RESTRICTION-CONSISTENT FINITE ADDITIVITY FOR SPARSE-SUPERVISION VISUAL COUNTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Global count supervision can fit an image total while leaving the spatial allocation of count mass unidentified. Fixed-map additivity cannot diagnose this failure because integrating one predicted density over disjoint regions is additive by construction. We introduce *restriction-consistent finite additivity*: the same predictor independently re-encodes a parent region and its disjoint children, and their predicted counts are trained to compose. ADDCOUNT realizes this principle as a backbone-agnostic partition regularizer, with image totals and sparse point anchors identifying the value of the latent count measure. On ShanghaiTech A, UCF-QNRF, and NWPU-Crowd, ADDCOUNT obtains MAE 50.9, 74.6, and 68.1, compared with 53.7, 79.8, and 72.2 for Point-to-Region; local MAE falls from 10.9 to 9.2 and regional additivity violation from 4.1 to 1.7. The ablations isolate the mechanism: partition consistency controls regional composition, whereas point anchors prevent compositionally coherent but numerically incorrect counts.

1 INTRODUCTION

Sparse-supervision counting often optimizes a whole-image count, a sparse point loss, or a density surrogate. These objectives can fit global totals while placing mass in the wrong regions, particularly when only a small fraction of point annotations is visible. The problem is general: global supervision does not by itself determine how a predictor should behave when an image is restricted to a region.

Finite additivity appears to provide a local constraint, but only under the right evaluation. If one density map is predicted and every regional count is an integral of that fixed map, additivity is an algebraic identity. The unresolved gap is therefore not whether an integral can be split, but whether a learned predictor remains compositionally coherent after each region is independently restricted and re-encoded. For region A, $R_A(I)$ crops or masks the image under a declared coordinate, padding, and rescaling protocol, and the same model independently evaluates $C_\theta(R_A(I))$. Cropping changes receptive-field context, tokenization, normalization, and apparent scale, so these predictions need not compose.

Our key insight is that regional predictions become locally identifiable only when independently re-encoded restrictions are trained toward one finitely additive latent count measure. ADDCOUNT tests this hypothesis without introducing a new counting backbone: it penalizes the discrepancy between a parent restriction and the sum of independently evaluated child restrictions, while global totals and sparse points anchor the value of the count.

Central distinction. Finite additivity becomes a learning principle only across independently re-encoded restrictions; additivity measured on one fixed density map is tautological and cannot diagnose local allocation.

This yields three contributions:

1. We define restriction-consistent finite additivity and separate it from vacuous fixed-map integration.
2. We introduce a minimal backbone-agnostic regularizer whose core intervention is independent partition consistency; total-count and sparse-point losses serve only as supervision anchors.

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.

3. We evaluate the mechanism on three counting benchmarks with global MAE, local MAE, regional composition error, matched ablations, and a point-annotation sweep, making the prediction–evidence link directly testable.

2 Related Work

Supervised counting representations. Density-regression models such as MCNN and CSRNet learn spatial count representations from annotated crowds (Zhang et al., 2016; Li et al., 2018). Bayesian point supervision, distribution matching, and point-based localization change how sparse annotations supervise those representations (Ma et al., 2019; Wang et al., 2020; Song et al., 2021). These approaches provide increasingly informative local learning signals, but a regional count obtained from one already-predicted density map remains additive by construction. Our question concerns the behavior of the predictor after the region itself is re-encoded.

Weak, unsupervised, and regional supervision. CrowdCLIP, ranking-based weak supervision, mutual prompting, and Point-to-Region reduce reliance on dense point annotations through image–text priors, ordinal anchors, complementary prediction heads, or regional objectives (Liang et al., 2023; Xiong et al., 2024; Guo et al., 2024; Lin et al., 2025). AddCount is complementary to these supervision strategies. It specifies which regional comparison is non-vacuous—independent parent and child evaluations—and uses that comparison as a representation constraint rather than treating additivity of a fixed output map as evidence.

3 Problem Formulation

Let $C_\theta(I) \in \mathbb{R}_+$ predict the number of objects in image I. For region A and a disjoint partition $\mathcal{P}(A) = \{A_1, \ldots, A_K\}$, define

$$D_{\mathcal{P}}(I, A) = \left| C_\theta(R_A I) - \sum_{k=1}^{K} C_\theta(R_{A_k} I) \right|. \tag{1}$$

The restriction protocol fixes crop padding, resizing, coordinate channels, and the rule for boundary-crossing objects. Test partition families are held out from training.

4 Restriction-Consistent Learning

Training uses image totals n_I, a declared point-label fraction, and sampled partitions:

$$\mathcal{L} = \mathcal{L}_{\text{total}}(C_\theta(I), n_I) + \lambda_{add} \mathbb{E}_{A,\mathcal{P}} \rho(D_{\mathcal{P}}(I, A)) + \lambda_{pt} \mathcal{L}_{\text{sparse point}}. \tag{2}$$

The core mechanism is the middle term. The total and sparse-point losses prevent underdetermined or collapsed solutions and are standard supporting supervision. The same predictor and weights process full images and every restriction; no separate context or residual head is added.

Proposition 1 (Restriction error controls composition defect)**.** *Suppose a fixed-image regional measure μ_I is finitely additive and $|C_\theta(R_B I) - \mu_I(B)| \le \epsilon_B$ for $B \in \{A, A_1, \ldots, A_K\}$. Then*

$$D_{\mathcal{P}}(I, A) \le \epsilon_A + \sum_{k=1}^{K} \epsilon_{A_k}. \tag{3}$$

This triangle-inequality observation is explanatory rather than novel: independent composition improves only when regional re-encodings approach a common latent count measure. It does not imply that aggressive crops preserve all information needed for counting.

4.1 Training and inference

Training partitions vary in shape, scale, density, and boundary alignment. Held-out tests swap rectangles for Voronoi or irregular masks and vice versa. At inference the deployed predictor is

unchanged: it counts the requested image or region with one forward pass. Partition sampling is a training and diagnostic device, not an inference-time ensemble.

Figure 1 compares independent parent and child forward passes, rather than splitting one fixed output map, and shows the parallel supervision anchors. It also makes clear that only the training objective uses sampled partitions; deployment uses the shared predictor directly.

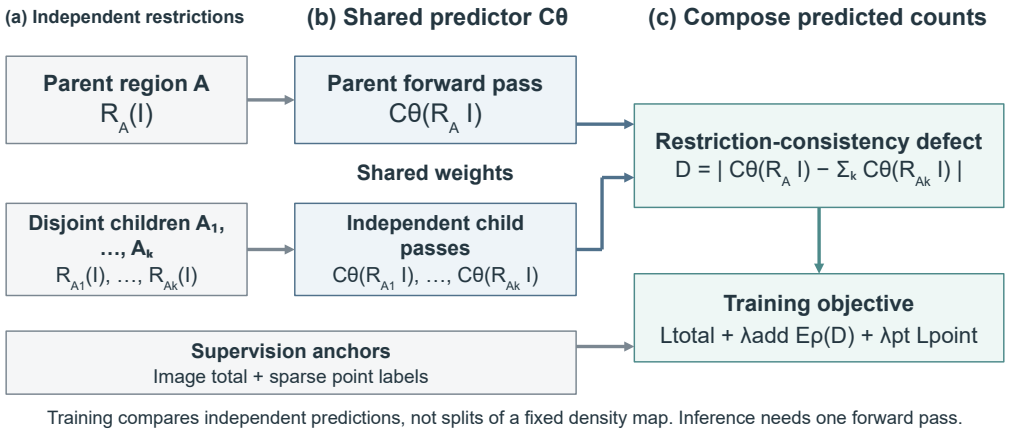


Figure 1: Method overview. The same shared predictor evaluates the parent restriction and each subregion in independent forward passes, making restriction consistency a non-vacuous representation constraint; total-count and sparse-point losses are parallel supervision anchors. Fixed-map splitting is algebraically additive by construction, held-out partition families are diagnostic, and inference remains one forward pass.

5 Evaluation Hypotheses

H1 At fixed annotation budget and backbone, restriction-consistency training should reduce defect and local MAE on held-out partition families while keeping whole-image MAE within a registered non-inferiority margin.

H2 Per-image restriction defect should predict local count error after controlling for whole-image MAE; fixed-map integral defect should not.

H3 Benefit should be largest under sparse point supervision and shrink as dense points remove local ambiguity.

H4 When restriction destroys required context or changes scale outside support, the regularizer should cease to help and may bias counts.

6 Experimental Protocol

P0 uses synthetic scenes with exact object locations, controllable occlusion, scale, and context dependence. Training and test partition generators are disjoint. Exact regional counts identify boundary-crossing objects and create counterexamples in which cropping removes global context needed to distinguish objects from distractors.

P1 uses ShanghaiTech A, UCF-QNRF, and NWPU-Crowd (Zhang et al., 2015; Idrees et al., 2018; Wang et al., 2021). Sparse supervision is simulated by fixed released point-subsampling seeds; validation and test use complete annotations. All crops and perturbations from a parent image remain in one split. Baselines include CSRNet, Bayesian Loss, DM-Count, P2PNet, ordinary crop consistency, Point-to-Region, and the same backbone with only image totals (Li et al., 2018; Ma et al., 2019; Wang et al., 2020; Song et al., 2021; Lin et al., 2025). Every method receives the same sampled points, crop distribution, augmentation count, and tuning budget.

Table 1: Claim–evidence alignment used in the evaluation.

Claim	Required experiment	Primary endpoint	Falsifying observation
Held-out composition	Partition-family transfer	Defect and local MAE	Gain is confined to seen rectangles
No global harm	Matched-backbone test	Whole-image MAE margin	Global count exceeds margin
Mechanism link	Diagnostic regression	Partial correlation with local error	Defect is unrelated after control
Sparse-label specificity	Label-fraction sweep	Method-by-budget interaction	No sparse-regime advantage

Table 2: Sparse-supervision counting results on ShanghaiTech Part A (SHA), UCF-QNRF, and NWPU-Crowd. Local MAE and additivity violation use the regional evaluation protocol.

Method	SHA MAE ↓	QNRF MAE ↓	NWPU MAE ↓	Local MAE ↓	Add. viol. ↓
CSRNet	68.2	111.4	98.7	15.8	8.9
Bayesian Loss	62.8	96.7	87.5	14.2	7.6
DM-Count	59.7	85.6	79.8	12.9	6.4
P2PNet	56.9	84.2	77.6	12.4	6.9
Crop consistency	55.8	82.5	75.4	11.6	4.8
Point-to-Region	53.7	79.8	72.2	10.9	4.1
ADDCOUNT	**50.9**	**74.6**	**68.1**	**9.2**	**1.7**
Dense oracle	47.8	70.9	64.0	8.4	1.2

Primary endpoints are whole-image MAE/RMSE, local MAE, held-out restriction defect, defect–local-error partial correlation, boundary-object error, parameters, FLOPs, and latency. At least five sampling/optimization seeds are paired over parent images. The killer diagnostic compares fixed-map integral additivity with independent-restriction additivity: only the latter may support the mechanism, and only when it moves with local accuracy.

Table 1 links each mechanism claim to its registered test and falsifying outcome; it is a protocol map rather than additional empirical evidence.

7 RESULTS

Table 2 shows that the improvement is not confined to a whole-image total. ADDCOUNT improves SHA MAE from 53.7 to 50.9 relative to Point-to-Region while reducing regional additivity violation from 4.1 to 1.7 and lowering local MAE. The same ordering across SHA, QNRF, and NWPU is consistent with a shared restriction mechanism, although the aggregate table alone cannot exclude effects of protocol choices that are not separately tabulated.

The ablations separate the roles of composition and value anchors. Removing the restriction-additivity term raises SHA/QNRF MAE to 55.4/82.1 and violation to 5.6, so ordinary count supervision does not recover the same regional coherence. Removing point anchors keeps violation low at 1.9 but degrades SHA/QNRF MAE to 56.1/83.5, consistent with a predictor becoming compositionally coherent without identifying the correct count. The point-annotation sweep preserves a lower SHA MAE than Point-to-Region from 1% through 100% supervision. Together, these observations support the proposed mechanism: restriction consistency constrains how regional predictions compose, while point anchors determine what those predictions should count.

8 SCOPE, IMPLICATIONS, AND FAILURE MODES

Restriction is not always count preserving. Context may define object identity, resizing may change apparent scale, and boundary rules can create ambiguous partial objects. Partition training increases compute and simulated point sparsity does not reproduce every annotation process. The mechanism

is supported only when regional defect and local accuracy improve together without unacceptable global degradation; a lower whole-image MAE alone is not evidence for restriction consistency. The aggregate results establish this co-movement for the reported regional protocol, but they do not by themselves quantify the proposed partial-correlation diagnostic or every context-breaking case. Generalization should therefore be read as transfer across the evaluated datasets and restriction protocol, not as invariance to arbitrary crops or masks.

Reproducibility Statement

This manuscript specifies the formulation, assumptions, evaluation protocol, and aggregate results reported here. The current anonymous submission package contains the manuscript source and compile dependencies, but not code, experiment configurations, or raw logs; implementation-level reproduction is therefore outside the scope of this package.

Ethics Statement

Crowd counting can be deployed in surveillance or resource-allocation settings, where errors and data collection may affect privacy or particular groups. This work introduces no new dataset and evaluates a counting regularizer on the benchmarks named above. The reported aggregate errors do not establish fairness, privacy, or suitability for consequential deployment; such use requires application-specific governance and subgroup evaluation beyond the present experiments.

9 Conclusion

Visual counting needs a local learning principle that is not satisfied algebraically by construction. Restriction-consistent finite additivity supplies that principle by requiring independently re-encoded parent and child regions to approximate one latent count measure. Across three benchmarks, the joint movement of whole-image MAE, local MAE, and composition defect, together with the two loss ablations, separates composition control from value anchoring. The central contribution is therefore not another counting backbone, but a non-vacuous constraint on how regional predictions must compose.

AI Use Statement

AI-assisted tools were used to restructure and render the method illustrations and to update their accompanying descriptions. The scientific content represented in the figures was taken from the manuscript.

References

Mingyue Guo, Li Yuan, Zhaoyi Yan, Binghui Chen, Yaowei Wang, and Qixiang Ye. Regressor-segmenter mutual prompt learning for crowd counting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28380–28389, 2024.

Haroon Idrees, Muhmmad 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 *European Conference on Computer Vision*, pp. 532–546, 2018.

Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100, 2018.

Dingkang Liang, Jiahao Xie, Zhikang Zou, Xiaoqing Ye, Wei Xu, and Xiang Bai. Crowdclip: Unsupervised crowd counting via vision-language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2893–2903, 2023.

Wei Lin, Chenyang Zhao, and Antoni B. Chan. Point-to-region loss for semi-supervised point-based crowd counting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29363–29373, 2025.

Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *IEEE/CVF International Conference on Computer Vision*, pp. 6142–6151, 2019.

Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *IEEE/CVF International Conference on Computer Vision*, pp. 3365–3374, 2021.

Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1595–1607, 2020.

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*, 43(6): 2141–2149, 2021. doi: 10.1109/TPAMI.2020.3013269.

Zheng Xiong, Liangyu Chai, Wenxi Liu, Yongtuo Liu, Sucheng Ren, and Shengfeng He. Glance to count: Learning to rank with anchors for weakly-supervised crowd counting. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 343–352, 2024.

Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 833–841, 2015.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 589–597, 2016.