COUNTING THE UNCOUNTED: CORRECTING OCCLUSION-CAUSED BIAS IN CROWD COUNTING BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-based crowd counting has advanced considerably, yet its real-world reliability remains limited by occlusion, which existing methods treat as undifferentiated noise rather than as a structured scene element. We challenge a foundational assumption of current benchmarks: that ground-truth (GT) annotations reflect true crowd size. By auditing four widely used datasets (ShanghaiTech A, UCF-QNRF, JHU-Crowd++, and NWPU-Crowd), we introduce a three-class occlusion taxonomy: Atmospheric (AO), Barrier (BO), and Embedded Occlusions (EO). We identify EOs as the most consequential yet overlooked category. EOs are small, non-transparent objects embedded within the crowd, such as umbrellas, picket signs, and flags. Occlusions appear in 40.73% of images, and 78.46% of these involve EOs, yet no annotation protocol accounts for the individuals concealed behind them. We correct the GT through occluder-specific annotation and local density inference with a new annotation tool, Occlusion Density Fusion (ODFlow). The correction shows that existing GT undercounts EO-affected images by 6.3% on average, so published MAE understates real-world error by about 22 counts per image. Beyond data correction, we propose ODFNet, an occlusion-aware counting model. Its Density Reconstruction and Count Fusion (DRCF) module reconstructs human and EO density maps separately, then fuses them while removing redundancy introduced by overlapping EOs. Validation on both synthetic crowd scenarios and real crowd images shows that ODFNet recovers the ground-truth count regardless of EO type, number, or spatial pattern, whereas existing methods drift substantially. We release corrected GT for all four datasets, a comprehensive benchmark re-evaluation, a newly curated occlusion-aware dataset (ODF-A), and ODFNet.

1 INTRODUCTION

Reliable crowd counting is fundamental to public safety, urban planning, and large-scale event management, where accurate situational awareness can directly inform life-critical decisions Gao et al. (2020); Fiandeiro et al. (2023); Bai et al. (2022); Kothari et al. (2023); Elsepae et al. (2026); Liang et al. (2022b). Over the past decade, vision-based density estimation models have achieved substantial gains in state-of-the-art (SOTA) performance, systematically addressing challenges, including illumination variation Liang et al. (2022a); Wang et al. (2025), environmental changes Huang et al. (2023), and perspective distortion Ranjan et al. (2018), using from convolutional density-regression models (CSRNet (Li et al., 2018)) to point-based (P2R-Loss (Lin et al., 2025), P2PNet (Song et al., 2021)), pattern-based (DM-Count (Wang et al., 2020a)), distribution-matching (CLIP-EBC (Ma et al., 2025)), transformer-based (TMTB (Yang et al., 2025)), and vision-language-based counting methods (AdaSEEMCount (Wan et al., 2024)).

Despite these advances, practical reliability remains limited by a problem the field has yet to formally confront: **occlusion**. In crowd scenes, it occurs when objects obscure one another, among people (intra-class) or between people and environmental structures (inter-class), interweaving semantic features and making it difficult for both annotators and detectors to resolve object boundaries. In real-world settings such as political rallies, religious gatherings, and sporting events, structured and scattered obstructions, buildings, trees, umbrellas, flags, and picket signs, routinely conceal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Occlusion taxonomy and representative occlusion examples from benchmark datasets (Zhang et al., 2016; Idrees et al., 2018; Sindagi et al., 2022; Wang et al., 2020b).

Taxonomy	Description	Occlusion Examples
Atmospheric **O**cclusions **(AO)**	semi-transparent weather phenomena such as fog, rain, and snow	
Barrier **O**cclusions **(BO)**	large opaque environmental structures such as buildings, vehicles, trees, and walls	
Embedded **O**cclusions **(EO)**	small distributed objects within the crowd itself such as umbrellas, picket signs, and flags	
Mixed	Multiple occlusions	

portions of the crowd. Rather than modeling these as explicit scene components, existing methods treat them uniformly as unstructured noise(Almalki et al., 2021; Wang et al., 2018), and annotation protocols ignore the individuals hidden behind them entirely. The consequence is a silent, pervasive undercount baked into every major benchmark and every model trained on them.

We challenge a common assumption that published ground-truth (GT) annotations in crowd datasets faithfully reflect the crowd count. Through a rigorous per-image audit of four widely used datasets, ShanghaiTech A (Zhang et al., 2016), UCF-QNRF (Idrees et al., 2018), JHU-Crowd++ (Sindagi et al., 2022), and NWPU-Crowd (Wang et al., 2020b), we confirmed this assumption does not hold. To structure our analysis, as presented in Table 1, we introduce a three-class occlusion taxonomy grounded in physical properties and annotation impact: **Atmospheric Occlusions (AO)**, **Barrier Occlusions (BO)**, and **Embedded Occlusions (EO)**. While all three categories affect model performance, AO and BO are tractable. AO caused degradation can be partially mitigated through image enhancement Huang et al. (2023), and individuals concealed behind BOs typically fall outside the scene of crowd counting interest. However, EO is systematic and trackable, not negligible noise. The people it partially or fully conceals are physically present and countable in principle, yet existing benchmarks ignore them, yielding inaccurate GT. As illustrated in Figure 1 (a), EO scenes such as JHU 0010 feature geometrically structured occluders, such as pickets arranged in linear perspective, that partially or fully conceal individuals while maintaining complex spatial interactions and crowd structure. Unlike general multi-object detection (Ren et al., 2015; Redmon et al., 2016), EO involves spatially coherent occluding layers whose density and extent must be explicitly modeled. Existing GT annotation pipelines fail to account for this. As shown in Figure 1 (b), occluded regions are routinely ignored during labeling, yielding systematically underestimated counts. Current SOTA crowd counting models inherit this bias, lacking explicit representations of occluder geometry or spatial distribution, and consequently exhibit significantly elevated error rates under real-world EO conditions, undermining their reliability precisely where accurate counts matter most.

To address this limitation, we identify and correct EO-induced errors in existing datasets to build an EO-aware dataset and benchmark, and we design a new counting method that fuses human and EO density estimates to recover concealed individuals. First, guided by our three-class occlusion taxonomy, we establish a GT correction pipeline. Our analysis of four widely used crowd counting datasets (Table 2) shows that 40.73% of images contain occlusions and that EO is the dominant type, appearing in 78.46% of occluded images. Despite this prevalence, existing annotation protocols ignore EO regions. Our correction, which merges occluder-specific annotation with local density inference, shows that existing GT undercounts EO-affected images by 6.3% on average. Consequently, published MAE understates real-world error by about 22 counts per image. To support this correction, we develop an annotation tool, Occlusion Density Fusion (ODFlow) (Authors,

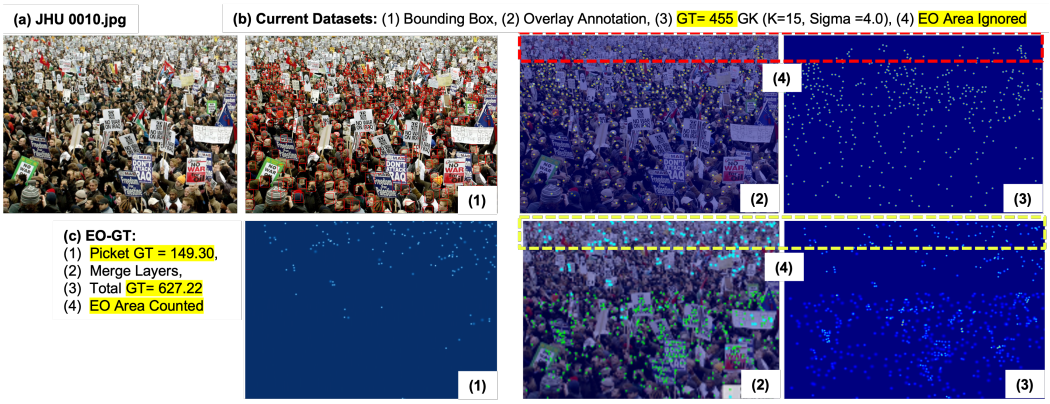

Figure 1: (a) An example scene from the JHU dataset containing occluding objects (e.g., pickets). (b) Conventional annotation marks visible individuals but ignores occluded regions (b-4), resulting in systematically underestimated GT counts. (c) The proposed EO-GT captures partially and fully occluded individuals behind the EO occluding layer, and merges the resulting occlusion map with the visible human density to produce corrected total counts.

2026a), and curate a dedicated dataset, ODF-A (Authors, 2026b). ODF-A contains collected and augmented images with structured EOs, occluder-specific annotations, and EO-aware density maps (Figure 1(c)), providing more reliable supervision for training and more accurate benchmarking for evaluation. Second, we propose the **Occlusion Density Fusion Network (ODFNet)**, a multi-stream density estimation architecture that explicitly models both human and EO densities. Rather than treating EOs as noise or merely detecting them, ODFNet learns separate feature representations for visible individuals and EOs such as picket signs and umbrellas. Its Density Reconstruction and Count Fusion (DRCF) module fuses these streams to recover the hidden crowd distribution while removing the redundancy introduced by overlapping EOs, so that concealed individuals are neither double-counted nor omitted. As a result, ODFNet yields counts that remain consistent regardless of EO type, number, or spatial pattern, whereas existing methods drift substantially. Because published MAE and RMSE scores are computed against GT that omits EO-concealed individuals, they overstate real-world performance. We therefore publicly release the corrected GT, benchmark re-evaluation, ODF-A, ODFlow, and ODFNet, establishing a corrected standard for evaluating future crowd counting methods. The main contributions of this work are as follows:

1. **Occlusion Taxonomy:** A three-class taxonomy (AO, BO, EO) that categorizes crowd occlusions by physical properties, spatial distribution, and impact on annotation integrity.
2. **Dataset Audit and Correction:** A per-image audit and annotation tool, ODFlow (Authors, 2026a), that produces corrected EO-aware GT for ShanghaiTech A, UCF-QNRF, JHU-Crowd++, and NWPU-Crowd.
3. **Occlusion-Aware Dataset:** ODF-A (Authors, 2026b), which combines real dense-crowd images containing EOs with synthetic scenes to support EO learning and validate the GT correction.
4. **Benchmark Re-evaluation:** A re-evaluation of published methods against corrected GT, quantifying how much reported MAE and RMSE understate real-world error.
5. **Redundancy-Aware Count Fusion:** ODFNet's DRCF module, which fuses separate human and EO density maps while removing EO-induced bias, yielding consistent counts across EO type, number, and spatial pattern, unlike existing methods.

2 Datasets and Benchmarks

We present our occlusion-aware crowd counting framework. We first audit four widely used datasets and re-annotate them with EO labels to produce corrected GT, then introduce ODF-A, a dedicated occlusion-aware dataset. Finally, we describe ODFNet, a multi-stream architecture trained on ODF-A that estimates human and EO density maps and uses multi-scale fusion with density reconstruction

Table 2: Occlusion prevalence across benchmark crowd counting datasets, including ShanghaiTech A Zhang et al. (2016), UCF-QNRF Idrees et al. (2018), JHU-Crowd++ Sindagi et al. (2022), and NWPU-Crowd Wang et al. (2020b), reporting the proportion of occluded images and the breakdown of occlusion types (EO, BO, and AO) among occluded scenes.

Datasets	Images	Occlusion images (%)	EO (%)	BO (%)	AO (%)
ShanghaiTech A	482	189 (39.20 %)	110 (58.20 %)	118 (62.43 %)	3 (1.60 %)
UCF-QNRF	1,535	761 (49.60 %)	504 (66.23 %)	387 (50.85 %)	12 (1.58 %)
JHU-Crowd++	4,822	1,772 (36.75 %)	1,607 (90.69 %)	1,348 (76.1 %)	57 (3.22 %)
NWPU-Crowd	5,109	2,145 (42 %)	1,598 (74.50 %)	2,043 (95.24 %)	89 (4.15 %)
Total	11,948	4,867 (40.73 %)	3,819 (78.46 %)	3,896 (80 %)	161 (3.30 %)

to keep spatial density maps consistent with total counts. We also use ODFNet to benchmark the existing datasets and assess how occlusion affects counting performance.

2.1 Existing Datasets

We select four foundational and widely adopted datasets, NWPU-Crowd (Wang et al., 2020b), ShanghaiTech A (Zhang et al., 2016), UCF-QNRF (Idrees et al., 2018), and JHU-Crowd++ (Sindagi et al., 2022), spanning diverse scenes, large-scale annotations, and broad density variation, as the basis for our EO annotation, corrected GT generation, and cross-dataset occlusion analysis.

To assess occlusion prevalence, we conducted a systematic per-image audit of all 11,948 images. As presented in Table 2, approximately 40.73% of images contain at least one form of occlusion, with EO being the dominant category at roughly 78.46% of occluded scenes. EO prevalence varies across datasets: JHU-Crowd++ exhibits the highest rate at 90.69% of its occluded images, followed by NWPU-Crowd (74.5%), UCF-QNRF (66.23%), and ShanghaiTech A (58.2%). Despite this pervasiveness, no existing annotation protocol accounts for individuals concealed behind EO objects, introducing a systematic, dataset-wide undercount. To quantify this annotation error, we apply our **EO-GT** correction procedure to all EO-affected 3,819 images. As shown in Table 3, approximately 32% of all images contain EO, ranging from 22.9% in ShanghaiTech A to 33.3% in JHU-Crowd++. Corrected EO-GT reveals consistent undercounting across all four datasets: average per-image counts increase by 3.3% in ShanghaiTech A (529 $\rightarrow$ 547), 3.8% in UCF-QNRF (555 $\rightarrow$ 577), 6.8% in JHU-Crowd++ (385 $\rightarrow$ 413), and 7.1% in NWPU-Crowd (184 $\rightarrow$ 198). In aggregate, this amounts to around 80,214 uncounted individuals, an average correction of 22 counts per EO-affected image, or 6.3% of the original GT. Datasets with denser, more structured EO scenes such as JHU-Crowd++ and NWPU-Crowd exhibit the largest relative corrections, demonstrating that benchmarking error is most severe precisely where accurate counting matters most. These findings confirm that published MAE and RMSE scores across all four benchmarks are evaluated against a systematically undercounted GT (in Table 6).

1. **NWPU-Crowd** is one of the largest crowd-counting benchmarks, comprising 5,109 high-resolution images with over 2 million annotated heads across a wide density range, including negative samples to improve model robustness. Our audit finds that 31.3% of images contain EO, a 7.1% underestimation and among the largest relative errors across the four benchmarks. The corrected GT subset is designated **EO-NWPU**.
2. **ShanghaiTech Part A** consists of 482 highly dense crowd scenes and is one of the most widely cited benchmarks in the crowd-counting literature. Although it has the lowest EO prevalence among the four datasets at 22.9%, corrected GT still reveals an average undercount of 18 people per EO-affected image, a 3.3% per-image error. The corrected GT subset is designated **EO-SHA**.
3. **UCF-QNRF** contains 1,535 high-resolution, extremely dense crowd images with up to 12,865 individuals per image. With 32.9% of images containing EO, corrected GT reveals an average undercount of 22 people per EO-affected image, a 3.8% correction amounting to 11,061 uncounted individuals across the dataset. The corrected GT subset is designated **EO-UCF**.
4. **JHU-Crowd++** comprises 4,822 images with 1.51 million annotations across diverse environmental conditions, including weather degradations. It has the highest EO prevalence among the four datasets, with 33.3% of images affected and 90.69% of occluded scenes containing EO,

Table 3: Occlusion prevalence across benchmark crowd counting datasets, reporting the proportion of occluded images and the breakdown of occlusion types (EO, BO, and AO) among occluded scenes.

Datasets	Images	EO-images (%)	Original-GT (Avg.)	EO-GT (Avg.)	Avg. Error (%)
ShanghaiTech A	482	110 (22.9 %)	58,210 (529)	60,134 (547)	18 (3.3 %)
UCF-QNRF	1,535	504 (32.9 %)	279,771 (555)	290,832 (577)	22 (3.8 %)
JHU-Crowd++	4,822	1,607 (33.3 %)	618,529 (385)	664,278 (413)	28 (6.8 %)
NWPU-Crowd	5,109	1,598 (31.3 %)	294,002 (184)	315,482 (198)	14 (7.1 %)
Total	11,948	3,819 (32 %)	1,250,512 (327)	1,330,726 (349)	22 (6.3 %)

Table 4: Statistics of the ODF-A Dataset.

Dataset	Image	Label	EO (%)	Avg.	Human (Min/Max)	EO (Min/Max)
ODF-A	491	85,455	7,406 (9.1%)	94	7 / 989	1 / 110

driven largely by protest and rally scenes where picket signs create the most severe annotation blind spots. The corrected GT reveals an average undercount of 28 people per EO-affected image, a 6.8% correction amounting to 45,749 uncounted individuals across the dataset. The corrected GT subset is designated **EO-JHU**.

2.2 Occlusion Density Fusion (ODF) Dataset

To measure the effect of embedded occluders on crowd counting, we construct ODF-All (ODF-A) Authors (2026b): a 491-image, EO-aware dataset with occluder-specific annotations and refined GT labels, serving as the primary benchmark for training and evaluating ODFNet on joint human and EO density modeling. Images are drawn from three sources: 161 from web search engines and royalty-free repositories, 25 synthetic images with known GT under controlled occlusion, and 305 sampled from JHU-Crowd++ (82), UCF-QNRF (126), NWPU-Crowd (45), and ShanghaiTech (52), and curated using ODFlow to ensure clear visibility of crowd regions and occluders. As shown in Table 4, ODF-A contains 85,455 annotations across 491 images, of which 7,406 (9.1%) are EO annotations, with per-image human counts ranging from 7 to 989 and EO counts from 1 to 110. This EO annotation volume is comparable in scale to the EO-affected subsets identified in the benchmark audit, confirming that ODF-A captures sufficient EO diversity and density to support robust occlusion-aware learning and cross-dataset validation.

2.3 ODFNet Architecture

The **ODFNet** is designed to achieve robust estimation under EOs. Unlike conventional density regression models that implicitly treat occlusions as unstructured noise, ODFNet explicitly models occlusion by separating spatial density estimation from global count prediction, allowing more principled handling of hidden regions. As illustrated in Figure 2, the architecture consists of three key components.

First, a hierarchical **ConvNeXt backbone** (Liu et al., 2022), leveraging its strong capability in dense prediction tasks and compatibility with modern training strategies. The backbone processes the input image through four hierarchical stages, progressing from shallow, high-resolution features to deep, low-resolution semantic representations.

$$\{P_k\}_{k=1}^{3} = \{\phi_k(I)\}, P_k \in \mathbb{R}^{\frac{H}{2^{k+2}} \times \frac{W}{2^{k+2}} \times C_k}, C_k \in \{192,\ 384,\ 768\} \tag{1}$$

Stages 2–4 are used to construct multi-scale feature maps P_1, P_2, and P_3, respectively, with progressively reduced spatial resolutions and increased channel dimensions, as defined in Eq. (1). The earlier stage retains higher spatial resolution and preserves fine-grained structural details necessary for localizing small or partially occluded objects. In contrast, the deeper stages capture increasingly abstract semantic information, encoding broader contextual cues such as global crowd layout and occlusion patterns. This hierarchical representation enables ODFNet to balance precise localization with robust contextual reasoning under structured occlusions.

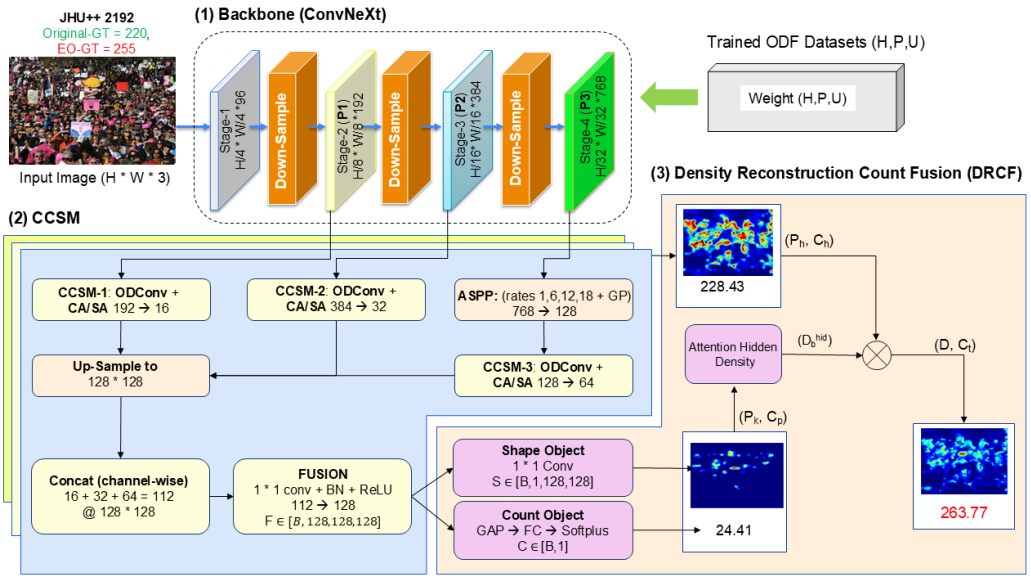

Figure 2: ODFNet architecture consists of (1) a hierarchical ConvNeXt backbone that extracts multi-scale features capturing both local details and global context; (2) a Context-aware Channel and Spatial Modulation (CCSM) module with ASPP to refine category-specific features using multi-scale context and adaptive attention; and (3) a Density Reconstruction and Count Fusion (DRCF) module that integrates density maps and enforces global consistency between spatial distributions and the total crowd count.

Second, a **Context-aware Channel and Spatial Modulation (CCSM)** enhances feature discrimination under occlusion. CCSM operates on the multi-scale feature maps P_1, P_2, and P_3 extracted from the backbone, combining local convolutional refinement with omni-dimensional dynamic convolution (ODConv) (Li et al., 2022) and dual attention mechanisms in both channel and spatial domains, following channel attention (Hu et al., 2018) and combined channel–spatial attention frameworks such as CBAM (Woo et al., 2018; Gao et al., 2019). Each CCSM consists of lightweight pointwise and depthwise convolutions followed by ODConv-based refinement for dynamic, context-dependent filtering, with batch normalization (BN) and ReLU activation applied throughout for stable optimization. Within each module, a squeeze-and-excitation style channel attention (CA) descriptor (Hu et al., 2018) is computed by aggregating pooled statistics and passing them through a shared multilayer perceptron. In parallel, a spatial attention (SA) mask is generated by applying average and max pooling across channels to capture location-specific importance (Zhou et al., 2025). The final CCSM output is produced by sequentially applying channel and spatial modulation and incorporating a residual connection to preserve original feature information. This mechanism amplifies crowd-relevant responses while suppressing features dominated by structured occluders, allowing the network to focus on meaningful human patterns under heavy occlusion. To further capture the wide variation in object scales and occlusion extents, the deepest feature map P_3 is enhanced with an Atrous Spatial Pyramid Pooling (ASPP) module (Chen et al., 2018). ASPP enriches high-level features with multi-scale contextual information by employing parallel dilated convolutions with dilation rates $6, 12, 18$, along with a 1×1 convolution branch and an image-level pooling branch followed by projection (Zhou et al., 2025).

$$\mathbf{F} = \mathrm{ReLU}\Big(\mathrm{BN}\big(\mathrm{Conv}_{1\times1}\big(\mathrm{Up}(\mathrm{CCSM}_i(\mathbf{P}_i))\big)\big)\Big), \quad i \in \{1,2,3\} \tag{2}$$

The outputs of all branches are concatenated and projected to a compact 128-channel representation. This design enlarges the effective receptive field while preserving spatial resolution, enabling the model to jointly capture compact head clusters and extended occlusion patterns (Liu et al., 2019). The enhanced multi-scale features are then aligned to a common resolution and aggregated through Multi-Scale ODF Fusion, which integrates information across semantic depths. The concatenated

Table 5: Quantitative comparison (MAE/RMSE) of ODFNet with AdaSEEMCount (ASEEM), TMTB, P2R-Loss, and CLIP-EBC using EO-GT datasets (ODF-A, EO-SHA, EO-UCF, EO-JHU, and EO-NWPU) and existing crowd-counting benchmarks (SHA, UCF, JHU, and NWPU). Red, Green, and Blue denote notable performance groups.

Datasets	ASEEM	TMTB	P2R-Loss	CLIP-EBC	**ODFNet**
SHA	102.6 / 176.3	72.4 / 126.4	69.9 / 119.5	52.5 / 85.9	104.3 / 198.42
EO-SHA	128.4 / 201.7	96.9 / 164.8	108.27 / 172.8	78.05 / 152.06	47.8 / 79.6
UCF	182.3 / 289.9	96.3 / 163.4	100.1 / 182.5	80.3 / 139.3	155.47 / 244.5
EO-UCF	214.5 / 332.7	128.8 / 218.6	137.4 / 236.9	108.9 / 194.5	72.6 / 126.8
JHU	102.7 / 360.7	60 / 240.3	77.8 / 293.5	58.4 / 129.6	158.42/ 285.54
EO-JHU	136.8 / 321.4	86.7 / 211.5	101.9 / 246.2	84.6 / 204.7	71.5 / 168.4
NWPU	104.3 / 251.8	73.6 / 136.4	88.9 / 180.2	38.6 / 90.3	162.65 / 298.91
EO-NWPU	134.3 / 281.8	103.6 / 186.4	128.9 / 210.2	61.7 / 143.5	52.4 / 108.9
ODF-A	137.2 / 218.5	112.6 / 188.4	103.8 / 174.9	89.6054 / 140.8509	22.161 / 40.741

feature maps are passed through a 1×1 convolution followed by BN and ReLU activation, producing a fused representation (Eq. (2)) that serves as a compact, context-rich feature map for the subsequent density reconstruction stage.

Third, a **Density Reconstruction and Count Fusion (DRCF)** mechanism factorizes the prediction into shape and count components. Then it fuses the visible-human representation with EO-derived hidden density into a single occlusion-compensated prediction. For each sample b, the fused feature representation $\mathbf{F}_b$ is passed to the **Shape Object** and **Count Object** branches. The Shape Object branch applies a 1×1 convolution to obtain the spatial representation $\mathbf{S}_b$, while the Count Object branch applies global average pooling followed by a fully connected layer and Softplus activation to estimate the non-negative count C_b (Eq. (3)).

$$\mathbf{S}_b = \mathrm{Conv}_{1\times1}(\mathbf{F}_b)\,, \qquad C_b = \mathrm{Softplus}(f_{\mathrm{cnt}}(\mathrm{GAP}(\mathbf{F}_b)))\,. \tag{3}$$

DRCF applies this reconstruction separately to humans and EOs (e.g., umbrellas and pickets), producing (P_h, C_h) and (P_k, C_p), respectively. The predicted picket and umbrella density maps form an occlusion-aware representation of the EO spatial distribution, from which the **Attention Hidden Density** module estimates hidden-human density $\mathbf{D}_b^{\mathrm{hid}}$ separately for each occluder type (Eq. (4)), where $\sigma(\cdot)$ produces the spatial attention associated with hidden-human evidence, while $\phi(\cdot)$ generates the corresponding non-negative density response. The element-wise modulation lets the recovered hidden density vary with the predicted occluder configuration, so each picket or umbrella region contributes its own amount of hidden-human density rather than a fixed correction factor. Finally, $\mathbf{D}_b^{\mathrm{hid}}$ is fused with (P_h, C_h) to obtain the final density map and total count (D, C_t).

$$\mathbf{D}_b^{\mathrm{hid}} = \sum_{k\in\{P,U\}} \sigma\Big(\mathcal{C}_k\left(\mathrm{Concat}\left(\hat{\mathbf{D}}_b^P, \hat{\mathbf{D}}_b^U\right)\right)\Big) \odot \phi\Big(\mathbf{W}_k * \mathrm{Concat}\left(\hat{\mathbf{D}}_b^P, \hat{\mathbf{D}}_b^U\right) + \mathbf{B}_k\Big) \tag{4}$$

3 Experimental Results

We compare ODFNet against four SOTA crowd-counting methods, AdaSEEMCount (Wan et al., 2024), TMTB (Yang et al., 2025), P2R-Loss (Lin et al., 2025), and CLIP-EBC (Ma et al., 2025), using four existing crowd-counting benchmarks (SHA, UCF, JHU, and NWPU) as well as five EO-GT datasets (ODF-A, EO-SHA, EO-UCF, EO-JHU, and EO-NWPU). All experiments are implemented in Python (PyTorch and TensorFlow) on an NVIDIA RTX 6000 Blackwell GPU.

3.1 Quantitative Evaluation and Benchmark Error Analysis

Table 6 reveals three key findings in MAE / RMSE. First (Red), ODFNet achieves the best performance on ODF-A with an MAE of 22.161 and RMSE of 40.741, directly reflecting the benefit of explicitly modeling EO-occluded regions. Second (Blue), a consistent performance gap emerges between original and EO-GT benchmarks across all four datasets. On UCF, AdaSEEMCount errors

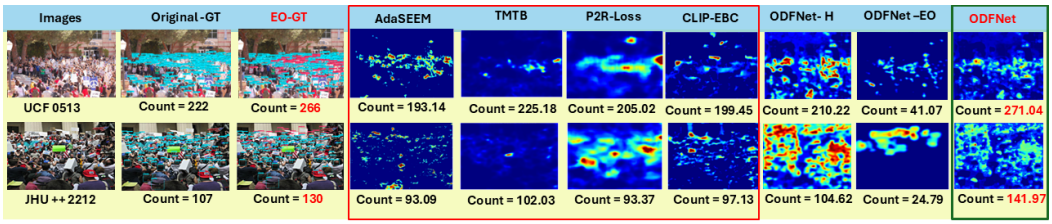

Figure 3: Qualitative evaluation of ODFNet on two representative samples. For each image, we visualize the Original-GT, the corrected EO-GT, and predicted density maps with corresponding counts from AdaSEEMCount (AdaSEEM), TMTB, P2R-Loss, CLIP-EBC, ODFNet-H, ODFNet-EO, and ODFNet. Existing methods undercount occluded crowds by focusing mainly on visible regions, whereas ODFNet produces predictions closest to the EO-GT, confirming the benefit of explicit occlusion-aware density modeling (DRCF).

increase from 182.3 / 289.9 to 214.5 / 332.7 when evaluated against EO-UCF, while CLIP-EBC errors increase from 80.3 / 139.3 to 108.9 / 194.5. Similar patterns hold across EO-SHA, EO-JHU, and EO-NWPU, confirming that published benchmark scores are systematically optimistic due to undercounted GT. Third, ODFNet (22.161 / 40.741) outperforms all benchmarks on ODF-A by a large margin, surpassing AdaSEEMCount (137.2 / 218.5), TMTB (112.6 / 188.4), P2R-Loss (103.8/174.9), and CLIP-EBC (89.605 / 140.851) (Green). This confirms that explicit EO density fusion recovers crowd counts that conventional methods fail to count. Conversely, ODFNet does not always achieve the best performance on the original benchmarks; for example, on SHA it obtains 104.3 MAE and 198.42 RMSE, while CLIP-EBC obtains 52.5 MAE and 85.9 RMSE. This behavior is expected because ODFNet is trained against EO-GT and therefore estimates individuals that original annotations may omit. This apparent overcount is not a weakness but further evidence of GT incompleteness. This is consistent with Table 4, where 7,406 out of 85,455 ODF-A annotations (9.1%) correspond to EO annotations that are systematically ignored by conventional annotation protocols.

3.2 Count-Fusion Property Validation

To verify that DRCF recovers concealed individuals without over- or undercounting, we insert umbrellas or pickets EOs into an image with a known GT count of 193 (H193), varying their number (30 or 70) and placement (scattered or clustered) in Figure 4. Variants are named H193-{U,P}-{30,70}-{S,C}, where U/P denotes umbrella/picket EOs, 30/70 the number of inserted EOs, and S/C scattered/clustered placement. For example, H193-U-70-S contains 70 unbrella EOs in a scattered pattern. Since the crowd is unchanged, GT remains 193 in every variant, so any deviation reflects sensitivity to EOs (Table 6).

On H193, which contains no EOs, all methods count within 5 of GT, and ODFNet is closest (191.25, error 1.75). This indicates that EO modeling does not compromise ODFNet's accuracy on unoccluded scenes. Baselines degrade sharply as EOs increase. Other baseline counts fall to 138.83–165.78 with 30 EOs and 103.10–128.80 with 70, an undercount of up to 47%. EO number dominates the baseline average, drops from 157.13 to 118.84, whereas type and placement shift it by less than 3 and 9 counts, respectively. In contrast, ODFNet stays within 6.1 counts (3.1%) of GT across all variants (186.93–196.59), largely unaffected by EO type, number, or placement. EO vs. Human decomposition explains this gap. ODFNet's visible-human counts (e.g., 122.46 on H193-U-70-S) match the baseline totals (120.78 for CLIP-EBC), indicating that existing methods count only the visible crowd. The EO stream recovers the remainder, scaling with EO number at roughly 0.8–1.1 concealed individuals per EO. Because the raw component sum can still deviate from GT, DRCF calibrates it. For example, it raises the sum on H193-P-70-C from 175.71 to 190.42, reducing the error from 17.29 to 2.58, while leaving accurate estimates nearly unchanged.

3.3 Validations with Real-World Datasets

Figure 3 evaluates ODFNet on two real-world images with EOs from UCF (UCF-0513) and JHU++ (JHU++2212). In the UCF example, the original GT count is 222, while the EO-GT increases to 266.

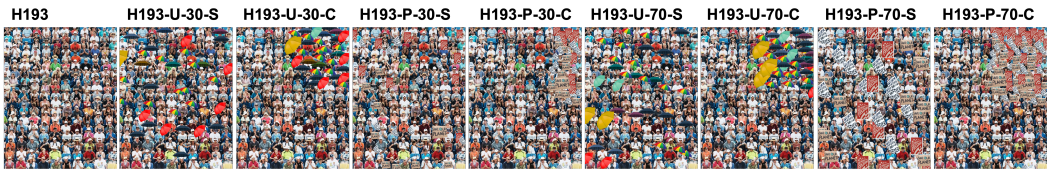

Figure 4: Synthetic occlusion scenarios built from a single image with a known count of 193 humans (H193). Variants are named H193-{U,P}-{30,70}-{S,C}, where U/P denotes umbrella/picket EOs, 30/70 the number of inserted EOs, and S/C scattered/clustered placement. For example, H193-P-30-C contains 30 pickets in a clustered pattern.

Table 6: DRCF counting results on image H193 (GT = 193) under synthetic EO scenarios. Red and blue mark the best and second-best counts (closest to GT), respectively. As EOs increase from 30 to 70, baseline counts fall to 103–129, whereas ODFNet stays within 6.1 counts (3.2%) of GT regardless of EO type, number, or spatial pattern.

Images	AdaSEEM	TMTB	P2R-Loss	CLIP-EBC	**ODFNet (EO : Human)**
H193	188.45	189.86	190.14	195.97	191.25 (0 : 191.25)
H193-U-30-S	147.37	138.83	157.45	164.98	190.25 (29.67 : 155.53)
H193-U-30-C	155.53	159.2	162.13	165.7	196.59 (32.53 : 163.77)
H193-P-30-S	150.15	158.26	153.4	156.85	190.78 (30.82 : 159.78)
H193-P-30-C	160.79	157.86	159.71	165.78	190.17 (26.39 : 155.78)
H193-U-70-S	114.9	116.6	113.6	120.78	186.93 (64.23 : 122.46)
H193-U-70-C	109.89	128.8	126.47	119.5	188.8 (61.73 : 123.81)
H193-P-70-S	103.1	114.9	119.25	122.19	190.68 (63.63 : 124.81)
H193-P-70-C	113.61	117.67	117.69	128.37	190.42 (58.12 : 117.59)

ODFNet predicts 271.04, which is significantly closer to the EO-GT compared with AdaSEEM-Count (193.14), TMTB (225.18), P2R-Loss (205.02), and CLIP-EBC (199.45). In the JHU example, the original GT count is 107, while the EO-GT increases to 130. ODFNet predicts 141.97, closer to the EO-GT count than AdaSEEMCount (93.09), TMTB (102.03), CLIP-EBC (97.13), and P2R-Loss (93.37). ODFNet-H recovers visible human crowd density on par with other SOTA methods, while ODFNet-EO models the spatial distribution of occluders. The full ODFNet fuses both streams to recover hidden individuals and maintain spatial coherence across heavily occluded regions. For example in USF-0513, ODFNet-H count is 210.22 and ODFNet-EO count is 41.07. A simple merge by adding those numbers only 251.29 which is not as good as the result of 271.04. In contrast, conventional methods exhibit severe and inconsistent undercounting, particularly under clustered and densely concentrated occluders. These results confirm that explicit EO density modeling is essential for accurate real-world crowd counting, and that the performance gap between ODFNet and existing SOTA methods grows precisely in the high-occlusion conditions where reliable counts matter most.

4 Conclusion

This work challenges the assumption that published GT annotations reflect true crowd size. By auditing four widely used datasets, we found that 40.73% of images contain occlusions and 78.46% of these involve EOs, yet no annotation protocol recovers individuals concealed behind them. To close this gap, we introduced a three-class occlusion taxonomy (AO, BO, EO), a per-image audit methodology, and an annotation tool, ODFlow, that corrects GT by combining occluder-specific annotation with local density inference. The correction revealed a 6.3% average undercount in EO-affected images, so published MAE understates real-world error by about 22 counts per image. We further proposed ODFNet, whose DRCF module reconstructs human and EO density maps separately and fuses them while removing EO-induced redundancy, yielding consistent counts across EO type, number, and spatial pattern where existing methods drift substantially. We release corrected GT for all four datasets (EO-GT), a 491-image occlusion-aware dataset (ODF-A), a benchmark re-evaluation, and ODFNet. Together, these contributions reframe EOs as structured, recoverable scene elements rather than incidental noise, establishing a corrected foundation for future crowd counting research.

REFERENCES

Khalid J. Almalki, Baek-Young Choi, Yu Chen, and Sejun Song. Characterizing scattered occlusions for effective dense-mode crowd counting. In *ICCV Workshops*, 2021.

Anonymous Authors. Anonymous Code Repository for ODFNet and ODFlow Occlusion-Aware Crowd Counting. Anonymous GitHub Repository and Anonymous Supplementary Repository, 2026a. Code and supplementary files are available at: https://drive.google.com/drive/folders/1UBWsyx7A97eVsFafMmJ8i5-G-E2pi4dt?usp=drive_link.

Anonymous Authors. ODF-A: An Anonymous Occlusion-Aware Crowd Counting Dataset. Anonymous Supplementary Dataset Repository, 2026b. Dataset and annotations are available at: https://drive.google.com/drive/folders/1XMWJTjXjGf47CdvgJQf4c7j_M4uq0bbx?usp=drive_link.

Haoyue Bai, Jiageng Mao, and S.-H. Gary Chan. A survey on deep learning-based single image crowd counting: Network design, loss function and supervisory signal. *Neurocomputing*, 508: 1–18, 2022. doi: 10.1016/j.neucom.2022.08.037. URL https://doi.org/10.1016/j.neucom.2022.08.037.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with atrous convolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. doi: 10.1109/TPAMI.2017.2699184.

Heba F Elsepae, Heba M El-Hoseny, Ehab KI Hamad, and El-Sayed M El-Rabaie. Deep learning for crowd counting in complex environments: challenges and novel trends. *Discover Computing*, 29(1):101, 2026.

Miguel Fiandeiro, Thanh Thi Nguyen, Hanting Wong, and Edbert B. Hsu. Modernized crowd counting strategies for mass gatherings-a review. *Journal of Acute Medicine*, 13(1):4–11, 2023. doi: 10.6705/j.jacme.202303_13(1).0002. URL https://doi.org/10.6705/j.jacme.202303_13(1).0002.

Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. CNN-based density estimation and crowd counting: A survey. 2020. URL https://arxiv.org/abs/2003.12783.

Junyu Gao, Qi Wang, and Yuan Yuan. Scar: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing*, 363:1–8, 2019.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Zhi-Kai Huang, Wei-Ting Chen, Yuan-Chun Chiang, Sy-Yen Kuo, and Ming-Hsuan Yang. Counting crowds in bad weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Haroon Idrees, Muhammad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Computer Vision – ECCV 2018*, pp. 544–559, 2018. doi: 10.1007/978-3-030-01216-8_33. URL https://arxiv.org/abs/1808.01050.

Rohan Kothari, Shubham Gupta, Ankit Singh, and Amit Kumar. Deep learning in crowd counting: A survey. *IET Computer Vision*, 17(5):379–402, 2023.

Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, 2010. URL https://papers.neurips.cc/paper/4043-learning-to-count-objects-in-images.pdf.

Chao Li, Aojun Zhou, and Anbang Yao. Omni-dimensional dynamic convolution. In *International Conference on Learning Representations (ICLR)*, 2022.

Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1091–1100, 2018. doi: 10.1109/CVPR.2018.00120. URL https://openaccess.thecvf.com/content_cvpr_2018/papers/Li_CSRNet_Dilated_Convolutional_CVPR_2018_paper.pdf.

Dingkang Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. Transcrowd: Weakly-supervised crowd counting with transformers. *Science China Information Sciences*, 65(6): 160104, 2022a. doi: 10.1007/s11432-021-3445-y. URL https://doi.org/10.1007/s11432-021-3445-y.

Dingkang Liang, Wei Xu, and Xiang Bai. End-to-end crowd counting via vision transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022b.

Wei Lin, Chenyang Zhao, and Antoni B. Chan. Point-to-region loss for semi-supervised point-based crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5099–5108, 2019.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

Yiming Ma, Victor Sanchez, and Tanaya Guha. Clip-ebc: Clip can count accurately through enhanced blockwise classification. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2025.

Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *European Conference on Computer Vision (ECCV)*, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788. IEEE, 2016. doi: 10.1109/CVPR.2016.91. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Vishwanath A. Sindagi, Rajeev Yasarla, and Vishal M. Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2594–2609, 2022. doi: 10.1109/TPAMI.2020.3035969. URL https://doi.org/10.1109/TPAMI.2020.3035969.

Qingyu Song, Changan Wang, and Zhaoxiang Jiang. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3365–3374, 2021.

Jia Wan, Qiangqiang Wu, Wei Lin, and Antoni B. Chan. Robust zero-shot crowd counting and localization with adaptive resolution sam. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8410–8419, 2020a.

Mingtao Wang, Xin Zhou, and Yuanyuan Chen. A comprehensive survey of crowd density estimation and counting. *IET Image Processing*, 2025.

Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8198–8207, 2019.

Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020b.

Xinlong Wang, Bin Xiao, Fangyin Wei, and Jian Sun. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7774–7783, 2018. doi: 10.1109/CVPR.2018.00811.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Maochen Yang, Zekun Li, Jian Zhang, Lei Qi, and Yinghuan Shi. Taste more, taste better: Diverse data and strong model boost semi-supervised crowd counting. pp. 24440–24451, 2025.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 589–597, 2016.

Jiarui Zhou, Jianming Zhang, and Yan Gui. Crowd counting in domain generalization based on multi-scale attention and hierarchy level enhancement. *Scientific Reports*, 15(1):155, 2025.

A Appendix

A.1 Dataset Availability

The ODF-A dataset introduced in this work is made available through an anonymous supplementary repository. The dataset contains the images, EO-aware annotations, and associated ground-truth files used for training and evaluation in this study.

ODF-A Dataset Repository:

`https://drive.google.com/drive/folders/1XMWJTjXjGf47CdvgJQf4c7j_M4uq0bbx?usp=drive_link`

The repository is provided anonymously for the review period and will be updated with permanent author and project information upon publication.

A.2 Preprocessing, Data Augmentation, and Ground-Truth Generation

ODF-A images are uniformly resized to 512×512 pixels prior to training to maintain a consistent spatial resolution that preserves fine-grained details while ensuring computational efficiency. To improve robustness against viewpoint variation, illumination changes, and sensor artifacts, we employ a controlled data augmentation pipeline during training. Each image is randomly rotated within $[-15°, 15°]$. Approximately 15% of training images are converted to grayscale, while the remaining samples undergo moderate color jittering in hue, saturation, and brightness. Additionally, low-amplitude impulse noise is injected into up to 0.1% of pixels to simulate sensor and compression noise. All geometric transformations are synchronously applied to the corresponding annotations to preserve spatial consistency between human and EO objects.

We introduce ODFlow Authors (2026a) (in Figure 5), a custom annotation tool built to support occlusion-aware GT generation. ODFlow enables manual labeling of both visible crowd regions and embedded occluders, such as umbrellas and picket signs, using bounding-box annotations, and exports structured JSON labels containing box coordinates, class counts, image metadata, and totals alongside the original image. Unlike general-purpose tools such as Roboflow, ODFlow is purpose-built for our corrected GT workflow, offering full control over annotation logic and output format — essential for systematically studying crowd members hidden by embedded occluders. All images are annotated with human, umbrella, and picket objects to generate corresponding GT density

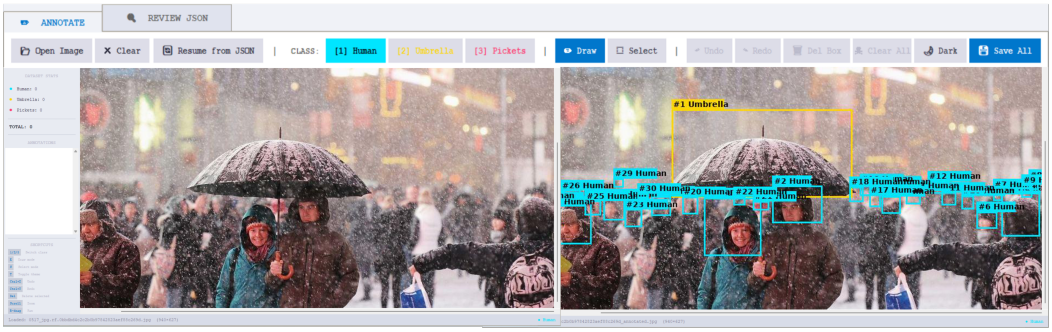

Figure 5: **ODFlow** interface showing original (left) and annotated (right) crowd scene with Human (cyan) and Umbrella (yellow) bounding-box labels for EO-GT generation.

maps. We adopt the geometry-adaptive Gaussian kernel method (Zhang et al., 2016), building on the density-estimation formulation of (Lempitsky & Zisserman, 2010). Specifically, the annotated object locations in the original image are transformed into a continuous GT density map $F(x)$ as defined in Eq. (5).

$$F(x) = \sum_{t=1}^{N} \delta(x - x_i) * G_{\sigma i}(x), \text{ with } \sigma_i = \beta \overline{\mathrm{d_i}} \tag{5}$$

where N denotes the total number of annotated objects in the image, x_i represents the spatial location of the i-th object, and $\overline{d}_i$ is the average distance to its k-nearest neighbors. The delta function $\delta(x - x_i)$ is convolved with a Gaussian kernel whose standard deviation is σ_i, producing a smooth density heatmap for each object. For the ODF-A dataset, we generate three distinct GT density maps and corresponding counts for different labeled object categories (Wang et al., 2019). In Eq. (6), given an image with N bounding boxes, the GT density map $D^{\mathrm{gt}} \in \mathbb{R}^{H_o \times W_o}$ is constructed by assigning uniform mass within each bounding-box region. Let R_n denote the set of output-grid pixels covered by the n-th bounding box.

$$D^{\mathrm{gt}}(i,j) = \sum_{n=1}^{N} \frac{1[(i,j) \in R_n]}{|R_n|}, \qquad \sum_{i,j} D^{\mathrm{gt}}(i,j) = N \equiv C^{\mathrm{gt}} \tag{6}$$

A.3 Evaluation Metrics: MAE and RMSE

For quantitative assessment, we adopt the standard metrics of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), as defined below:

$$MAE = \frac{1}{N} \sum_{b=1}^{N} |C_b - C_b^{\mathrm{gt}}| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{b=1}^{N} \left(C_b - C_b^{\mathrm{gt}}\right)^2} \tag{8}$$

where N denotes the total number of test images, C_b represents the predicted count for the b-th image, and C_b^{gt} denotes its corresponding ground-truth count.

A.4 Attention Hidden Density

To explicitly estimate human density concealed by embedded occluders (EO), we use the predicted Picket and Umbrella density maps represented by D_{EO}.

For each occluder category $k \in \{P, U\}$, a learnable convolutional operator predicts a spatial hidden-human attention map,

$$P_k^{\mathrm{hid}} = \sigma\left(C_k\left(D_{\mathrm{EO}}\right)\right), \qquad k \in \{P, U\}, \tag{9}$$

where $C_k(\cdot)$ denotes the occluder-specific convolutional predictor and $\sigma(\cdot)$ is the sigmoid activation function. The resulting P_k^{hid} provides a spatial attention response indicating the strength of hidden-human evidence associated with each EO type.

In parallel, a non-negative hidden-density intensity map is estimated as

$$\Lambda_k = \phi\left(W_k * D_{\mathrm{EO}} + B_k\right), \qquad k \in \{P, U\}, \tag{10}$$

where W_k and B_k denote learnable convolutional weights and bias terms, respectively, $*$ denotes convolution, and $\phi(\cdot)$ is a non-negative activation function.

The EO-specific hidden-human density is obtained by modulating the hidden-density intensity with the corresponding hidden-human attention map,

$$D_k^{\mathrm{hid}} = P_k^{\mathrm{hid}} \odot \Lambda_k, \qquad k \in \{P, U\}, \tag{11}$$

where $\odot$ denotes element-wise multiplication.

Substituting Eqs. (9) and (10) into Eq. (11) gives

$$D_k^{\mathrm{hid}} = \sigma\left(C_k\left(D_{\mathrm{EO}}\right)\right) \odot \phi\left(W_k * D_{\mathrm{EO}} + B_k\right), \qquad k \in \{P, U\}. \tag{12}$$

The total hidden-human density is then expressed as

$$D_H^{\mathrm{hid}} = \sum_{k \in \{P,U\}} D_k^{\mathrm{hid}}, \tag{13}$$

or equivalently,

$$D_H^{\mathrm{hid}} = \sum_{k \in \{P,U\}} \sigma\left(C_k\left(D_{\mathrm{EO}}\right)\right) \odot \phi\left(W_k * D_{\mathrm{EO}} + B_k\right). \tag{14}$$

The corresponding hidden-human count is obtained by integrating the estimated hidden-human density over the spatial domain,

$$C_H^{\mathrm{hid}} = \frac{1}{S} \sum_{x,y} D_H^{\mathrm{hid}}(x, y), \tag{15}$$

where S denotes the density normalization factor.

This formulation separates the spatial hidden-human attention from the magnitude of the recovered hidden density. The sigmoid-based attention term determines where EO-related human evidence should be emphasized, while the non-negative density-intensity term determines the corresponding density contribution. This allows Picket- and Umbrella-induced hidden-human density to adapt spatially across different EO configurations.

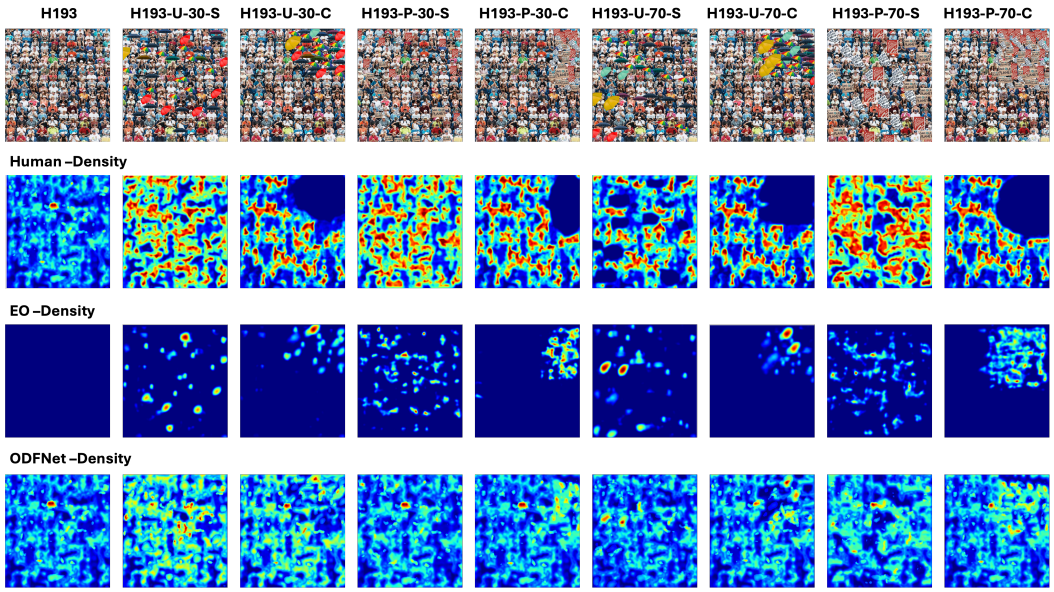

Figure 6: Qualitative analysis of ODFNet under controlled synthetic occlusion. Starting from the H193 scene with a fixed GT count of 193 humans, we insert umbrella (U) or picket (P) EOs in quantities of 30 or 70 using scattered (S) or clustered (C) layouts. The first row shows the resulting synthetic scenes, followed by the predicted Human-Density, EO-Density, and final ODFNet-Density maps. The human and EO branches capture complementary visible-crowd and occlusion-related spatial information, while their fusion produces a stable reconstructed crowd-density response across substantial changes in EO type, number, and spatial arrangement.