DECO: A MULTIMODAL DETECTION FRAMEWORK BASED ON DENSITY CONSENSUS LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal object detection aims to exploit complementary observations from heterogeneous sensors for accurate object localization. However, existing fusion approaches mainly rely on feature-level interaction without explicitly modeling object distribution and spatial organization, limiting their performance in dense scenes. This limitation is particularly evident in remote sensing applications such as optical and SAR ship detection. In this paper, we propose DeCo, a multimodal detection framework based on density consensus learning. DeCo formulates multimodal fusion as consensus estimation over a shared object distribution space, where feature interaction is guided by cross-modal spatial agreement and density consensus, enabling robust multimodal detection under high target density. Specifically, modality-specific encoders transform optical and SAR observations into a shared latent density space, where spatially varying modality reliability is estimated to construct a cross-modal density consensus representation. The learned consensus subsequently regulates multiscale feature fusion by adaptively allocating modality contributions and enhancing spatially consistent object evidence. To supervise density learning without manual density profile design, we introduce a box-aware Bayesian density objective that optimizes instance-level mass assignment through probabilistic ownership modeling. Extensive experiments on paired optical–SAR ship detection benchmarks demonstrate that DeCo achieves robust multimodal detection and accurate object counting, especially in densely packed scenes.

1 INTRODUCTION

Object detection in high-resolution remote sensing imagery is important for maritime surveillance, port management, and rescue. Compared with objects in natural images, remotely sensed targets often exhibit arbitrary orientations, extreme aspect ratios, and complex spatial distributions (Cheng & Han, 2016; Xia et al., 2018; Li et al., 2020b). These challenges are particularly severe for ships, which may occupy only a few pixels offshore or appear densely packed near ports and coastal structures. Most existing detectors, however, rely on a single sensing modality and are therefore sensitive to modality-specific degradation. Optical (OPT) imagery provides rich spectral, textural, and contextual cues but is vulnerable to clouds, haze, and poor illumination. Synthetic aperture radar (SAR) enables all-weather and day-and-night observation but suffers from speckle noise, complex scattering, and coastal clutter. Their complementary sensing characteristics make OPT–SAR fusion attractive for robust ship detection.

Existing multimodal detectors improve perception by exchanging or aligning information across modalities through attention, geometric alignment, or reliability-aware interaction (Hwang et al., 2015; Sun et al., 2022b; Fang et al., 2021; Shen et al., 2024; Yuan & Wei, 2024; Dong et al., 2025). However, OPT–SAR fusion is particularly challenging because corresponding ships are represented by fundamentally different physical responses: optical reflectance in OPT imagery and microwave scattering in SAR imagery. Consequently, feature similarity does not necessarily imply target consistency, and the reliability of each modality varies spatially under cloud contamination, weak backscatter, speckle noise, or coastal interference. Existing feature-level fusion does not explicitly model where the two modalities agree on the existence and spatial organization of ships. We therefore seek a target-centric intermediate representation. It should capture cross-modal agreement without relying on modality-specific appearance.

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.

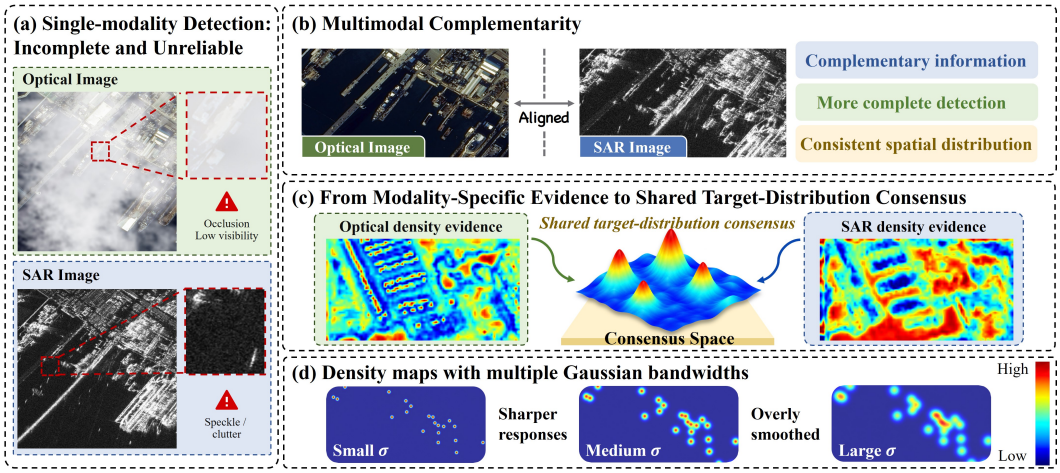

Figure 1: Motivation for density consensus learning. OPT and SAR observations provide heterogeneous and spatially varying evidence for the same underlying ship distribution. We establish their agreement in a shared density space to preserve mutually supported target responses while suppressing modality-specific interference. Different Gaussian bandwidths in (d) illustrate the dependence of conventional pixel-wise density targets on predefined kernel shapes.

Object density provides a natural candidate for this purpose. A density representation describes both target existence and spatial concentration, allowing heterogeneous OPT and SAR observations to be mapped into a shared target-distribution space. As illustrated in Fig. 1, the two modalities may exhibit different local degradations while still observing the same underlying ship distribution. This suggests establishing cross-modal consensus in density space rather than directly enforcing similarity between heterogeneous features. For detection-oriented density learning, the essential requirement is instead to preserve instance-wise target mass and suppress background responses without forcing the network to reproduce an artificial Gaussian shape. Distribution-based objectives such as Bayesian loss provide a principled way to achieve this through probabilistic mass assignment (Ma et al., 2019; Wan et al., 2021).

Based on these observations, we propose **DeCo**, a multimodal detection framework based on density consensus learning. DeCo first derives modality-specific density evidence from multi-scale OPT and SAR features and maps them into a shared density space. It then estimates spatially varying modality reliability to construct a consensus density representation, which regulates the contribution of each modality and guides feature fusion across detection scales. To learn density representations without predefined pixel-wise density targets, we introduce a box-aware Bayesian objective that assigns predicted density mass to individual objects and background according to box-conditioned probabilistic ownership. In this way, DeCo uses agreement in target distribution to organize heterogeneous cross-modal interaction. The main contributions are summarized as follows:

- We propose **DeCo**, a multimodal detection framework that formulates cross-modal fusion as consensus estimation over a shared target-distribution space, explicitly modeling cross-modal agreement on target existence and spatial organization.
- We develop a density-guided interaction mechanism that learns modality-specific density evidence and spatially varying modality reliability, and uses the resulting consensus representation to adaptively regulate feature fusion across multiple detection scales.
- We introduce a box-aware Bayesian density objective that constrains instance-wise density mass through probabilistic ownership assignment. It avoids predefined pixel-wise Gaussian targets and provides more flexible density supervision for dense target detection.

2 Related Work

Remote Sensing Ship Detection. This task has been extensively studied in both optical and SAR domains. Optical benchmarks such as DOTA (Xia et al., 2018), DIOR (Li et al., 2020b), FAIR1M

(Sun et al., 2022a), and HRSC2016 (Liu et al., 2017) have driven rapid advances in remote sensing object detection, while SAR datasets such as SSDD (Li et al., 2017) and HRSID (Wei et al., 2020) have advanced SAR ship detection under all-weather conditions. The complementary characteristics of optical and SAR imagery have motivated research on joint OPT-SAR detection.

Multimodal Detection. Multimodal object detection improves perception robustness by integrating complementary sensing information. Early methods mainly relied on direct input fusion or feature concatenation, while later approaches increasingly adopted adaptive cross-modal interaction. Representative methods employ attention-based exchange (Fang et al., 2021; Shen et al., 2024), transformer or state-space architectures (Yuan & Wei, 2024; Dong et al., 2025), conflict-aware learning (He et al., 2023), geometric alignment (Yuan et al., 2022), illumination-aware interaction (Hu et al., 2025), bidirectional alignment (Shi et al., 2026), and sparse feature exchange (Tian et al., 2026). These methods improve heterogeneous feature integration, but most estimate modality contributions from feature responses or cross-modal correlations.

Existing studies, including QXS-SAROPT (Huang et al., 2021), DKDNet (Yue et al., 2025), M4-SAR (Wang et al., 2026), and MHFNet (Zhang et al., 2026), investigate paired data construction, knowledge distillation, feature alignment, and uncertainty-aware fusion. However, feature-level interaction alone may not reliably indicate target consistency. Our work instead models cross-modal agreement in a shared target-distribution space.

Density-Based Learning for Object Detection and Counting. Density estimation represents object distributions by mapping discrete annotations into continuous spatial responses. Classical counting methods (Zhang et al., 2016; Li et al., 2018) regress density maps whose integrals approximate object counts while retaining spatial information. This paradigm has been extended to remote sensing counting under severe scale variation and nonuniform object distributions (Gao et al., 2021; 2022; Ding et al., 2022; Gao et al., 2024; Chen et al., 2026). Density information has also been incorporated into object detection as a spatial prior for crowded-region discovery, adaptive cropping, and dense-object refinement (Li et al., 2020a; Duan et al., 2021; Zhao et al., 2026).

Existing methods mainly treat density as a unimodal counting representation or auxiliary output. Most construct pixel-wise supervision by placing Gaussian kernels at object locations and optimizing regression losses. Fixed kernels are sensitive to object scale and spacing, while adaptive or learned kernels (Wan & Chan, 2019) still impose predefined density profiles. Distribution-based objectives provide an alternative by emphasizing instance-wise mass assignment rather than exact pixel correspondence.

3 METHODOLOGY

3.1 FRAMEWORK OVERVIEW

Given a spatially paired optical image $\mathbf{I}^o \in \mathbb{R}^{3\times H\times W}$ and SAR image $\mathbf{I}^s \in \mathbb{R}^{1\times H\times W}$, our goal is to exploit their complementary observations for ship detection. OPT reflectance and SAR scattering responses are inherently heterogeneous, making direct feature similarity an unreliable indicator of target consistency. We therefore formulate OPT-SAR fusion as consensus estimation over a shared target-distribution space, where the two modalities are encouraged to agree on spatial evidence of target existence rather than appearance responses.

As illustrated in Fig. 2, two identical but parameter-independent backbones extract modality-specific feature pyramids:

$$\left\{\mathbf{F}_m^l\right\}_{l=3}^{5} = \mathcal{B}_m\left(\mathbf{I}^m\right), \qquad m \in \{o, s\}, \tag{1}$$

where $\mathcal{B}_m$ denotes the backbone for modality m, and $l \in \{3, 4, 5\}$ indexes feature levels with strides of 8, 16, and 32. A density-consensus estimator $\mathcal{C}$ maps them into a shared latent space:

$$(\mathbf{Z}_o, \mathbf{Z}_s, \mathbf{Z}_c, \mathbf{W}) = \mathcal{C}\left(\left\{\mathbf{F}_o^l, \mathbf{F}_s^l\right\}_{l=3}^{5}\right), \tag{2}$$

where $\mathbf{Z}_o$ and $\mathbf{Z}_s$ denote modality-specific density evidence, $\mathbf{W} = [\mathbf{W}_o, \mathbf{W}_s]$ represents spatially varying modality reliability, and $\mathbf{Z}_c$ is the resulting density consensus. The consensus is then injected into the detection pathway:

$$\widetilde{\mathbf{F}}^l = \mathcal{G}_l\left(\mathbf{F}_o^l, \mathbf{F}_s^l; \mathbf{Z}_c, \mathbf{W}\right), \qquad l \in \{3, 4, 5\}, \tag{3}$$

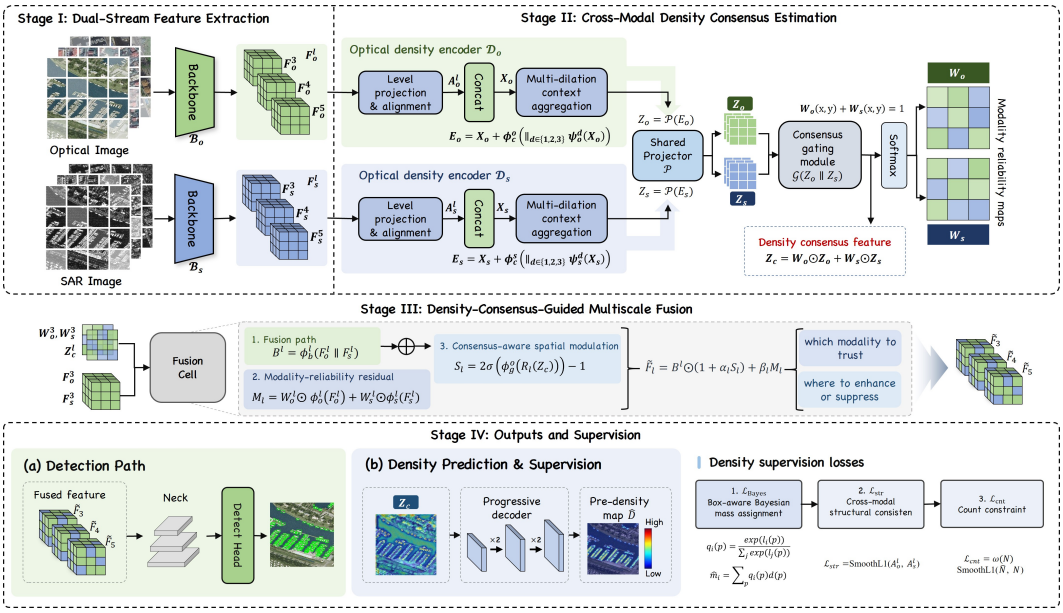

Figure 2: Overview of DeCo. Independent backbones extract multimodal, multiscale features, which are projected into a shared density space to estimate modality reliability and construct density consensus. The resulting consensus guides multiscale feature fusion for detection and is simultaneously decoded into a density map for density supervision.

where $\mathcal{G}_l$ denotes the fusion operator at level l. Target-distribution consensus directly governs cross-modal interaction for detection. The fused features $\widetilde{\mathbf{F}}^l$ are processed by the detection neck and head, while $\mathbf{Z}_c$ is decoded into a density map and supervised by the density-learning objectives introduced below.

3.2 Cross-Modal Density Consensus Estimation

Directly comparing OPT and SAR features is unreliable because optical reflectance and microwave scattering exhibit substantially different response statistics. Therefore, we transform modality-specific multiscale features into a shared latent density space, where cross-modal agreement is established according to target-distribution evidence.

Modality-Specific Density Encoding. For each modality $m \in \{o, s\}$, the multiscale backbone features $\{\mathbf{F}_m^l\}_{l=3}^5$ are projected to a common channel dimension and spatially aligned to the resolution of the finest level:

$$\mathbf{A}_m^l = \mathcal{U}_l\left(\phi_m^l\left(\mathbf{F}_m^l\right)\right), \qquad l \in \{3,4,5\}, \tag{4}$$

where ϕ_m^l denotes the level-specific projection and $\mathcal{U}_l$ resizes the feature to the resolution of $\mathbf{F}_m^3$. The aligned features are concatenated and aggregated as $\mathbf{X}_m$. To capture density evidence of target distributions at different spatial extents, we employ parallel context operators with multiple dilation rates:

$$\mathbf{E}_m = \mathbf{X}_m + \phi_m^c\left(\Big\|_{d\in\{1,2,3\}} \psi_m^d\left(\mathbf{X}_m\right)\right), \tag{5}$$

where ψ_m^d denotes a context operator with dilation rate d, and ϕ_m^c fuses the resulting features.

Shared-Space Projection and Consensus Construction. Using a shared projector $\mathcal{P}$, the modality-specific representations are mapped into a common latent density space as $\mathbf{Z}_o = \mathcal{P}(\mathbf{E}_o)$ and $\mathbf{Z}_s = \mathcal{P}(\mathbf{E}_s)$, with $\mathbf{Z}_o, \mathbf{Z}_s \in \mathbb{R}^{C_d\times H/8\times W/8}$. Since modality reliability varies spatially, we estimate competitive contribution maps from the joint representation:

$$[\mathbf{W}_o, \mathbf{W}_s] = \mathrm{Softmax}_m\left(\mathcal{G}\left(\mathbf{Z}_o \| \mathbf{Z}_s\right)\right), \tag{6}$$

where $\mathcal{G}$ denotes the consensus-gating function, and the softmax is applied along the modality dimension. Thus, $\mathbf{W}_o(x,y) + \mathbf{W}_s(x,y) = 1$ at each spatial location. The density consensus is then constructed as

$$\mathbf{Z}_c = \mathbf{W}_o \odot \mathbf{Z}_o + \mathbf{W}_s \odot \mathbf{Z}_s, \tag{7}$$

where $\odot$ denotes element-wise multiplication. The resulting $\mathbf{Z}_c$ aggregates target-distribution evidence according to spatially varying modality reliability and serves as the shared consensus representation for subsequent density-guided detection.

3.3 Density-Consensus-Guided Multiscale Fusion

The density consensus provides two complementary cues for multimodal fusion: the modality weights $\mathbf{W}_o$ and $\mathbf{W}_s$ indicate which modality is more reliable at each location, while $\mathbf{Z}_c$ describes where target-related responses should be enhanced or suppressed. We inject both cues into multiscale detection features through residual modulation.

For each feature level $l \in \{3,4,5\}$, a fused representation is first constructed by

$$\mathbf{B}^l = \phi_b^l\left(\mathbf{F}_o^l \| \mathbf{F}_s^l\right), \tag{8}$$

where ϕ_b^l denotes a 1×1 convolution followed by batch normalization and SiLU activation. The modality weights are resized to the corresponding feature resolution and used to aggregate modality-specific features:

$$\mathbf{M}^l = \mathcal{R}_l(\mathbf{W}_o) \odot \phi_o^l(\mathbf{F}_o^l) + \mathcal{R}_l(\mathbf{W}_s) \odot \phi_s^l(\mathbf{F}_s^l), \tag{9}$$

where $\mathcal{R}_l$ denotes spatial resampling and ϕ_o^l and ϕ_s^l are modality-specific projections.

Meanwhile, the consensus representation is transformed into a signed spatial modulation map:

$$\mathbf{S}^l = 2\sigma\left(\phi_g^l\left(\mathcal{R}_l(\mathbf{Z}_c)\right)\right) - 1, \tag{10}$$

where ϕ_g^l projects the consensus feature to a spatial response and $\mathbf{S}^l \in [-1,1]$. The final density-guided fusion is formulated as

$$\widetilde{\mathbf{F}}^l = \mathbf{B}^l \odot \left(1 + \alpha_l \mathbf{S}^l\right) + \beta_l \mathbf{M}^l, \qquad l \in \{3,4,5\}, \tag{11}$$

where $\alpha_l = \tanh(a_l)$ and $\beta_l = \tanh(b_l)$ are learnable coefficients that bound the influence of density-induced modulation. The resulting features $\widetilde{\mathbf{F}}^l$ combine spatial consensus guidance with reliability-aware modality allocation before being fed into the detection head.

3.4 Density Prediction and Supervision

The consensus representation $\mathbf{Z}_c$ is decoded into a full-resolution density map $\widehat{\mathbf{D}} \in \mathbb{R}^{1\times H\times W}$, which is constrained to be nonnegative. DeCo supervises $\widehat{\mathbf{D}}$ through instance-wise probabilistic mass assignment, without imposing a predefined pixel-wise density profile.

Box-Aware Bayesian Mass Supervision. For an image containing N annotated targets, the i-th box is represented by $\mathbf{b}_i = (c_{x,i}, c_{y,i}, w_i, h_i)$. Its spatial uncertainty is defined according to the box geometry:

$$\sigma_{x,i} = \max(\eta w_i, \sigma_{\min}), \qquad \sigma_{y,i} = \max(\eta h_i, \sigma_{\min}), \tag{12}$$

where η controls the box-to-scale conversion and $\sigma_{\min}$ avoids degenerate scales for small targets.

For a pixel $\mathbf{p} = (p_x, p_y)$, the box-conditioned likelihood of target i is defined as

$$\ell_i(\mathbf{p}) = -\frac{1}{2}\left[\frac{(p_x - c_{x,i})^2}{\sigma_{x,i}^2} + \frac{(p_y - c_{y,i})^2}{\sigma_{y,i}^2}\right]. \tag{13}$$

Together with a background component $\ell_0(\mathbf{p})$, the posterior ownership of each pixel is

$$q_i(\mathbf{p}) = \frac{\exp(\ell_i(\mathbf{p}))}{\sum_{j=0}^{N}\exp(\ell_j(\mathbf{p}))}, \qquad i \in \{0,\ldots,N\}. \tag{14}$$

The predicted mass assigned to target i is then

$$\widehat{m}_i = \sum_{\mathbf{p}\in\Omega} q_i(\mathbf{p}) \frac{\widehat{\mathbf{D}}(\mathbf{p})}{\kappa}, \tag{15}$$

where Ω denotes the image domain and κ is the density mass corresponding to one target. Each foreground instance is expected to receive unit mass, whereas the background receives zero mass. The Bayesian objective is therefore written as

$$\mathcal{L}_{\text{Bayes}} = \frac{1}{B}\sum_{b=1}^{B}\left[\sum_{i=1}^{N_b} e_i^{(b)}\,\text{sg}\left(e^{-e_i^{(b)}}\right) + e_0^{(b)}\right], \tag{16}$$

where $e_i = |m_i^* - \widehat{m}_i|$, B is the batch size, and $\text{sg}(\cdot)$ denotes stop-gradient.

Structural and Count Constraints. To encourage OPT and SAR density representations to preserve consistent spatial organization, we compare their normalized energy maps. For modality $m \in \{o, s\}$, $\mathbf{A}_m(x,y) = \frac{1}{C_d}\sum_{c=1}^{C_d}\mathbf{Z}_m^2(c,x,y)$, $\overline{\mathbf{A}}_m = \frac{\mathbf{A}_m}{\text{Mean}_{x,y}(\mathbf{A}_m)+\epsilon}$. The cross-modal structure loss is $\mathcal{L}_{\text{str}} = \text{SmoothL1}\left(\overline{\mathbf{A}}_o, \overline{\mathbf{A}}_s\right)$. We further constrain the image-level density integral using

$$\widehat{N}_b = \frac{1}{\kappa}\sum_{\mathbf{p}\in\Omega}\widehat{\mathbf{D}}_b(\mathbf{p}), \tag{17}$$

and define the count loss as

$$\mathcal{L}_{\text{cnt}} = \frac{1}{B}\sum_{b=1}^{B}\omega(N_b)\,\text{SmoothL1}\left(\widehat{N}_b, N_b\right), \tag{18}$$

where N_b is the number of annotated targets and $\omega(\cdot)$ assigns larger weights to denser scenes.

Overall Training Objective. The overall supervision objective of DeCo is formulated as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{den}}\left(\mathcal{L}_{\text{Bayes}} + \lambda_{\text{str}}\mathcal{L}_{\text{str}} + \lambda_{\text{cnt}}\mathcal{L}_{\text{cnt}}\right), \tag{19}$$

where $\mathcal{L}_{\text{det}}$ denotes the detection loss.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate DeCo on two paired OPT-SAR ship datasets. See Appendix B.2 for more details. **MOS-Ship** (You et al., 2026) contains 1,246 paired 1024×1024 scenes (996 training and 250 test pairs), with 8,992 ship instances annotated by refined horizontal bounding boxes. The test set additionally provides cloud-degraded optical observations for robustness evaluation, while the real paired optical and SAR observations enable complementary validation under real heterogeneous sensing conditions. **DenseShip6K** is constructed for evaluating multimodal detection under substantially denser target distributions. It contains 6,000 paired 512×512 OPT-SAR images with 279,897 annotated ships, split into 4,500 training and 1,500 test pairs. The dataset contains an average of 46.65 ships per image and a maximum of 648, covering isolated, clustered, and highly congested ship distributions.

Implementation. All experiments are implemented in PyTorch and conducted on an NVIDIA Tesla V100 GPU. All methods are trained for 100 epochs. Detailed implementation and training settings are provided in Appendix B.1.

Evaluation Metrics. Detection performance is evaluated using AP_{50}, and $AP_{50:95}$. Counting accuracy is measured by mean absolute error (MAE) and root mean squared error (RMSE).

Compared Methods. We compare DeCo with representative unimodal detectors, multimodal detectors, and density-estimation approaches. The baselines include TridentNet (Li et al., 2019), Sparse R-CNN (Sun et al., 2021), DINO (Zhang et al., 2022), YOLOv8 (Jocher et al., 2023), YOLO11 (Jocher & Qiu, 2024), YOLOv13 (Lei et al., 2025). The multimodal baselines include CFT (Fang et al., 2021), SuperYOLO (Zhang et al., 2023), ICAFusion (Shen et al., 2024),

DPAL (Liu et al., 2025), EI^{2}Det (Hu et al., 2025), Fusion-Mamba (Dong et al., 2025), and representative density estimators, including MCNN (Zhang et al., 2016), CSRNet (Li et al., 2018), SCAR (Gao et al., 2019), SFCN (Wang et al., 2019), DACount (Lin et al., 2022), MTCP (Zhu et al., 2024), MRC-Crowd (Qian et al., 2024), and TMTB (Yang et al., 2025).

Table 1: Comparison with unimodal and multimodal detectors on the **DenseShip6K** and **MOS-Ship** datasets. MAE and RMSE of detection-based methods are calculated based on the number of detected bounding boxes. ↑/↓ denotes that a higher/lower value is better. The best and second-best results in each column within each dataset are highlighted in **bold** and blue, respectively.

ID	Method	Year	DenseShip6K				MOS-Ship			
			AP_{50}↑	$AP_{50:95}$↑	MAE↓	RMSE↓	AP_{50}↑	$AP_{50:95}$↑	MAE↓	RMSE↓
OPT-based detectors										
#1	TridentNet	2019	63.0	38.3	13.82	33.71	55.3	38.1	3.46	12.23
#2	Sparse R-CNN	2021	59.1	34.5	12.94	32.28	43.3	25.5	3.14	6.61
#3	DINO	2022	84.5	46.4	9.03	17.42	78.8	47.0	1.66	7.83
#4	SuperYOLO	2023	77.9	48.1	6.82	13.39	74.3	42.7	1.43	3.34
#5	YOLOv8s	2023	82.7	52.7	6.66	12.74	70.4	44.1	2.59	9.88
#6	YOLO11s	2024	83.9	56.9	6.58	12.99	71.9	50.2	2.37	9.21
#7	YOLOv13s	2025	83.7	56.4	6.57	12.75	74.2	52.9	2.34	7.63
SAR-based detectors										
#8	TridentNet	2019	55.6	30.6	19.42	35.08	51.3	26.4	4.76	12.72
#9	Sparse R-CNN	2021	52.5	27.9	14.15	32.86	38.5	19.7	3.04	7.91
#10	DINO	2022	78.7	40.0	11.55	21.01	65.5	31.3	1.84	6.15
#11	SuperYOLO	2023	69.6	39.6	7.74	14.52	53.2	25.3	2.58	7.62
#12	YOLOv8s	2023	78.9	50.4	8.02	15.13	67.7	38.1	1.92	7.39
#13	YOLO11s	2024	78.3	49.5	7.77	14.39	66.2	36.5	1.49	4.06
#14	YOLOv13s	2025	78.2	49.1	7.78	14.49	65.5	38.5	2.67	8.48
Multi-modal detectors										
#15	CFT	2021	73.8	43.8	8.09	14.95	75.8	44.1	1.35	4.22
#16	Super-YOLO	2023	77.7	48.2	6.78	13.28	76.4	42.9	1.87	4.20
#17	ICAFusion	2024	69.1	49.7	8.55	19.10	69.9	41.7	1.22	3.41
#18	DPAL-P	2025	68.5	40.1	25.56	35.43	70.4	43.1	4.78	18.95
#19	DPAL-R	2025	74.2	46.3	12.61	21.02	78.4	51.4	1.89	7.73
#20	EI2Det	2025	67.8	42.3	10.50	24.53	69.3	44.0	2.08	10.69
#21	Fusion-Mamba	2025	75.6	46.5	7.75	15.12	79.0	48.9	1.52	5.50
#22	**DeCo (Ours)**	–	**85.0**	**58.5**	**5.85**	**10.51**	**81.9**	**54.7**	**1.03**	**2.02**

4.2 Comparison with State-of-the-Art Methods

Object Detection. Table 1 compares DeCo with representative unimodal and multimodal detectors on DenseShip6K and MOS-Ship. On DenseShip6K, DeCo achieves the best overall performance, reaching 85.0% AP_{50} and 58.5% $AP_{50:95}$, while obtaining the lowest MAE and RMSE of 5.85 and 10.51, respectively. Compared with the strongest unimodal baseline, DeCo improves AP_{50} and $AP_{50:95}$ by 0.5% and 1.6%. The larger margins over existing multimodal methods indicate that simply introducing an additional modality is insufficient under heterogeneous OPT-SAR observations. By using density consensus to regulate cross-modal interaction, DeCo improves both localization accuracy and counting consistency in densely populated scenes. On MOS-Ship, DeCo achieves 81.9% AP_{50} and 54.7% $AP_{50:95}$, outperforming the strongest multimodal baselines by 2.9% and 3.3%, respectively. It also obtains the lowest MAE and RMSE of 1.03 and 2.02. These results show that density-consensus-guided fusion remains effective beyond highly dense scenes and generalizes to multimodal observations with different target distributions and background conditions. Qualitative visualizations of detection results on both datasets are provided in Appendix C.1. For completeness, we report box-based counting results for DeCo and compare the two output paradigms in Appendix C.3.

Density Estimation. We further compare the density output of DeCo with representative density-estimation methods in Table 2. On DenseShip6K, DeCo achieves the lowest MAE of 5.85 and a competitive RMSE of 10.51. On MOS-Ship, it obtains the best MAE and RMSE of 1.03 and 2.02, respectively. Although the density branch is primarily designed to provide target-distribution guidance for multimodal detection, it maintains competitive counting accuracy across both datasets.

4.3 Ablation Studies

Unless otherwise specified, all ablation experiments are conducted on DenseShip6K under the same training setting. We organize the ablation studies around three questions: whether detection gains can be attributed to modality addition alone, whether density consensus provides guidance for the detection task or merely serves as auxiliary supervision, and whether the learned density representation must explicitly participate in cross-modal feature interaction. We further investigate the density supervision objective, the constituent regularization terms, the consensus-guidance mechanism, and the sensitivity to the associated loss weights, with the corresponding results provided in Appendix C.4.

Table 2: Comparison with density estimation methods on the DenseShip6K and MOS-Ship datasets. Lower MAE and RMSE are better. The best and second-best results in each column are highlighted in **bold** and blue, respectively.

Method	DenseShip6K		MOS-Ship	
	MAE↓	RMSE↓	MAE↓	RMSE↓
MCNN (2016)	11.23	18.52	4.04	7.70
CSRNet (2018)	6.38	10.49	4.46	10.60
SCAR (2019)	6.09	10.33	1.81	2.77
SFCN (2019)	5.86	**10.03**	7.29	17.69
DACount (2022)	6.16	10.97	1.07	2.19
MTCP (2023)	9.42	16.00	2.77	9.94
MRC-Crowd (2024)	7.27	13.26	6.67	19.97
TMTB (2025)	6.80	12.19	5.32	14.86
DeCo (Ours)	**5.85**	10.51	**1.03**	**2.02**

Overall Framework. We first examine the progressive construction of DeCo, including multimodal fusion, density learning, density consensus, and consensus-guided feature interaction. As shown in Table 3, neither additional SAR observations nor an auxiliary density task alone consistently improves detection; the main gains arise when the learned density consensus explicitly guides cross-modal feature fusion.

Table 3: Incremental ablation of the proposed density-consensus-guided detection framework on the DenseShip6K dataset. All variants contain the OPT stream. The best results are highlighted in **bold**.

SAR	Ordinary Fusion	Density Task	Density Consensus	Consensus Guidance	Pr↑	Re↑	AP_{50}↑	$AP_{50:95}$↑	MAE↓	RMSE↓
✗	✗	✗	✗	✗	81.1	79.9	83.6	57.1	–	–
✓	✓	✗	✗	✗	81.6	78.7	83.6	56.9	–	–
✓	✓	✓	✗	✗	78.6	77.5	82.7	55.1	7.88	12.96
✓	✓	✓	✓	✗	80.1	76.8	82.1	55.6	6.41	11.43
✓	✗	✓	✓	✓	**83.0**	**80.6**	**85.0**	**58.5**	**5.85**	**10.51**

Density Supervision. We compare conventional Gaussian MSE supervision with standard Bayesian and the proposed box-aware Bayesian objectives. As shown in Table 4, replacing predefined pixel-wise density targets with probabilistic mass assignment improves both detection and counting, while incorporating box geometry further improves the learned density representation.

Table 4: Comparison of different density-supervision objectives under the same complete network architecture. Pixel GT denotes supervision using a preconstructed pixel-level density map, Iso. denotes a fixed isotropic scale, Aniso. denotes a box-aware anisotropic posterior, and NS denotes noise suppression. The best results are highlighted in **bold**.

Density supervision	Pixel GT	Iso.	Aniso.	NS	AP_{50}↑	$AP_{50:95}$↑	MAE↓	RMSE↓
Fixed Gaussian MSE ($\sigma = 6$)	✓	✓	✗	✗	83.9	55.0	6.96	11.34
Standard Bayesian	✗	✓	✗	✗	84.4	58.0	6.29	11.17
Noise-suppressed Bayesian	✗	✓	✗	✓	84.7	58.2	6.03	10.80
Box-aware anisotropic Bayesian	✗	✗	✓	✓	**85.0**	**58.5**	**5.85**	**10.51**

Consensus Guidance. We separately evaluate modality-reliability guidance and consensus-aware spatial modulation. As shown in Table 5, both components improve the baseline, and their com-

bination provides the strongest performance, confirming their complementary roles in cross-modal fusion. Fig. 3 shows a consistent qualitative trend.

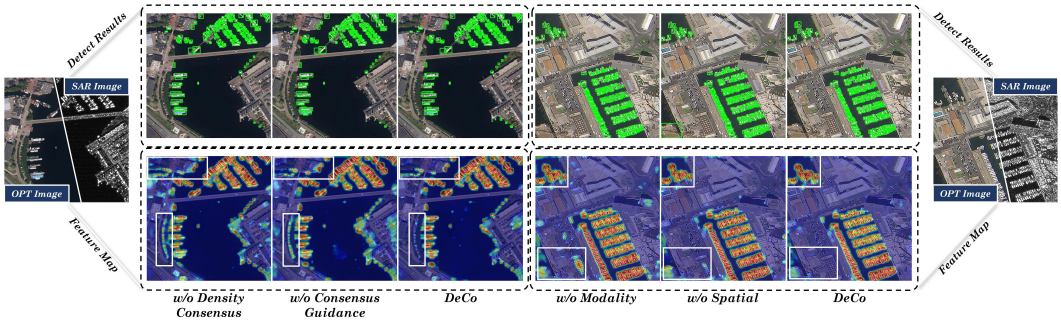

Figure 3: Feature visualization of the density consensus module. Detection results and corresponding feature activation maps are presented, with box regions highlighting the effect of density consensus on target enhancement and background suppression.

Additional Ablations. Additional experiments investigate the individual density-learning objectives, Gaussian-kernel bandwidth, regularization weights, density-aware count reweighting, and the detection-density task balance. The results show complementary benefits from structural and count constraints, sensitivity of fixed Gaussian supervision to kernel selection, and stable performance around the adopted loss configuration. Full results are provided in Appendix C.4.

5 Discussion

The effectiveness of DeCo does not arise from modality addition or density estimation alone, but from establishing cross-modal agreement in a shared target-distribution space and transferring this consensus to detection features, which is particularly beneficial in dense scenes.

Table 5: Ablation of multiplicative spatial guidance and additive modality-reliability guidance on DenseShip6K. The best results are highlighted in **bold**.

Spatial	Modality	$AP_{50}\uparrow$	$AP_{50:95}\uparrow$	MAE↓	RMSE↓
✗	✗	82.1	55.6	6.41	11.43
✗	✓	84.0	57.2	6.11	11.30
✓	✗	83.8	57.0	6.22	11.53
✓	✓	**85.0**	**58.5**	**5.85**	**10.51**

DenseShip6K uses synthesized SAR counterparts and therefore cannot fully capture the sensing characteristics of real SAR observations. Its primary role is to provide a controlled benchmark for analyzing multimodal detection and density estimation under highly dense scenes. Complementarily, MOS-Ship consists of real paired OPT-SAR observations and provides evidence for the effectiveness of DeCo under real heterogeneous sensing conditions.

Performance still degrades as target density increases, indicating that severe overlap and local ambiguity remain challenging. Future work will explore more fine-grained instance-aware consensus modeling and broader multimodal scenarios.

6 Conclusion

We presented DeCo, a multimodal detection framework based on density consensus learning. DeCo projects heterogeneous multimodal features into a shared target-distribution space, estimates spatially varying modality reliability, and uses the resulting density consensus to guide multiscale feature fusion. A box-aware Bayesian objective further enables instance-wise density-mass supervision without predefined pixel-wise Gaussian density targets. Experiments on DenseShip6K and MOS-Ship demonstrate improvements in detection and counting. Ablation studies showed that transferring density consensus to detection features is essential to realizing its benefits. These results support density consensus as an effective intermediate representation for coordinating heterogeneous multimodal observations.

AI USE STATEMENT

In this work, we used generative AI tools to assist with refining the experimental design and translating technical descriptions. Additionally, generative AI tools were used to edit auxiliary analysis and visualization code. All AI-assisted suggestions were manually reviewed by the authors, and the edited code was verified and tested before use. We did not use AI tools to prove mathematical claims; the remaining required-disclosure tasks involving formal proofs are not applicable to this work.

We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or personally identifiable information. DenseShip6K is constructed from remote sensing imagery obtained from publicly available research datasets, including DOTA and DIOR, together with satellite imagery collected from Google Earth for research purposes. All images are cropped into 512×512 patches and manually re-annotated with ship bounding boxes for the tasks studied in this work.

We respect the licenses and terms of use associated with the original data sources. All source data are used in accordance with their respective licenses and terms of use. The processed dataset is used solely for academic research in this work, with the aim of supporting research on remote sensing object detection and counting.

REPRODUCIBILITY STATEMENT

We make every effort to ensure the reproducibility of our results. The main paper and Appendix A provide the necessary methodological and implementation details, while Appendix B describes the dataset processing, data splits, training settings, hyperparameters, and evaluation protocols. Additional ablation and sensitivity analyses are reported in Appendix C.4. Our implementation is based on PyTorch. We will release the source code, configuration files, and trained model weights upon publication.

REFERENCES

Yaxiong Chen, Qijian Li, Qingwang Wang, and Guirong Feng. MSDCNet: A Multilevel Spatial Distribution-Enhanced Coupled Network for Ship Counting in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 64:4705414, 2026. doi: 10.1109/tgrs.2026.3708703.

Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:11–28, 2016. doi: 10.1016/j.isprsjprs.2016.03.014.

Guanchen Ding, Mingpeng Cui, Daiqin Yang, Tao Wang, Sihan Wang, and Yunfei Zhang. Object Counting for Remote-Sensing Images via Adaptive Density Map-Assisted Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. doi: 10.1109/tgrs.2022.3208326.

Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Guodong Guo, and Baochang Zhang. Fusion-Mamba for Cross-Modality Object Detection. *IEEE Transactions on Multimedia*, 27:7392–7406, 2025. doi: 10.1109/TMM.2025.3599020.

Chengzhen Duan, Zhiwei Wei, Chi Zhang, Siying Qu, and Hongpeng Wang. Coarse-grained Density Map Guided Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2789–2798, 2021. doi: 10.1109/iccvw54120.2021.00313.

Qingyun Fang, Dapeng Han, and Zhaokui Wang. Cross-Modality Fusion Transformer for Multispectral Object Detection. *arXiv preprint arXiv:2111.00273*, 2021.

Guangshuai Gao, Qingjie Liu, and Yunhong Wang. Counting From Sky: A Large-Scale Data Set for Remote Sensing Object Counting and a Benchmark Method. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):3642–3655, 2021. doi: 10.1109/tgrs.2020.3020555.

Guangshuai Gao, Qingjie Liu, Zhenghui Hu, Lu Li, Qi Wen, and Yunhong Wang. PSGCNet: A Pyramidal Scale and Global Context Guided Network for Dense Object Counting in Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. doi: 10.1109/tgrs.2022.3153946.

Junyu Gao, Qi Wang, and Yuan Yuan. SCAR: Spatial-/Channel-wise Attention Regression Networks for Crowd Counting. *Neurocomputing*, 363:1–8, 2019. doi: 10.1016/j.neucom.2019.08.018.

Junyu Gao, Liangliang Zhao, and Xuelong Li. NWPU-MOC: A Benchmark for Fine-Grained Multicategory Object Counting in Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/tgrs.2024.3356492.

Xiao He, Chang Tang, Xin Zou, and Wei Zhang. Multispectral Object Detection via Cross-Modal Conflict-Aware Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1465–1474, 2023. doi: 10.1145/3581783.3612651.

Ke Hu, Yudong He, Yuan Li, Jiayu Zhao, Song Chen, and Yi Kang. EI^{2}Det: Edge-Guided Illumination-Aware Interactive Learning for Visible-Infrared Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):7101–7115, 2025. doi: 10.1109/tcsvt.2025.3539625.

Meiyu Huang, Yao Xu, Lixin Qian, Weili Shi, Yaqin Zhang, Wei Bao, Nan Wang, Xuejiao Liu, and Xueshuang Xiang. The QXS-SAROPT Dataset for Deep Learning in SAR-Optical Data Fusion. *arXiv preprint arXiv:2103.08259*, 2021.

Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1037–1045, 2015. doi: 10.1109/cvpr.2015.7298706.

Glenn Jocher and Jing Qiu. Ultralytics YOLO11, 2024. URL `https://github.com/ultralytics/ultralytics`.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023. URL `https://github.com/ultralytics/ultralytics`.

Mengqi Lei, Siqi Li, Yihong Wu, Han Hu, You Zhou, Xinhu Zheng, Guiguang Ding, Shaoyi Du, Zongze Wu, and Yue Gao. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv preprint arXiv:2506.17733*, 2025.

Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. Density Map Guided Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 737–746, 2020a. doi: 10.1109/cvprw50498.2020.00103.

Jianwei Li, Changwen Qu, and Jiaqi Shao. Ship detection in SAR images based on an improved faster R-CNN. In *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*, pp. 1–6, 2017. doi: 10.1109/bigsardata.2017.8124934.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020b. doi: 10.1016/j.isprsjprs.2019.11.023.

Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-Aware Trident Networks for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6054–6063, 2019. doi: 10.1109/ICCV.2019.00615.

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.

Hui Lin, Zhiheng Ma, Xiaopeng Hong, Yaowei Wang, and Zhou Su. Semi-Supervised Crowd Counting via Density Agency. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1416–1426, 2022. doi: 10.1145/3503161.3547867.

Yanfeng Liu, Wei Guo, Chaojun Yao, and Lefei Zhang. Dual-Perspective Alignment Learning for Multimodal Remote Sensing Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. doi: 10.1109/TGRS.2025.3577046.

Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A High Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, pp. 324–331, 2017. doi: 10.5220/0006120603240331.

Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian Loss for Crowd Count Estimation With Point Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6141–6150, 2019. doi: 10.1109/iccv.2019.00624.

Yifei Qian, Xiaopeng Hong, Zhongliang Guo, Ognjen Arandjelović, and Carl R. Donovan. Semi-Supervised Crowd Counting With Contextual Modeling: Facilitating Holistic Understanding of Crowd Scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):8230–8241, 2024. doi: 10.1109/TCSVT.2024.3392500.

Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024. doi: 10.1016/j.patcog.2023.109913.

Yutian Shi, Guoquan Li, Zhilong Shen, Hongying Meng, and Yu Pang. BANet: Enhancing Weakly Aligned Multimodal Object Detection via Balanced Bidirectional Alignment Network. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–14, 2026. doi: 10.1109/tgrs.2026.3674946.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse R-CNN: End-to-End Object Detection With Learnable Proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14454–14463, 2021. doi: 10.1109/CVPR46437.2021.01422.

Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, Martin Weinmann, Stefan Hinz, Cheng Wang, and Kun Fu. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022a. doi: 10.1016/j.isprsjprs.2021.12.004.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection Via Uncertainty-Aware Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022b. doi: 10.1109/tcsvt.2022.3168279.

Chao Tian, Zikun Zhou, Chao Yang, Guoqing Zhu, and Zhenyu He. Efficient RGB-T Object Detection via Sparse Cross-Modality Fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 541–559, 2026. doi: 10.1007/978-3-032-37281-9_33.

Jia Wan and Antoni Chan. Adaptive Density Map Generation for Crowd Counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1139, 2019. doi: 10.1109/iccv.2019.00122.

Jia Wan, Ziquan Liu, and Antoni B. Chan. A Generalized Loss Function for Crowd Counting and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1974–1983, 2021. doi: 10.1109/cvpr46437.2021.00201.

Chao Wang, Wei Lu, Xiang Li, Jian Yang, and Lei Luo. M4-SAR: A Multi-Resolution, Multi-Polarization, Multi-Scene, Multi-Source Dataset and Benchmark for Optical-SAR Object Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2505.10931.

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. doi: 10.1109/CVPR.2019.00839.

Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access*, 8: 120234–120254, 2020. doi: 10.1109/access.2020.3005861.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3974–3983, 2018. doi: 10.1109/cvpr.2018.00418.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24440–24451, 2025.

Jianing You, Yixuan Lv, Shengyang Li, Silei Liu, Kailun Zhang, and Yuxuan Liu. A Multi-Modal Approach for Robust Oriented Ship Detection: Dataset and Methodology. *Remote Sensing*, 18 (2):274, 2026. doi: 10.3390/rs18020274.

Maoxun Yuan and Xingxing Wei. C^2Former: Calibrated and Complementary Transformer for RGB-Infrared Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. doi: 10.1109/tgrs.2024.3376819.

Maoxun Yuan, Yinyan Wang, and Xingxing Wei. Translation, Scale and Rotation: Cross-Modal Alignment Meets RGB-Infrared Vehicle Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 509–525, 2022. doi: 10.1007/978-3-031-20077-9_30.

Chenke Yue, Yin Zhang, Junhua Yan, Zhaolong Luo, Yong Liu, and Pengyu Guo. Diffusion Mechanism and Knowledge Distillation Object Detection in Multimodal Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025. doi: 10.1109/tgrs.2025.3561133.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection, 2022.

Jiaqing Zhang, Jie Lei, Weiying Xie, Zhenman Fang, Yunsong Li, and Qian Du. SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. doi: 10.1109/tgrs.2023.3258666.

Wensi Zhang, Xiaoling Zhang, Xiaowo Xu, Shunjun Wei, Jun Shi, Yunqi Wang, and Tianjiao Zeng. MHFNet: Multimodal hybrid fusion framework for misaligned SAR-Optical ship detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 231:151–166, 2026. doi: 10.1016/j.isprsjprs.2025.10.027.

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. doi: 10.1109/CVPR.2016.70.

Zhicheng Zhao, Xuanang Fan, Lingma Sun, Chenglong Li, and Jin Tang. Learning Where to Focus: Density-Driven Guidance for Detecting Dense Tiny Objects. *IEEE Transactions on Geoscience and Remote Sensing*, 64:5629615, 2026. doi: 10.1109/tgrs.2026.3708550.

Pengfei Zhu, Jingqing Li, Bing Cao, and Qinghua Hu. Multi-Task Credible Pseudo-Label Learning for Semi-Supervised Crowd Counting. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10394–10406, 2024. doi: 10.1109/TNNLS.2023.3241211.

A Additional Method Details

A.1 Detailed Density Encoding Architecture

For each modality $m \in \{o, s\}$, the backbone produces multiscale features $\{\mathbf{F}_m^l\}_{l=3}^5$. As described in the main text, each level is first projected to a common channel dimension and spatially aligned to the resolution of the finest feature level. The aligned features are then concatenated and compressed as

$$\mathbf{X}_m = \phi_m^a \left(\mathbf{A}_m^3 \| \mathbf{A}_m^4 \| \mathbf{A}_m^5\right), \tag{20}$$

where $\|$ denotes channel-wise concatenation and ϕ_m^a denotes the aggregation projection.

To accommodate variations in ship scale and spatial concentration, $\mathbf{X}_m$ is further processed by parallel context operators with different receptive fields (as shown in Fig. 4):

$$\mathbf{E}_m = \mathbf{X}_m + \phi_m^c \left(\Big\| _{d\in\{1,2,3\}} \psi_m^d \left(\mathbf{X}_m\right) \right), \tag{21}$$

where ψ_m^d denotes a context operator with dilation rate d, and ϕ_m^c integrates the outputs of the parallel branches. The residual connection preserves the original multiscale evidence while enriching the representation with context at different spatial extents. The OPT and SAR density encoders use independent parameters, allowing them to model density-related evidence according to their respective sensing statistics before entering the shared latent space.

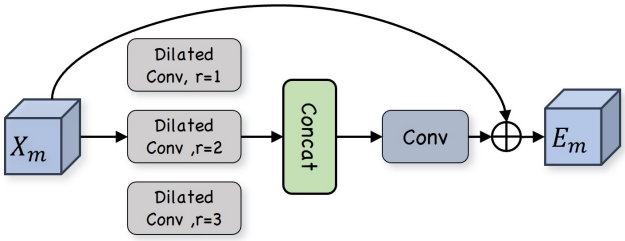

Figure 4: Multi-dilation context aggregation. The aggregated feature $\mathbf{X}_m$ is processed by three parallel context branches with dilation rates 1, 2, and 3. Their outputs are concatenated, fused, and residually added to $\mathbf{X}_m$ to obtain the density-aware representation $\mathbf{E}_m$.

A.2 Consensus Gating Module

Although $\mathbf{Z}_o$ and $\mathbf{Z}_s$ are mapped into the same latent density space, their local reliability can still differ substantially across the scene. A uniform average would therefore assume equal contributions from both modalities at all spatial locations, which is undesirable under spatially varying optical degradation and SAR interference.

We infer modality-wise logits from the concatenated shared-space representations (as shown in Fig. 5):

$$[\mathbf{L}_o, \mathbf{L}_s] = \mathcal{G}\left(\mathbf{Z}_o \| \mathbf{Z}_s\right), \tag{22}$$

followed by a modality-wise softmax:

$$[\mathbf{W}_o, \mathbf{W}_s] = \mathrm{Softmax}_m\left([\mathbf{L}_o, \mathbf{L}_s]\right). \tag{23}$$

Thus,

$$\mathbf{W}_o(x, y) + \mathbf{W}_s(x, y) = 1 \tag{24}$$

at each spatial location. Each scalar weight is shared across the latent channels at the corresponding position and represents the relative contribution of one modality to the local target-distribution interpretation.

The resulting competitive weighting allows the consensus to favor OPT evidence where texture and structural cues are reliable, while assigning greater weight to SAR where the optical observation is degraded. Conversely, SAR responses affected by speckle or coastal scattering can be suppressed when they are less consistent with the jointly inferred target distribution.

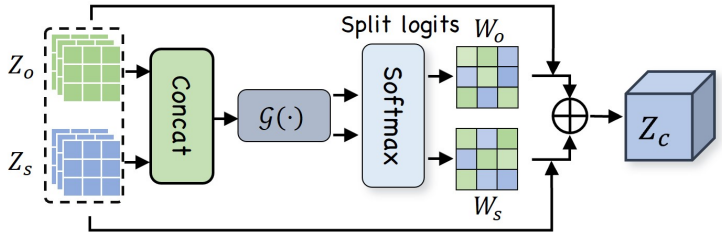

Figure 5: Consensus gating module. The shared-space representations $\mathbf{Z}_o$ and $\mathbf{Z}_s$ are concatenated and mapped to modality-wise logits. A modality-wise softmax produces competitive reliability maps $\mathbf{W}_o$ and $\mathbf{W}_s$, which are used to construct the density consensus representation $\mathbf{Z}_c$.

A.3 Additional Details of Consensus-Guided Fusion

The consensus-guided fusion module combines two complementary forms of target-level guidance. The modality weights determine *which modality* should contribute more strongly at each spatial location, while the consensus representation indicates *where* the fused feature should be enhanced or suppressed.

The reliability maps are initially estimated at the resolution of the latent density space and are resized to each detection level by bilinear interpolation:

$$\mathbf{W}_m^l = \mathcal{R}_l\left(\mathbf{W}_m\right), \qquad m \in \{o, s\}. \tag{25}$$

Since the original weights are normalized along the modality dimension, their competitive relationship is preserved after spatial resampling. The baseline concatenation pathway is retained as a stable information basis, while density-induced effects are introduced as bounded residual corrections. In particular, the spatial modulation map satisfies $\mathbf{S}^l \in [-1, 1]$, and the learnable coefficients are parameterized as

$$\alpha_l = \tanh(a_l), \qquad \beta_l = \tanh(b_l), \tag{26}$$

which bounds both coefficients to $[-1, 1]$. Therefore, the multiplicative factor $1 + \alpha_l \mathbf{S}^l$ remains bounded within $[0, 2]$, preventing excessively large consensus-induced modulation.

The scale coefficients are initialized with small magnitudes so that the fusion module begins close to conventional concatenation-based fusion. During training, the influence of density consensus is gradually increased only when supported by the joint detection and density objectives. By attenuating sensitivity to inaccurate density estimates during early optimization, the residual formulation permits the learned consensus to regulate modality allocation and spatial feature refinement in later stages.

A.4 Density Decoder and Nonnegative Prediction

The latent consensus representation $\mathbf{Z}_c$ is decoded into a single full-resolution density map through a progressive reconstruction pathway. Specifically, spatial upsampling is alternated with learnable feature reconstruction to gradually recover fine-grained spatial details from the low-resolution consensus representation. A final projection produces an unbounded density response:

$$\widetilde{\mathbf{D}} = \phi_{\text{out}}\left(\mathbf{H}^3\right), \tag{27}$$

where $\widetilde{\mathbf{D}} \in \mathbb{R}^{1\times H\times W}$. Since negative density values have no physical interpretation, we apply a smooth nonnegative mapping:

$$\widehat{\mathbf{D}} = \rho\left(\widetilde{\mathbf{D}}\right) = \frac{1}{2}\left(\widetilde{\mathbf{D}} + \sqrt{\widetilde{\mathbf{D}}^{\,2} + \epsilon}\right), \tag{28}$$

where ϵ is a small constant for numerical stability. The resulting $\widehat{\mathbf{D}}$ is the only explicit density prediction produced by DeCo. Its spatial integral defines the estimated ship count:

$$\widehat{N} = \frac{1}{\kappa}\sum_{\mathbf{p}\in\Omega}\widehat{\mathbf{D}}(\mathbf{p}), \tag{29}$$

where Ω denotes the image domain and κ is the density mass corresponding to one object. $\widehat{\mathbf{D}}$ is supervised through probabilistic assignment of predicted mass to annotated instances and background.

A.5 Background Ownership in Bayesian Mass Assignment

The main text defines the box-conditioned foreground likelihood $\ell_i(\mathbf{p})$ for each annotated ship. To distinguish foreground mass from spurious background responses, we introduce an explicit background component.

Let

$$d^2_{\min}(\mathbf{p}) = \min_i \|\mathbf{p} - \mathbf{c}_i\|_2^2 \tag{30}$$

denote the squared distance from pixel $\mathbf{p}$ to the nearest annotated ship center. We define a reference spatial scale as

$$\sigma_{\text{ref}} = \text{Median}_i \sqrt{\sigma_{x,i}\sigma_{y,i}}. \tag{31}$$

The background log-likelihood is then formulated as

$$\ell_0(\mathbf{p}) = -\frac{1}{2\sigma^2_{\text{ref}}} \frac{(\gamma \min(H, W))^2}{d^2_{\min}(\mathbf{p}) + \epsilon}, \tag{32}$$

where γ controls the transition between foreground and background regions. The proposed method assigns low background plausibility near annotated ship centers, with this plausibility progressively increasing as the pixel moves away from all foreground instances.

Along with the foreground likelihoods, the posterior ownership is computed as

$$q_i(\mathbf{p}) = \frac{\exp(\ell_i(\mathbf{p}))}{\sum_{j=0}^{N} \exp(\ell_j(\mathbf{p}))}, \qquad i \in \{0, 1, \ldots, N\}. \tag{33}$$

After normalizing the predicted density into object mass,

$$d(\mathbf{p}) = \frac{\widehat{\mathbf{D}}(\mathbf{p})}{\kappa}, \tag{34}$$

the expected mass assigned to component i is

$$\widehat{m}_i = \sum_{\mathbf{p}\in\Omega} q_i(\mathbf{p}) d(\mathbf{p}). \tag{35}$$

Each foreground instance is expected to receive one unit of mass, whereas the background is expected to receive zero:

$$m^*_i = 1, \quad i = 1, \ldots, N, \qquad m^*_0 = 0. \tag{36}$$

The foreground and background likelihoods therefore determine only the ownership of predicted density mass and do not prescribe the target value of any individual pixel.

A.6 Reliability-Weighted Bayesian Objective

Given the target mass m^*_i and the predicted mass $\widehat{m}_i$, the residual of component i is

$$e_i = |m^*_i - \widehat{m}_i|. \tag{37}$$

To reduce the influence of severely ambiguous or noisy foreground instances, we apply a detached reliability factor to the foreground residuals:

$$\mathcal{L}_{\text{Bayes}} = \frac{1}{B} \sum_{b=1}^{B} \left[\sum_{i=1}^{N_b} e_i^{(b)} \, \text{sg}\left(\exp\left(-e_i^{(b)}\right)\right) + e_0^{(b)} \right], \tag{38}$$

where B denotes the batch size and $\text{sg}(\cdot)$ is the stop-gradient operation.

The reliability factor decreases the contribution of foreground instances with large mass-assignment errors, preventing highly ambiguous samples from dominating the optimization. Because the weighting term is detached, it does not introduce an additional optimization pathway through the reliability estimate itself.

For images without annotated ships, the entire predicted density is treated as background mass. More generally, this formulation separates mass conservation from density-shape construction. The network is therefore free to learn scene-conditioned density responses, unconstrained by the need to reproduce a predefined Gaussian profile.

A.7 Exact Density-Aware Count Weight

The global count constraint in the main text applies a density-aware weight $\omega(N)$ to increase the contribution of densely populated training samples. The weighting function is defined as

$$\omega(N) = \begin{cases} 1.0, & N < 50, \\ 1.2, & 50 \le N < 100, \\ 1.5, & 100 \le N < 150, \\ 2.0, & 150 \le N < 200, \\ 2.5, & N \ge 200. \end{cases} \tag{39}$$

Accordingly, the global count loss is

$$\mathcal{L}_{\mathrm{cnt}} = \frac{1}{B} \sum_{b=1}^{B} \omega(N_b)\, \mathrm{SmoothL1}\left(\widehat{N}_b, N_b\right). \tag{40}$$

The weighting is motivated by the larger accumulation of local mass-assignment errors in dense scenes. While $\mathcal{L}_{\mathrm{Bayes}}$ constrains mass allocation at the instance level, $\mathcal{L}_{\mathrm{cnt}}$ calibrates the aggregate density mass at the image level. Upweighting densely populated samples therefore emphasizes scenes where small local deviations readily accumulate into substantial global count errors.

B Additional Experimental Details

B.1 Implementation Details

DeCo is implemented in PyTorch and trained on an NVIDIA Tesla V100 GPU for 100 epochs. The two modality-specific branches use GELAN as the backbone. The model is optimized using SGD with an initial learning rate of 0.01, a weight decay of 0.0005, and a batch size of 8.

The loss coefficients are set to $\lambda_{\mathrm{det}} = 0.025$, $\lambda_{\mathrm{den}} = 0.30$, $\lambda_{\mathrm{str}} = 0.01$, and $\lambda_{\mathrm{cnt}} = 0.001$. For the box-aware Bayesian supervision, we set $\eta = 0.25$, $\sigma_{\min} = 2.0$, $\gamma = 0.15$, and $\kappa = 255$. All hyperparameters were determined during model development based solely on training behavior. Unless otherwise specified, these settings are used for all experiments. Official implementations are adopted for the compared methods whenever available, with their recommended optimization settings retained.

B.2 Dataset Details

We conduct experiments on two complementary OPT-SAR ship datasets with substantially different target-density characteristics. MOS-Ship consists of real paired optical and SAR observations and is used to evaluate DeCo under real heterogeneous sensing conditions. DenseShip6K is constructed primarily for analysis of densely populated ship scenes. The basic statistics of the two datasets are summarized in Table 6, and additional construction and annotation details are provided below.

Table 6: Statistics of the datasets used in our experiments.

Dataset	Resolution	Train / Test	Instances	Avg.	Max
MOS-Ship	1024×1024	996 / 250	8,992	7.22	141
DenseShip6K	512×512	4,500 / 1,500	279,897	46.65	648

B.2.1 MOS-Ship Details

MOS-Ship is a spatially registered OPT-SAR ship dataset containing paired 1024×1024 images acquired over diverse maritime environments. The original dataset provides oriented bounding-box annotations. For the experiments in this work, we use a refined annotation version in which missing ship instances are recovered and the annotations are converted into horizontal bounding boxes.

The resulting dataset contains 996 paired scenes for training and 250 paired scenes for testing, with 7,076 and 1,916 annotated ships, respectively, yielding 8,992 instances in total. The test set

Table 7: Comparison of representative object counting datasets. The statistics include dataset categories, acquisition platforms, modalities, image numbers, annotation formats, and object-count distributions.

Dataset	Categories	Platform	Modal	Images	Resolution	Annotation Format	Count Statistics		
							Total	Average	Max
Olive Trees	Tree	UAV	RGB	10	4000×3000	circle	1,251	125.1	143
COWC	Car	Aerial	RGB	–	low	center point	32,716	–	–
CARPK	Car	Drone	RGB	1,448	1280×720	HBB	89,777	62	188
DLR-ACD	Crowd	Aerial	RGB	33	3619×5226	center point	226,291	6,857	24,368
DroneCrowd	Crowd	Drone	RGB	33,600	1920×1080	center point	4,864,280	144.8	455
RSOC	Building	Satellite	RGB	2,468	512×512	OBB	76,215	30.88	142
	Small-vehicle			280	2473×2339		148,838	531.56	8,531
	Large-vehicle			172	1552×1573		16,594	96.48	1,336
	Ship			137	2558×2668		44,892	327.68	1,661
KR-GRUIDAE	5 object categories	Camera	RGB	1,423	1554×2326	center point	30,779	21.63	299
MOCSE13	13 object categories	Unity	RGB	120	512×512	HBB	–	–	–
NWPU-MOC	13 object categories	Aerial	RGB&NIR	3,416	1024×1024	center point	383,195	112.18	3,582
DenseShip6K	Ship	Mixed	OPT&SAR	6,000	512×512	HBB	279,897	46.65	648

additionally provides cloud-degraded optical observations paired with the same SAR images and annotations. These degraded optical images correspond to the same 250 test scenes and therefore do not constitute additional independent samples or ship instances. They are used specifically to evaluate multimodal robustness when the optical modality becomes less reliable.

B.2.2 DENSESHIP6K DETAILS

As summarized in Table 7, existing remote sensing counting datasets are predominantly based on a single optical modality and are not designed for joint multimodal detection and density estimation. Existing paired OPT-SAR ship datasets also contain relatively few high-density scenes. Therefore, DenseShip6K is constructed as a controlled benchmark for studying multimodal ship detection and density estimation under densely populated scenes.

DenseShip6K contains 6,000 image pairs and 279,897 annotated ships in total, with an average of 46.65 ships per image and a maximum of 648 ships in a single image. We use 4,500 pairs for training and 1,500 pairs for testing. The scenes cover substantial variations in target number, scale, and spatial organization, including isolated ships, locally clustered formations, and highly congested arrangements.

The source optical imagery was collected from multiple sources. A subset was derived from publicly available remote sensing datasets, including DOTA and DIOR, while the majority was collected from Google Earth for research purposes. The corresponding SAR images were generated from the optical images using a separately trained OPT-to-SAR generative adversarial network, yielding spatially aligned OPT-SAR pairs. The OPT-to-SAR translation pipeline and the quality of the synthesized SAR imagery have been systematically evaluated in a prior peer-reviewed study. All image pairs were cropped into 512×512 patches and manually re-annotated with horizontal bounding boxes under a unified annotation protocol. To avoid spatial leakage between the training and test sets, the dataset split is performed at the source-scene level before patch extraction. Source images and geographic scenes assigned to the training and test sets are strictly disjoint, and the corresponding patches are cropped only after the split is determined. Therefore, no overlapping or near-duplicate patches originating from the same source scene are shared across the two subsets.

The primary purpose of DenseShip6K is to provide a broad and controllable range of target densities for analyzing density-aware multimodal detection. It is not intended to replace evaluation on real paired OPT-SAR observations. Accordingly, we also evaluate DeCo on MOS-Ship, whose optical and SAR images are obtained from real heterogeneous sensors. MOS-Ship evaluates generalization under real multimodal sensing conditions. Compared with MOS-Ship, DenseShip6K exhibits a substantially broader density range and more frequent close-proximity instances, making it particularly suitable for evaluating localization accuracy, density estimation, and cumulative counting errors in crowded scenes. Unless otherwise specified, component ablations and sensitivity analyses are conducted on DenseShip6K.

B.3 Evaluation Protocol

For object detection, we report average precision at an intersection-over-union threshold of 0.5 (AP_{50}) and average precision averaged over IoU thresholds from 0.5 to 0.95 with an interval of 0.05 ($AP_{50:95}$). Counting accuracy is evaluated using mean absolute error (MAE) and root mean squared error (RMSE):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{C}_i - C_i\right|, \tag{41}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{C}_i - C_i\right)^2}, \tag{42}$$

where N is the number of evaluation images, C_i is the ground-truth object count of the i-th image, and $\widehat{C}_i$ is the corresponding predicted count. MAE measures the average counting deviation, whereas RMSE places greater emphasis on large errors and is therefore more sensitive to failures in densely populated scenes.

The source of $\widehat{C}_i$ depends on the output paradigm of the evaluated method. For unimodal and multimodal object detectors, the predicted count is the number of retained bounding boxes after standard post-processing:

$$\widehat{C}_i^{\mathrm{det}} = \left|\widehat{\mathcal{B}}_i\right|, \tag{43}$$

where $\widehat{\mathcal{B}}_i$ denotes the final detection set.

For DeCo's density output, the estimated count is obtained by integrating the predicted consensus density:

$$\widehat{C}_i^{\mathrm{den}} = \frac{1}{\kappa}\sum_{\mathbf{p}\in\Omega}\widehat{\mathbf{D}}_i(\mathbf{p}). \tag{44}$$

The same density-integration protocol is used for conventional density-estimation methods. Therefore, the MAE and RMSE reported for detection methods quantify counting errors from detected bounding boxes, whereas those reported for density-estimation methods are computed from integrated density predictions.

C Additional Experimental Results

C.1 Qualitative Results

C.1.1 DenseShip6K Detection Visualization

Fig. 6 presents qualitative comparisons on representative dense harbor scenes from DenseShip6K. These scenes contain closely berthed small vessels, substantial variations in ship scale, and surrounding docks and buildings that complicate instance discrimination. The compared detectors exhibit different failure patterns, including missed ships within crowded groups, redundant detections, and responses to nearby harbor structures. Similar errors remain visible in several multimodal baselines, indicating that the availability of complementary OPT and SAR observations alone does not guarantee effective suppression of cross-modal interference.

In comparison, DeCo provides more complete coverage of densely arranged vessels while producing fewer conspicuous responses around surrounding infrastructure. The predicted counts are also more consistent with the ground truth. In the first illustrated scene containing 124 ships, DeCo predicts 124, whereas the compared methods produce estimates ranging from 101 to 160. These observations are consistent with the quantitative results and indicate that density-consensus-guided fusion improves both instance localization and count consistency in densely populated scenes.

C.1.2 MOS-Ship Detection Visualization

Fig. 7 further compares detection results on MOS-Ship. Compared with the highly crowded scenes in DenseShip6K, these examples emphasize interference from piers, shoreline structures, and large

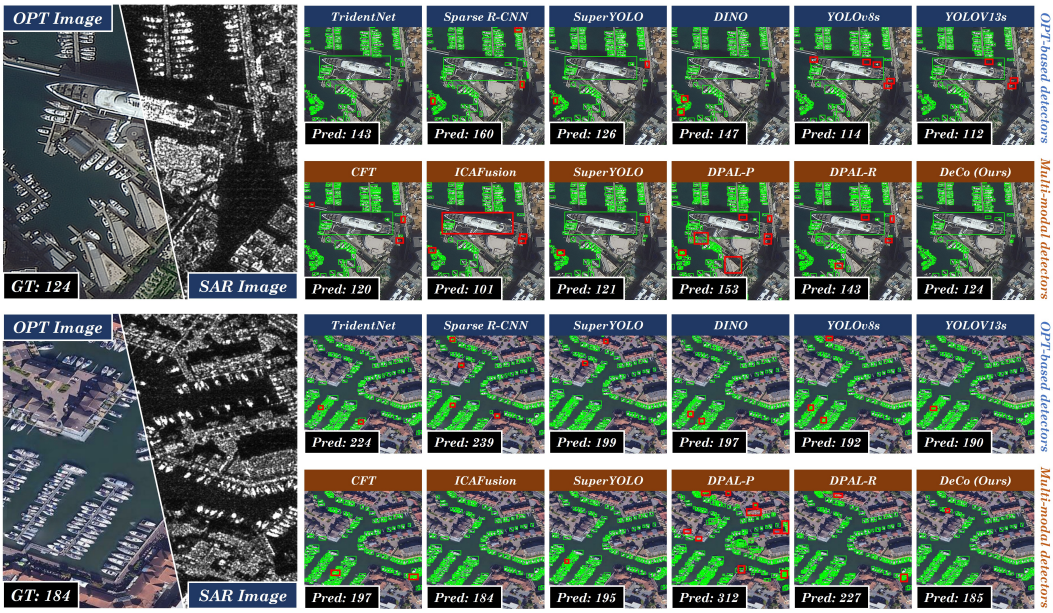

Figure 6: Visualization of ship detection results on the **DenseShip6K** dataset under dense scenes. The upper rows present unimodal detectors, and the lower rows compare multimodal detectors. Predicted ship counts are reported for each method. Green boxes denote predicted bounding boxes, while red boxes highlight missed detections and false positives.

variations in vessel size. Optical observations may contain ships with appearances similar to surrounding harbor facilities, while SAR images exhibit strong scattering responses from both ships and infrastructure.

Several unimodal and multimodal methods either miss vessels near piers or produce spurious detections in surrounding regions. DeCo retains more small and large vessels around the docks while reducing conspicuous background responses. In the two illustrated scenes, whose ground-truth counts are 24 and 27, DeCo predicts 21 and 25 ships, respectively. In comparison, YOLOv13s predicts only 9 and 12 ships, whereas DPAL-P overestimates the counts as 29 and 49. These examples further show that density-consensus-guided fusion remains effective in scenes with lower target density but substantial background interference.

C.1.3 Density Visualization

Fig. 8 compares density predictions on representative dense scenes from DenseShip6K. Most methods recover the overall spatial arrangement of ship groups, but differ substantially in response concentration and background suppression. DACount tends to produce relatively diffuse responses extending beyond densely occupied regions, while MTCP exhibits less regular responses within crowded ship groups. In contrast, the native density prediction of DeCo remains concentrated around ship locations while maintaining low responses over large portions of the surrounding background.

Importantly, the density response learned by DeCo is not constrained to reproduce a predefined Gaussian profile. The enlarged regions illustrate localized, instance-dependent responses learned through Bayesian mass supervision. For reference, **DeCo-G** denotes the DeCo response rendered using a predefined Gaussian kernel for visualization only and is not the supervision target used during training.

The predicted counts provide additional evidence of the quality of the learned density representation. In the first example containing 361 ships, DeCo estimates 352.55, compared with 293.25 for MCNN, 350.55 for CSRNet, 351.90 for SFCN, 318.71 for DACount, and 307.96 for MTCP. In the second example, DeCo predicts 219.82 ships for a ground-truth count of 219. These results show that the learned density representation preserves the spatial organization of crowded ship groups while providing accurate scene-level counts without requiring a manually predefined density shape.

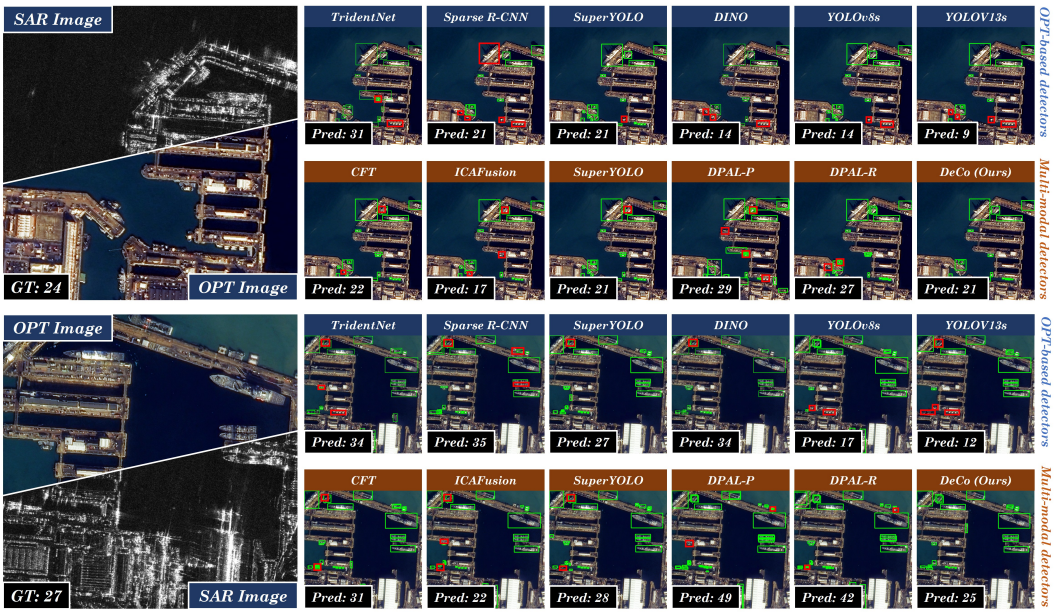

Figure 7: Visualization of ship detection results on the **MOS-Ship** dataset. The upper rows present unimodal detectors, and the lower rows compare multimodal detectors. Predicted ship counts are reported for each method. Green boxes denote predicted bounding boxes, while red boxes highlight missed detections and false positives.

C.2 Optical Degradation Robustness and Modality Reliability

To examine whether the learned modality reliability adapts to changes in observation quality, we evaluate DeCo on the cloud-degraded optical variants provided by MOS-Ship. These degraded images correspond to the same 250 test scenes as the original optical observations and share identical SAR images and annotations.

We evaluate both detection performance and the learned modality reliability under clean and degraded optical conditions. In addition to the standard detection metrics, we analyze the OPT and SAR reliability weights $\mathbf{W}_o$ and $\mathbf{W}_s$ to determine whether DeCo shifts its reliance toward SAR when the optical observation becomes less reliable.

Table 8 compares DeCo with representative multimodal detectors under clean and cloud-degraded optical observations. Because the two conditions share the same SAR images and annotations, the performance difference primarily reflects the sensitivity of each method to optical degradation. Under clean optical conditions, DeCo achieves the best detection performance, with 81.9% AP_{50} and 54.7% $AP_{50:95}$. When the optical observations are degraded by clouds, all compared methods experience performance drops. However, DeCo retains 60.1% AP_{50} and 38.8% $AP_{50:95}$, whereas the competing multimodal detectors achieve only 12.7-18.0% AP_{50} under the same degraded condition. Moreover, the AP_{50} drop of DeCo is 21.8 percentage points, substantially smaller than the 51.3-61.2 point drops of the compared methods. These results indicate that DeCo is considerably more robust to degradation of the optical modality.

We further investigate whether this robustness is associated with the intended adaptive modality-reliability mechanism. Since $\mathbf{W}_o$ and $\mathbf{W}_s$ determine the relative contributions of the optical and SAR modalities at each spatial location, their responses should vary according to the quality of the corresponding observations. In particular, when the optical modality is degraded, a reliable gating mechanism is expected to reduce the optical contribution and assign greater importance to the unaffected SAR observation.

$$\bar{W}_m^{\mathrm{obj}} = \frac{1}{|\Omega_{\mathrm{obj}}|} \sum_{\mathbf{p} \in \Omega_{\mathrm{obj}}} W_m(\mathbf{p}), \qquad m \in \{o, s\}, \tag{45}$$

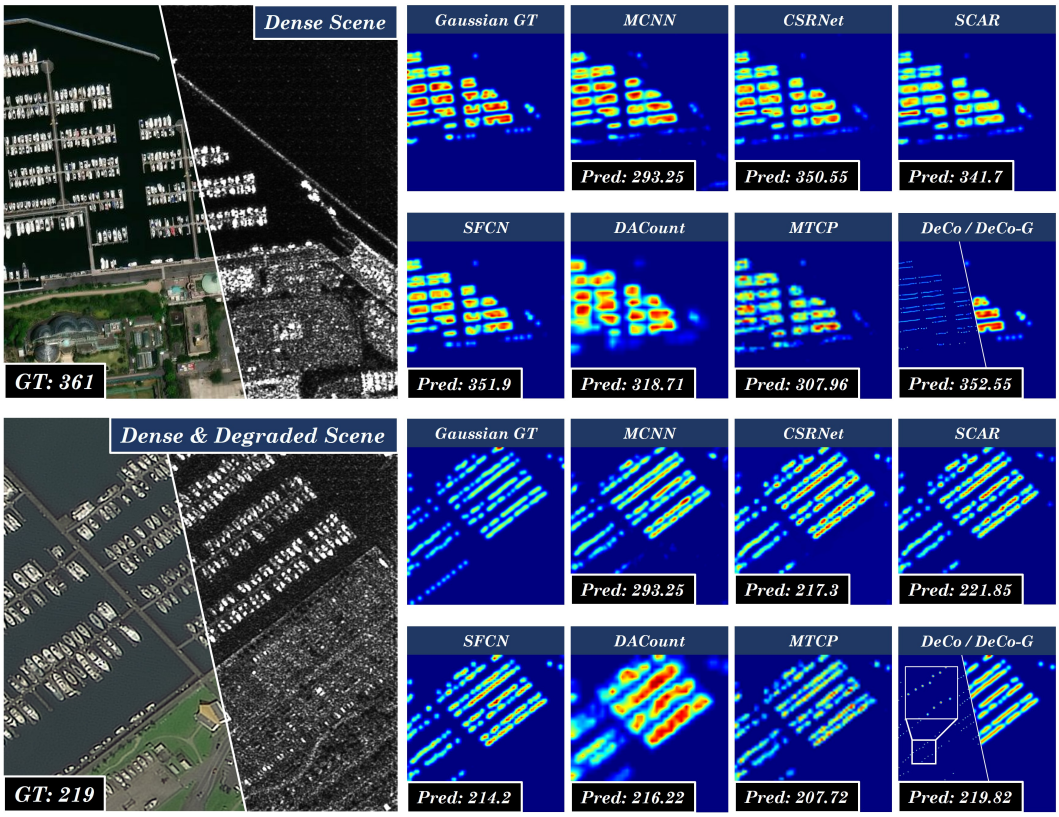

Figure 8: Visualization of density estimation results on the **DenseShip6K** dataset under dense and degraded scenes. Predicted density maps and ship counts from different methods are presented. DeCo is trained without predefined Gaussian density targets, while DeCo-G shows the corresponding response rendered with a Gaussian kernel for visualization only.

Table 8: Detection robustness under clean and cloud-degraded optical observations on the **MOSShip** dataset. ΔAP_{50} denotes the performance drop from clean to degraded optical conditions. The best results are highlighted in **bold**.

Method	Clean OPT		Degraded OPT		$\Delta AP_{50}\downarrow$
	$\mathbf{AP}_{50}\uparrow$	$\mathbf{AP}_{50:95}\uparrow$	$\mathbf{AP}_{50}\uparrow$	$\mathbf{AP}_{50:95}\uparrow$	
CFT	75.8	44.1	16.9	8.39	58.9
ICAFusion	69.9	41.7	12.7	6.66	57.2
EI2Det	69.3	44.0	18.0	10.2	51.3
Fusion-Mamba	79.0	48.9	17.8	9.93	61.2
DeCo	**81.9**	**54.7**	**60.1**	**38.8**	**21.8**

where Ω_{obj} denotes the union of ground-truth ship regions. Table 9 reports the average reliability weights under the two optical conditions. Under clean observations, $\bar{W}_o^{\text{obj}}$ and $\bar{W}_s^{\text{obj}}$ are 0.7580 and 0.2420, respectively. When cloud degradation is introduced, $\bar{W}_o^{\text{obj}}$ decreases to 0.5832, while $\bar{W}_s^{\text{obj}}$ increases to 0.4168. This 0.1748 shift in reliability toward SAR indicates that the gating mechanism adaptively adjusts the relative contributions of the two modalities according to observation quality rather than assigning fixed modality weights. The increased reliance on SAR is consistent with the substantially smaller detection-performance drop of DeCo in Table 8, providing direct evidence for its adaptive multimodal fusion mechanism.

Table 9: Variation of the learned modality reliability under clean and cloud-degraded optical conditions on **MOS-Ship** dataset. The reliability weights are averaged over ground-truth ship regions.

Condition	$\bar{W}_o^{\mathrm{obj}}$	$\bar{W}_s^{\mathrm{obj}}$
Clean optical	0.7580	0.2420
Cloud-degraded optical	0.5832	0.4168

C.3 Counting Evaluation under Different Output Paradigms

DeCo produces two outputs that can be used to estimate the number of ships: the retained detection boxes and the explicitly predicted density map. In the main experiments, the reported counting results of DeCo are obtained from its density prediction, since density estimation is an integral component of the proposed framework rather than an auxiliary post-processing procedure. For detection-only baselines that do not provide an explicit density output, the predicted count is naturally obtained from the number of retained detection boxes.

To further examine the effect of the count-extraction paradigm and provide a direct comparison under detection-based counting, we additionally compute the MAE and RMSE of DeCo using its detection boxes. Table 10 compares representative high-performing detection methods using box-based counts, together with representative density-estimation methods using integrated density predictions.

When counting directly from detection boxes, DeCo achieves an MAE/RMSE of 5.88/11.43 on DenseShip6K and 1.36/3.48 on MOS-Ship. Using its native density prediction further reduces these errors to 5.85/10.51 and 1.03/2.02, respectively. This comparison shows that the density output provides additional scene-level counting information beyond the discrete detection results. We therefore use the density-based count as the primary counting output of DeCo in the main experiments, while reporting the box-based results here for completeness and direct comparability with detection-only methods.

Table 10: Counting comparison under different output paradigms on DenseShip6K and MOS-Ship. Detection-based methods estimate object counts from retained bounding boxes, whereas density-based methods obtain counts by integrating the predicted density maps. The best results are highlighted in **bold**.

Method	**Count Source**	**DenseShip6K**		**MOS-Ship**	
		MAE↓	**RMSE↓**	**MAE↓**	**RMSE↓**
DINO	Detection boxes	9.03	17.42	1.66	7.83
YOLOv11s	Detection boxes	6.58	12.99	2.37	9.21
YOLOv13s	Detection boxes	6.57	12.75	2.34	7.63
ICAFusion	Detection boxes	8.55	19.10	**1.22**	**3.41**
Fusion-Mamba	Detection boxes	7.75	15.12	1.52	5.50
DeCo (Ours)	Detection boxes	**5.88**	**11.43**	1.36	3.48
CSRNet	Density map	6.38	10.49	4.46	10.60
SCAR	Density map	6.09	**10.33**	1.81	2.77
DACount	Density map	6.16	10.97	1.07	2.19
MTCP	Density map	9.42	16.00	2.77	9.94
DeCo (Ours)	Density map	**5.85**	10.51	**1.03**	**2.02**

C.4 Additional Ablation Studies

C.4.1 Ablation of Density-Learning Objectives

We further analyze the individual objectives used to optimize the density consensus representation. As shown in Table 11, the box-aware anisotropic Bayesian objective alone establishes instance-

level mass correspondence, but does not explicitly constrain either the global density integral or the structural agreement between modality-specific density representations.

Table 11: Ablation of the individual objectives used for density-consensus learning on the DenseShip6K dataset. Aniso. denotes the box-aware anisotropic Bayesian objective, Gcc. denotes the global count constraint, and Csc. denotes cross-modal structure consistency. The best results are highlighted in **bold**.

Aniso.	Gcc.	Csc.	$AP_{50}\uparrow$	$AP_{50:95}\uparrow$	MAE↓	RMSE↓
✓	✗	✗	84.5	58.3	7.25	12.22
✓	✓	✗	84.4	58.1	6.59	10.85
✓	✗	✓	84.6	58.4	6.83	11.49
✓	✓	✓	**85.0**	**58.5**	**5.85**	**10.51**

Adding the global count constraint reduces the MAE and RMSE from 7.25 and 12.22 to 6.59 and 10.85, respectively, while leaving the detection performance largely unchanged. This indicates that the count constraint mainly calibrates accumulated density mass at the image level. In contrast, introducing cross-modal structural consistency alone slightly improves $AP_{50:95}$ from 58.3% to 58.4% while also reducing the counting errors, suggesting that structural agreement between the OPT and SAR density representations provides more coherent spatial evidence.

Combining both constraints yields the strongest performance, reaching 85.0% AP_{50}, 58.5% $AP_{50:95}$, an MAE of 5.85, and an RMSE of 10.51. The results confirm that structural consistency regularizes the spatial organization of the modality-specific density representations, and the global count constraint calibrates the total predicted mass.

C.4.2 Gaussian Kernel Bandwidth Sensitivity

To examine the sensitivity of conventional pixel-wise density supervision to manually selected kernel parameters, we vary the Gaussian bandwidth σ while keeping the network architecture unchanged. As shown in Fig. 9, the counting performance changes non-monotonically with the Gaussian bandwidth.

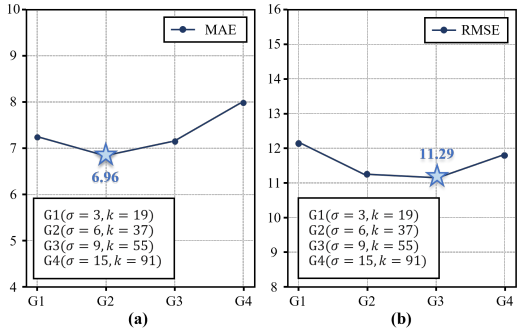

Figure 9: Sensitivity of fixed Gaussian density supervision to the Gaussian bandwidth σ and the corresponding kernel size k. MAE and RMSE vary non-monotonically with the predefined kernel parameters, illustrating the sensitivity of pixel-wise Gaussian supervision to manual scale selection.

A narrow kernel with $\sigma = 3$ results in an MAE of 7.29 and an RMSE of 12.15. Increasing the bandwidth to $\sigma = 6$ reduces the errors to 6.96 and 11.34, respectively. Further increasing the bandwidth does not consistently improve performance: $\sigma = 9$ achieves a slightly lower RMSE of 11.29 but a higher MAE of 7.20, while an excessively broad kernel with $\sigma = 15$ increases the MAE to 7.98.

These results illustrate a limitation of fixed-kernel density supervision. Small kernels produce highly concentrated targets and can impose large penalties for minor spatial deviations, whereas large kernels diffuse density mass over wider regions and may blur the separation between closely spaced instances. Moreover, no single bandwidth simultaneously optimizes both MAE and RMSE, since the

same isotropic scale is imposed on ships with heterogeneous sizes and aspect ratios. This sensitivity further motivates the box-aware Bayesian formulation, which uses annotation geometry for probabilistic mass assignment without requiring the network to reproduce a manually specified Gaussian density profile.

C.4.3 SENSITIVITY TO λ_{str}

We first analyze the sensitivity to the cross-modal structural consistency weight λ_{str}, while fixing the global count weight to $\lambda_{\text{cnt}} = 0.001$. As shown in Fig. 10, without structural regularization, the model achieves an $AP_{50:95}$ of 58.1%, an MAE of 6.59, and an RMSE of 10.85. Increasing λ_{str} to a moderate value improves both detection and counting performance, with $\lambda_{\text{str}} = 0.01$ yielding the best overall results of 58.5% $AP_{50:95}$, 5.85 MAE, and 10.51 RMSE.

Further increasing the structural weight degrades both tasks. This non-monotonic behavior reflects that an insufficient constraint cannot reliably establish agreement between the OPT and SAR density representations, whereas an excessively strong constraint may suppress modality-specific information that remains complementary for target perception. We therefore set $\lambda_{\text{str}} = 0.01$ in all other experiments.

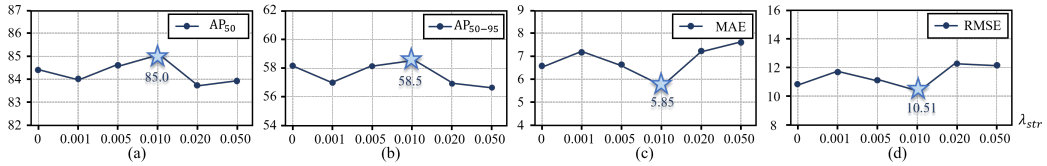

Figure 10: Sensitivity analysis of the cross-modal structural consistency weight λ_{str}. (a) AP_{50}, (b) $AP_{50:95}$, (c) MAE, and (d) RMSE under different λ_{str} values.

C.4.4 SENSITIVITY TO λ_{cnt}

Next, we investigate the global count constraint weight λ_{cnt} while fixing $\lambda_{\text{str}} = 0.01$. As shown in Fig. 11, introducing a small count weight of $\lambda_{\text{cnt}} = 0.001$ reduces the overall MAE and RMSE from 6.83 and 11.49 to 5.85 and 10.51, respectively, while the detection accuracy remains stable. Further increasing the weight does not provide additional improvements.

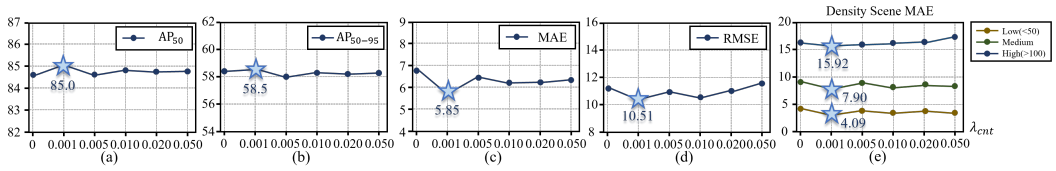

Figure 11: Sensitivity analysis of the global count constraint weight λ_{cnt}. (a) AP_{50}, (b) $AP_{50:95}$, (c) overall MAE, (d) overall RMSE, and (e) MAE across low-, medium-, and high-density scenes under different λ_{cnt} values.

These results suggest that the instance-wise Bayesian objective already provides the primary supervision for local density-mass assignment, and only a relatively weak global constraint is required to correct accumulated count deviations. Excessive count regularization may instead over-constrain the image-level density integral without providing additional information about spatial mass allocation.

The density-stratified results exhibit consistently larger errors as scene density increases. With the selected count weight, the MAE values for low-, medium-, and high-density scenes decrease from 4.34, 8.28, and 16.05 to 4.09, 7.90, and 15.92, respectively. The remaining error gap in highly dense scenes indicates that global mass calibration alone cannot fully resolve severe local ambiguity among closely spaced instances.

C.4.5 DENSITY-AWARE COUNT REWEIGHTING

To evaluate the effect of density-aware count reweighting, we divide the training samples into five count intervals: $[0, 50)$, $[50, 100)$, $[100, 150)$, $[150, 200)$, and $[200, +\infty)$. We compare three weighting strategies: uniform weighting $[1, 1, 1, 1, 1]$, mildly progressive weighting $[1, 1.1, 1.2, 1.3, 1.4]$, and the adopted density-aware weighting $[1, 1.2, 1.5, 2.0, 2.5]$.

As shown in Table 12, progressively increasing the contribution of dense samples consistently improves the results. Compared with uniform weighting, the adopted strategy improves $AP_{50:95}$ from 57.8% to 58.5% and reduces the overall MAE and RMSE from 6.71 and 11.24 to 5.85 and 10.51, respectively. Improvements are also observed across all three density regimes. This result indicates that uniform image-level weighting does not fully reflect the greater optimization difficulty of highly crowded scenes, which contain more instance-level mass assignments and stronger interactions among adjacent density responses.

Table 12: Comparison of different density-aware count reweighting strategies. The best results are highlighted in **bold**.

Bin weights	$AP_{50:95}\uparrow$	Overall		Density Scene (MAE↓)		
		MAE↓	RMSE↓	Low (<50)	Medium	High (>100)
$[1, 1, 1, 1, 1]$	57.8	6.71	11.24	4.37	8.22	16.13
$[1, 1.1, 1.2, 1.3, 1.4]$	58.0	6.60	11.06	4.29	8.16	16.08
$[1, 1.2, 1.5, 2.0, 2.5]$	**58.5**	**5.85**	**10.51**	**4.09**	**7.90**	**15.92**

C.4.6 DETECTION-DENSITY TASK BALANCE

The coefficients λ_{det} and λ_{den} control the contributions of the detection and density objectives during joint training. They are not constrained to sum to one, and their ratio does not directly represent the relative gradient contributions of the two tasks, which also depend on the underlying loss functions and shared network parameters. We therefore analyze the sensitivity to each coefficient by varying one while fixing the other.

Table 13: Effect of the density-task weight under $\lambda_{\text{det}} = 0.025$. The best results are highlighted in **bold**.

λ_{den}	Ratio $(\lambda_{\text{den}}/\lambda_{\text{det}})$	AP_{50}	$AP_{50:95}$	MAE	RMSE
0.025	1	83.8	57.1	7.96	12.13
0.050	2	84.1	57.7	7.38	11.97
0.100	4	83.4	56.7	8.75	13.00
0.200	8	83.5	56.8	7.32	12.64
0.300	12	**85.0**	**58.5**	**5.85**	**10.51**
0.400	16	83.8	57.2	7.06	11.93

Table 14: Effect of the detection-task weight under $\lambda_{\text{den}} = 0.30$. The best results are highlighted in **bold**.

λ_{det}	Ratio $(\lambda_{\text{den}}/\lambda_{\text{det}})$	AP_{50}	$AP_{50:95}$	MAE	RMSE
0.0125	24	82.9	55.7	8.02	12.47
0.0250	12	**85.0**	**58.5**	**5.85**	**10.51**
0.0375	8	83.7	57.2	6.62	11.84
0.0500	6	83.9	57.4	7.98	12.28
0.0750	4	83.8	56.4	7.61	11.99

With $\lambda_{\text{det}} = 0.025$, Table 13 shows that increasing λ_{den} does not lead to monotonic improvements. Among the evaluated settings, $\lambda_{\text{den}} = 0.30$ achieves the strongest overall performance, with 58.5%

$AP_{50:95}$, an MAE of 5.85, and an RMSE of 10.51. Increasing λ_{den} further to 0.40 reduces $AP_{50:95}$ to 57.2% and increases the MAE to 7.06, indicating that excessively emphasizing density supervision does not necessarily benefit either task.

Table 14 presents the complementary analysis with $\lambda_{\text{den}} = 0.30$. Increasing λ_{det} from 0.0125 to 0.025 improves $AP_{50:95}$ from 55.7% to 58.5% and reduces the MAE from 8.02 to 5.85. Larger detection weights do not provide further improvements in either detection or counting. Overall, $(\lambda_{\text{det}}, \lambda_{\text{den}}) = (0.025, 0.30)$ provides the best joint detection and counting performance among the evaluated configurations, indicating that the two objectives require a moderate balance during joint optimization.