

Media Practice-Driven Technology Iteration and Value Realization: A Framework Study on Geometry-Guided UAV Crowd Counting

1st Ziqing He

School of Media Practice

The University of Sydney

Sydney, NSW 2050, Australia

Hangzhou Institute of Artificial Intelligence

Zhejiang Normal University

Hangzhou, Zhejiang, China

13845457268@163.com

2nd Longfei Wang

School of Computer Science and Technology

Zhejiang Normal University

Jinhua, Zhejiang, China

flying252019@zjnu.edu.cn

3rd Huiying Xu*

School of Computer Science and Tech.

Zhejiang Normal University

Jinhua, Zhejiang, China

xhy@zjnu.edu.cn

4th Xinzhang Zhu

School of Computer Science and Tech.

Zhejiang Normal University

Jinhua, Zhejiang, China

Hangzhou Institute of Artificial Intelligence

Zhejiang Normal University

Hangzhou, Zhejiang, China

zxz@zjnu.edu.cn

5th Jiaxiao Xiong

School of Computer Science and Tech.

Zhejiang Normal University

Jinhua, Zhejiang, China

2585665322@qq.com

Abstract—In the pursuit of compelling, data-driven narratives, media communication increasingly leverages Unmanned Aerial Vehicles (UAVs) to capture large-scale events. However, realizing the full potential of this aerial perspective—transforming raw footage into verifiable, journalistic assets—is critically hampered by severe perspective distortion and scale variation inherent in the imagery. To bridge this media-technology gap, we developed the Geometry-Guided Point-to-Point Network (G²P²-Net). This framework is not merely a technical exercise but a direct answer to the media’s need for accuracy, synergizing physical principles with deep learning. Its core Perspective-Aware Attention Pyramid (PAAP) embeds a camera-derived perspective map to counteract scale distortion, markedly improving counting precision and enabling reliable data journalism. The deployment of this accurate tool enables our primary contribution: a media practice-driven framework for technology iteration. This symbiotic loop sees our algorithm transforming crowd data into powerful visual narratives, while diverse media deployments—from covering rural markets to festivals—provide rich, in-the-wild validation that fuels continuous technological evolution. This paper’s contribution is therefore twofold: 1) an advanced, geometry-aware algorithm that empowers data-driven aerial storytelling, and 2) a replicable model demonstrating how media needs can steer and accelerate technological innovation. Our work pioneers a synergistic ecosystem uniting cutting-edge technology with the future of media communication.

Index Terms—UAV Crowd Counting, Media Practice, Geometric Priors, Technology Iteration, Synergistic Ecosystem

I. INTRODUCTION

The capacity for accurate crowd estimation, long recognized as a critical requirement in urban governance and public safety [1], is now pivotal for data-driven journalism’s pursuit of verifiable narratives. Unmanned Aerial Vehicles (UAVs) have emerged as a transformative tool in this endeavor [2] [3], granting an unparalleled aerial perspective to document the scale and dynamics of large-scale events, from public festivals to civic demonstrations [4]. Yet, the promise of translating this raw aerial footage into credible narrative assets is fundamentally challenged by the unique physics of aerial imaging. The top-down or oblique geometry inherent to UAVs introduces severe perspective distortion and scale variation, creating a significant domain gap that renders conventional perception algorithms largely ineffectual in this distinct aerial domain [5].

This domain shift fundamentally challenges the established suite of crowd-counting algorithms, which are largely engineered under the assumption of fixed, near-horizontal viewpoints [6]. Consequently, their application to this emergent media context is fraught with limitations. Early detection-based methods, while effective in sparse settings, falter catastrophically amid the dense occlusions characteristic of large-scale public events [7]. In response, density map regression became the dominant paradigm for over a decade; however, it is plagued by an intrinsic blurring artifact that precludes the

*Corresponding author.

precise localization essential for verifiable journalistic claims [8] [9]. Even recent point-based methods—a promising step toward unifying counting and localization [10] [11]—exhibit marked performance degradation when subjected to the severe geometric distortions of aerial imagery, where their underlying viewpoint assumptions are violated [12].

The intractability of this challenge is rooted in the fundamental physics of UAV-based imaging, namely the severe perspective distortion and scale variance induced by fluctuations in flight altitude [13]. A single aerial frame can exhibit a pronounced, nonlinear scale gradient absent in ground-level surveillance, where individuals at the nadir may span dozens of pixels while those at the periphery diminish to a handful [14]. Conventional multi-scale architectures, such as Feature Pyramid Networks (FPNs), attempt to model this variance; however, their data-driven nature limits their efficacy, as they struggle to infer the underlying geometric projection rules from finite training samples [15]. The work of Zhao et al. [16], for example, leverages sophisticated feature engineering to this end, yet inadvertently illuminates a core limitation: purely data-driven methods lack the explicit physical reasoning required to master aerial geometry. This underscores a critical need, driven by the demands of media practice, for a more principled paradigm that integrates such geometric priors directly into the learning process [17].

To bridge this emergent media-technology gap, we propose the Geometry-Guided Point-to-Point Network (G²P²-Net), a framework conceived as a direct response to the media’s need for reliable aerial data journalism. Architected around the Perspective-Aware Attention Pyramid (PAAP), our network performs explicit spatial correction at the feature level by generating perspective weight maps from camera parameters, thereby endowing the model with a priori geometric awareness to counteract scale distortion. This core mechanism is further fortified by an adaptive receptive field backbone and a context-aware prediction head to ensure robustness against complex crowd morphologies [18]. The efficacy of this integrated design is empirically validated by its superior performance on public benchmarks [19], providing the technical foundation for our primary conceptual contribution: a replicable model for media practice-driven technology iteration.

Accordingly, the contribution of this paper is twofold. First, we present an advanced, geometry-aware algorithm that empowers reliable, data-driven aerial storytelling. Second, we propose a replicable framework that delineates how the real-world demands of media practice can steer and accelerate technological innovation. In doing so, this work moves beyond offering a mere tool for static analysis; it pioneers a synergistic, closed-loop ecosystem uniting computer vision with the future of media communication, thereby laying the groundwork for subsequent research into dynamic, motion-aware applications [20].

II. RELATED WORKS

The evolution of crowd counting methodologies has followed a clear trajectory, marked by three dominant paradigms.

Early research primarily focused on detection-based methods [21], which excel in sparse scenes. Subsequently, to address the severe occlusion in dense crowds, density map-based regression emerged as the mainstream paradigm and has dominated the field for nearly a decade [6]. Recently, point-based methods, which aim to unify counting and localization, have become the new research frontier [10]. This section provides a systematic review of these three technical routes.

A. Detection-Based Methods

This line of work formulates crowd counting as a standard object detection task, aiming to localize and identify every individual. Early approaches relied on handcrafted features, while the advent of deep learning led to the adoption of advanced, general-purpose detectors, such as Faster R-CNN [22] and YOLO [23], which significantly improved performance. For instance, Liu et al. [7] employed a curriculum learning strategy to progressively optimize the model progressively, enhancing bounding box prediction accuracy in complex scenes. However, these methods are inherently limited by severe occlusion. In high-density crowds, extensive overlapping among individuals causes a sharp decline in detector performance. Furthermore, providing precise bounding box annotations for these regions is extremely challenging, which in turn hampers effective model supervision and training.

B. Density Map-Based Methods

Density map-based methods represent a significant evolution from detection-based approaches, specifically designed to overcome their limitations in densely crowded scenes. This paradigm shifts the focus from direct instance localization to regressing a spatial density map, wherein the integral over the mapped area yields the total count. The foundational work by Lempitsky and Zisserman [24] introduced this concept by applying Gaussian kernels to point annotations, thereby generating continuous density maps as supervisory signals. By circumventing explicit detection of individuals, this method effectively mitigates issues caused by severe occlusion, establishing density regression as the dominant strategy in high-density crowd counting. Subsequent research has extended this framework through various architectural innovations. For instance, MCNN [8] utilizes a multi-column structure to handle scale variations, while CSRNet [25] incorporates dilated convolutions to expand the receptive field and enhance contextual reasoning. SANet [9] further advances this line of work with a sparse aggregation module for more efficient feature representation. Later improvements have integrated multi-scale feature fusion [15], attention mechanisms [16], and refined network designs [17], collectively boosting robustness and accuracy. Despite these advancements, the density map paradigm suffers from two inherent limitations. First, the Gaussian smoothing process inherently sacrifices precise spatial information, restricting its applicability to tasks requiring accurate localization, such as tracking. Second, the kernel size—typically set either heuristically or via a separate pre-

diction network—introduces either manual bias or additional complexity, which can adversely affect counting performance.

C. Point-Based Methods

To address the inherent localization ambiguities of conventional density map-based techniques, point-based methods have arisen as a powerful alternative. Instead of estimating crowd density, these methods directly predict the explicit coordinates of each person’s head center, eliminating the intermediary steps involved in generating density maps or bounding boxes. This end-to-end, image-to-points pipeline offers a significant advantage by mirroring natural human annotation processes, enabling both accurate head counts and precise spatial localization while providing a reliable basis for higher-level tasks. Tracing their lineage back to the seminal work of Song et al. [10], who established P2P-Net as a definitive benchmark for direct point supervision, the field has witnessed a rapid succession of architectural innovations. Significant progress includes the introduction of a fuzzy generation module in FGNet [26], the application of transformer architectures for superior feature encoding in CLTR [27], the use of attention priors to better model spatial context in APGCC [28], and, most recently, the integration of state space models for dynamic scene understanding in VMambaCC [29] and VMamba-Crowd [30]. Through these continuous advancements in network design, loss formulation, and training methodologies, the boundaries of performance in this domain are constantly being pushed forward.

A critical limitation common to all three paradigms is their implicit design assumption: the input imagery originates from a near-horizontal, ground-level perspective. When applied to UAV-captured aerial imagery, these models suffer from significant performance degradation due to the substantial domain shift. They lack an explicit mechanism to handle the systematic geometric distortions and extreme scale variations introduced by the top-down or oblique viewing angles. Addressing this fundamental challenge is the primary focus of our work.

III. OUR WORKS

The architecture of our proposed G²P²-Net embodies a first-principles approach to resolving the challenges that impede reliable aerial data journalism. As depicted in Fig. 1, the framework channels information through a pipeline of three synergistic stages. A robust backbone network first processes the visual complexity inherent in real-world media-gathering scenarios. The novel Perspective-Aware Attention Pyramid (PAAP), the core of our innovation, then embeds a priori geometric knowledge to rectify the feature representation, restoring a measure of physical truth distorted by the aerial viewpoint. Finally, a context-aware prediction head yields the precise coordinates of each individual, furnishing the granular data essential for verifiable analysis. This principled pipeline—from raw feature processing to geometrically-rectified representation—culminates in a prediction that is not only accurate but also physically grounded, thereby meeting the stringent demands of data-driven journalism.

A. Perspective-Aware Attention Pyramid

The cornerstone of G²P²-Net is the PAAP, a novel module designed to explicitly model and compensate for the severe scale variations induced by aerial perspectives. Unlike conventional data-driven approaches that passively learn scale invariance, the PAAP module deeply integrates geometric priors—derived from camera parameters such as UAV altitude, pose, and focal length—into the feature learning process. As illustrated in Fig. 2, its mechanism unfolds in two key steps.

First, the module leverages the known camera geometry to generate an initial perspective map. This map quantifies the expected scale factor at each pixel location, effectively encoding the geometric projection from the 3D ground plane to the 2D image plane. To enhance robustness against noisy or imprecise camera parameters, this initial map is then refined by a lightweight, self-correcting sub-network. This sub-network learns to adjust the geometric priors by correlating them with semantic features from the image, producing a more accurate and context-aware perspective map. Each value in this map serves as a pixel-wise weight, representing the degree of spatial compression or expansion due to perspective effects. This map provides the network with explicit, physically grounded knowledge of the imaging process.

Second, this refined Perspective Map is employed to spatially recalibrate feature maps extracted from an FPN backbone [15]. Rather than using simple element-wise multiplication, we introduce a dynamic attention mechanism. This mechanism treats the perspective map not as a static multiplier, but as a conditional input to a small gating network. For each spatial location, the gate learns a non-linear recalibration function that modulates the feature activations based on both the geometric scale factor and the local feature content. This sophisticated modulation selectively amplifies features corresponding to distant, smaller individuals and attenuates those from closer, larger individuals with greater precision. This active spatial correction ensures that features across all scales are normalized with respect to their geometric context before fusion. The subsequent bottom-up fusion within the FPN structure is thus guided by this more nuanced geometric awareness, enabling the network to achieve enhanced spatial consistency and robustly handle extreme scale variations in a principled, rather than purely empirical, manner. This adaptive design endows the network with superior geometric generalization, resilience to imperfect camera data, and interpretability.

B. Synergistic Backbone and Prediction Head

To ensure robust feature representation and precise localization, particularly in densely packed and occluded scenes, we designed a highly synergistic backbone and prediction head. This subsystem is engineered to effectively capture complex crowd morphologies and enhance feature discriminability, as shown in Fig. 3.

The backbone network is built upon a multi-scale feature extractor that employs dilated (atrous) convolutions [31] with varying dilation rates (e.g., 1, 3, 5, 7). This design creates an adaptive receptive field, allowing the network to capture

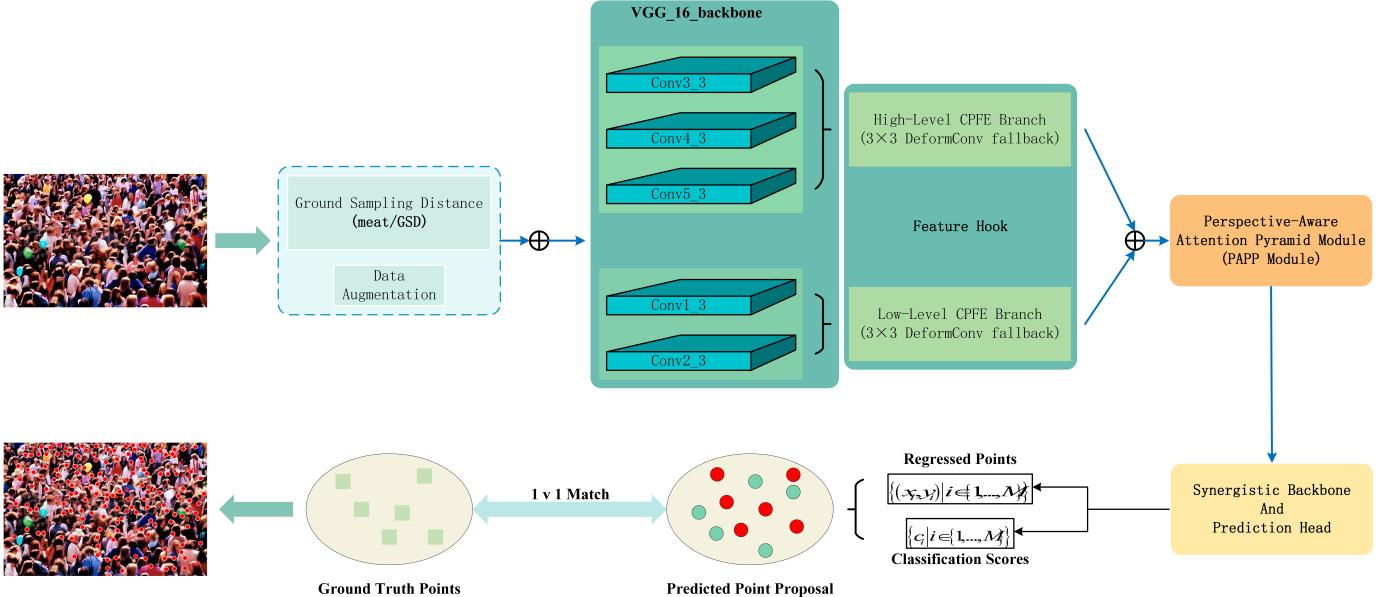


Fig. 1. The overall architecture of G²P²-Net.

contextual information over multiple scales without sacrificing spatial resolution. It proves highly effective for distinguishing individuals across a wide spectrum of crowd densities, from sparse peripheral regions to dense central clusters, while mitigating interference from background clutter.

To further enhance the extracted features, a dual-stream attention mechanism—comprising spatial and channel branches—is introduced [32]. The Spatial Attention Module (SAM) emphasizes salient spatial regions associated with crowd aggregations while suppressing non-essential background areas. In parallel, the Channel Attention Module (CAM) adaptively recalibrates channel-wise feature responses to prioritize those most relevant for head localization. Through the synergistic operation of SAM and CAM, the network effectively concentrates its representational capacity on discerning “what” and “where” matters most for accurate prediction.

Finally, the architecture incorporates multi-level residual connections [33], which facilitate unimpeded gradient flow throughout the deep network. This not only accelerates convergence during training but also allows the network to learn more effective feature representations by adaptively combining features from different depths. The refined feature maps are then fed into the prediction head, which performs the final regression to output the precise coordinates of each detected individual, ensuring both high accuracy in counting and fidelity in localization.

IV. EXPERIMENTS

A. Implementation Details

Data Preparation and Augmentation. To ensure a comprehensive evaluation, we conduct extensive experiments on six challenging crowd counting datasets: ShanghaiTech Part A/B [34], UCF_CC_50 [35], UCF-QNRF [36], NWPU-Crowd

[37], and JHU-Crowd [38]. These benchmarks collectively cover diverse challenges, including extreme crowd densities, varied scene types, perspective changes, and complex environmental conditions, thereby thoroughly assessing the robustness and generalization ability of our method for UAV-based aerial surveillance. All input images are uniformly resized and cropped to 224×224 pixels. During training, we apply standard online data augmentation—random horizontal flipping, random scaling (0.5–1.3×), and color jittering—with all spatial transformations applied consistently to both images and point annotations to maintain geometric alignment. For datasets providing camera parameters, we compute the Ground Sampling Distance (GSD) per image and feed it directly into our Perspective-Aware Attention Pyramid module to explicitly incorporate physical scene scale.

Training and Environment Configuration. All experiments were implemented in PyTorch [39] and conducted on a server with four NVIDIA GeForce RTX 3080 GPUs running Ubuntu 20.04. We trained our models using the AdamW optimizer [40] with a weight decay of 1e-4, applying a cosine annealing learning rate schedule following a 5-epoch linear warm-up. A differential learning rate strategy was adopted: the prediction head used a base rate of 1e-3, while the pre-trained FPN backbone was fine-tuned at 1e-5. The overall loss combines Focal Loss [41] for classification to handle severe foreground-background imbalance and L1 loss for point localization. The Hungarian algorithm [42] was employed to establish one-to-one matching between predictions and ground-truth points, enabling end-to-end optimization.

Evaluation Metrics. To ensure a comprehensive and multi-faceted evaluation, we assess model performance using standard metrics for both counting and localization tasks.

For counting performance, we report the two most widely

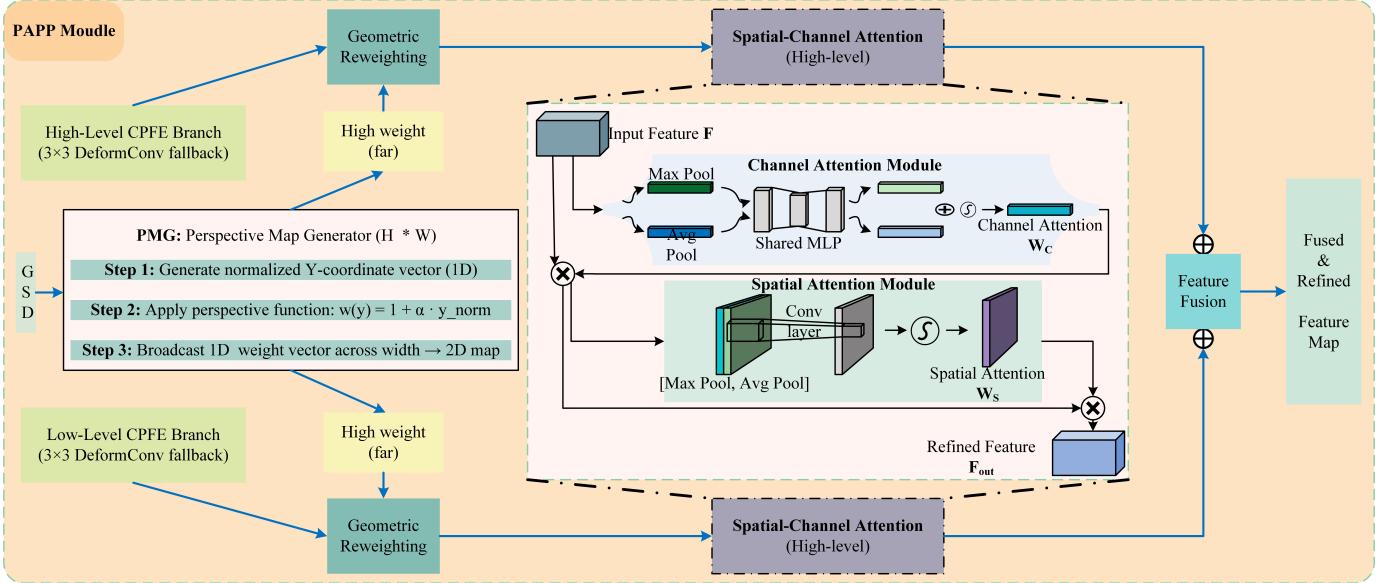


Fig. 2. Detailed architecture of the PAAP module.

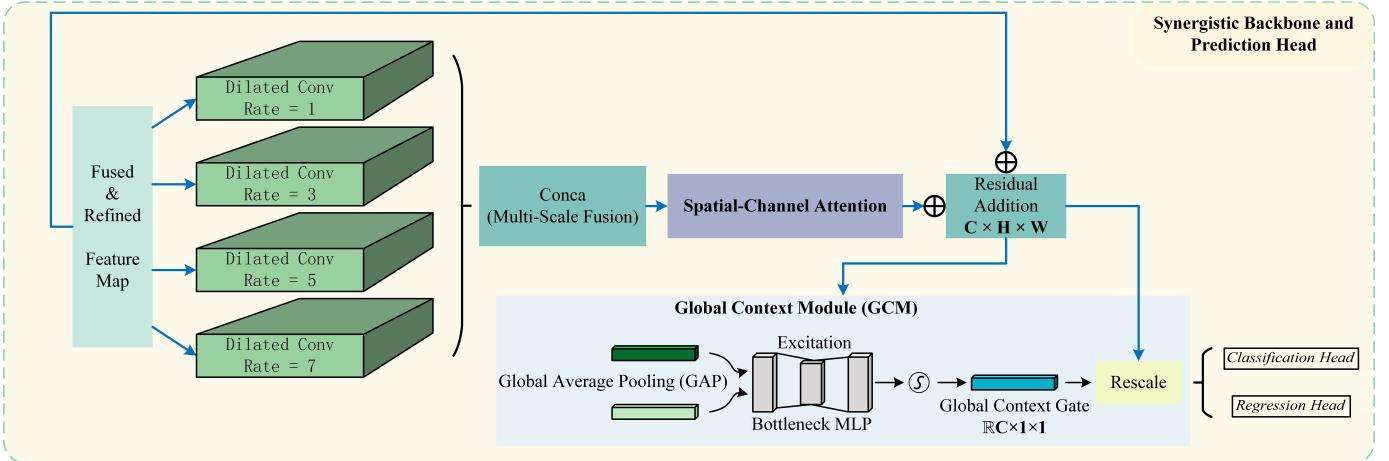


Fig. 3. The structure of the synergistic backbone and prediction head.

used metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE). MAE quantifies the average accuracy of the count, while MSE is more sensitive to large errors, thus reflecting the model's stability and robustness. They are defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^N |C_i - \hat{C}_i| \quad (1)$$

$$MSE = \sqrt{\frac{1}{N} \sum i = 1^N (C_i - \hat{C}_i)^2} \quad (2)$$

where N is the total number of test images, C_i is the ground-truth count for the i -th image, and \hat{C}_i is the count predicted by the model.

For localization performance, we adopt standard detection metrics: Precision (P), Recall (R), and their harmonic mean,

the F1-score. A prediction is considered a true positive (TP) if its distance to the nearest ground-truth point is within a predefined threshold; otherwise, it is classified as a false positive (FP). Ground-truth points that are not matched with any prediction are counted as false negatives (FN). The metrics are formulated as:

$$\text{Precision (P)} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall (R)} = \frac{TP}{TP + FN} \quad (4)$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (5)$$

This multi-faceted approach provides a holistic assessment of our model's capabilities in both counting and precise localization.

TABLE I
COMPREHENSIVE COMPARISON OF COUNTING ACCURACY BETWEEN OUR PROPOSED METHOD AND STATE-OF-THE-ART APPROACHES ON SIX PUBLIC BENCHMARK DATASETS. THE BEST RESULTS FOR EACH METRIC ARE HIGHLIGHTED IN BOLD.

Methods	SHT_A		SHT_B		UCF_QNRF		UCF_CC_50		JHU-Crowd		NWPU-Crowd	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [34]	110.2	-	26.4	-	-	-	377.6	509.1	160.6	377.7	232.5	714.6
CSRNet [25]	68.2	115.0	10.6	16.0	120.3	208.5	266.1	397.5	72.2	249.9	121.3	387.8
CAN [43]	62.3	100.0	7.8	12.2	107.0	183.0	212.2	243.7	100.1	314.0	110.0	495.3
S-DCNet [45]	58.3	95.0	6.7	10.7	104.4	176.1	204.2	301.3	-	-	-	-
ADSCNet [46]	55.4	97.7	6.4	11.3	71.3	132.5	198.4	267.3	-	-	-	-
ASNet [47]	57.78	90.13	-	-	91.59	159.71	174.84	251.63	-	-	-	-
DM-Count [48]	59.7	95.7	7.4	11.8	85.6	148.3	211.0	291.5	-	-	88.4	388.6
SCAR [44]	-	-	-	-	-	-	-	-	-	-	106.3	386.5
MAN [51]	56.8	90.3	-	-	77.3	131.5	-	-	-	-	-	-
GauNet [50]	54.8	89.1	6.2	9.9	81.6	153.7	186.3	256.5	69.4	262.4	-	-
P2PNet [10]	52.7	85.1	6.3	9.9	85.32	154.5	172.72	256.18	-	-	-	-
VMambaCC [29]	51.87	81.3	7.48	12.47	88.42	144.73	-	-	54.41	201.93	-	-
CLTR [49]	56.9	95.2	6.5	10.6	85.8	141.3	-	-	59.5	240.6	-	-
FGENet [26]	51.66	85.0	6.34	10.53	85.2	158.76	142.56	215.87	-	-	-	-
Ours	52.1	83.7	6.1	9.5	79.8	140.2	150.3	225.6	58.7	210.4	86.1	358.4

B. Counting Evaluation

To rigorously benchmark our proposed G²P²-Net, we conducted a comprehensive evaluation against a suite of established methods across six challenging public datasets. The quantitative results, presented in Table I, are corroborated by the qualitative visualizations in Fig. 4. Together, they demonstrate the robust generalization of our geometry-aware framework, particularly its effectiveness in compensating for the severe perspective distortion and dense occlusions characteristic of aerial imagery.



Fig. 4. Qualitative Results of G²P²-Net on Crowd Scenes.

Notably, on the large-scale NWPU-Crowd dataset, known for its vast diversity in scenes and densities, our model achieves a compelling MAE of 86.1 and an MSE of 358.4. This performance in a complex, unconstrained environment substantiates our central hypothesis: that explicitly embedding geometric priors enhances generalization. This capability is further evidenced on ShanghaiTech Part B (SHTB), where the model attains a competitive MAE of 6.1 and MSE of 9.5, effectively compensating for the pronounced perspective shifts common in urban aerial scenes.

In scenarios of extreme crowd density, such as the congested ShanghaiTech Part A (SHTA) dataset, G²P²-Net maintains highly robust performance. It achieves a competitive MAE of 52.1 while securing a more favorable MSE of 83.7. This lower MSE suggests enhanced stability and a reduced susceptibility to large estimation errors, a critical attribute for reliable performance in volatile conditions. Consistent, competitive results are also observed on the UCF-QNRF and JHU-Crowd++ datasets.

Collectively, the performance across these varied benchmarks validates the effectiveness of our geometry-aware approach. The findings affirm that integrating physical imaging principles into the learning framework, rather than relying solely on data-driven feature extraction, is an effective strategy for advancing aerial crowd analysis.

C. Localization Evaluation

In addition to counting, the framework's localization performance was evaluated using P, R, and F1-score, with results presented in Table II. The model achieves robust localization on the ShanghaiTech Part B (SHTB) dataset, attaining a balanced F1-score of 80.0%. This effective performance on a benchmark characterized by prominent perspective shifts underscores the direct contribution of the PAAP module's geometric reasoning to spatial accuracy. Despite the extreme occlusions in more congested datasets like SHTA and UCF-QNRF, the model maintains effective localization. A key observation across all benchmarks is the consistent balance between precision and recall, indicating that G²P²-Net identifies a high proportion of individuals without introducing a significant number of false positives. This dual capability validates G²P²-Net as not only an accurate counter but also a reliable localization framework, critical for enabling downstream analysis tasks such as tracking and behavior modeling.

TABLE II
LOCALIZATION PERFORMANCE OF G²P²-NET ACROSS SIX BENCHMARK DATASETS. THE BEST RESULT IN EACH METRIC COLUMN IS HIGHLIGHTED IN BOLD.

Dataset	Localization Metrics		
	P(%)	R(%)	FI-score(%)
SHT_B	80.1	79.8	80.0
SHT_A	77.2	76.9	77.0
UCF-QNRF	68.8	67.3	68.0
JHU-Crowd	63.0	61.1	62.0

D. Ablation Studies

To validate the contribution of each key component within our G²P²-Net framework, we conducted a comprehensive ablation study on the NWPU-Crowd dataset. Starting from the full model, we systematically removed or replaced individual components and measured the resulting impact on performance. The results, detailed in Table III, confirm the integral role of each designed module.

TABLE III
ABLATION STUDY ON THE NWPU-CROWD DATASET. WE DEMONSTRATE THE CONTRIBUTION OF EACH COMPONENT BY STARTING WITH OUR FULL MODEL AND REMOVING IT INDIVIDUALLY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Components				NWPU-Crowd	
PAAP	DCN	GCM	Clustering	MAE ↓	MSE ↓
✓	✓	✓	✓	86.1	358.4
✗	✓	✓	✓	91.5	410.2
✓	✗	✓	✓	89.2	390.5
✓	✓	✗	✓	87.8	375.1
✓	✓	✓	✗	87.2	365.3

Ablation studies confirm the critical importance of our core designs: removing the PAAP module leads to the most substantial degradation (MAE +5.4, MSE +51.8), strongly validating that embedding geometric priors is essential for handling severe scale variations in aerial imagery. The architectural refinements—DCN for non-rigid morphologies and the GCM for feature refinement—also prove highly effective, with their absence causing clear performance drops. Furthermore, replacing our GSD-guided physical radius clustering with a fixed-pixel approach reduces accuracy, underscoring the value of scale-aware post-processing. These results collectively demonstrate that all components of G²P²-Net operate synergistically to achieve robust and accurate crowd analysis.

E. Limitations and Future Directions

While this study establishes a robust foundation, its limitations delineate clear avenues for future research. Primarily, the framework's static design precludes the dynamic narrative generation essential for deeper journalistic insight. Furthermore, the model's out-of-distribution robustness remains unverified for the unpredictable conditions of live news gathering, necessitating systematic cross-domain validation. Perhaps most

critically, this work has deferred the crucial layer of interpretation and ethics required to bridge the gap from accurate data to credible public narratives. Future research must therefore focus on developing tools for uncertainty quantification, contextual visualization, and human-in-the-loop validation, all situated within a robust ethical framework.

V. CONCLUSION

In this work, we explored the critical gap between the promise of aerial data journalism and the limitations of conventional computer vision by proposing a principled integration of physical geometry. Our empirically validated G²P²-Net, with its core Perspective-Aware Attention Pyramid (PAAP) module, provides the technical foundation for a central focus of this work: the proposal of a replicable framework for media practice-driven technology iteration. We illustrated a synergistic, closed-loop ecosystem where our algorithm can empower new journalistic narratives, while the demands of real-world media deployment can, in turn, guide its continuous evolution. Ultimately, this research not only lays the groundwork for future dynamic, motion-aware analysis but also suggests a promising model for cross-disciplinary innovation, highlighting how impactful advancements can emerge from the crucible of real-world practice.

ACKNOWLEDGMENT

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

REFERENCES

- [1] J. S. Jacques Junior, S. Musse, and C. Jung, Crowd Analysis Using Computer Vision Techniques," IEEE Signal Processing Magazine, vol. 27, no. 5, pp. 66–77, Sep. 2010.
- [2] M. Mahdi and B. Morris, Looking at Intersections: A Survey of Intersection Monitoring, Behavior and Safety Analysis of Recent Studies," IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 1, pp. 4–24, Jan. 2017.
- [3] T. Ahmad *et al.*, Future UAV/Drone Systems for Intelligent Active Surveillance and Monitoring," ACM Computing Surveys, vol. 58, no. 2, pp. 1–37, Aug. 2025.
- [4] A. A. B. Abdelnabi and G. Rabadi, Human Detection from Unmanned Aerial Vehicles' Images for Search and Rescue Missions: A State-of-the-Art Review," IEEE Access, vol. 12, pp. 152009–152035, Jan. 2024.
- [5] D. P. S. Henrique, S. Marsico, J. Rodrigo, and L. A. Barbosa, Human Detection in UAV Imagery Using Deep Learning: A Review," Neural Computing and Applications, vol. 37, pp. 18109–18150, Jul. 2025.
- [6] 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, vol. 107, pp. 3–16, May 2018.
- [7] Y.-T. Liu, M. Shi, Q. Zhao, and X. Wang, Point in, Box Out: Beyond Counting Persons in Crowds," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6469–6478.
- [8] R. M. Sharma, Single-Image Crowd Counting Using Multi-Column Neural Network," International Journal of Computer Applications, vol. 175, no. 11, pp. 31–35, Aug. 2020.
- [9] X. Cao, Z. Wang, Y. Zhao, and F. Su, Scale Aggregation Network for Accurate and Efficient Crowd Counting," in Lecture Notes in Computer Science, Jan. 2018, pp. 734–750.
- [10] Q. Song *et al.*, Rethinking Counting and Localization in Crowds: A Purely Point-Based Framework," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 3365–3374.

- [11] Q. Wang, J. Gao, W. Lin, and Y. Yuan, Learning From Synthetic Data for Crowd Counting in the Wild,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8198–8207.
- [12] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, Drone-Based Object Counting by Spatially Regularized Regional Proposal Network,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4145–4153.
- [13] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, Dense Dilated Convolutions’ Merging Network for Land Cover Classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.
- [14] M. Barekatain *et al.*, Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, Jul. 2017, pp. 28–35.
- [15] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature Pyramid Networks for Object Detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2117–2125.
- [16] H. Zhao, W. Min, X. Wei, Q. Wang, Q. Fu, and Z. Wei, MSR-FAN: Multi-Scale Residual Feature-Aware Network for Crowd Counting,” IET Image Processing, vol. 15, no. 14, pp. 3512–3521, Mar. 2021.
- [17] J. Ma, Y. Dai, and Y.-P. Tan, Atrous Convolutions Spatial Pyramid Network for Crowd Counting and Density Estimation,” Neurocomputing, vol. 350, pp. 91–101, Jul. 2019.
- [18] J. Dai *et al.*, Deformable Convolutional Networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 764–773.
- [19] J. Wan and A. Chan, Adaptive Density Map Generation for Crowd Counting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1130–1139.
- [20] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, Vision Meets Drones: A Challenge,” arXiv.org, Apr. 23, 2018.
- [21] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, Privacy Preserving Crowd Monitoring: Counting People Without People Models or Tracking,” in Proc. 2008 IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2008, pp. 1–7.
- [22] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.
- [24] V. Lempitsky and A. Zisserman, Learning To Count Objects in Images,” in Advances in Neural Information Processing Systems 23 (NIPS 2010), 2010.
- [25] Y. Li, X. Zhang, and D. Chen, CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 1091–1100.
- [26] H.-Y. Ma, L. Zhang, and X.-Y. Wei, FGENet: Fine-Grained Extraction Network for Congested Crowd Counting,” in Lecture Notes in Computer Science, Jan. 2024, pp. 43–56.
- [27] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, TransCrowd: Weakly-Supervised Crowd Counting with Transformers,” Science China Information Sciences, vol. 65, no. 6, p. 160104, Apr. 2022.
- [28] I.-H. Chen, W.-T. Chen, Y.-W. Liu, M.-H. Yang, and S.-Y. Kuo, Improving Point-Based Crowd Counting and Localization Based on Auxiliary Point Guidance,” in Computer Vision – ECCV 2024, ser. Lecture Notes in Computer Science, Nov. 2024, pp. 428–444.
- [29] H.-Y. Ma, L. Zhang, and S. Shi, VMambaCC: A Visual State Space Model for Crowd Counting,” arXiv.org, May 2024.
- [30] Z. Huo, C. Yuan, K. Zhang, Y. Qiao, and F. Luo, VMamba-Crowd: Bridging Multi-Scale Features from Visual Mamba for Weakly-Supervised Crowd Counting,” Pattern Recognition Letters, vol. 197, pp. 297–303, Nov. 2025.
- [31] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, Rethinking Atrous Convolution for Semantic Image Segmentation,” arXiv.org, Jan. 2017.
- [32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, CBAM: Convolutional Block Attention Module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 3–19.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
- [34] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, Single-Image Crowd Counting via Multi-Column Convolutional Neural Network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 589–597.
- [35] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 2547–2554.
- [36] H. Idrees *et al.*, Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 532–546.
- [37] 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, vol. 43, no. 6, pp. 2141–2149, Jul. 2020.
- [38] 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, vol. 44, no. 5, pp. 2594–2609, Nov. 2020.
- [39] A. Paszke *et al.*, PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019, pp. 8026–8037.
- [40] I. Loshchilov and F. Hutter, Decoupled Weight Decay Regularization,” in Proc. Int. Conf. Learn. Represent., Sep. 2018.
- [41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, Focal Loss for Dense Object Detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.
- [42] H. W. Kuhn, The Hungarian Method for the Assignment Problem,” Naval Research Logistics Quarterly, vol. 2, no. 1–2, pp. 83–97, Mar. 1955.
- [43] W. Liu, M. Salzmann, and P. Fua, Context-Aware Crowd Counting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5099–5108.
- [44] J. Gao, Q. Wang, and Y. Yuan, SCAR: Spatial-/Channel-Wise Attention Regression Networks for Crowd Counting,” Neurocomputing, vol. 363, pp. 1–8, Oct. 2019.
- [45] H. Xiong, H. Lu, C. Liu, L. Liu, Z. Cao, and C. Shen, From Open Set to Closed Set: Counting Objects by Spatial Divide-and-Conquer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8362–8371.
- [46] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, Adaptive Dilated Network With Self-Correction Supervision for Counting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 4594–4603.
- [47] X. Jiang *et al.*, Attention Scaling for Crowd Counting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 4706–4715.
- [48] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, Distribution Matching for Crowd Counting,” in Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, pp. 1595–1607.
- [49] D. Liang, W. Xu, and X. Bai, An End-to-End Transformer Model for Crowd Localization,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Oct. 2022, pp. 38–54.
- [50] Z.-Q. Cheng, Q. Dai, H. Li, J. Song, X. Wu, and A. G. Hauptmann, Rethinking Spatial Invariance of Convolutional Networks for Object Counting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 19638–19648.
- [51] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, Boosting Crowd Counting via Multifaceted Attention,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 19628–19637.