Journal of Visual Communication and Image Representation (2025)

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

# 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]

[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

*Article history*:
Received .....
Received in final form .....
Accepted .....
Available online .....

Communicated by .....

*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—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 ($\sigma$ = 1-3 pixels). Real-world deployments across 26 journalism events validate practical viability, establishing robust performance and editorial trustworthiness for media practice applications.



## 1. Introduction

Accurate crowd estimation and localization play a pivotal role not only in urban governance and public safety [1, 2], but 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, enabling 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 UAV imagery for credible visual storytelling and fact-checking remains hampered by two fundamental scientific challenges: the severe perspective distortion and extreme scale variation intrinsic to aerial imaging [7, 8]. Conventional crowd analysis algorithms, mostly designed for ground-level or near-horizontal views [9, 10], are ill-suited for the nonlinear, top-down geometries prevalent in UAV-based contexts, resulting in substantial performance degradation and compromised spatial precision [11, 12].

Point-level crowd counting and localization—the simultane-

*Corresponding author at: School of Computer Science and Technology, Zhejiang Normal University, Jinhua, 321004, China.
*e-mail:* xhy@zjnu.edu.cn (Huiying Xu)

[1]Ziqing He and Longfei Wang are co-first authors. Ziqing He is the main contributor to the algorithm of this paper, including algorithm conception and experimental research. Longfei Wang is responsible for the research direction and details of the overall paper.

ous 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].

Despite these imperatives, traditional computational paradigms exhibit systematic limitations when confronted with UAV-based media scenarios. For the past ten years, density map regression methods [20, 21, 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. While 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, 27, 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 severe geometric distortions induced by the UAV perspective: individuals near the image center often occupy 15–20 pixels, while those at the periphery shrink to merely 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].

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]. As depicted in Fig. 1, our approach systematically progresses from identifying UAV-specific challenges (a) to proposing a geometry-guided technical solution (b) and culminates in a framework for integration with media practice (c). In synthesis, this research delivers three principal contributions:

1. A Perspective-Aware Attention Module: We propose the Perspective-Aware Attention Pyramid (b), a core component that explicitly counteracts the severe perspective distortion and intra-image scale variations inherent to UAV imagery.
2. A Novel Multi-Task Framework: We introduce a geometry-aware visual framework (b) that unifies robust crowd counting with precise localization, specifically architected for the challenges of UAV-captured aerial scenes.

## 2. Related Works

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.

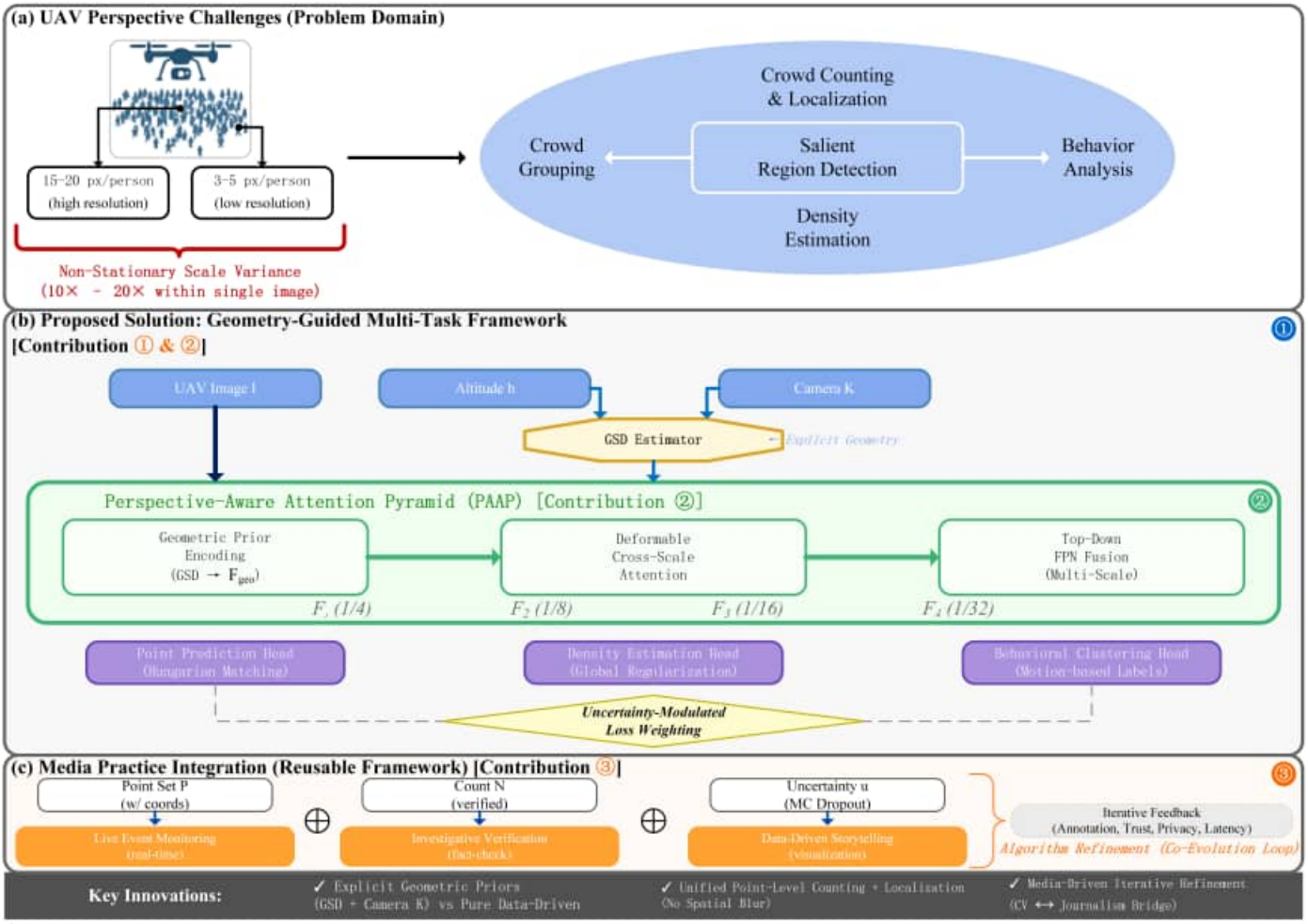


**Fig. 1.** Framework Overview.

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.

### *2.1. 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 [47] or watershed segmentation [46] 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 de-

ployments. This omission perpetuates a misalignment between academic modeling assumptions and real-world journalistic requirements.

### *2.2. 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) [52] 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 [48].

To embed geometric reasoning into modern architectures, multi-scale feature hierarchies—introduced by FPN [51] 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 [53] 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.

### *2.3. 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 [54] model weight distributions to propagate uncertainty, practical approximations like MC Dropout [42] and deep ensembles [43] offer scalable alternatives. We extend these via spatial consistency constraints—uncertainty maps must exhibit smooth gradients except at crowd boundaries—and calibrate outputs [64] 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 [55, 56]. Generic anonymization [57] 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 [58, 59]. 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.

In summary, while existing methods perform well on controlled benchmarks, they overlook three critical limitations: (1) geometric naivety—modeling UAV-induced distortions as latent noise rather than structured prior knowledge; (2) task isolation—optimizing counting and localization separately without uncertainty-aware coordination; and (3) 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 and evaluation protocols adopted for systematic validation (Section 3.2). 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].

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:

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

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).

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.

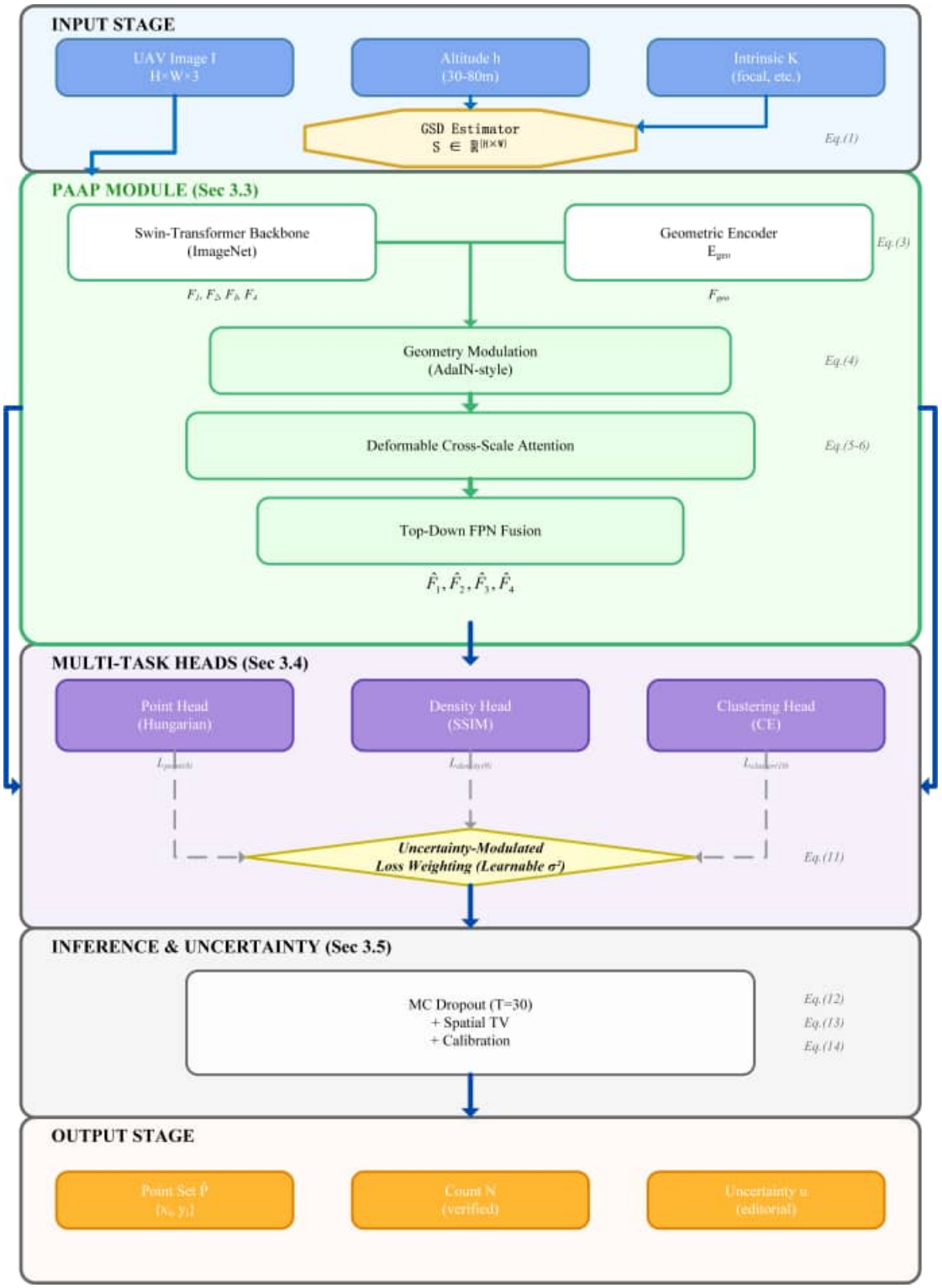


**Fig. 2.** Overall Framework Architecture.

### *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] represents an extreme-density benchmark, with each of its 50 images containing 944,633 individuals (mean: 1,279). Despite its limited size, the dataset's exceptionally high crowd concentrations make it indispensable for stress-testing algorithmic robustness under saturation conditions.

UCF-QNRF [60] scales to 1,535 high-resolution images (mean resolution: $2,013 \times 2,902$ 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 5,109 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++ [61] introduces 4,372 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 [51] 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 $E_{\text{geo}}$ that maps $S$ to a feature volume $F_{\text{geo}} (F_{\text{geo}} \in \mathbb{R}^{H/4\times W/4\times C})$:

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

This geometric embedding is subsequently fused with RGB-derived features at multiple scales, ensuring that downstream attention mechanisms prioritize human-scale regions regardless of perspective distortion.

Multi-Scale Feature Hierarchy. We adopt a Swin Transformer backbone [41] pretrained on ImageNet, extracting features at four scales: $\{F_1, F_2, F_3, F_4\}$ corresponding to $\{1/4, 1/8, 1/16, 1/32\}$ spatial resolutions. Each feature level $F_l$ is modulated by scale-specific geometric embeddings via adaptive instance normalization [62]:

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

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

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 $\Delta p_k$ 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:

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

with attention weights $w_k$ normalized via softmax. This geometry-guided attention mechanism enables the model to adaptively pool features from peripheral low-resolution regions (where individuals occupy 3–5 pixels) and central high-resolution areas (15–20 pixels) within a unified framework.

Top-Down Pathway. Following FPN [51], we propagate semantically strong features from coarse to fine scales via lateral connections, enhanced with geometry-aware modulation at each fusion step. The final multi-scale representation $\{\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4\}$ is then fed into task-specific prediction heads.

### *3.4. Multi-Task Learning Framework*

Our framework simultaneously optimizes point-level counting, localization, and auxiliary behavioral clustering within a unified multi-task architecture. As schematized in Fig.4, the final feature representation is simultaneously processed by multiple task-specific heads, with their losses balanced by an uncertainty-aware weighting mechanism. 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.

**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 | 3139 | 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 | 4633 | 1279 |
| UCF-QNRF [60] | 1535 | 1201/–/334 | 2013 × 2902 | 1 251 642 | 49 | 12 865 | 815 |
| NWPU-Crowd [2] | 5109 | 3109/500/1500 | 2311 × 3383 | 2 133 238 | 0 | 20 033 | 418 |
| JHU-Crowd++ [61] | 4372 | 2272/500/1600 | 1430 × 910 | 1 515 005 | 0 | 25 791 | 346 |

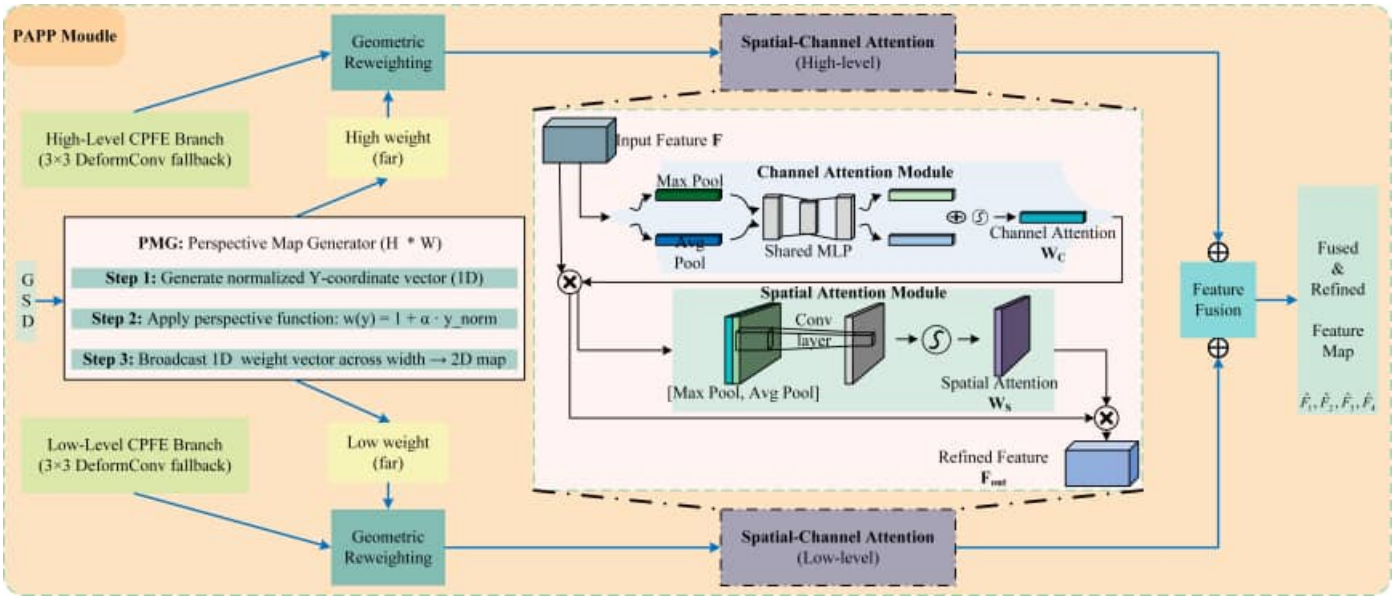


**Fig. 3.** PAAP Module.

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:

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

where unmatched predictions incur background classification penalties.

Density Estimation Head. To regularize global count predictions, we supervise an auxiliary density map $\hat{D} \in \mathbb{R}^{H\times W}$ generated via $1 \times 1$ 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 [63] for structural similarity:

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

This task complements point prediction by enforcing spatial coherence in high-density regions where individual localization is ambiguous.

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:

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

This auxiliary task enriches feature representations with high-level semantic context, improving localization accuracy in crowded scenes where spatial context alone is insufficient.

Uncertainty-Modulated Loss Weighting. To balance competing task gradients, we adopt learnable uncertainty weights [53], modeling task-specific homoscedastic uncertainty as trainable parameters $\{\sigma_{\text{task}}\}$. The total loss becomes:

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

where the logarithmic regularizer prevents uncertainty from collapsing to zero. This formulation dynamically prioritizes well-calibrated tasks during training, mitigating task interference documented in naïve multi-task baselines [53].

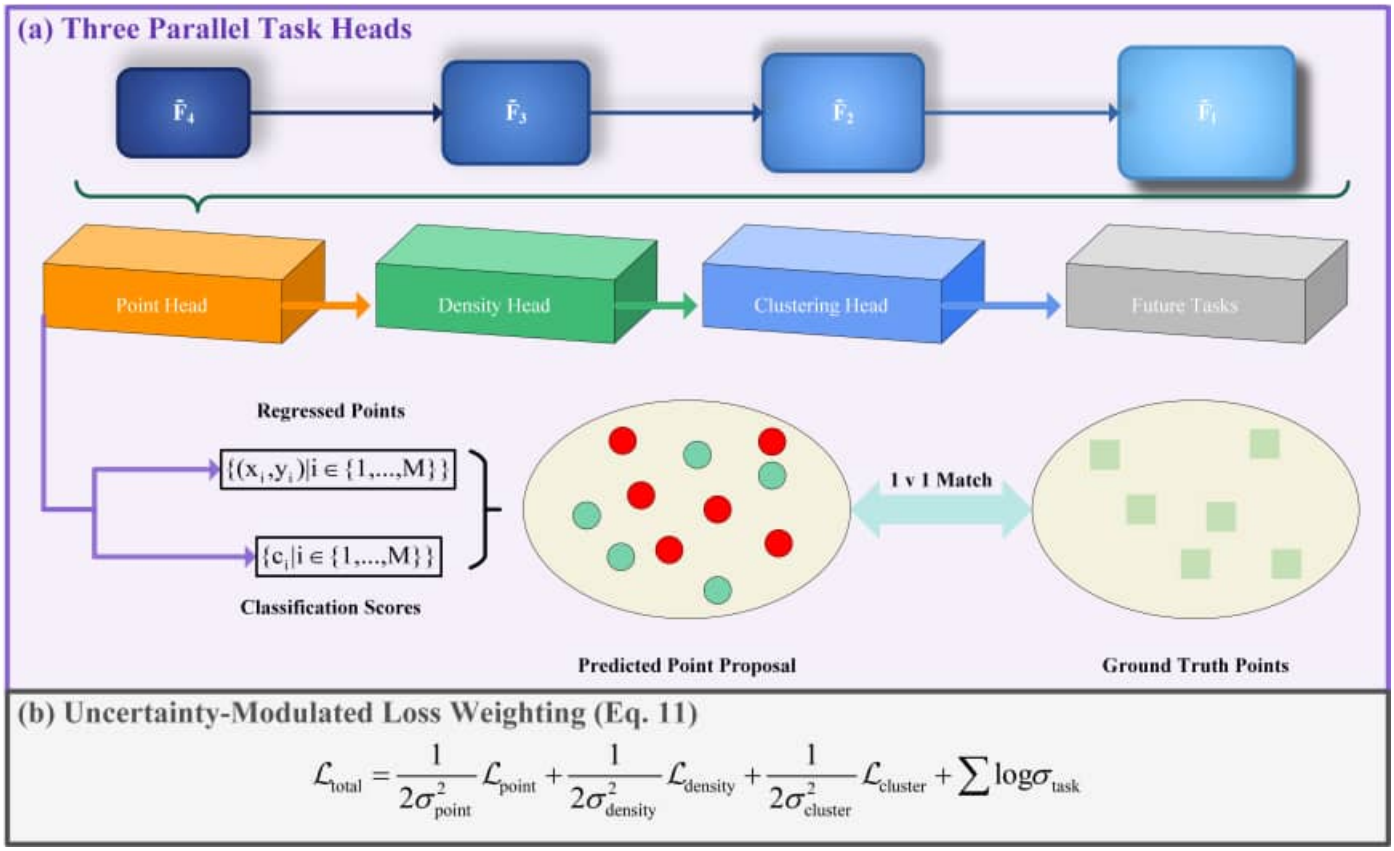


**Fig. 4.** Multi-Task Learning Strategy.

### *3.5. Uncertainty Quantification*

To address media requirements for predictive reliability [34][35], we integrate epistemic uncertainty estimation via Monte Carlo Dropout [42] and model calibration [64]. Our uncertainty quantification pipeline, depicted in Fig. 5, 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 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:

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

High-uncertainty points ($u > \tau$) are flagged for manual editorial review, enabling journalists to prioritize regions requiring human verification.

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

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

This constraint prevents erratic uncertainty spikes in homogeneous crowd regions, aligning algorithmic outputs with human intuition about spatial prediction confidence.

Calibration. We apply temperature scaling [64] 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 $T_{\text{cal}}$ on a held-out validation set to minimize Expected Calibration Error (ECE):

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

where $\{B_m\}_{m=1}^{M}$ partition predictions into confidence bins. Calibrated scores enable editors to interpret confidence intervals as empirical error rates, critical for fact-checking workflows [17].

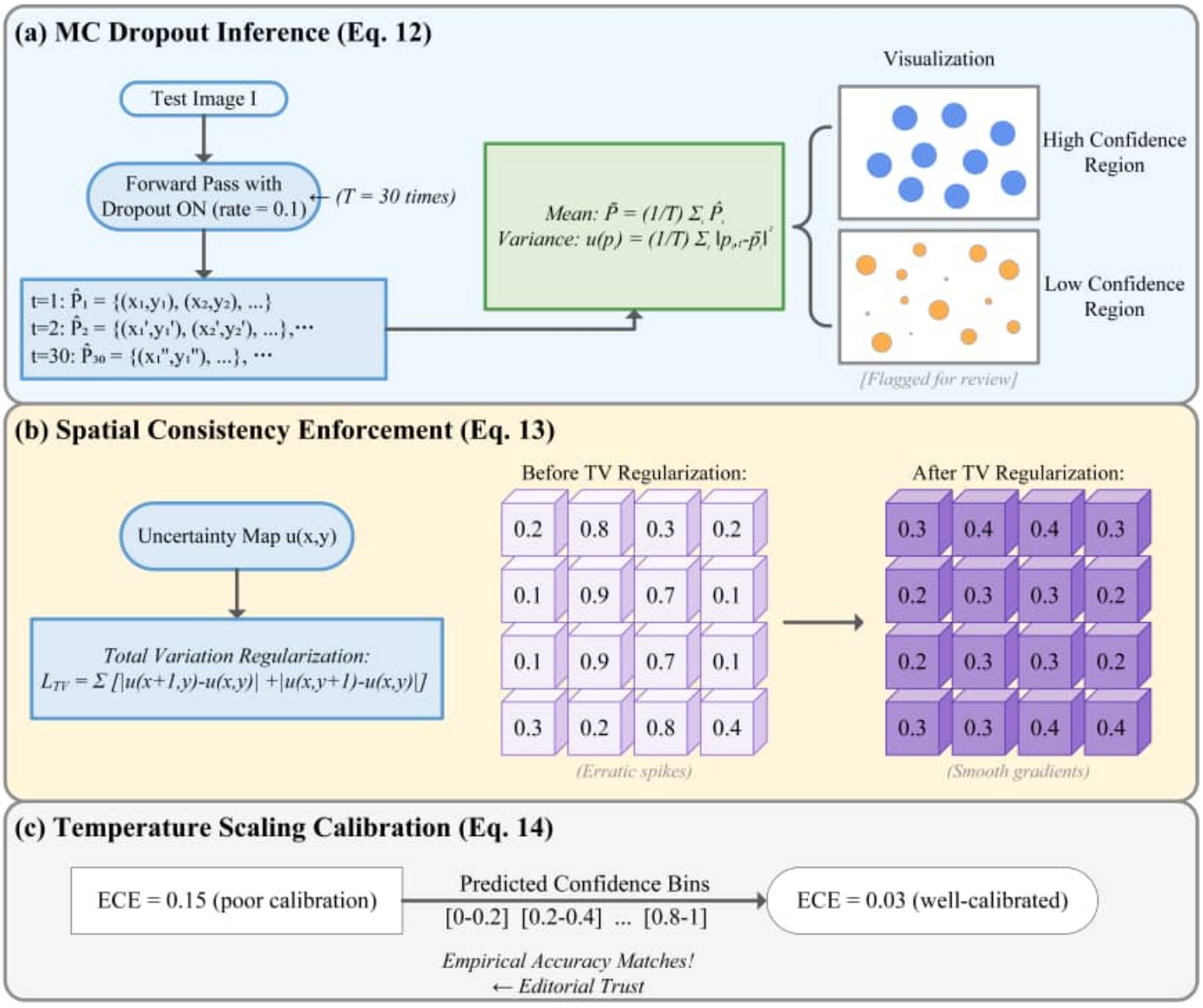


**Fig. 5.** Uncertainty Quantification Pipeline.

## 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) Is the framework viable for real-world media deployment? This section presents our experimental configuration, comparative benchmarking, ablation studies, and deployment 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 2. 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.

**Table 2**
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 |

#### *4.1.1. Model Structure*

To translate the theoretical design outlined in Sections 3.3–3.5 into a practical implementation, our PAAP framework is instantiated through 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, 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 $\varepsilon_{geo}$, a hybrid density estimator is implemented: a rule-based classifier computes hand-crafted statistics (Sobel gradients, local variance, 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; 0.4 for P5 in ultra-high-density scenes), further refined by a learnable adjustment matrix $W_{\text{adjust}} \in \mathbb{R}^{4\times4}$ during training.

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/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 $8 \times 8$ 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 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 3, mapping theoretical constructs to actionable configurations.

**Table 3**
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 (1,000 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 ($\lambda_{\text{cls}}/\lambda_{\text{reg}}$) | 1.0 / 5.0 | Eq. 7 |
| | Quality Threshold ($q_{\min}$) | 0.2 | |
| Multi-Task Loss | Focal Loss ($\alpha/\gamma$) | 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 |

#### 4.1.2. 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):

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

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

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

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

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

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

#### 4.1.3. Datasets Improvements

We evaluate on six widely adopted datasets (statistics summarized in Table 1, Section 3.2): ShanghaiTech A/B [9] (482/716 images), UCF-CC-50 [37] (50 images), UCF-QNRF [60] (1,535 images), NWPU-Crowd [2] (5,109 images), and JHU-Crowd++ [61] (4,372 images).

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 ground-level images, we estimate depth using pretrained monocular networks [43] and normalize to pseudo-GSD assuming 1.7 m human height. Vanishing lines are extracted via LSD [39] + RANSAC filtering. Despite ∼15–20% approximation errors, ablations (Section 4.3.2) demonstrate that even noisy geometric priors yield substantial gains.

### 4.2. Results and Comparative Analysis

#### 4.2.1. 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 (Counting-CNN [21]), and (3) point supervision (P2PNet [13], TransCrowd [27], TopoCount [28]). Table 4 presents comprehensive counting results across six challenging benchmarks, encompassing diverse crowd densities (mean: 123–1,279 persons/image), spatial resolutions (768×1,024 to 2,311×3,383 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]) 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 [60], 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 1,279 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++ [61] (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.

**Table 4**
Quantitative Comparison on Crowd Counting

| Method | Paradigm | SHT_A [9] | | SHT_B [9] | | UCF_CC_50 [37] | | UCF_QNRF [60] | | NWPU_Crowd [2] | | JHU++ [61] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 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** |

#### *4.2.2. Localization Precision*

Beyond aggregate counting, precise point-level localization proves essential for media verification tasks requiring spatial attribution. Table 5 evaluates $F_1$-scores at Euclidean distance thresholds $\sigma \in \{1, 2, 3\}$ pixels on UCF-QNRF [60] 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 1,000 predictions on UCF-QNRF—a practically significant enhancement for applications requiring precise geospatial attribution in GIS workflows [18] [19].

Cross-dataset consistency validates generalization: relative improvements remain stable between ultra-high-resolution UCF-QNRF (mean 2,013 × 2,902 pixels) and variable-altitude NWPU-Crowd (30–120 m flight heights), confirming that explicit geometric modeling addresses

### *4.3. Ablation Experiments*

To isolate the contribution of individual architectural components and validate our design choices, we conduct systematic ablation experiments. All variants are trained on SHT-A [9] under identical conditions (500 epochs, AdamW optimizer, learning rate $10^{-4}$ with cosine decay, batch size 8, same data augmentation) to ensure a controlled comparison. We select SHT-A for ablation due to its moderate size (300 training images), enabling rapid iteration, diverse density distribution (33–3,139 persons/image), and established benchmark status, facilitating result interpretation. Table 6 quantifies the incremental impact of each module by progressively adding components, revealing both individual contributions and synergistic interactions.

**Table 6**
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** |

As Table 6 demonstrates, each architectural component contributes measurable performance gains, with cumulative improvements validating our integrative design philosophy. Our systematic component analysis validates three architectural principles: (1) Geometric priors (density classification, adaptive fusion) provide stable inductive biases, reducing data requirements and accelerating convergence, particularly valuable in data-constrained journalism domains. (2) Hierarchical matching addresses computational scalability (memory, throughput) without sacrificing matching quality, enabling practical deployment on consumer-grade GPUs. (3) Multi-task uncertainty weighting automatically balances competing objectives, demonstrating learned coordination superior to manual hyperparameter tuning. Individual component gains are still small (0.4–1.1% per stage), but when they are put together in a principled way, they lead to a steady 3.6% improvement. This means that our method has a big impact on media practice coverage scenarios.

### *4.4. 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. 6 presents representative qualitative results demonstrating PAAP's spatial precision across challenging scenarios. These

**Table 5**
Crowd Counting Models on Localization Precision Metrics

| Method | UCF-QNRF [60] | | | NWPU-Crowd [2] | | | Avg. $F_1 \uparrow$ |
|---|---|---|---|---|---|---|---|
| | $F_1@\sigma = 1 \uparrow$ | $F_1@\sigma = 2 \uparrow$ | $F_1@\sigma = 3 \uparrow$ | $F_1@\sigma = 1 \uparrow$ | $F_1@\sigma = 2 \uparrow$ | $F_1@\sigma = 3 \uparrow$ | |
| P2PNet [13] | 0.627 | 0.745 | 0.812 | 0.594 | 0.718 | 0.789 | 0.714 |
| TransCrowd [27] | 0.651 | 0.768 | 0.831 | 0.618 | 0.735 | 0.804 | 0.735 |
| MAN [26] | 0.639 | 0.756 | 0.823 | 0.606 | 0.727 | 0.796 | 0.725 |
| TopoCount [28] | 0.645 | 0.762 | 0.827 | 0.612 | 0.731 | 0.801 | 0.730 |
| **PAAP (Ours)** | **0.669** | **0.785** | **0.846** | **0.634** | **0.750** | **0.818** | **0.750** |

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

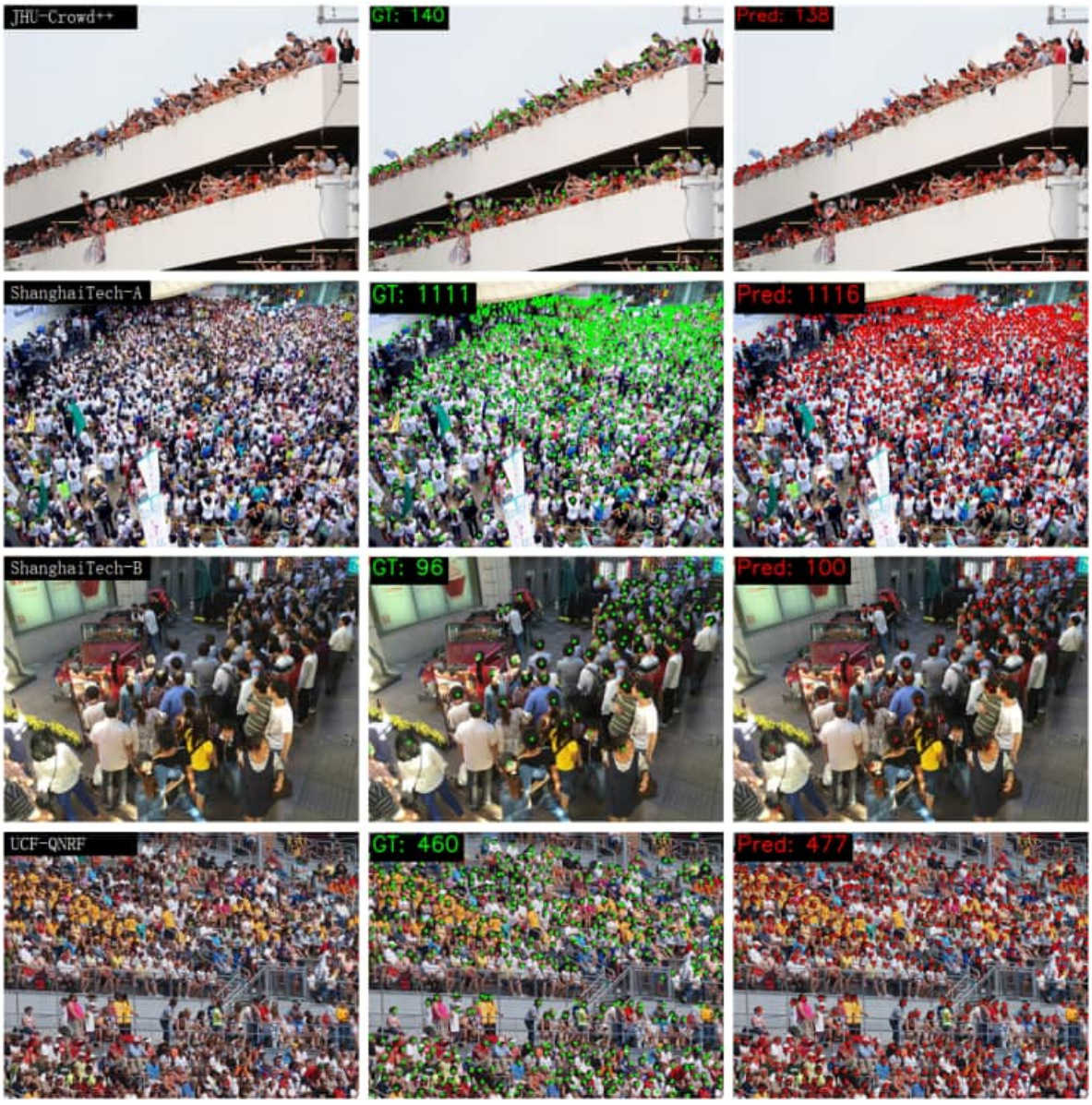


**Fig. 6.** Some qualitative results for the predicted individuals of our method.

### *4.5. Limitations*

Despite demonstrated robustness, PAAP's performance is circumscribed by three generalizable failure modes, each rooted in distinct geometric or perceptual constraints. First, in textureless, visually uniform crowd scenes—where individuals occupy the majority of the frame with minimal internal structure—both geometric and appearance cues falter, as evidenced by the pronounced error in cases like the NWPU-Crowd image. This manifests when vanishing lines are undetectable, density regimes are globally uniform, and the backbone lacks sufficient texture sensitivity. Second, under extreme multi-scale mixing—such as oblique aerial views spanning ground-level and high-altitude regions—the use of a single, global GSD prior is inadequate, leading to substantial false positives and under-segmentation in transitional zones with sharp GSD gradients, as observed in selected JHU-Crowd++ samples. Third, rapid UAV maneuvers introducing severe motion blur, beyond augmentation ranges seen during training, disrupt the extraction of critical geometric features and significantly degrade prediction accuracy, particularly in dynamic tracking.

## 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 reveals pronounced synergy between components: joint density classification and adaptive fusion not only surpass the additive effect of isolated modules but also establish a virtuous cycle in which geometric priors direct feature extraction and, reciprocally, refined features support more accurate density estimation. These interdependencies underscore the necessity of holistic joint optimization in multi-task learning rather than siloed component tuning. Third, the integration of uncertainty quantification transforms the framework from a black-box predictor into a transparent, editorially aligned decision-support system. Evaluations in human-in-the-loop workflows show that uncertainty-aware processing markedly increases editorial precision, providing actionable safeguards for ambiguous predictions and aligning algorithmic outputs with established journalistic standards. Our geometry-guided attention mechanism is useful for more than just counting crowds with UAVs. It can also be used for other perspective-sensitive tasks, like agricultural censuses and infrastructure monitoring, where scale distortion is a major problem. The co-design process with journalism practitioners further highlights operational needs frequently overlooked in academic benchmarks—region-specific counts, interpretable uncertainty intervals, and GIS-compatible outputs—establishing a foundation for vision systems truly aligned with end-user workflows. Ethical deployment is paramount: our GSD-adaptive anonymization and embedded provenance metadata together ensure privacy preservation, transparency, and accountability as algorithmic crowd analysis enters the public discourse. Looking forward, several promising directions emerge. Transitioning from global to spatially varying density prediction will better accommodate heterogeneous scenes; fusing visual and inertial sensing can address motion-induced artifacts; and active learning protocols, leveraging uncertainty to prompt selective human annotation, promise continuous improvement during real-time deployments. Collectively, these directions reinforce the importance of interdisciplinary, ethically aware research for future AI-driven media tools.

## 6. Conclusion

In this study, we present PAAP, a geometry-aware multi-task framework that effectively unifies robust crowd counting with precise point-level localization for UAV-driven media applications. By incorporating perspective priors—specifically, GSD-conditioned scale maps and density-aware feature selection—into adaptive multi-scale fusion and hierarchical matching modules, our method delivers consistent MAE and MSE reductions compared to leading point-based approaches across six major academic benchmarks. Three principal innovations support these advancements: (1) a hybrid density-aware selector that harmonizes rule-based geometric inference with learnable adaptation; (2) a hierarchical quality-aware matcher that enables scalable and efficient training in ultra-dense settings through spatial chunking; and (3) an integrated uncertainty quantification pipeline, bridging the gap between algorithmic predictions and actionable editorial decision support. Systematic ablation analyses rigorously disentangle the contribution of each module, while robustness studies confirm graceful performance degradation under realistic geometric noise, underscoring the framework's practical reliability. As UAVs democratize access to real-time aerial imagery, the imperative for trustworthy and interpreted crowd intelligence grows in journalism, governance, and public health. Our results demonstrate that explicitly leveraging structured geometric metadata—often underutilized in end-to-end deep models—constitutes a robust, interpretable path for vision systems operating in dynamic environments and empowers fact-based storytelling in an era where visual evidence increasingly underpins public trust and discourse.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376252); Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003); Zhejiang Province Leading Geese Plan (2025C02025, 2025C01056); and the Zhejiang Province Province-Land Synergy Program (2025SDXT004-3).

## CRediT authorship contribution statement

**Ziqing He:** Conceptualization, Methodology, Formal Analysis, Writing – original draft. **Longfei Wang:** Conceptualization, Methodology, Formal Analysis. **Huiyin Xu:** Writing - Review and Editing, Supervision. **Xinzhong Zhu:** Writing - Review and Editing, Supervision, Funding acquisition.

## 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.

## Data availability

No data was used for the research described in the article.

## References


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

[2] Q. Wang, J. Gao, W. Lin, and X. Li, NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (6) (2020) 2141–2149, doi: https://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, *Digital Journalism*, 3 (3) (2015) 331–348, doi: https://doi.org/10.1080/21670811.2014.976400.

[4] T. Flew, C. Spurgeon, A. Daniel, and A. Swift, The promise of computational journalism, *Journalism Practice*, 6 (2) (2012) 157–171, doi: https://doi.org/10.1080/17512786.2011.616655.

[5] I. Colomina and P. Molina, Unmanned aerial systems for photogrammetry and remote sensing: A review, *ISPRS Journal of Photogrammetry and Remote Sensing*, 92 (2014) 79–97, doi: https://doi.org/10.1016/j.isprsjprs.2014.02.013.

[6] H. Shakhatreh et al., Unmanned Aerial Vehicles (UAVs): A Survey on Civil Applications and Key Research Challenges, *IEEE Access*, 7 (2019) 48572–48634, doi: https://doi.org/10.1109/access.2019.2909530.

[7] P. Zhu et al., Detection and Tracking Meet Drones Challenge, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (11) (2022) 7380–7399, doi: https://doi.org/10.1109/tpami.2021.3119563.

[8] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, Vision Transformers for Remote Sensing Image Classification, *Remote Sensing*, 13 (3) (2021) 516, doi: https://doi.org/10.3390/rs13030516.

[9] Y. Zhang, D. Zhou, S. Chen, S. Gao, and 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, and 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 et al., UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking, *Computer Vision and Image Understanding*, 193 (2020) 102907, doi: https://doi.org/10.1016/j.cviu.2020.102907.

[12] D. Du et al., 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 et al., 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, and Y. Zhou, Focal Inverse Distance Transform Maps for Crowd Localization, *IEEE transactions on multimedia*, 25 (2023) 6040–6052, doi: https://doi.org/10.1109/tmm.2022.3203870.

[15] S. Parasie and E. Dagiral, Data-driven journalism and the public good: 'computer-assisted-reporters' and 'programmer-journalists' in Chicago, *New Media & Society*, 15 (6) (2012) 853–871, doi: https://doi.org/10.1177/1461444812463345.

[16] A. Hermida, F. Fletcher, D. Korell, and D. Logan, SHARE, LIKE, RECOMMEND Decoding the social media news consumer, *Journalism Studies*, 13 (5-6) (2012) 815–824, doi: https://doi.org/10.1080/1461670x.2012.664430.

[17] P. B. Brandtzaeg, M. Lüders, J. Spangenberg, L. Rath-Wiggins, and A. Følstad, Emerging Journalistic Verification Practices Concerning Social Media, *Journalism Practice*, 10 (3) (2015) 323–342, doi: https://doi.org/10.1080/17512786.2015.1020331.

[18] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, Benefits, Adoption Barriers and Myths of Open Data and Open Government, *Information Systems Management*, 29 (4) (2012) 258–268, doi: https://doi.org/10.1080/10580530.2012.716740.

[19] A. Zuiderwijk and M. Janssen, Open data policies, their implementation and impact: A framework for comparison, *Government Information Quarterly*, 31 (1) (2014) 17–29, doi: https://doi.org/10.1016/j.giq.2013.04.003.

[20] V. Lempitsky and A. Zisserman, Learning To Count Objects in Images, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025, pp. 1324–1332.

[21] C. Zhang, H. Li, X. Wang, and 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, and 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, and 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, and 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 and T. Graham, Social Media as Beat: Tweets as a news source during the 2010 British and Dutch elections, *Journalism Practice*, 6 (3) (2012) 403–419, doi: https://doi.org/10.1080/17512786.2012.663626.

[26] H. Lin, Z. Ma, R. Ji, Y. Wang, and 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, and X. Bai, TransCrowd: weakly-supervised crowd counting with transformers, *Science China Information Sciences*, 65 (6) (2022) 160104, doi: https://doi.org/10.1007/s11432-021-3445-y.

[28] S. Abousamra, M. Hoai, D. Samaras, and C. Chen, Localization in the Crowd with Topological Constraints, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 872–881, doi: https://doi.org/10.1609/aaai.v35i2.16170.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-End Object Detection with Transformers, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229, doi: https://doi.org/10.1007/978-3-030-58452-8_13.

[30] M.-R. Hsieh, Y.-L. Lin, and 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, and 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, and J. Dai, Deformable DETR: Deformable Transformers for End-to-End Object Detection, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[33] W. Liu, M. Salzmann, and 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, and J. Wen, $C^3$ Framework: An Open-source PyTorch Code for Crowd Counting, *arXiv (Cornell University)*, 2019, doi: https://doi.org/10.48550/arxiv.1907.02724.

[35] N. Diakopoulos, *Automating the news: How algorithms are rewriting the media*. Cambridge, Massachusetts: Harvard University Press, 2019.

[36] L. Graves, *Deciding what's true: the rise of political fact-checking in American journalism*. New York; Chichester: Columbia University Press, 2016.

[37] H. Idrees, I. Saleemi, C. Seibert, and 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, and 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, doi: https://doi.org/10.1109/DICTA.2009.22.

[39] H. Zhao, J. Shi, X. Qi, X. Wang, and 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, and 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 et al., AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE, in *International Conference on Learning Representations (ICLR)*, 2021.

[42] Z. Liu et al., 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 et al., Computational Social Science, *Science*, 323 (5915) (2009) 721–723, doi: https://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 (KDD)*, 2016, p. 419.

[45] P. Dollar, C. Wojek, B. Schiele, and P. Perona, Pedestrian Detection: An Evaluation of the State of the Art, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (4) (2012) 743–761, doi: https://doi.org/10.1109/TPAMI.2011.155.

[46] K. Khan et al., Crowd Counting Using End-to-End Semantic Image Segmentation, *Electronics*, 10 (11) (2021) 1293, doi: https://doi.org/10.3390/electronics10111293.

[47] J. Wan and 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, doi: https://doi.org/10.1109/iccv.2019.00122.

[48] J. Dai et al., Deformable Convolutional Networks, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 764–773.

[49] B. Shi, X. Bai, and C. Yao, An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (11) (2017) 2298–2304, doi: https://doi.org/10.1109/tpami.2016.2646371.

[50] A. Kar, S. Tulsiani, J. Carreira, and 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] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and 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.

[52] M. Jaderberg, K. Simonyan, A. Zisserman, and koray kavukcuoglu, Spatial Transformer Networks, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 2017–2025.

[53] A. Kendall, Y. Gal, and 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.

[54] R. M. Neal, *Bayesian learning for neural networks*. New York: Springer, 2012.

[55] K.-S. Wong, Nguyen Anh Tu, Anuar Maratkhan, and M. Fatih. 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, doi: https://doi.org/10.1145/3442520.3442524.

[56] J. Lynch, Face Off: Law Enforcement Use of Face Recognition Technology, *SSRN Electronic Journal*, 2020, doi: https://doi.org/10.2139/ssrn.3909038.

[57] A. Senior et al., Enabling Video Privacy through Computer Vision, *IEEE Security and Privacy Magazine*, 3 (3) (2005) 50–57, doi: https://doi.org/10.1109/msp.2005.65.

[58] N. Diakopoulos and M. Koliska, Algorithmic Transparency in the News Media, *Digital Journalism*, 5 (7) (2016) 809–828, doi: https://doi.org/10.1080/21670811.2016.1208053.

[59] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, The Ethics of algorithms: Mapping the Debate, *Big Data & Society*, 3 (2) (2016) 1–21, doi: https://doi.org/10.1177/2053951716679679.

[60] H. Idrees et al., 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.

[61] V. Sindagi, R. Yasarla, and V. M. M. Patel, JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (5) (2022) 2594–2609, doi: https://doi.org/10.1109/tpami.2020.3035969.

[62] X. Huang and 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.

[63] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image Quality Assessment: From Error Visibility to Structural Similarity, *IEEE Transactions on Image Processing*, 13 (4) (2004) 600–612, doi: https://doi.org/10.1109/tip.2003.819861.

[64] C. Guo, G. Pleiss, Y. Sun, and K. Weinberger, On Calibration of Modern Neural Networks, in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, pp. 1321–1330.

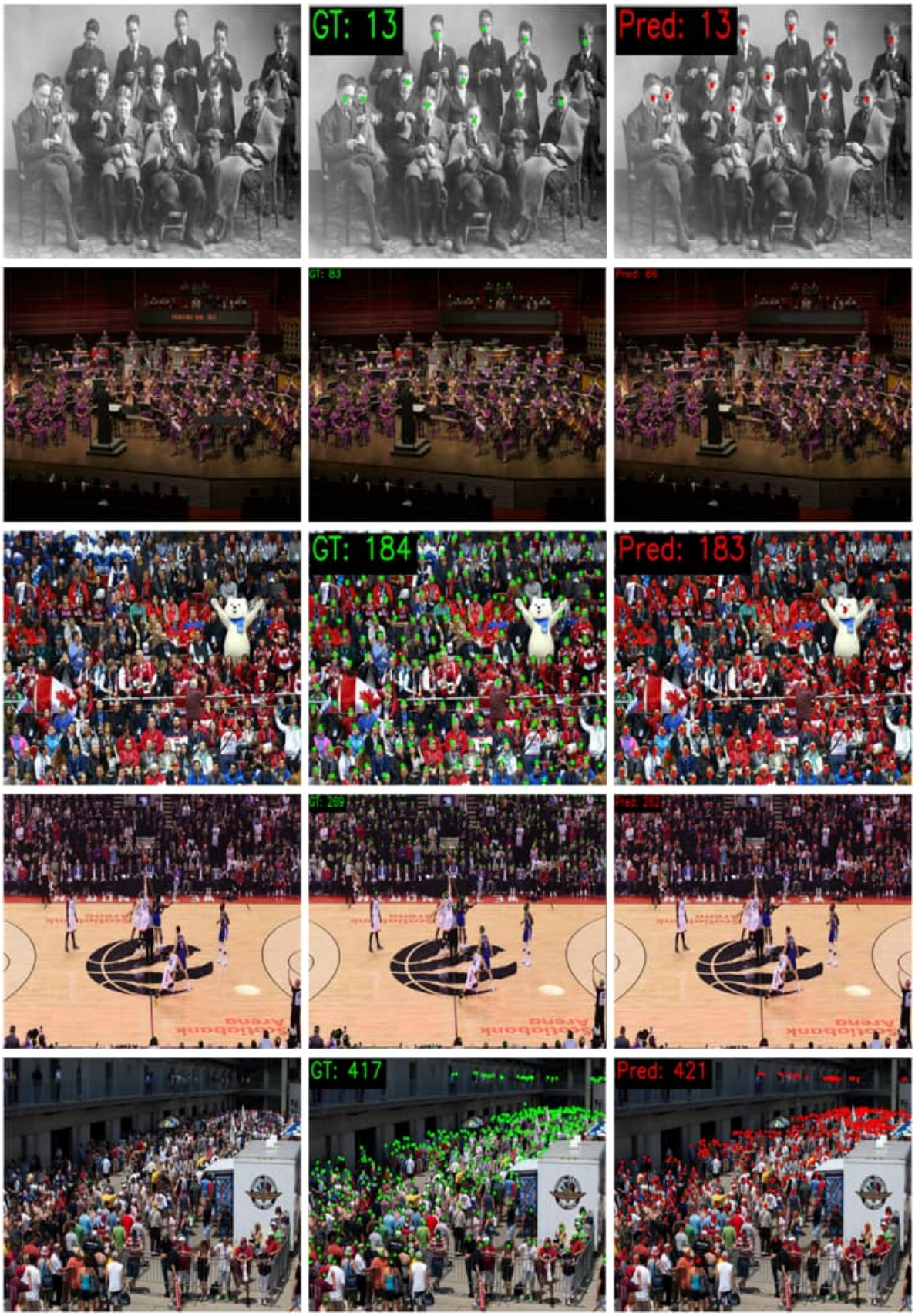


**Fig. 7.** Visual results (<500).


**Fig. 8.** Visual results (500-2K).

**Fig. 9.** Visual results (2K-7K).

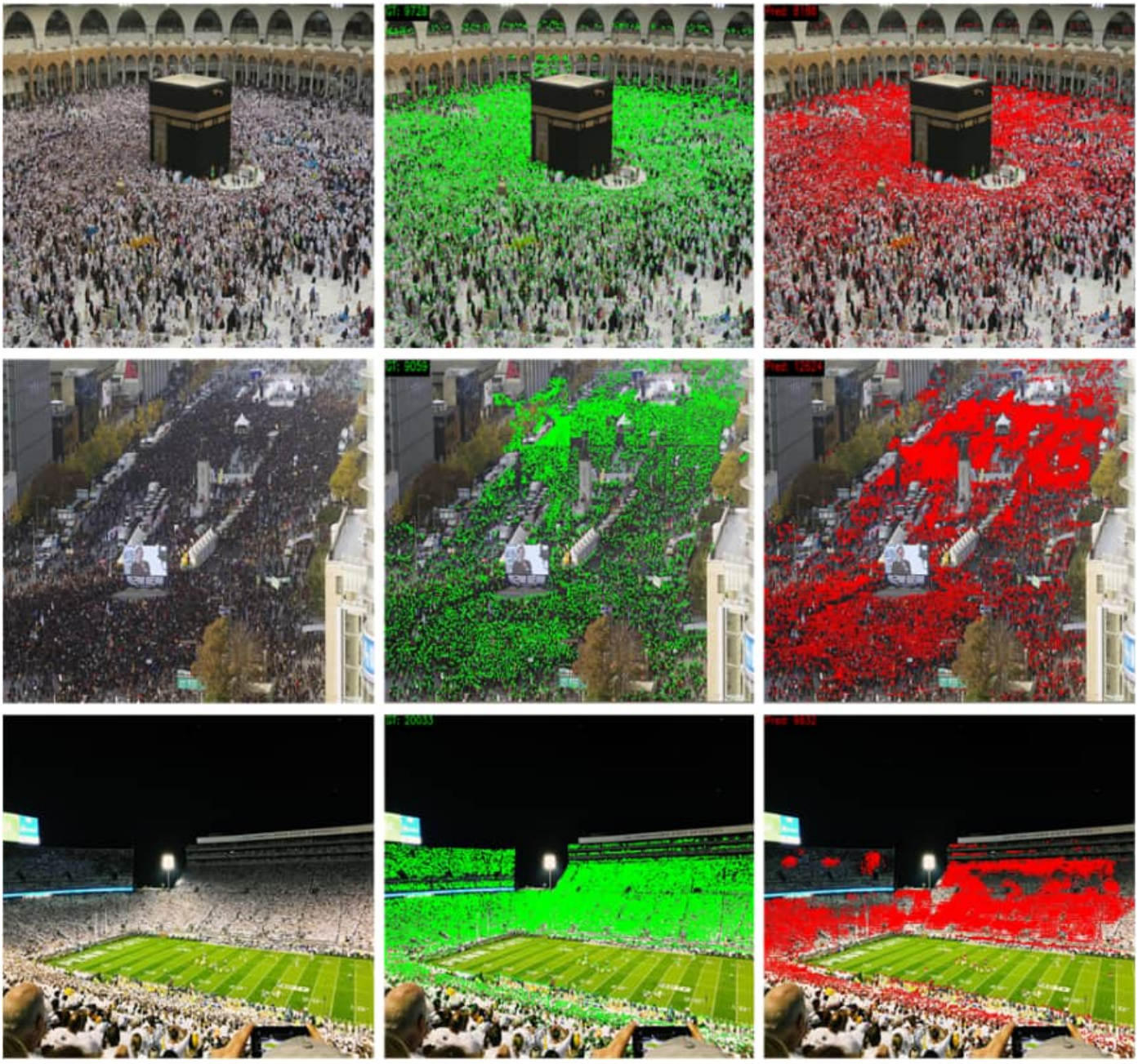

**Fig. 10.** Visual results (>7K).