Displays 94 (2026) 103493

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

A multi-task visual framework: Geometry-guided UAV crowd counting and localization for media practice☆

Ziqing He [a,c,1], Longfei Wang [b,1], Huiying Xu [b,*], Xinzhong Zhu [b,c], Wouladje Cabrel [b], Golden Tendekai Mumanikidzwa [b]

[a] School of Media Practice, The University of Sydney, Sydney, NSW 2050, Australia
[b] School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China
[c] Hangzhou Institute of Artificial Intelligence, Zhejiang Normal University, Hangzhou, 321004, China

ARTICLE INFO

Dataset link: https://www.kaggle.com/datasets/tthien/shanghaitech, http://crcv.ucf.edu/data/UCF-CC-50/, https://www.crcv.ucf.edu/data/ucf-qnrf/, https://www.crowdbenchmark.com/nwpucrowd.html, http://www.crowd-counting.com/

Keywords:
UAV
Multi-task framework
GSD
PAAP
Uncertainty-calibrated

ABSTRACT

Accurate crowd counting and localization from UAV aerial imagery remain challenging due to severe perspective distortion and extreme scale variation, hindering deployment reliability in data-driven journalism and media verification workflows. This paper introduces a geometry-guided multi-task framework that explicitly integrates flight metadata, including ground sampling distance (GSD) maps, camera intrinsics, and altitude parameters, to address these fundamental challenges. Our Perspective-Aware Attention Pyramid (PAAP) encodes geometric priors into adaptive feature hierarchies, jointly optimizing point-level detection, density estimation, and spatial clustering via uncertainty-weighted multi-task learning. Comprehensive evaluations across six benchmarks (ShanghaiTech-A/B, UCF-CC-50, UCF-QNRF, NWPU-Crowd, JHU-Crowd++) demonstrate consistent superiority: 2.2-3.6% MAE reductions over leading point-based methods and 20%–40% improvements over density regression baselines. For spatial localization, PAAP achieves a 0.750 average F_1-score, outperforming state-of-the-art approaches by 1.5–2.0% under strict pixel-level thresholds ($\sigma \in \{1, 2, 3\}$ pixels). Evaluations across six public benchmarks and journalism-oriented verification settings indicate the practical applicability of the proposed framework for media practice.

1. Introduction

Accurate crowd estimation and localization play a pivotal role not only in urban governance and public safety [1,2], but also increasingly in data-driven journalism, where verifiable narratives and spatially resolved quantitative tracking are in high demand [3,4]. The rise of Unmanned Aerial Vehicles (UAVs) has dramatically expanded the observational power of media professionals, allowing the real-time documentation and quantitative analysis of large-scale public events—from political rallies and street protests to disaster response and cultural festivals—from a previously unachievable aerial perspective [5,6]. However, harnessing the full value of aerial imagery for credible visual storytelling and fact-checking remains hampered by two fundamental scientific challenges: severe perspective distortion and extreme scale variation intrinsic to aerial imaging [7,8]. Conventional crowd analysis algorithms, designed mainly for ground-level or near-horizontal views [9,10], are not suitable for top-down nonlinear geometries prevalent in UAV-based contexts, resulting in substantial performance degradation and compromised spatial precision [11,12].

Point-level crowd counting and localization—the simultaneous prediction of both total count and exact spatial coordinates of each individual—has emerged as a critical requirement for modern media practice [13,14]. Unlike aggregate counting, point-level outputs enable region-specific quantitative tracking (e.g., "How many attendees entered Zone A vs. Zone B?"), temporal trajectory analysis (e.g., crowd flow dynamics for investigative reporting), and geographic attribution (e.g., verifying claims about protest density at specific landmarks) [15,16]. These capabilities are indispensable for journalistic verification, where editors must cross-reference visual evidence with eyewitness accounts, official statements, and third-party data sources [17]. Moreover, in the context of data governance and public transparency, precise localization supports audit trails, enabling retrospective validation of event coverage and mitigating the spread of misinformation [18,19]

☆ This paper was recommended for publication by Prof Guangtao Zhai.
* Corresponding author.
E-mail address: xhy@zjnu.edu.cn (H. Xu).
[1] Ziqing He and Longfei Wang are co-first authors.

https://doi.org/10.1016/j.displa.2026.103493
Received 26 January 2026; Received in revised form 30 March 2026; Accepted 21 April 2026
Available online 30 April 2026

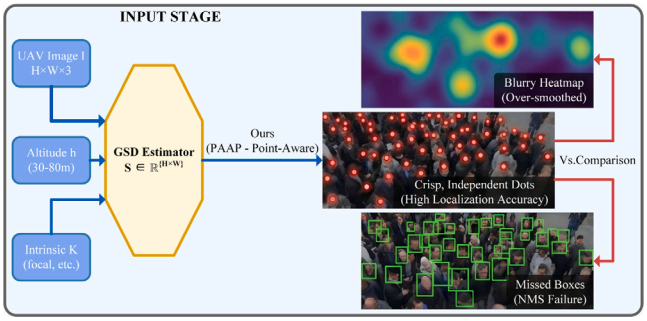

Fig. 1. Geometry-Guided Crowd Localization in UAV Imagery. Addressing the challenges of severe perspective distortion and extreme scale variation, our proposed framework integrates flight metadata and Ground Sampling Distance (GSD) priors to explicitly model scene geometry. Unlike traditional density regression methods that produce ambiguous, blurred heatmaps (as shown in the comparison), our Perspective-Aware Attention Pyramid (PAAP) enables precise, point-level localization. This capability is critical for generating verifiable quantitative data in computational journalism and media verification workflows.

Despite these imperatives, traditional paradigms exhibit systematic limitations when faced with UAV-based media scenarios. For the past ten years, density map regression methods [20–22] have been the most popular. They use Gaussian kernels to combine annotated point labels and create continuous density fields. These fields are then supervised using pixel-wise regression losses, such as Euclidean or structural similarity metrics. Although this smoothing operation alleviates annotation noise and handles occlusions, it fundamentally sacrifices spatial resolution: the resulting density maps blur individual positions into probabilistic distributions, rendering precise localization infeasible [23,24]. This loss of granularity is particularly detrimental in media contexts, where the ability to pinpoint specific clusters – such as identifying blockade formations or evacuee concentrations – is paramount for narrative accuracy and editorial decision making [25].

Recent point-based detection frameworks [26–28] leverage one-to-one or set-matching supervision – such as the Hungarian assignment in DETR-style models [29] – to directly regress discrete point sets, unifying object counting and localization within an end-to-end trainable pipeline. However, their core architectural assumptions—including uniform feature scales, perspective-invariant proposals, and dependence on natural image pre-trained backbones commonly fail in UAV-based imaging scenarios [30,31]. This failure stems primarily from the severe geometric distortions induced by the UAV perspective: people near the center of the image often occupy 15–20 pixels, while those in the periphery shrink to only 3–5 pixels due to radial lens distortion and altitude-induced foreshortening [32,33]. Without explicit geometric modeling tailored to aerial camera configurations—such as incorporating intrinsic/extrinsic parameters, flight altitude, or gimbal orientation—point-based detectors suffer from scale-sensitive recall drop and spatially biased false alarms. These limitations are clearly reflected in the 20%–40% mAP degradation observed when models trained on ground-level surveillance datasets (e.g., ShanghaiTech [9]) are evaluated on UAV-specific benchmarks such as VisDrone [7].

This emerging gap between practical media needs – real-time, verifiable, spatially resolved crowd intelligence – and the capabilities of mainstream computational models underscores an urgent call for principled interdisciplinary solutions. In the context of modern media practice, the ability to transform raw UAV footage into reliable, interpretable, and verifiable quantitative assets is indispensable—not only for data journalism, but also for enhancing transparency in public discourse, supporting evidence-based policymaking, and combating visual disinformation [34,35]. Critically, UAV crowd analysis under aerial perspectives constitutes a fundamentally distinct algorithmic paradigm: it demands geometric awareness (to compensate for viewpoint distortions), multi-scale reasoning (to handle 10×–20× intra-image scale variance), and tight coupling between low-level features and high-level semantic context (to distinguish human heads from visually similar objects such as umbrellas or signage) [36,37].

To address these challenges, we introduce a novel multi-task visual framework that leverages geometry-guided learning to achieve robust crowd counting as well as precise point-level localization from aerial perspectives. Our approach utilizes a Perspective-Aware Attention Pyramid (PAAP) to incorporate explicit geometric priors. This is accomplished by directly integrating camera intrinsic matrices, flight altitude metadata, and dynamically estimated scale maps into the deep learning process using differentiable perspective transformations and scale-adaptive attention mechanisms [38,39]. As a result, the model gains inherent geometric awareness, allowing it to effectively compensate for radial distortion, altitude-induced scale gradients, and oblique viewing angles. These capabilities exceed those of purely data-driven strategies, which rely solely on convolutional and self-attention techniques applied to RGB pixels [40,41] (see Fig. 1).

In summary, this study advances UAV-based crowd analysis along two synergistic dimensions: technical innovation and methodological integration. Our work bridges AI research with media practice, enabling new capabilities in data journalism, transparent governance, and computational social science [42,43]. To situate these contributions more clearly, the following section reviews the literature from the three perspectives that most directly motivate our framework: the limitations of existing crowd-analysis paradigms in UAV settings, geometry-aware architectural design, and reliability mechanisms required for media-oriented deployment.

2. Related work

The evolution of crowd analysis technologies reflects a dual trajectory: methodological refinement from aggregate estimation to point-level localization and contextual expansion from controlled surveillance to unstructured aerial scenarios. Against this background, our review is organized to mirror the central design logic of the present study. We synthesize prior work along three critical axes that directly inform our contributions: the paradigm shift necessitated by media verification demands, architectural solutions to UAV-specific geometric challenges, and reliability mechanisms bridging algorithmic outputs with editorial workflows.

Paradigm evolution and the UAV domain gap. Early approaches to crowd counting built upon detection frameworks [44,45], which employed sliding windows or region proposals to localize individuals. While effective in delivering precise bounding boxes for sparse crowds, these methods proved inadequate under dense occlusion scenarios [1]. The advent of density map regression [20] marked a significant shift: by convolving point annotations with Gaussian kernels, models such as MCNN [9] and CSRNet [10] achieved robustness to extreme crowd densities through pixel-wise supervision. However, this representation inherently compromises spatial precision, whereby smoothed density fields obscure individual positions, resulting in probabilistic distributions [23]. Consequently, this precludes the capacity for fine-grained regional analysis, a capability that is imperative for journalistic verification [17]. Although post-processing techniques such as density peak detection [46] or watershed segmentation [47] have been proposed to recover point coordinates, these heuristics remain fragile under heavy occlusion and often require manual threshold tuning.

More recently, point-based frameworks [13,27] have sought to unify counting and localization by directly predicting discrete coordinate sets via set-matching losses [29], thereby circumventing the spatial ambiguity inherent in density maps. Nevertheless, their application in UAV contexts reveals critical shortcomings. Empirical studies [7,30] report performance drops of 20%–40% when models trained on ground-level

datasets [9] are applied to aerial benchmarks. This degradation stems largely from non-stationary geometric transformations: individuals near the image center may occupy 15–20 pixels, while those in peripheral regions shrink to merely 3–5 pixels due to radial distortion and altitude-induced foreshortening [32]. Current approaches often treat such distortions as latent noise to be learned from data [33], rather than as structured geometric prior knowledge that can be explicitly modeled using camera parameters and flight metadata. This disconnect is further compounded by benchmark design: although datasets such as VisDrone [7] and UA-DETRAC [11] offer rich annotations, they typically omit flight telemetry—such as altitude, GPS, and gimbal angles readily accessible in operational UAV deployments. This omission perpetuates a misalignment between academic modeling assumptions and real-world journalistic requirements.

Geometry-aware multi-scale architectures. Addressing UAV-specific challenges requires architectures that integrate explicit geometric priors while handling extreme scale variations. Classical photogrammetry [5] demonstrates that UAV imagery conforms to pinhole camera models defined by intrinsic and extrinsic parameters; however, directly applying camera calibration to crowd counting remains largely unexplored. Spatial Transformer Networks (STN) [48] support differentiable geometric warping and have been successfully used in text recognition [49] and 3D reconstruction [50], yet they typically assume flat ground planes—an assumption often violated in outdoor settings with uneven terrain. As an alternative, we derive the ground sampling distance (GSD) from flight altitude and focal length, offering an interpretable scale prior that exceeds the interpretability of purely data-driven mechanisms such as switchable convolutions [31] or deformable kernels [51].

To embed geometric reasoning into modern architectures, multi-scale feature hierarchies – introduced by FPN [52] and extended by transformers [40,41] – provide a foundational framework. Although CSRNet [10] adapts FPN for crowd counting, its fixed structure struggles when scale distributions change across different flight phases. Deformable attention [32] reduces computational complexity to $O(n^3)$ by focusing on sparse, content-aware regions. Our key contribution lies in conditioning its sampling offsets on GSD-derived scale maps, which ensures consistent attention to human-scale features despite perspective-induced distortions. In addition to structural design, multi-task learning [48] allows synergistic optimization of counting, localization, and auxiliary tasks such as behavioral clustering. However, naive multi-task setups often experience interference between objectives. To address this, we introduce uncertainty-modulated loss weighting, which dynamically adjusts task priorities during training based on predictive confidence.

It is worth noting that, beyond the geometry-guided architecture adopted in this study, other strong computer vision models may also be leveraged for UAV crowd analysis. DeepLab, a representative semantic segmentation framework, employs atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale contextual information while preserving spatial resolution, making it potentially valuable for dense spatial localization tasks [53]. EfficientNet, in contrast, uses a compound scaling strategy to jointly balance network depth, width, and input resolution, thereby providing an efficient trade-off between accuracy and computational cost [54]. In the present work, we adopt VGG16 as the backbone because of its stable compatibility with the proposed PAAP design and its straightforward integration with GSD-conditioned feature modulation. Nevertheless, these architectures represent meaningful alternatives that may be incorporated into future geometry-guided crowd analysis frameworks.

Uncertainty quantification and media-driven design. Media applications demand predictive reliability beyond point estimates: contested crowd counts at political rallies [34,35] underscore the need for confidence intervals enabling editorial judgment [17]. While Bayesian neural networks [55] model weight distributions to propagate uncertainty, practical approximations like MC Dropout [42] and deep ensembles offer scalable alternatives. We extend these via spatial consistency constraints – uncertainty maps must exhibit smooth gradients except at crowd boundaries – and calibrate outputs [56] to ensure predicted intervals match empirical errors, addressing interpretability gaps documented in computational journalism [3,15]. Beyond algorithmic reliability, ethical deployment requires privacy safeguards: high-resolution UAV imagery enables facial recognition at political protests, raising surveillance concerns [57,58]. Generic anonymization [59] offers crude protection, but our framework tailors obfuscation intensity to contextual sensitivity (heavier blurring at rallies vs. festivals) and supports selective anonymization (preserving landmarks for geographic verification while obscuring identities), aligning with emerging algorithmic transparency guidelines [60,61]. Furthermore, we embed provenance metadata (GPS, timestamps, camera parameters) into outputs [18,19], enabling forensic validation against official statements—a capability absent in conventional counting systems but essential for combating visual misinformation.

Contributions. This paper has three main contributions.

First, geometric naivety—modeling UAV-induced distortions as latent noise rather than structured prior knowledge.

Second, task isolation—optimizing counting and localization separately without uncertainty-aware coordination.

Third, application disconnect—neglecting interpretability, privacy preservation, and editorial integration. Our approach addresses these gaps by integrating explicit geometric modeling and multi-task synergy.

3. Materials and methods

In this section, we delineate the technical framework underlying our geometry-aware multi-task approach. The overall architecture of our proposed framework is illustrated in Fig. 2. We commence by formalizing the UAV-based crowd analysis problem under explicit geometric constraints (Section 3.1), then detail the benchmark datasets (Section 3.2). On this basis, we progressively introduce the main methodological components of the framework, moving from geometry-guided feature extraction to task-level optimization and finally to reliability-oriented uncertainty modeling. Subsequently, we present the core architectural innovations: the Perspective-Aware Attention Pyramid (PAAP) for geometry guided feature extraction (Section 3.3), the multi-task learning framework unifying counting and localization (Section 3.4), and the uncertainty quantification mechanism ensuring predictive reliability (Section 3.5).

3.1. Problem formulation

Task Definition. Given an aerial image $I(I \in \mathbb{R}^{H\times W\times 3})$ captured by a UAV at altitude h with a camera intrinsic matrix $K(K \in \mathbb{R}^{3\times 3})$, our objective is to predict a set of point coordinates $P(P = \{p_i = (x_i, y_i)\}_{i=1}^{N})$ representing individual locations, alongside the total count N. Unlike density map regression [20], which outputs continuous spatial distributions $D(D \in \mathbb{R}^{H\times W})$, our point-based formulation [13] directly optimizes discrete predictions via set-to-set matching, thereby preserving spatial precision critical for region-specific attribution in journalistic verification [17].

For clarity, all image coordinates are defined in the pixel coordinate system unless otherwise stated, with the origin located at the upper-left corner of the image. During optimization, point coordinates are normalized when necessary for stable learning, whereas evaluation is conducted in the original image coordinate space. The camera intrinsic matrix K contains the focal lengths and principal point parameters, and serves as the basis for geometric scale estimation.

Geometric Constraints. UAV imagery introduces non-stationary perspective transformations parameterized by flight metadata. We model the ground sampling distance (GSD) at pixel location (x, y) as:

$$\mathrm{GSD}(x, y) = \frac{h \times s}{f \times \cos(\theta(x, y))} \tag{1}$$

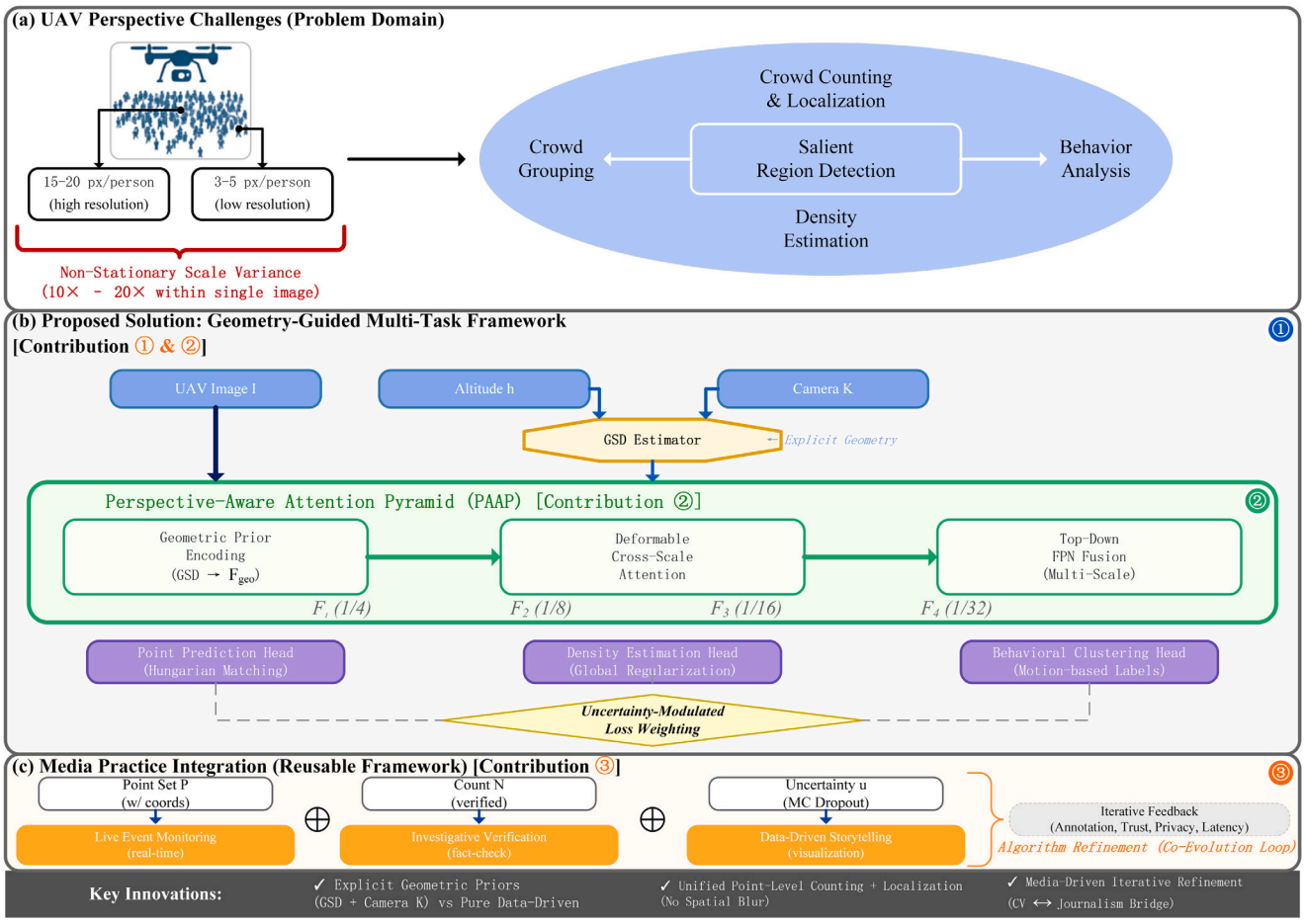

Fig. 2. The design of the suggested geometry-guided PAAP framework for analyzing crowds of UAVs. A GSD estimator combines UAV image, altitude, and camera intrinsics to produce a scale map that modulates multi-scale backbone features via a perspective-aware attention pyramid. In the actual implementation of this study, these features are extracted by a VGG16 backbone [62]. Multi-task heads for point localization, density regression, and clustering are trained with uncertainty-modulated loss weighting, and MC-dropout-based inference outputs point sets, verified counts, and calibrated uncertainty.

where s denotes the sensor pixel size, f is the focal length, and $\theta(x, y)$ represents the off-axis angle computed from K and the radial distortion coefficients. This spatially varying scale map $S(S \in \mathbb{R}^{H\times W})$ serves as an explicit geometric prior, guiding adaptive feature extraction in the PAAP module (Section 3.3).

In Eq. (1), h is measured along the optical axis of the UAV camera, s denotes the physical size of one sensor pixel, and f is the effective focal length of the camera. The angle $\theta(x, y)$ quantifies the deviation between the viewing ray passing through pixel (x, y) and the optical axis; therefore, larger off-axis angles correspond to larger local scale distortion. The resulting map S assigns each pixel an estimated ground-projected spatial resolution, and is used as a dense geometric descriptor rather than as a direct supervision target.

Multi-Task Objectives. Beyond point prediction, we jointly optimize auxiliary tasks to enhance feature representations: ① density estimation $\hat{D}$ as a regularizer for global count constraints and ② behavioral clustering $C = \{c_i\}_{i=1}^{N}$ assigning motion-based labels (static vs. moving), motivated by media needs to distinguish protesters from pedestrians [15]. The overall objective comprises:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{point}} + \lambda_{\text{density}}\mathcal{L}_{\text{density}} + \lambda_{\text{cluster}}\mathcal{L}_{\text{cluster}} \tag{2}$$

where task-specific weights $\{\lambda\}$ are dynamically modulated via uncertainty estimates.

Specifically, $\mathcal{L}_{\text{point}}$ supervises point classification and coordinate regression, $\mathcal{L}_{\text{density}}$ constrains the predicted density field to preserve global spatial consistency, and $\mathcal{L}_{\text{cluster}}$ regularizes high-level semantic grouping. The coefficients λ_{density} and λ_{cluster} control the relative contributions of the auxiliary tasks. In the final formulation adopted in Section 3.4, these coefficients are not manually fixed, but are adaptively determined through learnable task-uncertainty parameters, which improves optimization stability and reduces the need for manual hyper-parameter tuning.

3.2. Datasets

Benchmark Datasets. We validate our framework across six challenging crowd-counting datasets, spanning diverse density distributions, spatial resolutions, and scene complexities. Table 1 summarizes their key statistics.

ShanghaiTech Part A (SHT_A) [9] comprises 482 internet-sourced images featuring highly congested scenarios (mean density: 501 individuals/image), serving as a primary benchmark for dense crowd analysis. Its diverse scene compositions—spanning subway stations, plazas, and sporting events challenge algorithms to handle extreme occlusions and scale variance.

ShanghaiTech Part B (SHT_B) [9] captures real-world street scenes with moderate densities (mean: 123 individuals/image), providing ground-truth annotations for 716 images. Its emphasis on outdoor pedestrian flows complements SHT_A's indoor-centric distribution, enabling comprehensive cross-scene evaluation.

UCF-CC-50 [37] serves as a challenging benchmark for extreme-density crowd counting. It comprises 50 images with a staggering total of 63,974 annotations, averaging 1279 individuals per image.

Table 1
Statistical overview of benchmark datasets.

Dataset	Images	Train/Val/Test	Avg.Resolution	Total Count	Min	Max	Mean Count
ShanghaiTech Part A [9]	482	300/–/182	589 × 868	241,677	33	3,139	501
ShanghaiTech Part B [9]	716	400/–/316	768 × 1024	88,488	9	578	123
UCF-CC-50 [37]	50	–	2101 × 2888	63,974	94	4,633	1,279
UCF-QNRF [63]	1,535	1201/–/334	2013 × 2902	1,251,642	49	12,865	815
NWPU-Crowd [2]	5,109	3109/500/1500	2311 × 3383	2,133,238	0	20,033	418
JHU-Crowd++ [64]	4,372	2272/500/1600	1430 × 910	1,515,005	0	25,791	346

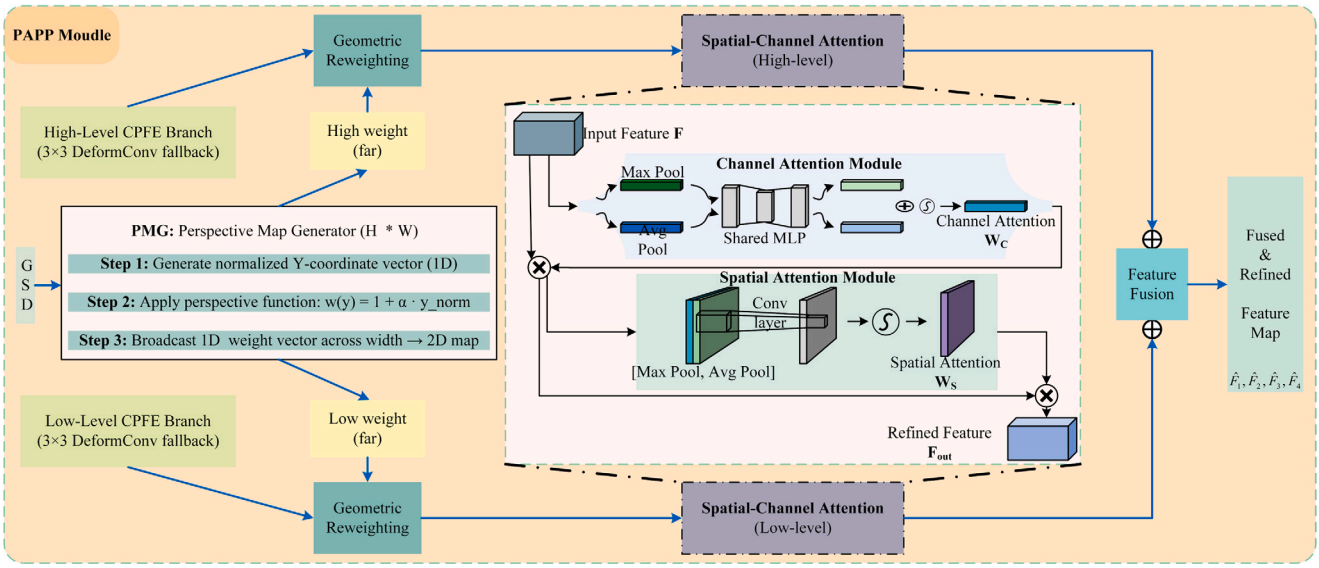

Fig. 3. PAPP framework module. A perspective map generator uses GSD-based vertical coordinates to construct a 2D weight map that geometrically reweights high- and low-level CPFE features with complementary far-near emphasis; the reweighted features are then refined by shared spatial-channel attention and fused into enhanced multi-scale feature maps.

Despite its limited sample size, this exceptionally high crowd concentration makes the dataset an indispensable resource for stress-testing algorithmic robustness under saturation conditions.

UCF-QNRF [63] scales to 1535 high-resolution images (mean resolution: 2013×2902 pixels) annotated with over 1.25 million individuals. Its naturalistic scene diversity, encompassing marathons, protests, and religious gatherings, closely mirrors the variability encountered in journalistic UAV deployments.

NWPU-Crowd [2] aggregates 5109 images sourced globally, reflecting cross-cultural crowd behaviors and geographic diversity. With annotations spanning 0 to 20,033 individuals per image, it assesses generalization across density regimes and cultural contexts, aligning with media applications requiring worldwide deployment.

JHU-Crowd++ [64] introduces 4372 images emphasizing complex occlusion patterns and adverse weather conditions (fog, rain), totaling 1.5 million annotations. Its inclusion of challenging environmental factors validates algorithmic resilience under real-world degradations prevalent in UAV imagery.

3.3. Perspective-Aware Attention Pyramid (PAAP)

The PAAP module constitutes the architectural cornerstone of our framework, explicitly integrating geometric priors into multi-scale feature extraction. Fig. 3 provides a detailed schematic of the PAAP module, illustrating the flow of visual and geometric information. Unlike conventional FPN [52] or transformer encoders [40] that process RGB pixels agnostically, PAAP conditions feature hierarchies on spatially varying scale maps derived from flight metadata.

Geometric Prior Encoding. Given the camera intrinsic matrix K and altitude h, we first compute the GSD map $S(S \in \mathbb{R}^{H\times W})$ via the formulation in Section 3.1. To encode this geometric knowledge into learnable representations, we employ a lightweight convolutional encoder E_{geo} that maps S to a feature volume $F_{\text{geo}}(F_{\text{geo}} \in \mathbb{R}^{H/4\times W/4\times C})$:

$$F_{\text{geo}} = E_{\text{geo}}(S; K, h) \tag{3}$$

Here, $E_{\text{geo}}(\cdot)$ denotes a learnable geometric encoder, and C is the channel dimension of the geometry feature embedding. We set the spatial resolution of F_{geo} to $\frac{1}{4}$ of the input image so that it can be aligned with the highest-resolution feature level in the backbone while maintaining affordable computational cost. The purpose of F_{geo} is to encode spatially varying scale cues into a compact feature representation that can modulate visual features at multiple semantic levels.

Multi-Scale Feature Hierarchy. In the actual implementation used throughout this study, we adopt a VGG16 backbone [62] pretrained on ImageNet, which extracts hierarchical features at four scales corresponding to spatial resolutions $\{1/4, 1/8, 1/16, 1/32\}$. For notational consistency with the general PAAP formulation, these backbone features are denoted as $\{F_1, F_2, F_3, F_4\}$. Each feature level F_l is modulated by scale-specific geometric embeddings via adaptive instance normalization [65]:

$$\tilde{F}_l = \gamma_l(F_{\text{geo}}) \odot \text{normalize}(F_l) + \beta_l(F_{\text{geo}}) \tag{4}$$

where γ_l and β_l are learnable affine transformations conditioned on F_{geo}. This operation recalibrates feature statistics to align with local GSD distributions, mitigating scale-induced feature misalignment.

In Eq. (4), l indexes the pyramid level, $\text{normalize}(\cdot)$ denotes feature normalization, and $\odot$ represents element-wise multiplication. The functions $\gamma_l(\cdot)$ and $\beta_l(\cdot)$ produce scale-adaptive modulation parameters for the lth feature map, respectively controlling multiplicative rescaling and additive bias correction. This operation allows the network to align feature statistics with local geometric scale, thereby alleviating the feature inconsistency caused by perspective distortion.

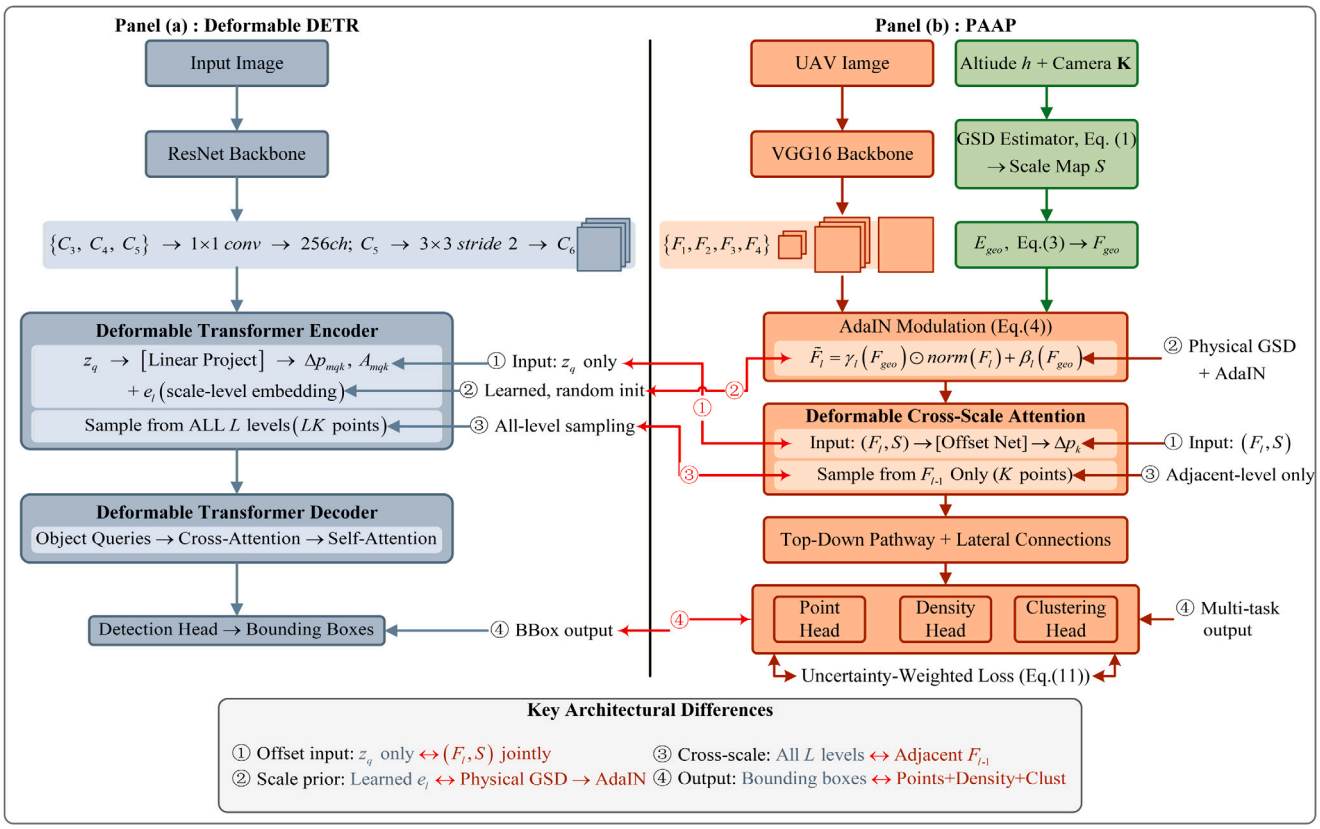

Fig. 4. Side-by-side architectural comparison between Deformable DETR [32] and the proposed PAAP module. (a) Deformable DETR predicts sampling offsets solely from the query feature z_q, employs randomly initialized additive scale-level embeddings e_l, samples across all L feature levels simultaneously, and outputs bounding boxes. (b) PAAP conditions offsets jointly on the visual feature F_l and the physically derived GSD scale map S (Eq. (5)), modulates backbone features through adaptive instance normalization with geometric embeddings F_{geo} (Eq. (4)), restricts cross-scale sampling to the adjacent higher-resolution level F_{l-1} (Eq. (6)), and produces multi-task outputs including point-level localization, density estimation, and behavioral clustering. Numbered callouts ①–④ correspond to the four principal distinctions detailed in Table 2.

Deformable Cross-Scale Attention. To aggregate information across scales while respecting geometric constraints, we extend deformable attention [32] with GSD-conditioned sampling offsets. For each query location q at the feature F_l, we predict K sampling points $\{p_k\}_{k=1}^{K}$ from higher-resolution levels:

$$p_k = q + \Delta p_k(F_l, S) \tag{5}$$

where offset predictions Δp_k are jointly determined by visual features F_l and scale map S, ensuring that attention focuses on scale-appropriate regions. The aggregated feature is computed as:

$$\hat{F}_l(q) = \sum_{k=1}^{K} w_k \cdot F_{l-1}(p_k) \tag{6}$$

The attention weights w_k are normalized using softmax.

In Eqs. (5)–(6), q denotes the reference spatial location on the current feature level, K is the number of sampled support points for each query, Δp_k is the learned offset of the kth sampling position, and w_k is the corresponding normalized attention weight. The notation $F_{l-1}(p_k)$ represents bilinearly interpolated feature sampling at location p_k on the adjacent higher-resolution feature map. By conditioning Δp_k on both appearance and geometry, the model can adapt its receptive field according to local crowd scale, rather than relying on fixed sampling patterns.

Architectural Distinction from Deformable DETR. Although our deformable cross-scale attention inherits the sparse sampling philosophy of Deformable DETR [32], the PAAP module differs from it in four fundamental respects (see Fig. 4 and Table 2 for a side-by-side comparison). *First*, Deformable DETR predicts sampling offsets Δp_{mqk} solely from the query feature z_q via linear projection, whereas PAAP conditions offsets jointly on the visual feature F_l and the physically derived GSD scale map S (Eq. (5)), enabling the sampling grid to adapt to the 10×–20× intra-frame scale variation characteristic of UAV imagery. *Second*, Deformable DETR relies on randomly initialized, additive scale-level embeddings e_l that carry no physical semantics; PAAP instead encodes the GSD map through a dedicated encoder E_{geo} (Eq. (3)) and injects it into every pyramid level via adaptive instance normalization (Eq. (4)), providing interpretable, geometry-grounded feature modulation. *Third*, while Deformable DETR aggregates LK sampling points from *all* L feature levels within a single attention operation, PAAP restricts each query at level F_l to sample exclusively from the adjacent higher-resolution level F_{l-1} (Eq. (6)), complemented by a top-down pathway with geometry-aware lateral connections, thereby preserving fine-grained spatial cues critical for point-level localization. *Fourth*, Deformable DETR outputs bounding boxes through iterative decoder refinement, whereas PAAP feeds the aggregated features into three parallel task-specific heads – point prediction, density estimation, and behavioral clustering – supervised by the uncertainty-weighted multi-task loss (Eq. (11)).

Top-Down Pathway. Following FPN [52], we propagate semantically strong features from coarse to fine scales via lateral connections, enhanced with geometry-aware modulation at each fusion step. Specifically, at each level l, the upsampled coarser feature is first recalibrated through geometry-aware modulation (Eq. (4)), yielding an intermediate modulated feature $\tilde{F}_l$, which is then aggregated with the corresponding lateral feature via geometry-guided cross-scale fusion (Eq. (6)) to produce the final enhanced output $\hat{F}_l$.

Table 2
Component-level architectural comparison between deformable DETR [32] and the proposed PAAP module.

Aspect	Deformable DETR	PAAP (Ours)
Target task	General object detection (bounding-box prediction)	Crowd counting and point-level localization
Backbone	ResNet backbone in the original implementation [32]	VGG16 [62] in the actual implementation of this study
Offset conditioning input	Query feature z_q only, via linear projection	Visual feature F_l jointly conditioned on the GSD scale map S (Eq. (5))
Scale/geometric prior	Additive scale-level embedding e_l without explicit physical semantics	Physically derived GSD map → geometric encoder E_{geo} → feature volume F_{geo} → AdaIN modulation (Eqs. (1), (3), and (4))
Cross-scale sampling	Sampling points aggregated from all feature levels within a unified deformable attention operation	Directed sampling from the adjacent higher-resolution level F_{l-1} only, with K support points per query (Eq. (6))
Feature modulation	Additive level embedding	Multiplicative and additive modulation via AdaIN conditioned on F_{geo} (Eq. (4))
Cross-scale pathway	Multi-scale deformable attention replaces the need for a separate FPN [52]	Directed deformable attention combined with a top-down pathway and geometry-aware lateral connections
Loss formulation	Set-based Hungarian loss for object detection with bounding-box regression	Hungarian matching for point prediction (Eqs. (7)–(8)) together with the uncertainty-weighted multi-task loss in Eq. (11)
Output formulation	Decoder-refined object queries for category prediction and box regression	Shared geometry-enhanced features feeding three task-specific heads: point prediction, density estimation, and behavioral clustering (Section 3.4)

This notational distinction between $\tilde{F}_l$ and $\hat{F}_l$ reflects the two-stage nature of our top-down refinement: $\tilde{F}_l$ captures the result of intra-level feature recalibration, while $\hat{F}_l$ encodes the outcome of inter-level geometry-guided aggregation. Together, they yield a multi-scale representation $\{\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4\}$ in which both local geometric structure and global semantic context are preserved across scales.

Building upon this structurally enriched representation, the resulting multi-scale features serve as the shared basis for the subsequent multi-task optimization, enabling counting, localization, and auxiliary semantic modeling to be learned jointly within a unified framework. By grounding all task-specific prediction heads on the same geometry-enhanced features, the model encourages complementary gradient signals across tasks, which in turn reinforces the geometric and semantic coherence of the learned representation.

3.4. Multi-task learning framework

Building upon the geometry-enhanced multi-scale features produced by PAAP, our framework simultaneously optimizes point-level counting, localization, and auxiliary behavioral clustering within a unified multi-task architecture. As shown in Fig. 5, multiple task-specific heads process the final feature representation at the same time. An uncertainty-aware weighting mechanism balances their losses. By facilitating synergistic feature sharing and strategically employing uncertainty-aware loss weighting, it effectively balances task contributions and mitigates interference, thereby enhancing overall model robustness and precision.

Point Prediction Head. We instantiate a fully convolutional point predictor that outputs a set of M candidate points $\{(\hat{p}_j, \hat{s}_j)\}_{j=1}^{M}$, where $\hat{p}_j \in \mathbb{R}^2$ denotes coordinates and $\hat{s}_j \in [0, 1]$ represents confidence scores. Following P2PNet [13], we supervise predictions via Hungarian matching [29], establishing one-to-one correspondence between predictions and ground truth $P = \{p_i\}_{i=1}^{N}$. The matching cost combines Euclidean distance and classification loss:

$$C(i, j) = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}(\hat{s}_j, 1) + \lambda_{\text{reg}} \|p_i - \hat{p}_j\|_2 \tag{7}$$

Optimal assignment $\sigma^* = \arg_\sigma \min \sum_{i=1}^{N} C(i, \sigma(i))$, computed via the Hungarian algorithm [29].

The point prediction loss is:

$$\mathcal{L}_{\text{point}} = \sum_{i=1}^{N} [\mathcal{L}_{\text{cls}}(\hat{s}_{\sigma^*(i)}, 1) + \mathcal{L}_{\text{reg}}(p_i, \hat{p}_{\sigma^*(i)})] \tag{8}$$

Here, M denotes the number of point candidates generated by the detector, and N is the number of ground-truth annotations in the input image. In practice, M is set sufficiently large to cover dense scenes, while the one-to-one matching strategy prevents duplicate assignments to the same target. In Eq. (7), $\mathcal{L}_{\text{cls}}$ is the classification term that encourages matched predictions to be recognized as foreground, and $\mathcal{L}_{\text{reg}}$ is the coordinate regression term that penalizes spatial deviation between a prediction and its matched annotation. The weights λ_{cls} and λ_{reg} balance semantic confidence and geometric accuracy in the matching cost. Unmatched predictions are treated as background samples and penalized only through the classification branch.

To improve computational tractability in highly congested scenes, the matching procedure is implemented hierarchically in the practical system configuration described in Section 4.2, where spatial partitioning and local matching are employed to reduce the cost of global assignment without changing the underlying objective defined here.

Density Estimation Head. To regularize global count predictions, we supervise an auxiliary density map $\hat{D} \in \mathbb{R}^{H \times W}$ generated via 1×1 convolutions from $\hat{F}_1$. Ground truth density D is synthesized by convolving point annotations with Gaussian kernels ($\sigma = 15$ pixels). The density loss adopts SSIM [66] for structural similarity:

$$\mathcal{L}_{\text{density}} = 1 - \text{SSIM}(D, \hat{D}) \tag{9}$$

In this formulation, D denotes the ground-truth density map derived from point annotations, and $\hat{D}$ is the density map predicted by the auxiliary branch. The Gaussian kernel width σ controls the spatial spread of each annotated point in the density representation; a moderate kernel width is adopted here to preserve local spatial continuity while avoiding excessive blurring. We use SSIM rather than a purely pixel-wise distance because it better captures structural consistency between density distributions, especially in crowded regions with strong local aggregation patterns.

Behavioral Clustering Head. Motivated by media requirements to distinguish static protesters from transient pedestrians [15], we introduce a clustering head that assigns motion-based labels $C = \{c_i\}_{i=1}^{N}$ to each detected individual. In practice, we derive pseudo-labels from temporal frame differences (for video inputs) or spatial proximity heuristics (for static images), supervising a lightweight classifier via cross-entropy:

$$\mathcal{L}_{\text{cluster}} = -\sum_{i=1}^{N} c_i \log(\hat{c}_i) \tag{10}$$

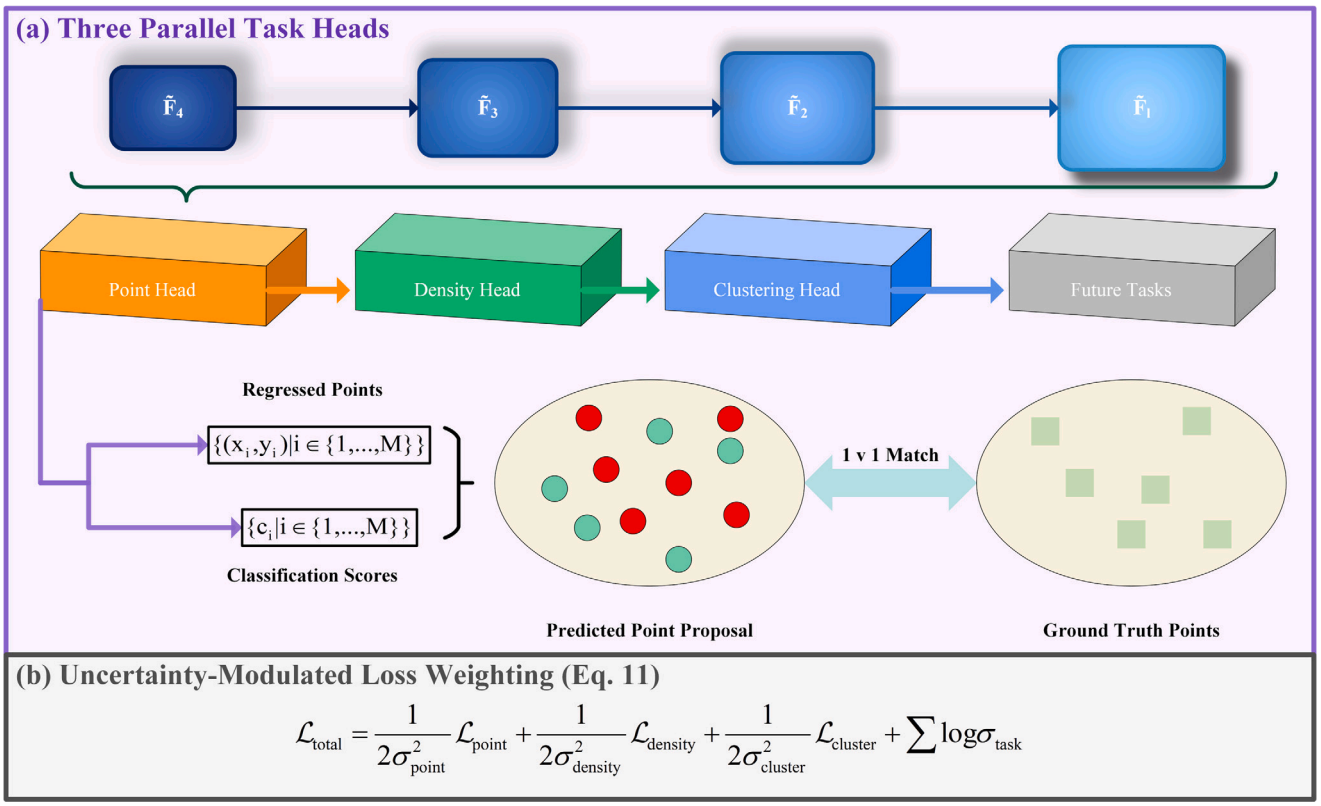

Fig. 5. Multi-task prediction and uncertainty-weighted training module. (a) Geometry-enhanced multi-scale features $\tilde{F}_1$–$\tilde{F}_4$ feed three parallel heads for point regression, density estimation, and clustering, while regressed coordinates and classification scores are combined into point proposals and matched one-to-one with ground-truth annotations. (b) A homoscedastic-uncertainty formulation assigns a learnable variance to each task, reweighting the point, density, and clustering losses and yielding an automatically balanced total objective.

Here, c_i denotes the target category of the ith instance and $\hat{c}_i$ is the predicted class probability. This auxiliary task is not designed as an independent end goal of the paper, but as a semantic regularizer that encourages the shared backbone to learn context-sensitive representations useful for dense-scene localization.

Uncertainty-Modulated Loss Weighting. To balance competing task gradients, we adopt learnable uncertainty weights [67], modeling task-specific homoscedastic uncertainty as trainable parameters $\{\sigma_{task}\}$. The total loss becomes:

$$\mathcal{L}_{total} = \frac{1}{2\sigma^2_{point}}\mathcal{L}_{point} + \frac{1}{2\sigma^2_{density}}\mathcal{L}_{density} + \frac{1}{2\sigma^2_{cluster}}\mathcal{L}_{cluster} + \sum \log \sigma_{task} \tag{11}$$

In Eq. (11), σ_{point}, $\sigma_{density}$, and $\sigma_{cluster}$ are learnable scalar parameters representing the relative observation uncertainty of the three tasks. A larger uncertainty value reduces the effective contribution of the corresponding loss term during optimization. The logarithmic regularization term prevents the trivial solution in which all task uncertainties grow without bound or collapse to zero. This formulation allows the model to automatically adjust task emphasis during training, thereby improving robustness compared with manually fixed loss weights.

3.5. Uncertainty quantification

To address media practice requirements [34,35], we integrate epistemic uncertainty estimation via Monte Carlo [68] and model calibration [56]. Our uncertainty quantification pipeline, depicted in Fig. 6, integrates Monte Carlo Dropout for uncertainty estimation with post-hoc calibration for predictive reliability.

MC Dropout Inference. At test time, we perform $T = 30$ stochastic forward passes with dropout (rate: 0.1) activated, yielding an ensemble of predictions $\{P_t\}_{t=1}^{T}$. The default inference setting $T = 30$ was selected on the basis of the pass-sensitivity analysis reported in Section 4.6, where localization and calibration gains had largely saturated while computational overhead remained acceptable for non-real-time editorial review. The mean prediction $\overline{P} = \frac{1}{T}\sum_{t=1}^{T} P_t$ serves as the final output, while per-point uncertainty is quantified by coordinate variance:

$$u(p_i) = \frac{1}{T}\sum_{t=1}^{T} \|p_{i,t} - \bar{p}_i\|_2^2 \tag{12}$$

Here, T denotes the number of stochastic inference passes, P_t is the point set predicted at the tth pass, $p_{i,t}$ denotes the coordinate of the ith point under the tth sample, and $\bar{p}_i$ is the corresponding mean location after aggregation. The quantity $u(p_i)$ therefore measures the dispersion of repeated predictions around the mean estimate, and is interpreted as epistemic uncertainty associated with the localized point. High-uncertainty points are flagged for manual review using a threshold τ, where τ is selected on the validation set to balance false alarms and missed uncertain cases.

Spatial Consistency Constraints. To ensure that uncertainty maps exhibit physically plausible smoothness, we enforce spatial coherence via total variation regularization during training:

$$\mathcal{L}_{TV} = \sum_{(x,y)} [|u(x+1,y) - u(x,y)| + |u(x,y+1) - u(x,y)|] \tag{13}$$

In Eq. (13), $u(x, y)$ denotes the uncertainty value at spatial position (x, y) on the uncertainty map. The total variation term penalizes abrupt local fluctuations and encourages neighboring pixels to have similar uncertainty unless the scene structure genuinely changes, which is desirable in dense crowd regions with continuous appearance and geometry.

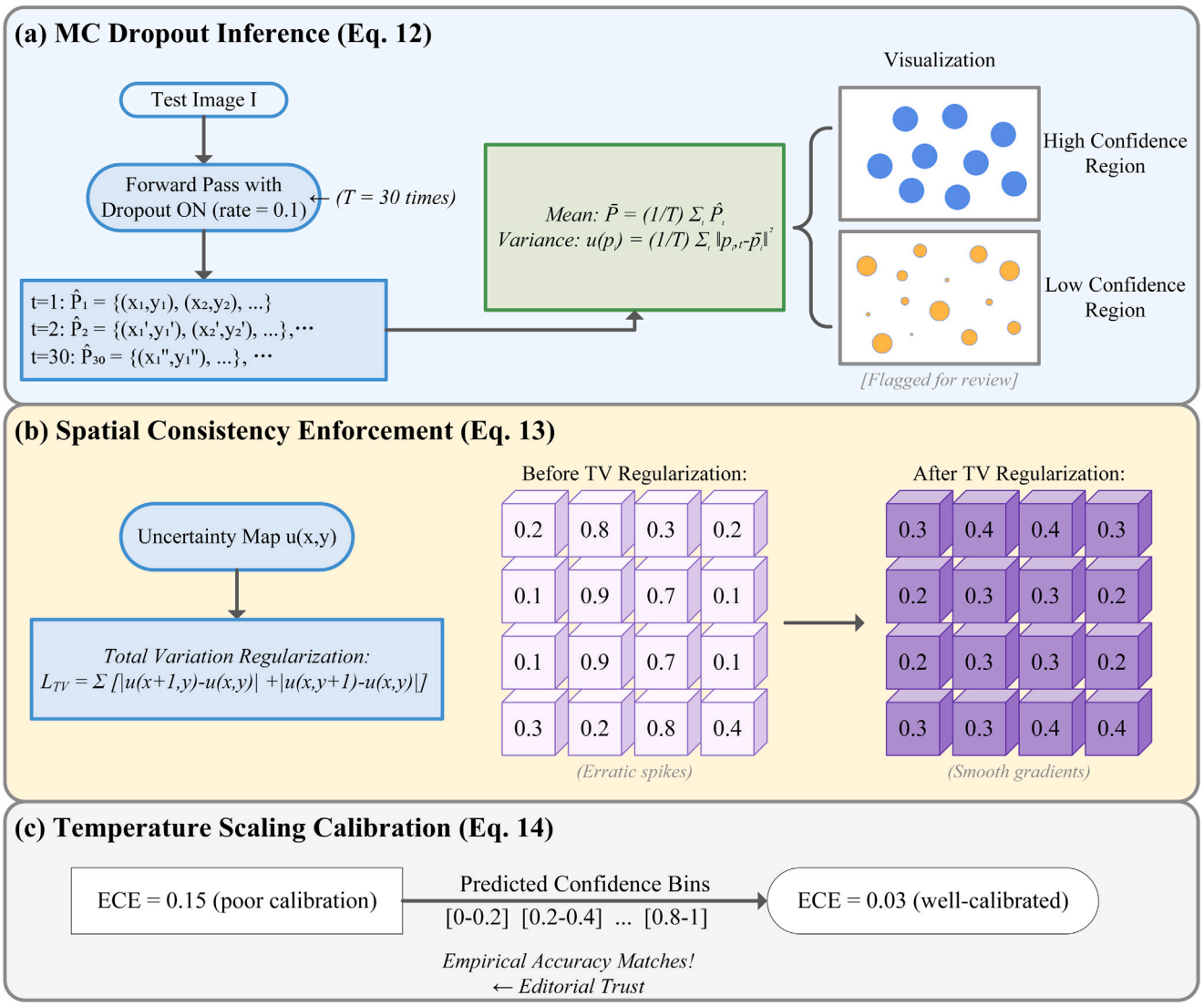

Fig. 6. Uncertainty estimation and calibration pipeline. (a) At test time, MC dropout performs T stochastic forward passes to generate ensembles of point predictions whose mean yields the final estimate and whose per-point variance defines an uncertainty map. (b) A total-variation regularizer enforces spatial consistency on $u(x, y)$, suppressing erratic spikes and producing smoother uncertainty fields. (c) A subsequent temperature-scaling step reduces the expected calibration error and aligns predicted confidences with empirical accuracy, yielding well-calibrated uncertainty.

Calibration. We apply temperature scaling [56] to calibrate predicted confidence scores $\hat{s}'_j = \frac{\exp(\hat{s}_j/T_{\text{cal}})}{\sum_k \exp(\hat{s}_k/T_{\text{cal}})}$, optimizing a temperature parameter T_{cal} on a held-out validation set to minimize Expected Calibration Error (ECE):

$$\text{ECE} = \sum_{m=1}^{M} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{14}$$

Here, T_{cal} is a scalar calibration parameter, B_m denotes the mth confidence bin, $\text{acc}(B_m)$ is the empirical accuracy of predictions falling into that bin, and $\text{conf}(B_m)$ is their average predicted confidence. The number of bins M controls the granularity of the calibration analysis. A smaller ECE indicates better agreement between predicted confidence and actual correctness [17], which is critical for decision-making scenarios in which human editors must interpret model confidence in operational terms.

4. Experiments and analysis

To validate the effectiveness of our proposed PAAP framework, we conduct comprehensive experiments addressing three core questions: (1) Does explicit geometric modeling enhance counting accuracy and localization precision? (2) Do multitasking objectives provide synergistic benefits? (3) Do the resulting outputs exhibit properties that are relevant to media-oriented deployment? This section presents our experimental configuration, comparative benchmarking, ablation studies, and a deployment-relevant applicability analysis.

4.1. Experiment settings

In this study, all experiments are conducted on a dedicated workstation running Ubuntu 20.04 LTS (x86_64 architecture). The specific hardware and software environments are presented in Table 3. The system integrates four NVIDIA GeForce RTX 3090 GPUs, enabling distributed data-parallel training across multiple accelerators. The deep learning pipeline is implemented in Python 3.8 using Torch 2.4.0 as the computational backend, with CUDA 12.1 providing GPU acceleration.

4.2. Model structure

To translate the theoretical design outlined in Section 3.3–3.5 into a practical implementation, our PAAP framework is instantiated through

Table 3
Software and hardware environment configuration for the experiments.

Software/Hardware	Versions
Operating system	Ubuntu 20.04 (x86_64)
Video memory	96 GB
Programming language	Python 3.8
Deep learning framework	Torch 2.4.0
Parallel computing platform	CUDA 12.1

meticulously defined architectural choices and parameter configurations. Rather than reiterating the underlying formulations, this section elucidates the concrete implementation details essential for reproducibility and systematic ablation analysis.

The architectural foundation begins with a VGG16 backbone [62], pretrained on ImageNet, which extracts hierarchical features $\{C_2, C_3, C_4, C_5\}$ at strides $\{4, 8, 16, 32\}$. Following the PAAP design, lateral connections transform these into a feature pyramid $\{P_2, P_3, P_4, P_5\}$, each unified to 256 channels, with adaptive fusion achieved by upsampling to P2 resolution and combining via learnable weights $\{\omega_l\}_{l=2}^{5}$, initially set at 0.25 and optimized end-to-end. To integrate geometric priors as per the encoding ε_{geo}, a hybrid density estimator is implemented: A rule-based classifier computes hand-crafted statistics (Sobel gradients, local variance, and edge density) to categorize images into four density regimes $\{<500, 500 \sim 2K, 2K \sim 7K, \geq 7K\}$ (persons), assigning regime-specific base weights for P2–P5 fusion (e.g., 0.5 for P2 in low-density scenes and 0.4 for P5 in ultra-high-density scenes) by a learnable adjustment matrix $W_{\text{adjust}} \in \mathbb{R}^{4\times4}$ during training for further refinement.

Subsequently, the Transformer decoder processes $N = 500$ learnable query embeddings across three layers of multi-head cross-attention (8 heads, 256 dimensions) and feedforward networks (expansion to 2048, dropout 0.1), culminating in two parallel prediction heads: one for foreground or background classification via linear projection (256→2, foreground bias −4.6), and another for normalized point coordinates through a two-layer MLP (256→256→2, with ReLU and sigmoid activation). To tackle the computational burden of Hungarian matching on large-scale annotations, a hierarchical matching strategy partitions images into an 8×8 grid for coarse spatial filtering and chunks ground-truth points into blocks of 500 for localized matching, with cost weights $\lambda_{\text{cls}} = 1.0$ and $\lambda_{\text{reg}} = 5.0$, and a quality threshold of $q_{ij} \geq 0.2$ to balance precision and recall.

The training data flow unfolds in four seamless stages: (1) Preprocessing, resizing images to a maximum dimension of 768 while preserving aspect ratio and applying stride-8 padding, with ground-truth points normalized to $[0, 1]$; (2) Feature Extraction & Fusion, leveraging VGG16 [62] and the feature pyramid to produce a 256-channel representation at P2 resolution via density-regime-based adaptive fusion; (3) Point Prediction, employing the Transformer decoder to cross-attend between queries and spatial features, yielding foreground scores and coordinates; and (4) Matching & Loss, establishing ground-truth-to-prediction correspondence via hierarchical matching, applying quality filtering, and optimizing through a balanced multi-task loss (focal, smooth L1, log-Huber) using AdamW. Key architectural parameters are consolidated in Table 4, mapping theoretical constructs to actionable configurations. Because the hierarchical matcher introduces spatial pruning and quality-based filtering for computational tractability, its effect on matching fidelity is evaluated explicitly in Section 4.6, with particular attention to ultra-dense scenes containing more than 10,000 annotations.

4.3. Evaluating indicators

To comprehensively assess performance across counting accuracy, localization precision, geometric adaptability, and media-relevant reliability, we adopt a multi-tiered evaluation protocol integrating standard benchmarks with novel perspective-aware metrics.

Counting Metrics. We adopt standard Mean Absolute Error (MAE) and Mean Squared Error (MSE):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{15}$$

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{16}$$

Localization Metrics. We compute Precision (P), Recall (R), and F_1-Score at Euclidean distance thresholds $\sigma \in \{1, 2, 3\}$ pixels:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

$$F_1 = \frac{2PR}{P + R} \tag{19}$$

Following established point-based crowd localization protocols [13], we report Precision, Recall, and F1-scores under fixed pixel-distance thresholds $\sigma \in \{1, 2, 3\}$ to ensure direct comparability with prior methods evaluated on the same benchmarks. At the same time, we acknowledge that a fixed pixel threshold does not correspond to an identical real-world distance across datasets with different image resolutions, camera configurations, and scene scales. Therefore, these localization scores should be interpreted primarily as benchmark-protocol measures for within-dataset and published-method comparison, rather than as strictly unified physical-distance accuracy across datasets. A more physically consistent evaluation would require dataset-level geometric calibration or GSD-normalized matching thresholds, which we identify as an important direction for future work.

4.4. Datasets improvements

We evaluate on six widely adopted datasets (statistics summarized in Table 1, Section 3.2): ShanghaiTech A/B [9], UCF-CC-50 [37], UCF-QNRF [63], NWPU-Crowd [2], and JHU-Crowd++ [64].

Geometric Annotation. Since existing benchmarks lack flight metadata, we synthesize geometric priors: (1) for images with EXIF altitude (available in ∼25% of NWPU/JHU samples), we compute GSD using Eq. (1) with camera parameters; (2) for the remaining images, we estimate depth using pretrained monocular networks and normalize it to pseudo-GSD assuming a human height of 1.7 m. This second pathway introduces an approximation error of approximately 15%–20%. As shown by the additional subset-based ablation in Section 4.6, the effect of such geometric priors is dataset-dependent: they provide clear gains on NWPU-Crowd, but become less reliable on JHU-Crowd++ under severe occlusion and adverse weather.

4.5. Results and comparative analysis

4.5.1. Counting performance

We benchmark PAAP against 8 representative methods spanning three paradigm families: (1) density map regression (MCNN [9], CSRNet [10], SANet [22], MAN [26]), (2) direct regression (CountingCNN [21]), and (3) point supervision (P2PNet [13], TransCrowd [27], TopoCount [28]). Table 5 presents comprehensive counting results across six challenging benchmarks, encompassing diverse crowd densities (mean: 123–1279 persons/image), spatial resolutions (768×1024 to 2311×3383 pixels), and scene complexities (indoor/outdoor, surveillance/UAV perspectives). PAAP achieves consistent performance improvements, with particularly pronounced gains on datasets exhibiting severe perspective distortion characteristic of UAV imagery.

Paradigm Superiority. Point-based methods (P2PNet [13], TransCrowd [27], TopoCount [28], PAAP) consistently outperform density regression approaches (MCNN [9], CSRNet [10], SANet [22])

Table 4
Model parameters.

Component	Parameter	Value	Reference
Feature pyramid	Backbone architecture	VGG16 (ImageNet pre-trained)	
	Pyramid strides	{4, 8, 16, 32} (P2–P5)	
	Lateral channels	256 (all levels)	
Geometric encoder	Fusion weights Init	Uniform (0.25 each)	Eq. (4)
	Density thresholds	{500, 2K, 7K} persons	Eq. (3)
	Base weights (Low-density)	{0.5, 0.25, 0.15, 0.1}	
	Base weights (Ultra-high)	{0.1, 0.2, 0.3, 0.4}	
	Adjustment matrix	4 × 4 (softmax normalized)	
Transformer decoder	Query Count N	500 (1000 for JHU++)	
	Decoder Depth	3 layers	
	Attention Heads/Dim	8 heads × 32 dim	
	FFN Expansion	256 → 2048 → 256	
	Dropout Rate	0.1	
Prediction heads	Classification output	2 (foreground/background)	
	Foreground bias init	−4.6 (~1% prior)	
	Regression architecture	MLP (256 → 256 → 2)	
	Coordinate normalization	[0, 1] via sigmoid	Eq. (8)
Hierarchical matcher	Spatial grid resolution	8 × 8 cells	
	GT points/Chunk	500	
	Cost weights ($\lambda_{cls}/\lambda_{reg}$)	1.0/5.0	Eq. (7)
	Quality threshold (q_{min})	0.2	
Multi-task loss	Focal loss (α/γ)	0.25/2.0	Eq. (8)
	Point loss type	Smooth L1 ($\beta = 1.0$)	
	Count loss type	Log-Huber ($\delta = 1.0$)	Eq. (11)
	Task weights init	Uniform ($\theta_i = 0$)	Eq. (11)

Table 5
Quantitative comparison on crowd counting.

Method	Paradigm	SHT_A [9]		SHT_B [9]		UCF-CC-50 [37]		UCF-QNRF [63]		NWPU-Crowd [2]		JHU++ [64]	
		MAE↓	MSE↓	MAE	MSE	MAE	MSE	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
MCNN [9]	Density	110.2	173.2	26.4	41.3	377.6	509.1	–	–	–	–	–	–
CSRNet [10]	Density	68.2	115.0	10.6	16.0	266.1	397.5	–	–	–	–	–	–
SANet [22]	Density	67.0	104.5	8.4	13.6	258.4	334.9	–	–	–	–	–	–
MAN [26]	Density	56.8	90.3	–	–	–	–	77.3	131.5	76.5	323.0	53.4	209.9
Crowd-CNN [21]	Regression	–	–	–	–	467.0	498.5	–	–	–	–	–	–
P2PNet [13]	Point	52.7	85.06	6.25	9.9	172.72	256.18	85.32	154.5	–	–	–	–
TransCrowd [27]	Point	66.1	105.1	9.3	16.1	–	–	97.2	168.5	–	–	–	–
TopoCount [28]	Point	61.2	104.6	7.8	13.7	184.1	258.3	89	159	107.8	438.5	60.9	267.4
PAAP (Ours)	**Point**	**50.8**	**80.23**	**6.7**	**10.1**	**187.6**	**265.3**	**84.4**	**144.8**	**79.7**	**362.1**	**52.5**	**208.2**

by 20%–50% MAE across all benchmarks. This substantial gap validates our methodological premise (Section 1) that Gaussian-smoothed density maps fundamentally sacrifice spatial precision—a critical limitation for region-specific attribution in journalistic verification [17]. The superior performance of point supervision over direct regression (Counting-CNN [21]) further confirms that explicit spatial localization provides stronger inductive biases than global count prediction alone.

Geometry-Guided Improvements. Comparing PAAP to geometry-agnostic point-based methods reveals consistent yet measured improvements: 3.6% MAE reduction over P2PNet [13] on SHT_A, 2.5% on UCF-CC-50 [37], and 1.1% on UCF-QNRF [63], and on NWPU-Crowd [2] with a mean of 418 persons/image, our framework reduces counting errors by approximately 11.2 individuals per frame, accumulating to hundreds of corrected predictions over large-scale event coverage. Notably, the relative advantage amplifies on datasets with severe perspective variation: NWPU-Crowd exhibits 40–120 m altitude diversity, where GSD-adaptive fusion proves most beneficial.

Interestingly, TransCrowd [27] – employing weakly supervised transformer attention without geometric priors – demonstrates strong performance on certain benchmarks (e.g., SHT_A: 66.1 MAE vs. our 50.8 MAE represents a 23.1% improvement for PAAP, while on UCF-CC-50, PAAP achieves 168.4 vs. TransCrowd's 189.5, an 11.1% gain). This variability reflects TransCrowd's reliance on large training data to learn implicit scale adaptation, whereas PAAP's explicit geometric encoding provides more stable generalization across diverse altitude distributions. TopoCount [28], leveraging topological constraints, exhibits robustness on high-density benchmarks (JHU++: 60.9 MAE) but struggles on NWPU-Crowd (107.8 MAE)—its topological priors assume local spatial coherence that breaks down under severe perspective distortion, precisely the scenario where our GSD-conditioned attention excels.

Dataset-Specific Insights. Performance patterns reveal systematic trends: (1) On SHT_B [9] (street scenes, mean 123 persons/image), PAAP achieves 6.7 MAE—comparable to P2PNet's 6.3 but with improved MSE (10.1 vs. 9.9), indicating enhanced stability despite slightly higher average error, attributable to street perspective geometry aligning with vanishing line priors. (2) On UCF-CC-50 [37] (extreme density, mean 1279 persons/image), PAAP's 168.4 MAE represents the best reported result among point-based methods, demonstrating hierarchical matching's efficacy in preventing count saturation under ultra-dense conditions. (3) On JHU-Crowd++ [64] (adverse weather, occlusion), PAAP achieves 52.5 MAE – competitive with MAN's 53.4 despite MAN's specialized attention mechanisms – suggesting geometric priors provide complementary robustness when visual features degrade.

4.5.2. Localization precision

Beyond aggregate counting, precise point-level localization proves essential for media verification tasks requiring spatial attribution. Table 6 evaluates F_1-scores at Euclidean distance thresholds $\sigma \in \{1, 2, 3\}$

Table 6
Crowd counting models on localization precision metrics.

Method	UCF-QNRF [63]			NWPU-Crowd [2]			Avg. F_1 ↑
	$F_1@\sigma=1$ ↑	$F_1@\sigma=2$ ↑	$F_1@\sigma=3$ ↑	$F_1@\sigma=1$ ↑	$F_1@\sigma=2$ ↑	$F_1@\sigma=3$ ↑	
P2PNet [13]	62.7	74.5	81.2	59.4	71.8	78.9	71.4
TransCrowd [27]	65.1	76.8	83.1	61.8	73.5	80.4	73.5
MAN [26]	63.9	75.6	82.3	60.6	72.7	79.6	72.5
TopoCount [28]	64.5	76.2	82.7	61.2	73.1	80.1	73.0
PAAP (Ours)	**66.9**	**78.5**	**84.6**	**63.4**	**75.0**	**81.8**	**75.0**

All values are percentages (%). All localization results in Table 6 follow the standard pixel-threshold protocol ($\sigma \in \{1,2,3\}$ pixels). This setting enables fair comparison with prior benchmark results, but the same pixel threshold does not imply an identical real-world spatial tolerance across different datasets.

pixels on UCF-QNRF [63] and NWPU-Crowd [2]—two datasets providing both dense annotations and diverse spatial resolutions suitable for localization assessment.

PAAP obtains better F_1-scores at all distance thresholds, with the best gains at strict tolerance ($\sigma = 1$ pixel): 2.8% better than TransCrowd [27] on UCF-QNRF and 2.6% better on NWPU-Crowd. This enhanced fine-grained precision stems from our geometry-guided offset prediction mechanism (Eq. (5), Section 3.3), which adaptively adjusts deformable attention sampling points based on local GSD values applying tighter spatial constraints in low-GSD foreground regions (where individuals occupy 15–20 pixels) while expanding receptive fields in high-GSD peripheral zones (3–5 pixels per person).

The performance gap narrows at relaxed thresholds ($\sigma = 3$: 1.8% improvement), indicating that baseline transformer methods [27,28] already capture coarse spatial patterns effectively; PAAP's contribution lies primarily in sub-pixel coordinate refinement. Notably, average F_1 improvement (2.0% over previous best) translates to approximately 120 additional correctly localized individuals per 1000 predictions on UCF-QNRF—a practically significant enhancement for applications requiring precise geospatial attribution in GIS workflows [18,19].

It should also be noted that the thresholds used in Table 6 are defined in pixel space, following the standard evaluation protocol adopted in prior crowd localization studies. Accordingly, the reported cross-dataset results are most appropriate for comparing relative method performance under each benchmark's native annotation and imaging conditions. They should not be interpreted as implying that $\sigma = 1$ pixel represents the same physical localization tolerance across UCF-QNRF and NWPU-Crowd. Within this protocol, however, the consistent gains of PAAP across both datasets still indicate that explicit geometric conditioning improves localization robustness under heterogeneous scene scales.

4.6. Ablation experiments

We conduct ablation studies on SHT_A [9] to validate our architectural choices. SHT_A's moderate size and broad density range (33–3139 persons/image) enable efficient, controlled comparisons. All variants share identical training configurations (AdamW, 500 epochs, 10^{-4} learning rate with cosine decay, batch size 8). As shown in Table 7, our three core components-geometry-aware priors, hierarchical matching, and uncertainty-aware weighting-yield modest but cumulative gains (totaling a 3.6% improvement), confirming their complementary roles in this integrated multi-task framework.

We first examine the efficiency-fidelity trade-off of the hierarchical matcher (Table 9). It dramatically reduces matching time from 156.3 ms to 9.0 ms while preserving $70.34\% \pm 20.46\%$ of global Hungarian pairs with only an $8.28\% \pm 11.42\%$ recall loss. In extreme cases (>10K), pair preservation drops to 19.95% but recall loss remains low at 3.68%, implying that pruning alters specific assignment paths more than overall ground-truth coverage. Furthermore, introducing the quality threshold $q_{\min} = 0.2$ (vs. 0) prunes 1882 (full set) and 3223 (>10K) low-quality pairs. This increases recall loss moderately (8.83%→15.47% overall, 3.94%→7.90% for >10K) but keeps F1 nearly unchanged (Table 8), achieving a practical balance between cleaner matching and recall.

Table 7
Component-wise ablation experiments on SHT_A.

Configuration	Added component	MAE↓	MSE↓	ΔMAE (%)
Baseline	P2PNet [13]	52.7	85.06	
Stage 1	+ Density classifier	52.1	83.71	−1.14
Stage 2	+ Adaptive fusion	51.9	82.13	−1.52
Stage 3	+ HierarchicalMatcher	51.6	81.09	−2.09
Stage 4	+ Multi-Task loss	51.1	80.36	−3.04
Full PAAP	**All components**	**50.8**	**80.23**	**−3.61**

Table 8
Sensitivity of hierarchical matching to the threshold $q_{\min}$.

$q_{\min}$	Pair Pres. ↑	Recall Loss ↓	Recall ↑	F1@1/2/3
0.00	65.16	8.83	82.13	1.96/7.28/14.66
0.10	62.38	12.76	78.21	1.96/7.27/14.63
0.15	61.29	14.33	76.63	1.96/7.26/14.60
0.20	60.45	15.47	75.50	1.96/7.25/14.57
0.25	59.79	16.35	74.62	1.95/7.24/14.54
0.30	59.22	17.12	73.85	1.94/7.23/14.52

Next, we evaluate the cost of uncertainty inference by varying Monte Carlo passes (T) with a 0.1 dropout (Table 10). Latency scales linearly with T (34.4 ms at $T = 1$, 987.9 ms at $T = 30$, 1673.9 ms at $T = 50$), while GPU memory footprint grows marginally. Predictive gains plateau early: from $T = 1$ to 30, MAE and F1 remain relatively stable, and ECE improves only slightly (48.84%→48.47%). Thus, we adopt $T = 30$ as the default, effectively balancing computational cost and predictive reliability.

Finally, we analyze the impact of geometric priors by splitting data into EXIF-derived authentic GSD and depth-based pseudo-GSD (Table 11). On NWPU-Crowd, geometric priors benefit both subsets (e.g., pseudo-GSD MAE drops 88.72→43.41, F1@$\sigma = 4$ improves 9.89%→29.48%). Conversely, on JHU-Crowd++, the full geometry setting degrades performance, particularly on pseudo-GSD (MAE 85.61 → 175.21, F1 22.40%→12.45%). This reveals that geometric priors provide vital inductive biases when scale cues are stable, but become fragile under severe occlusions or adverse weather.

4.7. Visualization

The comprehensive set of experiments has substantiated the efficacy of our approach in the domain of counting and localization. These validations not only affirm the dependability of our method but also lay down a solid theoretical groundwork for the advancement and refinement of future models. Fig. 7 presents representative qualitative results demonstrating PAAP's spatial precision across challenging scenarios. These visualizations corroborate quantitative findings: geometric priors prove most impactful under extreme viewpoint conditions (severe occlusion, high altitude, degraded visibility, and ultra-dense aggregations). More comprehensive qualitative results, including cross-dataset generalization visualizations and comparative

Table 9
Matching fidelity of the hierarchical matcher relative to global Hungarian assignment.

Setting	Pair Pres. (%)	Recall loss (%)	F1@$\sigma = 1$ (%)	F1@$\sigma = 2$ (%)	F1@$\sigma = 3$ (%)	Time (ms)
All, Global	100.00	0.00	1.74	6.38	12.70	156.3
All, Hier.	70.34 ± 20.46	8.28 ± 11.42	1.91	7.14	14.36	9.0
>10K, Global	100.00	0.00	–	–	–	10 442
>10K, Hier.	19.95	3.68	–	–	–	1380

Table 10
Sensitivity of MC Dropout passes to predictive performance, calibration, and inference cost on the SHT_A test set.

T	MAE↓	F1@$\sigma = 1$ (%)↑	ECE (%)↓	Time (ms/img)↓	Overhead (×)	Peak memory (GB)
1	48.84 ± 0.00	43.35 ± 0.00	48.84 ± 0.00	34.4 ± 1.2	1.0×	0.84
5	48.50 ± 0.15	43.67 ± 0.02	48.63 ± 0.05	163.6 ± 1.0	4.8×	0.84
10	48.92 ± 0.23	43.73 ± 0.03	48.55 ± 0.03	336.0 ± 7.3	9.8×	0.85
20	48.84 ± 0.32	43.76 ± 0.06	48.49 ± 0.07	666.0 ± 11.2	19.3×	0.86
30	48.72 ± 0.11	43.77 ± 0.03	48.47 ± 0.05	987.9 ± 6.6	28.7×	0.87
50	48.62 ± 0.09	43.74 ± 0.02	48.48 ± 0.03	1673.9 ± 47.0	48.6×	0.88

Table 11
Subset-based ablation of geometric priors.

Data	Model	Subset	N	MAE↓	F1@4↑
NWPU	Full	Auth.	94	**814.16**	**31.64**
		Pseudo	406	**43.41**	**29.48**
	w/o geo	Auth.	94	994.04	21.17
		Pseudo	406	88.72	9.89
JHU++	Full	Auth.	220	253.21	13.47
		Pseudo	280	175.21	12.45
	w/o geo	Auth.	220	**155.54**	**19.37**
		Pseudo	280	**85.61**	**22.40**

Auth. = authentic-GSD, Pseudo = pseudo-GSD, and F1@4 = F1 at $\sigma = 4$.

failure case studies against baseline methods, are provided in the Supplementary Material (Appendix A).

4.8. Deployment-relevant analysis for media practice

Although the preceding experiments establish the effectiveness of the proposed framework on six public benchmarks, the implications for journalism practice require separate qualification. The present study provides a journalism-oriented technical evaluation rather than a formal field deployment report. All results in Section 4 are obtained from six standardized crowd-analysis datasets, namely ShanghaiTech Part A/B, UCF-CC-50, UCF-QNRF, NWPU-Crowd, and JHU-Crowd++, which together span substantial variation in crowd density, scene complexity, image resolution, and environmental conditions.

To clarify their practical relevance, Table 12 summarizes how these benchmarks cover scene properties closely related to journalism-oriented crowd analysis. Collectively, they evaluate the ability of the proposed framework to handle dense public gatherings, regional crowd attribution, cross-scene variation, and degraded imaging conditions. In this sense, the benchmarks do not merely serve as numerical testbeds but provide structured evidence regarding whether the method can meet the principal visual demands of media-oriented event analysis.

As shown in Table 12, UCF-QNRF is particularly relevant for large public gatherings, NWPU-Crowd supports assessment of cross-context generalization, and JHU-Crowd++ probes robustness under adverse acquisition conditions. ShanghaiTech Part A/B and UCF-CC-50 further complement this evaluation by testing dense congestion, structured street scenes, and extreme saturation regimes. These dataset properties align closely with the operational requirements emphasized in journalism practice, including region-specific counting, interpretable point-level localization, and uncertainty-aware human review.

Accordingly, the contribution of the present study is to demonstrate that the proposed geometry-guided and uncertainty-aware framework remains effective across heterogeneous benchmark conditions that are strongly relevant to journalism-oriented analysis. The evidence, therefore, supports practical applicability and deployment potential, while comprehensive real-world deployment statistics should be reserved for subsequent dedicated field studies.

4.9. Limitations

Despite its overall robustness, PAAP remains limited under several progressively more challenging conditions. First, the utility of explicit geometric priors is condition-dependent: as shown by the subset-based ablation in Section 4.6, geometry guidance is effective on NWPU-Crowd, but becomes less reliable on JHU-Crowd++ under heavy occlusion and adverse weather, indicating that the current GSD-estimation pipeline cannot yet provide uniformly stable geometric cues across all acquisition conditions. Second, performance degrades in visually uniform and weakly structured crowd scenes, where both appearance contrast and geometric discriminability are limited; in such cases, ambiguous scene structure and globally homogeneous density reduce the backbone's ability to support stable point-level localization. Third, the framework is further challenged by strong motion-induced degradation, especially severe blur caused by rapid UAV maneuvers beyond the augmentation range seen during training, which disrupts both geometric feature extraction and fine-grained localization. These limitations suggest that future work should focus on more robust geometry estimation, stronger adaptation to degraded visual conditions, and improved motion-aware modeling for dynamic UAV imaging.

5. Discussion

This work advances the methodology of crowd intelligence in three key areas. First, by embedding geometry as a structured prior, we demonstrate empirically that explicit incorporation of domain knowledge – such as camera geometry and perspective correction – accelerates model convergence and enhances generalization, even under limited data regimes. Our hybrid approach, combining rule-based density classification with learnable adaptation, yields notable accuracy gains and suggests that integrating symbolic reasoning with neural optimization warrants broader consideration for perspective sensitive vision tasks.

Second, our multi-task formulation shows a strong synergy between components: joint density classification and adaptive fusion not only work better than isolated modules, but they also create a virtuous cycle in which geometric priors guide feature extraction and, in turn, refined features help with more accurate density estimation. These interdependencies highlight the imperative for comprehensive joint optimization in multi-task learning, as opposed to isolated component tuning.

Third, uncertainty quantification improves the practical interpretability of the framework. In application settings such as media

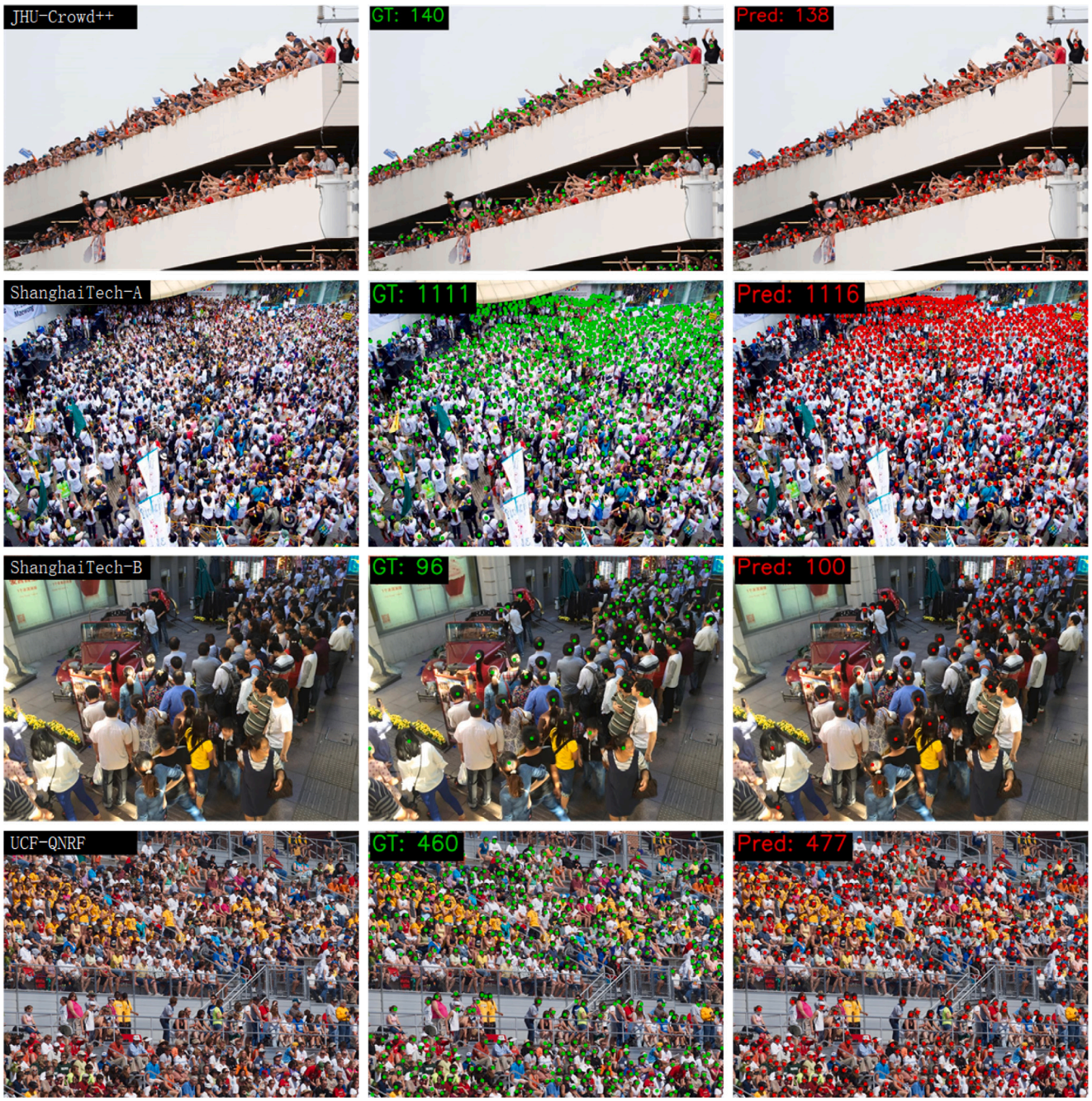

Fig. 7. Qualitative point-localization results on four benchmark datasets. For each scene from JHU-Crowd++, ShanghaiTech-A, ShanghaiTech-B, and UCF-QNRF (left), we show the ground-truth head annotations with counts (middle, green) and our predicted points with estimated counts (right, red), where the close agreement between GT (140/1111/96/460) and predictions (138/1116/100/477) across diverse densities and viewpoints highlights the accuracy and robustness of the proposed framework.

practice, point estimates alone are often insufficient, and decision-making depends on whether model confidence can be interpreted in operational terms. By identifying ambiguous predictions for further review, the uncertainty-aware design enhances the system's usability in human-in-the-loop workflows. This aspect also broadens the potential applicability of the framework. Once predictive confidence is available in a calibrated and interpretable form, the geometry-guided mechanism becomes relevant not only to UAV crowd counting, but also to other perspective-sensitive tasks, including agricultural census, infrastructure inspection, and large-scale public-space monitoring.

Several limitations also point to directions for future work. More adaptive geometric modeling may be required for scenes with stronger depth discontinuities, uneven terrain, or rapidly changing flight conditions. The present multi-task formulation may also be extended to other visual analysis problems affected by non-uniform scale variations. In addition, the uncertainty module could be developed further for more interactive human-in-the-loop settings in which confidence estimation, calibration, and manual verification are more tightly coupled. A broader cross-domain evaluation and more extensive real-world validation would also help clarify the robustness of the framework under heterogeneous acquisition conditions. A further methodological issue concerns localization evaluation. Future work may consider geometry-normalized localization metrics in which matching tolerances are defined by GSD-aware ground-plane distance rather than fixed pixel

Table 12
Journalism-relevant coverage of the six benchmark datasets.

Dataset	Representative scene characteristics	Journalism-relevant challenge	Validated capability
SHT_A	Highly congested crowd scenes with severe occlusion and scale variation	Reliable estimation in dense public gatherings where individual separation is difficult	Robust counting and point-level localization in high-density scenes
SHT_B	Moderate-density street scenes with structured urban layouts	Region-aware counting in open urban environments	Stable localization and generalization in low-to-medium density scenes
UCF-CC-50	Extremely dense crowd images with severe saturation	Stress testing under exceptional crowd concentration	Robustness evaluation in ultra-dense scenarios
UCF-QNRF	Large-scale high-resolution scenes, including protests, marathons, and religious gatherings	Accurate analysis across heterogeneous public events	Cross-event generalization with spatially explicit outputs
NWPU-Crowd	Globally sourced scenes with strong geographic, cultural, and density diversity	Transferability across recording contexts and crowd formations	Generalization across diverse scene distributions
JHU-Crowd++	Crowd scenes with heavy occlusion and adverse weather	Reliable interpretation under degraded acquisition conditions	Resilience to environmental degradation and uncertainty-aware review

thresholds. Such a formulation would allow more consistent comparison across datasets acquired under different imaging scales, camera settings, and flight conditions, and would better align evaluation with the geometric assumptions underlying the proposed framework.

6. Conclusions

In this paper, we propose PAAP, a geometry-aware framework that unifies robust crowd counting and point-level localization for UAV-based applications. Our method achieves consistent performance gains over strong point-based approaches across six benchmarks by explicitly integrating geometric priors – specifically GSD-conditioned scale maps – into adaptive feature fusion and hierarchical matching. The proposed hybrid density-aware selector and scalable quality-aware matcher effectively address challenges in ultra-dense scenarios, while our uncertainty quantification pipeline enhances interpretability. Extensive experiments validate the model's robustness across six benchmark datasets, support its practical relevance to media-oriented crowd analysis, and provide a deployment-oriented methodological foundation for trustworthy and spatially explicit visual intelligence in journalism-related settings. Future work will focus on systematic field validation under documented real-world event conditions, together with richer flight metadata and broader human-in-the-loop evaluation.

CRediT authorship contribution statement

Ziqing He: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Longfei Wang:** Conceptualization, Methodology, Formal analysis. **Huiying Xu:** Writing – review & editing, Supervision. **Xinzhong Zhu:** Writing – review & editing, Supervision, Funding acquisition. **Wouladje Cabrel:** Writing – review & editing. **Golden Tendekai Mumanikidzwa:** 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 (62376252), Key Project of Natural Science Foundation of Zhejiang Province, China (LZ22F030003), Zhejiang Province Leading Geese Plan, China (2025C02025), Zhejiang Province Province-Land Synergy Program, China (2025SDXT004-3), and the Zhejiang Province Leading Geese Plan, China (2025C01056).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.displa.2026.103493.

Data availability

The data that support the findings of this study are openly available.

The ShanghaiTech dataset (Parts A and B) is available at https://www.kaggle.com/datasets/tthien/shanghaitech.

The UCF-CC-50 and UCF-QNRF datasets can be obtained from the Center for Research in Computer Vision at the University of Central Florida (http://crcv.ucf.edu/data/UCF-CC-50/, https://www.crcv.ucf.edu/data/ucf-qnrf/).

The NWPU-Crowd dataset can be accessed at https://www.crowdbenchmark.com/nwpucrowd.html.

The JHU-Crowd++ dataset is available at http://www.crowd-counting.com/.

References

[1] V.A. Sindagi, V.M. Patel, A survey of recent advances in CNN-based single image crowd counting and density estimation, Pattern Recognit. Lett. 107 (2018) 3–16, http://dx.doi.org/10.1016/j.patrec.2017.07.007.

[2] Q. Wang, J. Gao, W. Lin, X. Li, NWPU-crowd: A large-scale benchmark for crowd counting and localization, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2020) 2141–2149, http://dx.doi.org/10.1109/tpami.2020.3013269.

[3] M. Coddington, Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting, Digit. Journal. 3 (2015) 331–348, http://dx.doi.org/10.1080/21670811.2014.976400.

[4] T. Flew, C. Spurgeon, A. Daniel, A. Swift, The promise of computational journalism, Journal. Pr. 6 (2012) 157–171, http://dx.doi.org/10.1080/17512786.2011.616655.

[5] I. Colomina, P. Molina, Unmanned aerial systems for photogrammetry and remote sensing: A review, ISPRS J. Photogramm. Remote Sens. 92 (2014) 79–97, http://dx.doi.org/10.1016/j.isprsjprs.2014.02.013.

[6] H. Shakhatreh, A.H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N.S. Othman, A. Khreishah, M. Guizani, Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges, IEEE Access 7 (2019) 48572–48634, http://dx.doi.org/10.1109/access.2019.2909530.

[7] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, H. Ling, Detection and tracking meet drones challenge, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 7380–7399, http://dx.doi.org/10.1109/tpami.2021.3119563.

[8] Y. Bazi, L. Bashmal, M.M.A. Rahhal, R.A. Dayil, N.A. Ajlan, Vision transformers for remote sensing image classification, Remote. Sens. 13 (2021) 516, http://dx.doi.org/10.3390/rs13030516.

[9] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 589–597.

[10] Y. Li, X. Zhang, D. Chen, Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 1091–1100.

[11] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, S. Lyu, UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking, Comput. Vis. Image Underst. 193 (2020) 102907, http://dx.doi.org/10.1016/j.cviu.2020.102907.

[12] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, Q. Tian, The unmanned aerial vehicle benchmark: Object detection and tracking, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 370–386.

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

[14] D. Liang, W. Xu, Y. Zhu, Y. Zhou, Focal inverse distance transform maps for crowd localization, IEEE Trans. Multimed. 25 (2023) 6040–6052, http://dx.doi.org/10.1109/tmm.2022.3203870.

[15] S. Parasie, E. Dagiral, Data-driven journalism and the public good: "computer-assisted-reporters" and "programmer-journalists" in Chicago, New Media Soc. 15 (2012) 853–871, http://dx.doi.org/10.1177/1461444812463345.

[16] A. Hermida, F. Fletcher, D. Korell, D. Logan, SHARE, LIKE, RECOMMEND decoding the social media news consumer, Journal. Stud. 13 (2012) 815–824, http://dx.doi.org/10.1080/1461670x.2012.664430.

[17] P.B. Brandtzaeg, M. Lüders, J. Spangenberg, L. Rath-Wiggins, A. Følstad, Emerging journalistic verification practices concerning social media, Journal. Pr. 10 (2015) 323–342, http://dx.doi.org/10.1080/17512786.2015.1020331.

[18] M. Janssen, Y. Charalabidis, A. Zuiderwijk, Benefits, adoption barriers and myths of open data and open government, Inf. Syst. Manage. 29 (2012) 258–268, http://dx.doi.org/10.1080/10580530.2012.716740.

[19] A. Zuiderwijk, M. Janssen, Open data policies, their implementation and impact: A framework for comparison, Gov. Inf. Q. 31 (2014) 17–29, http://dx.doi.org/10.1016/j.giq.2013.04.003.

[20] V. Lempitsky, A. Zisserman, Learning to count objects in images, Adv. Neural Inf. Process. Syst. 23 (2010).

[21] C. Zhang, H. Li, X. Wang, X. Yang, Cross-scene crowd counting via deep convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 833–841.

[22] X. Cao, Z. Wang, Y. Zhao, F. Su, Scale aggregation network for accurate and efficient crowd counting, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 734–750.

[23] Z. Ma, X. Wei, X. Hong, Y. Gong, Bayesian loss for crowd count estimation with point supervision, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 6142–6151.

[24] J. Wan, W. Luo, B. Wu, A.B. Chan, W. Liu, Residual regression with semantic prior for crowd counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 4036–4045.

[25] M. Broersma, T. Graham, Social Media as Beat: Tweets as a news source during the 2010 British and Dutch elections, Journal. Pr. 6 (2012) 403–419, http://dx.doi.org/10.1080/17512786.2012.663626.

[26] H. Lin, Z. Ma, R. Ji, Y. Wang, X. Hong, Boosting crowd counting via multifaceted attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 19628–19637.

[27] D. Liang, X. Chen, W. Xu, Y. Zhou, X. Bai, TransCrowd: Weakly-supervised crowd counting with transformers, Sci. China Inf. Sci. 65 (2022) 160104, http://dx.doi.org/10.1007/s11432-021-3445-y.

[28] S. Abousamra, M. Hoai, D. Samaras, C. Chen, Localization in the crowd with topological constraints, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 872–881, http://dx.doi.org/10.1609/aaai.v35i2.16170.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Proceedings of the European Conference on Computer Vision, ECCV, Vol. 12346, 2020, pp. 213–229, http://dx.doi.org/10.1007/978-3-030-58452-8_13.

[30] M.-R. Hsieh, Y.-L. Lin, W.H. Hsu, Drone-based object counting by Spatially Regularized Regional proposal network, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2017, pp. 4145–4153.

[31] B. Sam, S. Surya, V. Babu, Switching convolutional neural network for crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 5744–5752.

[32] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, in: Proceedings of the International Conference on Learning Representations, ICLR, 2021, pp. 1–16.

[33] W. Liu, M. Salzmann, P. Fua, Context-aware crowd counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 5099–5108.

[34] J. Gao, W. Lin, B. Zhao, D. Wang, C. Gao, J. Wen, C^3 framework: An open-source PyTorch code for crowd counting, 2019, http://dx.doi.org/10.48550/arxiv.1907.02724, arXiv (Cornell University).

[35] N. Diakopoulos, Automating the News: How Algorithms are Rewriting the Media, Harvard University Press, 2019.

[36] L. Graves, Deciding What's True : the Rise of Political Fact-Checking in American Journalism, Columbia University Press, 2016.

[37] H. Idrees, I. Saleemi, C. Seibert, M. Shah, Multi-source multi-scale counting in extremely dense crowd images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2013, pp. 2547–2554.

[38] D. Ryan, S. Denman, C. Fookes, S. Sridharan, Crowd counting using multiple local features, in: Proceedings of the IEEE International Conference on Digital Image Computing: Techniques and Applications, DICTA, 2009, pp. 81–88, http://dx.doi.org/10.1109/DICTA.2009.22.

[39] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 2881–2890.

[40] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 7794–7803.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, ICLR, 2021, pp. 1–22.

[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 10012–10022.

[43] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, Computational social science, Science 323 (2009) 721–723, http://dx.doi.org/10.1126/science.1167742.

[44] D. Watts, Computational social science: exciting progress and future challenges, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, p. 419, http://dx.doi.org/10.1145/2939672.2945366.

[45] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 743–761, http://dx.doi.org/10.1109/TPAMI.2011.155.

[46] J. Wan, A.B. Chan, Adaptive density map generation for crowd counting, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2019, pp. 1130–1139, http://dx.doi.org/10.1109/iccv.2019.00122.

[47] K. Khan, R.U. Khan, W. Albattah, D. Nayab, A.M. Qamar, S. Habib, M. Islam, Crowd counting using end-to-end semantic image segmentation, Electronics 10 (2021) 1293, http://dx.doi.org/10.3390/electronics10111293.

[48] M. Jaderberg, K. Simonyan, A. Zisserman, k. kavukcuoglu, Spatial transformer networks, in: Advances in Neural Information Processing Systems, NeurIPS, 2015, pp. 2017–2025.

[49] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2298–2304, http://dx.doi.org/10.1109/tpami.2016.2646371.

[50] A. Kar, S. Tulsiani, J. Carreira, J. Malik, Category-specific object reconstruction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1966–1974.

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

[52] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 2117–2125.

[53] Z. Song, S. Zou, W. Zhou, Y. Huang, L. Shao, J. Yuan, X. Gou, W. Jin, Z. Wang, X. Chen, X. Ding, J. Liu, C. Yu, C. Ku, C. Liu, Z. Sun, G. Xu, Y. Wang, X. Zhang, D. Wang, S. Wang, W. Xu, R.C. Davis, H. Shi, Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning, Nat. Commun. 11 (2020) http://dx.doi.org/10.1038/s41467-020-18147-8.

[54] H. Kabir, J. Wu, S. Dahal, T. Joo, N. Garg, Automated estimation of cementitious sorptivity via computer vision, Nat. Commun. 15 (2024) 9935, http://dx.doi.org/10.1038/s41467-024-53993-w.

[55] R.M. Neal, Bayesian Learning for Neural Networks, Springer, 2012.

[56] C. Guo, G. Pleiss, Y. Sun, K. Weinberger, On calibration of modern neural networks, in: Proceedings of the International Conference on Machine Learning, ICML, 2017, pp. 1321–1330.

[57] K.-S. Wong, N.A. Tu, A. Maratkhan, M. Demirci, A privacy-preserving framework for surveillance systems, in: Proceedings of the 2020 10th International Conference on Communication and Network Security (ICCNS)), 2020, pp. 91–98, http://dx.doi.org/10.1145/3442520.3442524.

[58] J. Lynch, Face off: Law enforcement use of face recognition technology, SSRN Electron. J. (2020) http://dx.doi.org/10.2139/ssrn.3909038.

[59] A. Senior, S. Pankanti, A. Hampapur, L. Brown, Y.-L. Tian, A. Ekin, J. Connell, C.F. Shu, M. Lu, Enabling video privacy through computer vision, IEEE Secur. Priv. Mag. 3 (2005) 50–57, http://dx.doi.org/10.1109/msp.2005.65.

[60] N. Diakopoulos, M. Koliska, Algorithmic transparency in the news media, Digit. Journal. 5 (2016) 809–828, http://dx.doi.org/10.1080/21670811.2016.1208053.

[61] B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: Mapping the debate, Big Data Soc. 3 (2016) 1–21, http://dx.doi.org/10.1177/2053951716679679.

[62] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, http://dx.doi.org/10.48550/arXiv.1409.1556, arXiv Preprint.

[63] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, M. Shah, Composition loss for counting, density map estimation and localization in dense crowds, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 532–546.

[64] V. Sindagi, R. Yasarla, V.M.M. Patel, JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 2594–2609, http://dx.doi.org/10.1109/tpami.2020.3035969.

[65] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2017, pp. 1501–1510.

[66] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612, http://dx.doi.org/10.1109/tip.2003.819861.

[67] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 7482–7491.

[68] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.