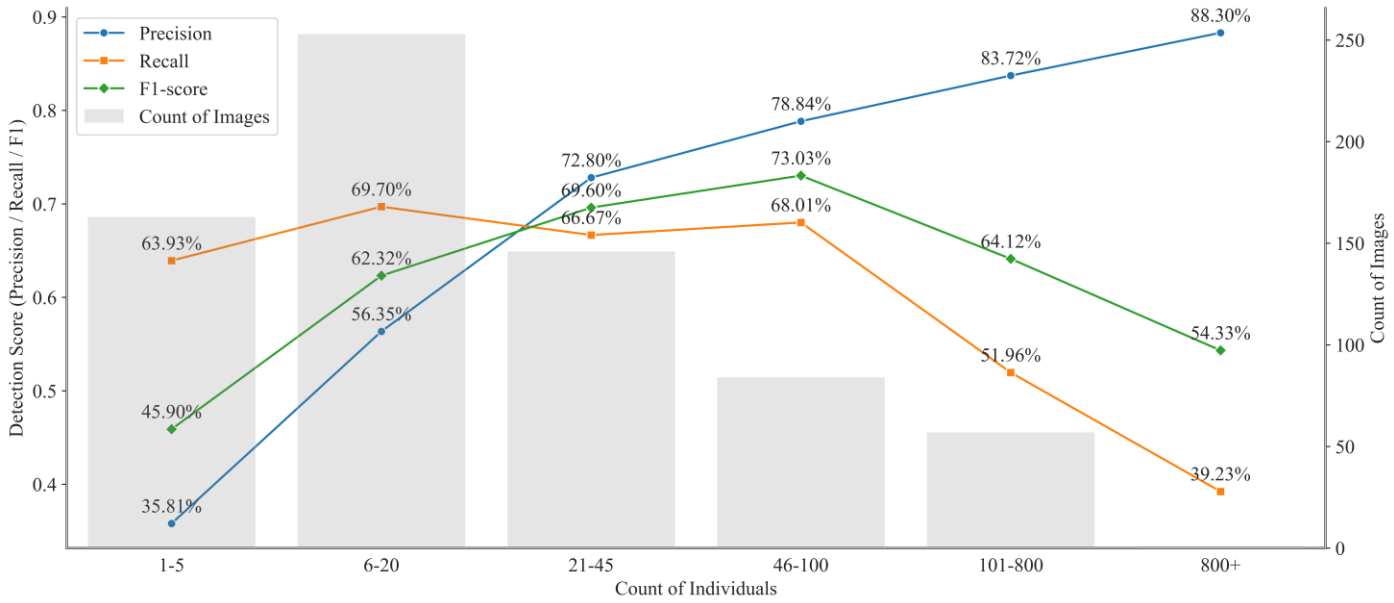

Fig. 13. Performance of CrowdSat-Net on CrowdSat across different crowd densities. The lines show the Precision, Recall and F1-score evaluated on subsets of images grouped by crowd densities, while the grey bars in the background indicate the number of images in each group.

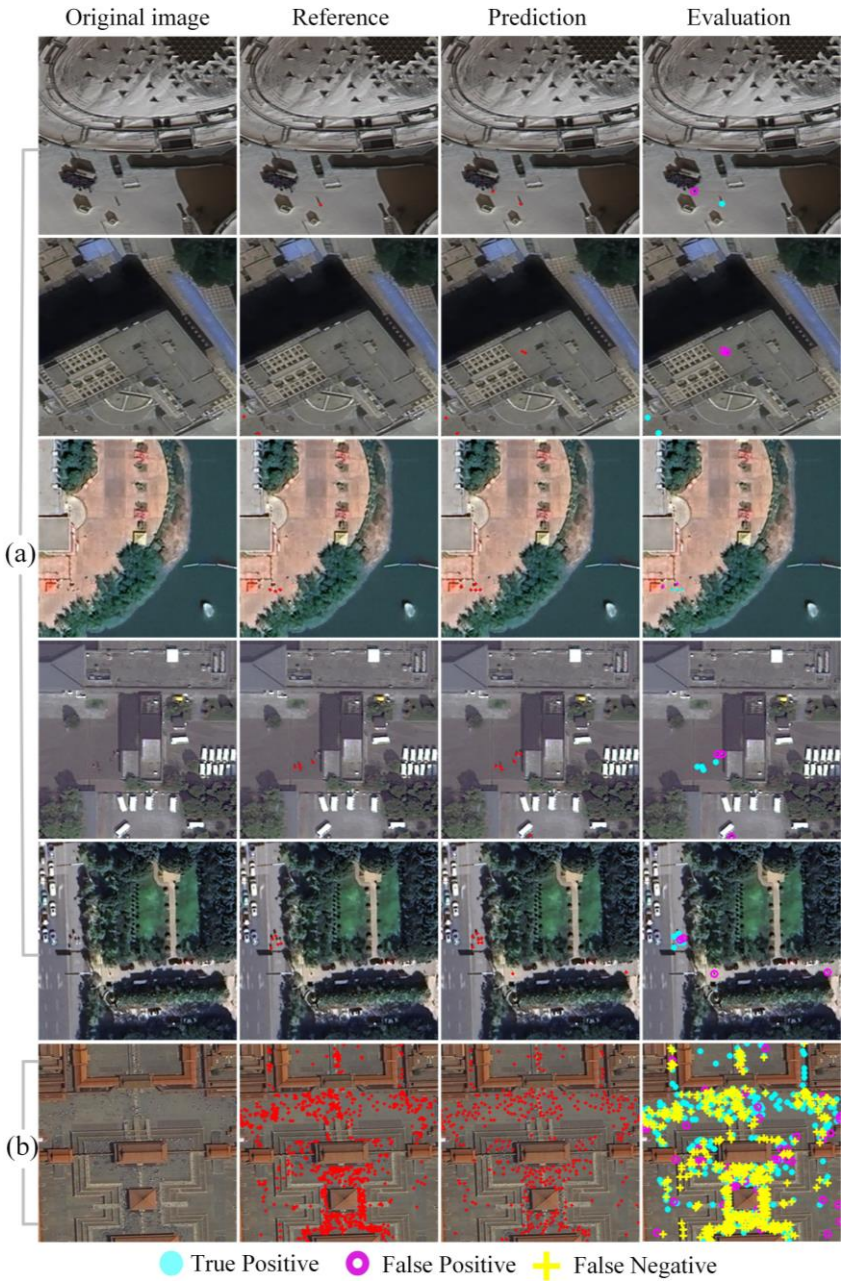

Fig. 14. Examples of the localization performance of CrowdSat-Net in extreme scenarios: (a) Extremely sparse. (b) Extremely dense.

4.8. Cross-Regional Generalization of CrowdSat-Net

In this research, we used CrowdSat to cover a broad range of scenes, e.g., snowy regions, vegetated areas, beaches, desert regions, etc., so that the generalization ability of CD models can be examined under diverse conditions. To further assess cross-regional generalization, we additionally tested CrowdSat-Net and the

benchmark methods (HerdNet, UNE, SCALNet, P2PNet, FIDTMCL and APGCC) on six unseen global regions that are not included in CrowdSat. These regions are Red Square (Moscow, Russia), Metropolitan Cathedral (Mexico City, Mexico), Phra Nakhon (Bangkok, Thailand), Djemaa el Fna (Marrakesh, Morocco), India Gate (New Delhi, India) and National Mall (Washington, D.C., USA). Also, these regions exhibit distinct characteristics. Red square and Djemaa el Fna are large plazas surrounded by tall buildings, where shadows and open areas with dark colors make individuals present low contrast. Metropolitan Cathedral and India Gate contain complex traffic junctions, with moving vehicles mixed with pedestrians on roads. Phra Nakhon and the National Mall include extensive vegetated areas in which dark green grass shows similar texture characteristics to individuals. All images were annotated with the same point-based protocol as CrowdSat. Their image sizes, collection dates, sources, spatial resolution, count of individuals and scenes included are summarized in Table 7.

Table 7 Description of cross-regional test regions.

Regions	Image Size (pixels)	Date	Source	Spatial Resolution	Count	Scenes
Red Square	512 × 512	Sep. 27, 2023	Google Earth	0.3 m	394	Open public areas; built-up areas
Metropolitan Cathedral	1280 × 1405	Oct. 11, 2024	Google Earth	0.3 m	898	Built-up areas; traffic junctions
Phra Nakhon	1402 × 1754	Jan. 11, 2025	Google Earth	0.3 m	725	Open public areas; vegetated areas; traffic junctions
Djemaa el Fna	2003 × 1514	Mar. 14, 2024	Google Earth	0.3 m	773	Open public areas; built-up areas; traffic junctions
India Gate	2337 × 677	Oct. 26, 2024	Google Earth	0.3 m	519	Open public areas; vegetated areas; traffic junctions
National Mall	2718 × 2079	Mar. 3, 2024	Google Earth	0.3 m	575	Vegetated areas; traffic junctions

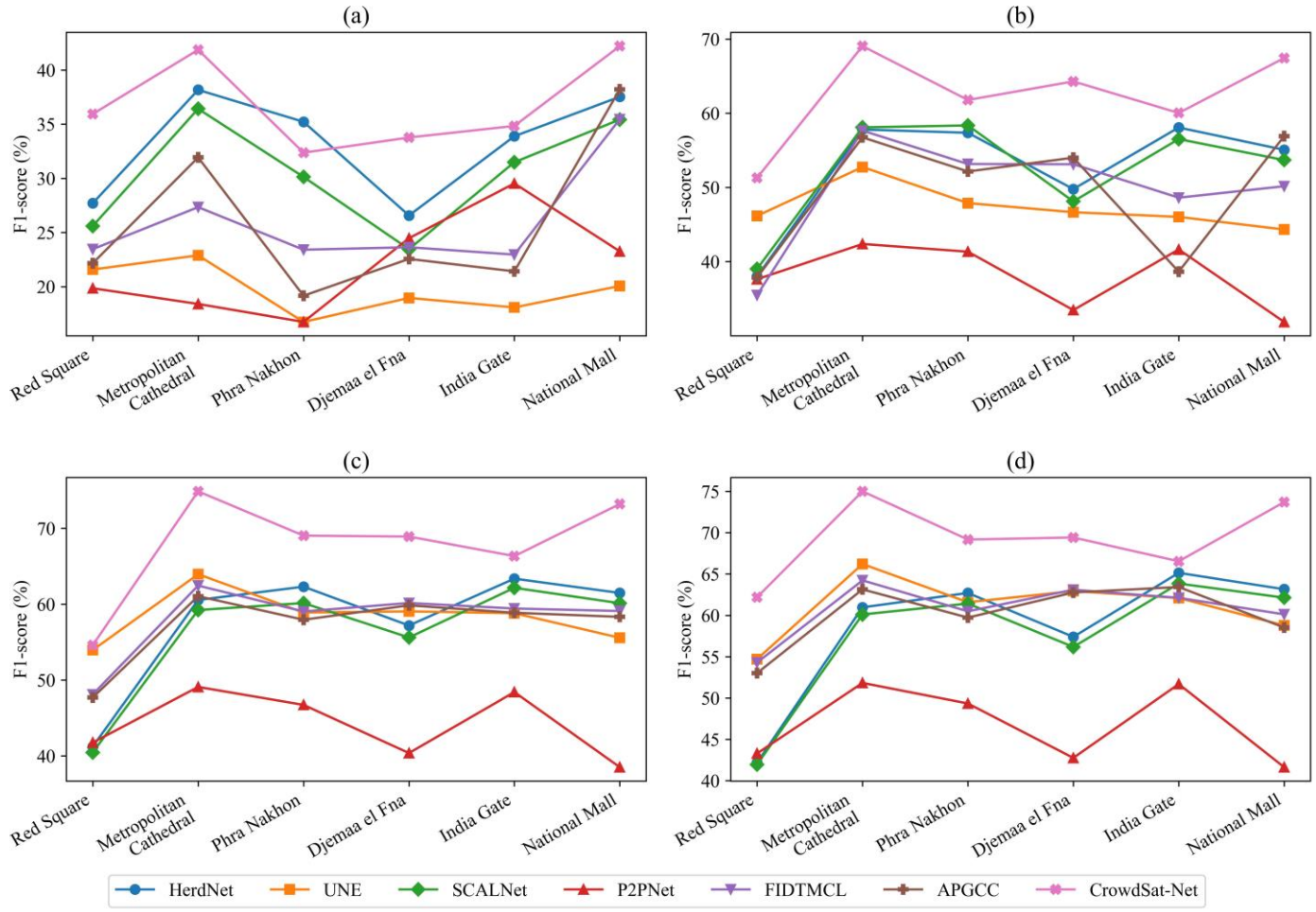

Fig. 15. F1-scores of different CD methods for the unseen regions under varying point-matching thresholds. (a) γ=0. (b) γ=1. (c) γ=1.5. (d) γ=2.

As displayed in Fig. 15, across all six unseen regions, CrowdSat-Net achieves the largest or second-largest F1-score for all point-matching thresholds. CrowdSat-Net produces mean F1-scores of 36.83%, 62.32%, 67.83% and 69.34% for γ=0.0, 1.0, 1.5 and 2.0, respectively. Figs. 16 and A.2 show the predictions of CrowdSat-Net in the six unseen regions. For low-contrast plazas, CrowdSat-Net detects most individuals successfully in open areas and along building fa ades. At complex traffic junctions, CrowdSat-Net is able to separate most of the individuals from nearby vehicles. In vegetated areas, CrowdSat-Net can still localize a large proportion of isolated individuals that are scattered across the grass. However, several obvious false positives and negatives are also observed in these regions: false positives mostly occur on background objects that exhibit visual characteristics similar to individuals, such as road asphalt, street lamps, stone pillars, etc..

False negatives are concentrated in highly dense clusters where adjacent individuals overlap. These limitations will be further discussed in Section 5.3.

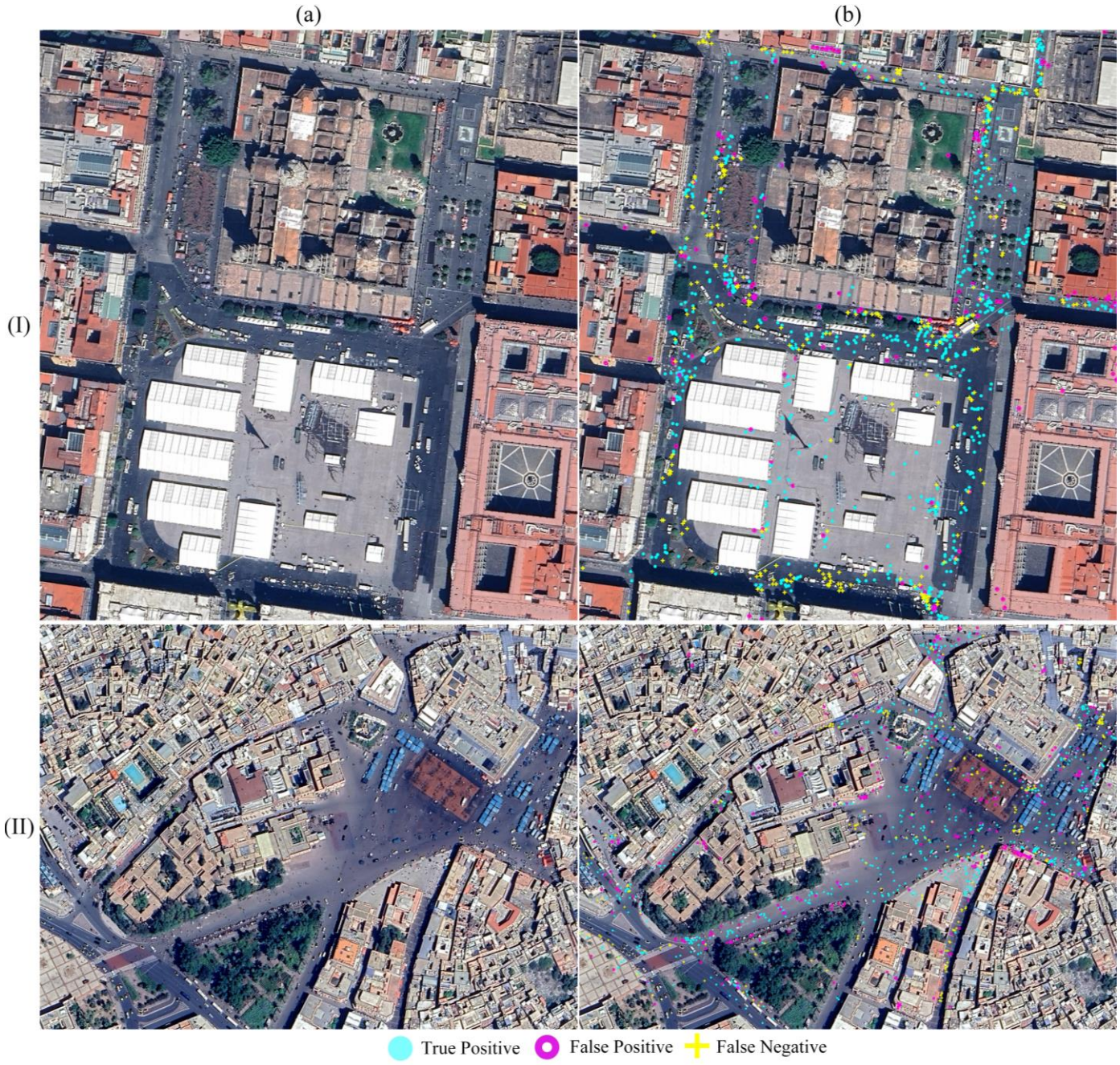

Fig. 16. The visual localization performance of CrowdSat-Net in unseen foreign regions. (a) Metropolitan Cathedral. (b) Djemaa el Fna. (I) Original image. (II) Evaluation. The images in (a)-(b) were obtained from the Google Earth platform (© Google Earth 2025).

5. Discussion

5.1. Limitations of the CrowdSat Dataset

While CrowdSat demonstrates advantages for CD via VFR satellite imagery, several inherent limitations must be acknowledged to ensure its appropriate application and interpretation, which are as follows:

1) Limited Coverage Area. While CrowdSat was collected in regions with diverse environmental conditions, its coverage range is within China, which may limit the generalizability of models in extremely rare global environments, such as polar regions and dense tropical forests.

2) Occlusion Constraints. The reliance of CrowdSat on optical imagery limits its effectiveness inherently in occluded environments. Buildings, dense tree canopies and their shadows can obscure individuals, making crowds only partially detectable or undetectable in some scenes.

3) Minimum Detectable Crowd Size. At a spatial resolution of 0.3 m, more than one individual may fall within a single pixel in extremely high-density gatherings, but such cases are very rare in CrowdSat. To compete with current CD models, which can also not resolve sub-pixel individuals, we adopt a labelling protocol that restricts annotations to at most one individual per pixel. This strategy may lead to an underestimation of the number of individuals.

4) Temporal Resolution Constraints. Although VFR satellites offer relatively frequent revisit times (e.g., BJ3N: every 5 days), the short-lived and time-specific nature of crowd activities makes them difficult to capture reliably. Cloud cover and atmospheric conditions often extend the effective revisit interval.

5.2. Applicability of CrowdSat-Net

In this research, CrowdSat-Net demonstrated reliable localization performance of individuals, and it is expected that CrowdSat-Net and its two innovative modules (DCPAN and HFGDO) have the potential for broader applications.

5.2.1. Generalized Small Object Detection

While CrowdSat-Net was proposed for large-scale CD, its core design supports generalized small-object detection for satellite imagery. Many non-human objects share similar attributes with human crowds: low pixel occupancy (typically 4-10 pixels per object), irregular spatial distributions and high contextual heterogeneity. Potential applications include 1) wildlife monitoring, such as detection of migratory ungulates, penguin colonies or marine species from VFR satellite imagery; 2) transportation surveillance, including identification of vehicles in road networks, parking lots or logistics hubs; and 3) environmental mapping, such as recognition of vegetation clusters or deforestation patterns in ecological studies. For example, CrowdSat-Net achieves an F1-score of 83.16%, a Precision of 83.69% and a Recall of 82.67% on the AED dataset (Naude et al., 2019), which was collected in Africa and contains 1,649 RGB images with 12,455 manually annotated elephants (point annotation), showing a great localization performance in more general small-object detection tasks.

5.2.2. Modular Integration with Existing Remote Sensing Frameworks

DCPAN and HFGDO can be potentially integrated into various backbone architectures. Specifically, these two modules can be embedded into widely used coarse and fine spatial resolution fusion architectures, such as Feature Pyramid Network (FPN) (Lin et al., 2017a), U-Net (Ronneberger et al., 2015) and HRNet (Sun et al., 2019). For example, in U-Net, each standard skip connection can be embedded with DCPAN, and the traditional upsampling method (bilinear interpolation) can be replaced with HFGDO.

5.3. Limitations of CrowdSat-Net

To explore the limitations of CrowdSat-Net, we selected several representative scenes, including stadiums, parks, bridges, beaches, road junctions and desert tourist areas, where CrowdSat-Net exhibits noticeable localization errors, as shown in Fig. 17. First, a non-negligible fraction of false negatives occurs in shadows or under occlusions by trees and buildings (row 2). In these regions, CrowdSat-Net tends to miss individuals or to produce only a few sparse detections along trees or buildings. Second, in highly dense scenes, such as bridges and beaches, the signals of adjacent individuals overlap heavily, which makes CrowdSat-Net difficult to separate all of them into distinct instances and leads to false negatives within compact groups. Third, false positives are frequently observed on high-contrast background patterns, including stadium markings, beach umbrellas, E-bikes and road markings. The textures of these objects are highly similar to those of individuals, causing the network to overestimate the number of individuals in these scenes. Finally, in sparsely populated regions, isolated objects such as small kiosks or shadows of infrastructure in desert regions may also be misclassified as individuals.

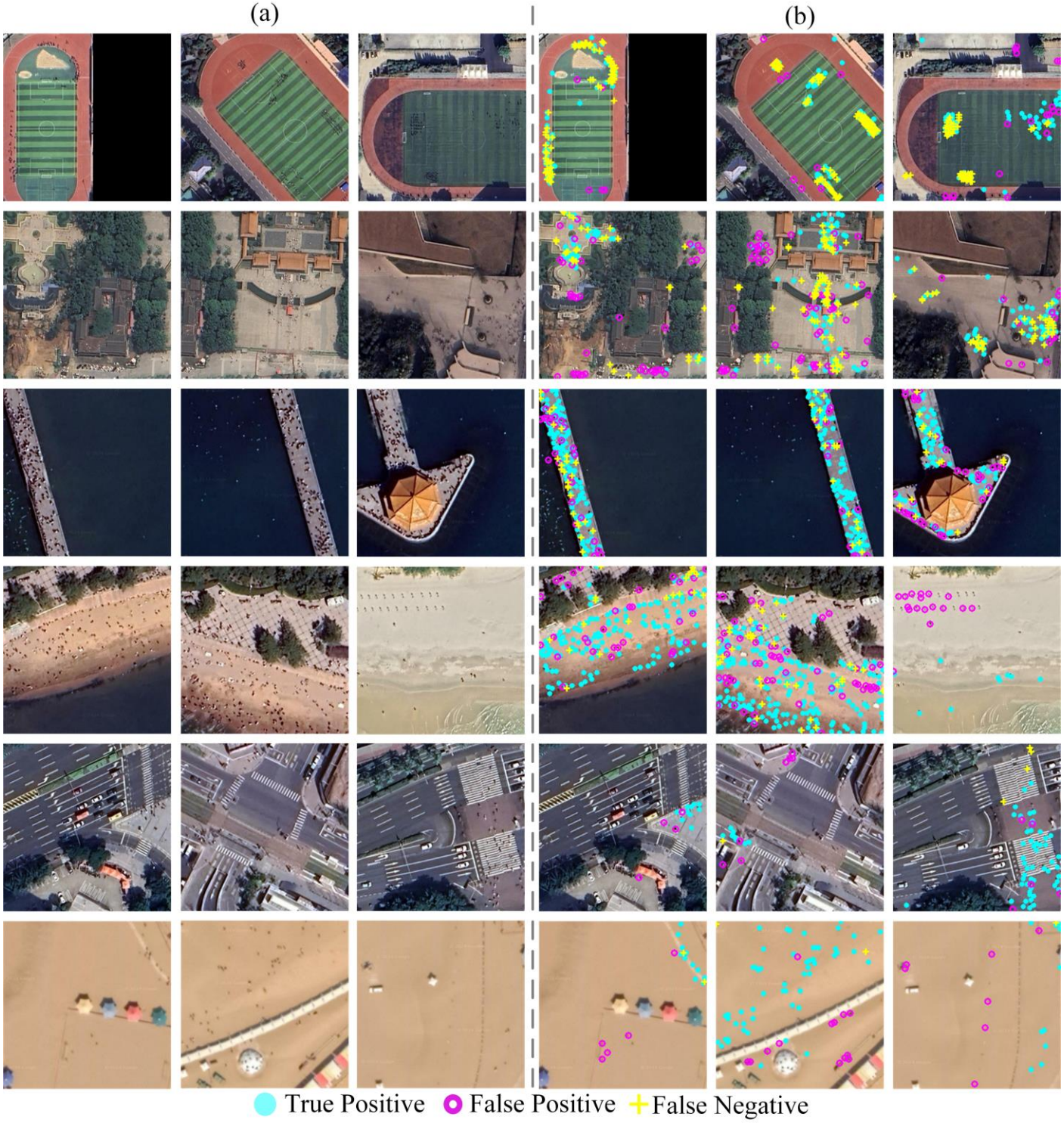

Fig. 17. Error maps of CrowdSat-Net on representative scenes. (a) Original images. (b) Evaluation. Some of the images were obtained from the Google Earth platform (© Google Earth 2025).

5.4. Future Research

Based on the limitations identified in CrowdSat and CrowdSat-Net, we propose two key directions for future research:

1) Large-Scale and Multi-Source Dataset Expansion. The CrowdSat dataset primarily covers Chinese regions, which may limit its geographical generalizability, especially for extremely rare global environments. To enhance model robustness, future datasets should incorporate global crowd patterns (by leveraging multi-source optical satellite imagery (0.3 m spatial resolution) and SAR constellations) and encompass more complex background environments (such as dense urban environments, mixed land cover regions and varied climatic conditions) to improve model adaptability across diverse scenarios.

2) Temporal-Aware and Super-Resolution-Based Crowd Detection Frameworks. To decrease the false detection rates in extremely sparse crowd regions, future models may integrate multi-temporal image sequences with dynamic change detection mechanisms. One possible solution is a dual-branch design, where one part of the model focuses on detecting individuals in an image while the other tracks how they persist over time, reducing false detections due to static, small-sized background objects. On the other hand, to alleviate the individual signal occlusion in extremely dense crowd regions, super-resolution techniques may be employed to enhance image clarity, making it easier to distinguish individuals.

5.5. Ethical Considerations

While this research focuses on the use of VFR satellite imagery for constructive and socially beneficial applications, such as urban planning, public safety monitoring, post-disaster response, etc., we acknowledge the broader discussion of the dual-use potential of CD using VFR satellite imagery. Recent research (Ghamisi et al., 2025) highlights the need to consider the privacy risks associated with fine spatial resolution remote sensing data carefully. It is important to note that this research utilizes historical satellite imagery, not real-time data. Additionally, while this research can detect the presence of individuals, the individuals are represented as indistinct black or white dot-like shapes in VFR satellite imagery, making it impossible to

identify them. Nevertheless, we support future research to develop privacy protection frameworks to ensure the ethical and responsible use of AI-driven geospatial analysis.

6. Conclusion

This research investigated CD from VFR satellite imagery and introduced both a new dataset and a tailored CD framework. Specifically, we constructed CrowdSat, a large-scale CD dataset derived from multiple VFR satellite sensors with over 120k labeled individuals across heterogeneous scenes, together with a multi-temporal background-removal strategy to reduce point-level mislabeling. Building on this dataset, we proposed CrowdSat-Net with the innovative DCPAN and HFGDU modules to enhance small-object feature representation and recover high-frequency details during upsampling. The key findings of this research are summarized as follows.

1) VFR satellite imagery overcomes the limitations of small-scale (both spatially and temporally) ground and aerial imagery, offering a promising pathway for large-scale crowd analysis in various applications.
2) CrowdSat-Net outperforms five advanced CD methods (designed for ground or aerial imagery) based on CrowdSat, achieving the largest F1-score of 66.12%and Precision of 73.23%.
3) The DCPAN and HFGDO modules are effective in increasing CD accuracy, increasing the F1-score by 1.70% and Precision by 4.69%.
4) CrowdSat-Net performs reliably in moderate and relatively high crowd density scenarios, with F1-scores of 69.60% and 73.03%, respectively.
5) CrowdSat-Net demonstrates great cross-regional generalization in regions across the globe, with F1-scores of 75.00% and 73.70% in the Metropolitan Cathedral, Mexico and the National Mall, USA, respectively.

Overall, for large-scale CD, VFR satellite imagery offers an appropriate source, and CrowdSat-Net provides an effective solution. Future research will focus on constructing larger-scale CD datasets and developing more refined and generalized model architectures.

CRediT Authorship Contribution Statement

Tong Xiao: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing–original draft, Writing–review & editing. **Qunming Wang:** Conceptualization, Data curation, Funding acquisition, Supervision, Writing–original draft, Writing–review & editing. **Ping Lu:** Formal analysis, Validation, Writing–review & editing. **Tenghai Huang:** Data curation. **Xiaohua Tong:** Conceptualization, Formal analysis, Writing–review & editing. **Peter M. Atkinson:** Supervision, Conceptualization, Writing–review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant Nos. 42222108, 42221002 and 42171345. The authors are grateful to Google Earth Platform for its satellite data support.

References

Abousamra, S., Hoai, M., Samaras, D., Chen, C., 2021. Localization in the crowd with topological constraints. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 872–881.

Adam, K.D.B.J., et al., 2014. A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980.

Al-Nami, W.T., 2023. Ranking and analysis the strategies of crowd management to reduce the risks of crushes and stampedes in crowded environments and ensure the safety of passengers. *Neutrosophic Systems with Applications* 8, 61–78.

Bahmanyar, R., Vig, E., Reinartz, P., 2019. MRCNet: Crowd counting and density map estimation in aerial and ground imagery. *arXiv preprint* arXiv:1909.12743.

Bashir S.M.A., Wang Y., 2021. Small object detection in remote sensing images with residual feature aggregation-based super-resolution and object detector network. *Remote Sensing* 13(9), 1854.

von Borstel, M., Kandemir, M., Schmidt, P., Rao, M.K., Rajamani, K., Hamprecht, F.A., 2016. Gaussian process density counting from weak supervision. In: *European Conference on Computer Vision*. Springer, pp. 365–380.

Chen, I.H., Chen, W.T., Liu, Y.W., Yang, M.H., Kuo, S.Y., 2024. Improving point-based crowd counting and localization based on auxiliary point guidance. In: *European Conference on Computer Vision*. Springer, pp. 428–444.

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing* 12(10), 1662.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 764–773.

Delplanque, A., Foucher, S., Théau, J., Bussiere, E., Vermeulen, C., Lejeune, P., 2023. From crowd to herd counting: How to precisely detect and count African mammals using aerial imagery and deep learning? *ISPRS Journal of Photogrammetry and Remote Sensing* 197, 167–180.

Feliciani, C., Shimura, K., Nishinari, K., 2022. *Introduction to Crowd Management: Managing Crowds in the Digital Era: Theory and Practice*. Springer Nature.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154.

931 Gao, G., Liu, Q., Wang, Y., 2020a. Counting dense objects in remote sensing images. In: *ICASSP 2020-2020 IEEE International*
932 *Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4137–4141.
933 Gao, G., Liu, Q., Wang, Y., 2020b. Counting from sky: A large-scale dataset for remote sensing object counting and a benchmark
934 method. *IEEE Transactions on Geoscience and Remote Sensing* 59, 3642–3655.
935 Gao, J., Han, T., Wang, Q., Yuan, Y., Li, X., 2020c. Learning independent instance maps for crowd localization. *arXiv preprint*
936 arXiv:2012.04164.
937 Gazzawe, F., Albahar, M., 2024. Reducing traffic congestion in Makkah during Hajj through the use of AI technology. *Heliyon*
938 10(1).
939 Ghamisi, P., Yu, W., Marinoni, A., Gevaert, C.M., Persello, C., Selvakumaran, S., Girotto, M., Horton, B.P., Rufin, P., Hostert, P.,
940 Pacifici, F., Atkinson, P.M., 2025. Responsible Artificial Intelligence for Earth Observation: Achievable and realistic paths
941 to serve the collective good. *IEEE Geoscience and Remote Sensing Magazine*, 2–26.
942 Guo, Y., Jiang, H., Wu, C., Zhang, L., Du, B., 2022. Density Map-based vehicle counting in remote sensing images with limited
943 resolution. *ISPRS Journal of Photogrammetry and Remote Sensing* 189, 201–217.
944 Han, T., Bai, L., Liu, L., Ouyang, W., 2023. Steerer: Resolving scale variations for counting and localization via selective
945 inheritance learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21848–21859.
946 Hebei Provincial People's Government, 2011. Regulations of Hebei Province on aerial surveying and photogrammetry. URL:
947 https://www.hebei.gov.cn/columns/67c6d89b-b5b0-4492-98d6-997dac63a29b/202309/12/0808f9e0-84c9-4580-82e9-bfe
948 3b734c875.html.
949 Joiner, A., McFarlane, C., Rella, L., Uriarte-Ruiz, M., 2024. Problematising density: Covid-19, the crowd, and urban life. *Social &*
950 *Cultural Geography* 25, 181–198.
951 Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multi-source building extraction from an open aerial and satellite
952 imagery dataset. *IEEE Transactions on Geoscience and Remote Sensing* 57(1), 574-586.
953 Khan, M.A., Menouar, H., Hamila, R., 2023. Revisiting crowd counting: State-of-the-art, trends, and future perspectives. *Image and*
954 *Vision Computing* 129, 104597.
955 Kramer, O., 2013. K-nearest neighbors. In: *Dimensionality Reduction with Unsupervised Nearest Neighbors*. Springer, pp. 13–23.
956 Laradji, I.H., Rostamzadeh, N., Pinheiro, P.O., Vazquez, D., Schmidt, M., 2018. Where are the blobs: Counting by localization with
957 point supervision. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 547–562.
958 Lempitsky, V., Zisserman, A., 2010. Learning to count objects in images. *Advances in Neural Information Processing Systems* 23.
959 Li, X., Zhang, J., Yang, Y., Cheng, G., Yang, K., Tong, Y., Tao, D., 2024. SFNet: Faster and accurate semantic segmentation via
960 semantic flow. *International Journal of Computer Vision* 132, 466–489.

Lian, D., Li, J., Zheng, J., Luo, W., Gao, S., 2019. Density map regression guided detection network for RGB-D crowd counting and localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1821–1830.

Liang, D., Xu, W., Zhu, Y., Zhou, Y., 2022a. Focal inverse distance transform maps for crowd localization. *IEEE Transactions on Multimedia* 25, 6040–6052.

Liang, D., Xu, W., Bai, X., 2022b. An end-to-end transformer model for crowd localization, in: European Conference on Computer Vision. Springer, pp. 38–54.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988.

Liu, C., Lu, H., Cao, Z., Liu, T., 2023a. Point-query quadtree for crowd counting, localization, and more. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1676–1685.

Liu, C., Weng, X., Mu, Y., 2019. Recurrent attentive zooming for joint crowd counting and precise localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1217–1226.

Liu, W., Lu, H., Fu, H., Cao, Z., 2023b. Learning to upsample by learning to sample. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6027–6037.

Liu, Y., Sun, P., Wergeles, N., Shang, Y., 2021. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications* 172, 114602.

Li, Z., He, W., Cheng, M., Hu, J., Yang, G., Zhang, H., 2023. SinoLC-1: the first 1 m resolution national-scale land-cover map of China created with a deep learning framework and open-access data. *Earth System Science Data* 15(11), 4749–4780.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440.

Luo, Z., Marchi, L., Chen, F., Zhang, Y., Gaspari, J., 2025. Correlating urban spatial form and crowd spatiotemporal behavior: A case study of Lhasa, China. *Cities* 160, 105812.

Meynberg, O., Cui, S., Reinartz, P., 2016. Detection of high-density crowds in aerial images using texture classification. *Remote Sensing* 8, 470.

Mliki, H., Arous, O., Hammami, M., 2019. Abnormal crowd density estimation in aerial images. *Journal of Electronic Imaging* 28, 013047.

Mo, H., Zhang, X., Tan, J., Yang, C., Gu, Q., Hang, B., Ren, W., 2024. Countformer: Multi-view crowd counting transformer. In: *European Conference on Computer Vision*. Springer, pp. 20–40.

Naude, J., Joubert, D., 2019. The aerial elephant dataset: A new public benchmark for aerial object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 48-55.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: *European Conference on Computer Vision*. Springer, pp. 483–499.

Ni, W., Li, H., Wang, C., Chen, Y., Lin, D., Tao, S., Xi, X., Wang, Y., Hu, J., Li, X., 2024. Forest height extraction using GF-7 very high-resolution stereoscopic imagery and Google Earth multi-temporal historical imagery. *Journal of Remote Sensing* 4, 0158.

Pokhrel, S., Chhetri, R., 2021. A literature review on impact of covid-19 pandemic on teaching and learning. *Higher Education for the Future* 8, 133–141.

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

Rezaee, K., Rezakhani, S.M., Khosravi, M.R., Moghimi, M.K., 2024. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. *Personal and Ubiquitous Computing* 28, 135–151.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241.

Rumora, L., Gašparović, M., Miler, M., Medak, D., 2020. Quality assessment of fusing Sentinel-2 and WorldView-4 imagery on Sentinel-2 spectral band values: A case study of Zagreb, Croatia. *International Journal of Image and Data Fusion* 11, 77–96.

Sam, D.B., Peri, S.V., Sundararaman, M.N., Kamath, A., Babu, R.V., 2021. Locate, size, and count: Accurately resolving people in dense crowds via detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43(8), 2739–2751.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1874–1883.

Shi, Q., Liu, M., Marinoni, A., Liu, X., 2023. UGS-1m: fine-grained urban green space mapping of 31 major cities in China based on the deep learning framework. *Earth System Science Data* 15(2), 555–577.

Shu, W., Wan, J., Tan, K.C., Kwong, S., Chan, A.B., 2022. Crowd counting in the frequency domain. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19618–19627.

Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Wu, Y., 2021. Rethinking counting and localization in crowds: A purely point-based framework. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3365–3374.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703.

Tong, K., Wu, Y., 2022. Deep learning-based detection from the perspective of small or tiny objects: A survey. *Image and Vision Computing* 123, 104471.

Wan, J., Liu, Z., Chan, A.B., 2021. A generalized loss function for crowd counting and localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1974–1983.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual attention network for image classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164.

Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D., 2019. CARAFE: Content-aware reassembly of features. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3007–3016.

Wang, Q., Gao, J., Lin, W., Li, X., 2020a. NWPU-Crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 2141–2149.

Wang, Q., Shi, W., Atkinson, P.M., Zhao, Y., 2015. Downscaling MODIS images with area-to-point regression kriging. *Remote Sensing of Environment* 166, 191–204.

Wang, Q., Shi, W., Li, Z., Atkinson, P.M., 2016. Fusion of Sentinel-2 images. *Remote Sensing of Environment* 187, 241–252.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020b. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11534–11542.

Wang, Y., Hou, X., Chau, L.P., 2021. Dense point prediction: A simple baseline for crowd counting and localization. In: *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, pp. 1–6.

Wei, W., Cheng, Y., He, J., Zhu, X., 2024. A review of small object detection based on deep learning. *Neural Computing and Applications* 36, 6283–6303.

Weng, W., Wang, J., Shen, L., Song, Y., 2023. Review of analyses on crowd-gathering risk and its evaluation methods. *Journal of Safety Science and Resilience* 4, 93–107.

Wu, Z., Zhang, C., Gu, X., Duporge, I., Hughey, L. F., Stabach, J. A., Skidmore, A. K., Hopcraft, J. G. C., Lee, S. J., Atkinson, P. M., McCauley, D. J., Lamprey, R. H., Ngene, S. M., Wang, T., Wang, J., 2023. Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscapes. *Nature Communications* 14, 3072.

Xu, C., Liang, D., Xu, Y., Bai, S., Zhan, W., Bai, X., Tomizuka, M., 2022. AutoScale: Learning to scale for crowd counting. *International Journal of Computer Vision* 130, 405–434.

Yi, J., Pang, Y., Zhou, W., Zhao, M., Zheng, F., 2023a. A perspective-embedded scale-selection network for crowd counting in public transportation. *IEEE Transactions on Intelligent Transportation Systems* 25, 3420–3432.

Yi, J., Shen, Z., Chen, F., Zhao, Y., Xiao, S., Zhou, W., 2023b. A lightweight multiscale feature fusion network for remote sensing object counting. *IEEE Transactions on Geoscience and Remote Sensing* 61, 1–13.

Yuan, X., Chakravarty, A., Gu, L., Wei, Z., Lichtenberg, E., Chen, T., 2025. An empirical study of methods for small object detection from satellite imagery. *arXiv preprint* arXiv:2502.03674.

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032.

Zhang, Q., Lin, W., Chan, A.B., 2021. Cross-view cross-scene multi-view crowd counting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 557–567.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239.

Zhou, X., Koltun, V., Krähenbühl, P., 2020. Tracking objects as points. In: *European Conference on Computer Vision*. Springer, pp. 474–490.

Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as points. *arXiv preprint* arXiv:1904.07850.

Appendix

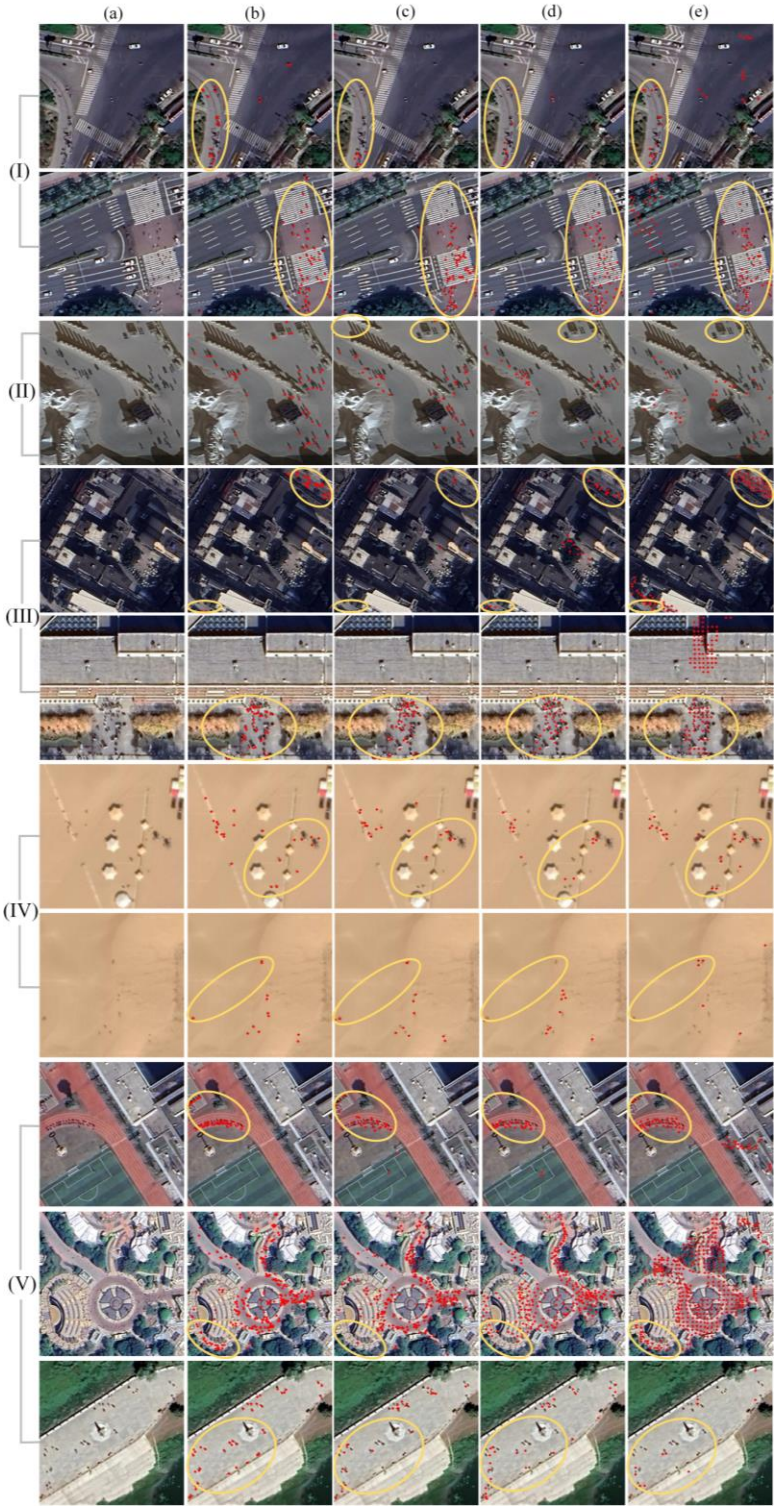

Fig. A.1 Visual comparisons between representative CD methods (P2PNet, CLTR and PET) for the CrowdSat dataset. (a) Original image. (b) Reference. (c) P2PNet. (d) CLTR. (e) PET. (I) Traffic junctions. (II) Snowfields. (III) Dense urban regions. (IV) Desert regions. (V) Other common impervious regions. Some of the images in panels (I)-(V) were obtained from the Google Earth platform (© Google Earth 2025).

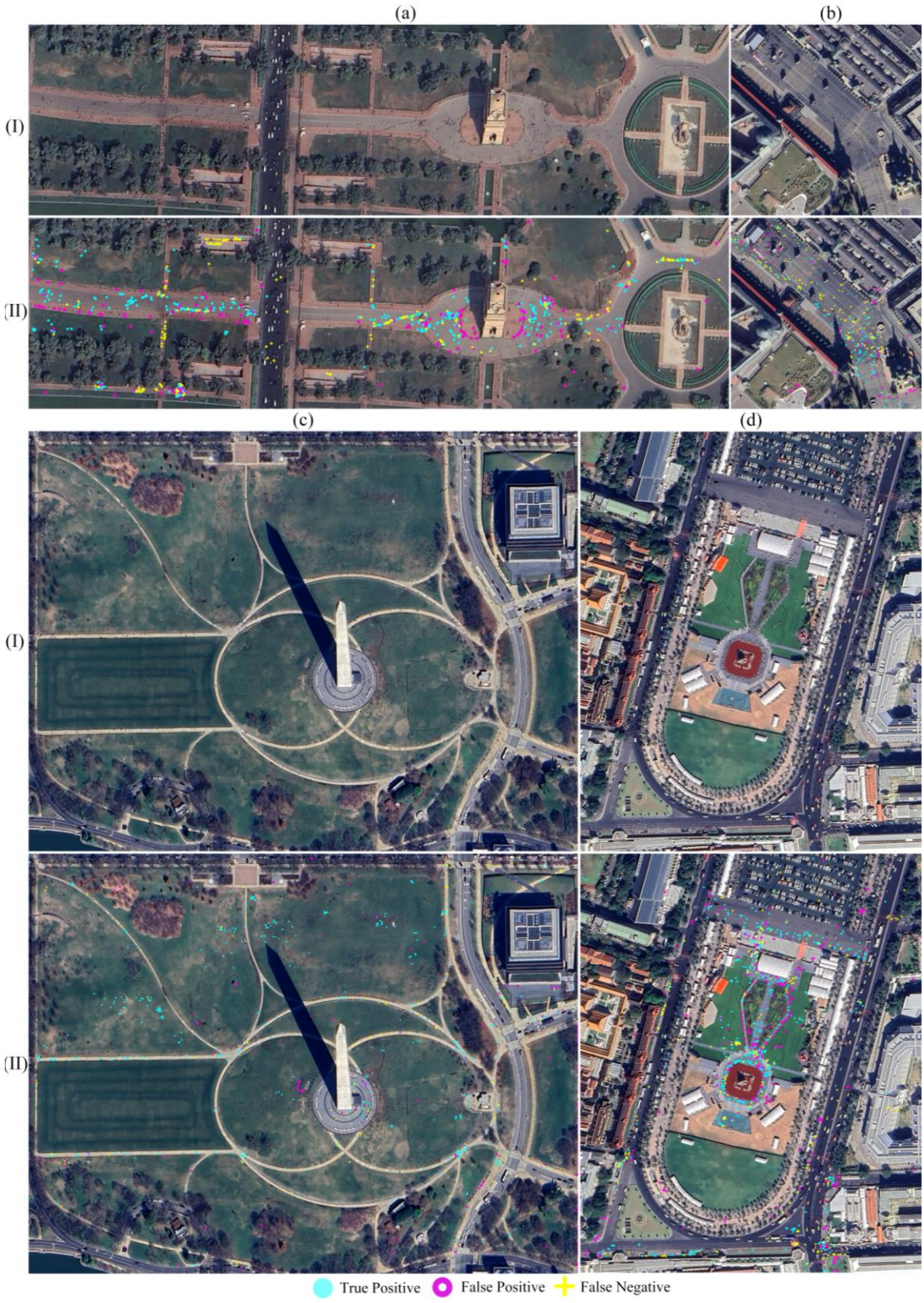

Fig. A.2. The visual localization performance of CrowdSat-Net in other unseen foreign regions. (a) India Gate. (b) Red Square. (c) National Mall. (d) Phra Nakhon. (I) Original image. (II) Evaluation. The images in (a)-(d) were obtained from the Google Earth platform (© Google Earth 2025).