DualCount: Structurally Consistent Density and Point Modeling for Zero-Shot Object Counting

Xuan Cuong Ngo

University of Arkansas, USA
cngo@uark.edu

Abstract. Zero-shot object counting aims to estimate the number of objects specified by a text query without category-specific training. Recent approaches primarily rely on density regression or detection-style instance prediction. While effective, density-based models often suffer from spatial ambiguity and background leakage due to weakly regulated mass allocation, leading to fragmented or part-biased representations that increase counting error in complex scenes. In this work, we propose an instance-aware dual-decoder framework that structurally couples density and point representations for zero-shot object counting. Instead of treating density estimation as independent pixel-wise regression, we interpret it as a structured mass allocation problem over a latent set of object instances. Predicted instance centers induce a soft instance-wise decomposition of the density map, upon which we enforce two geometric constraints: (1) per-instance mass conservation, ensuring each object contributes approximately one unit of density mass, and (2) center-of-mass alignment, encouraging each density component to concentrate around its corresponding predicted center. These constraints introduce instance-level geometric consistency and lead to more accurate mass allocation, thereby reducing counting error. Extensive experiments on FSC-147, PUCPR+, and CARPK show that our approach consistently reduces counting error and establishes new state-of-the-art performance in zero-shot object counting.

Keywords: Object Counting · Point Prediction · Zero-shot Learning

1 Introduction

Object counting traditionally focuses on estimating the number of instances belonging to a predefined class within an image, such as crowd counting [24], vehicle counting [5], or animal counting [18]. These approaches rely on category-specific supervision and struggle to generalize beyond trained object categories. To improve generalization, class-agnostic few-shot counting has been introduced, where the target object is specified at inference time using visual exemplars, typically user-provided bounding boxes highlighting a few instances. Although such methods enable open-world counting, they require manual intervention at test time, limiting full automation and scalability.

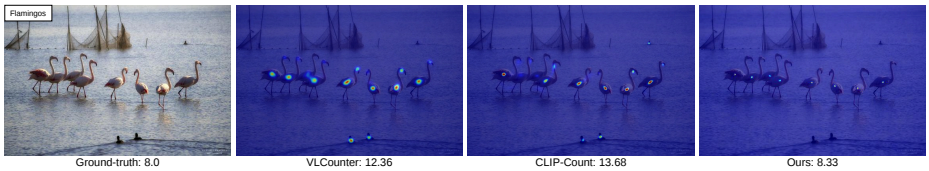

Fig. 1: Density map visualizations. CLIP-Count [8] and VLCounter [9] are prone to false positive predictions, while our model mitigates this issue by leveraging spatial constraints from the point prediction branch, resulting in more controlled density estimates.

To address this limitation, recent work explores zero-shot object counting, where the object of interest is specified using a textual description instead of visual exemplars [3,4,28]. By leveraging vision–language models, these approaches allow arbitrary object categories to be queried at inference time using class names. Most existing zero-shot counting methods adopt density regression, estimating object count by integrating a predicted density map conditioned on text prompts. While effective, this formulation has a key limitation: density supervision provides limited guidance on instance-level spatial structure. Although the predicted density map is encouraged to match the target density at each location, it does not explicitly regulate how the density mass should be distributed across individual object instances.

This limited instance-level guidance leads to spatial ambiguity. In practice, density mass may accumulate on correlated background regions or concentrate on highly discriminative object parts rather than stable instance centers. For objects with complex shapes, heavy occlusion, or cluttered surroundings, such ambiguous mass allocation results in fragmented instance representations and increased counting error. This failure case can be observed in Fig. 1. Importantly, this issue arises not from weak semantic alignment alone, but from the absence of instance-level structural constraints within the density formulation itself.

Our key insight is that density maps implicitly represent a mixture of object-level components, yet existing approaches treat them as unconstrained continuous fields. If each object contributes approximately one unit of mass, then density estimation can be reformulated as a structured mass allocation problem over a latent set of object instances. Enforcing geometric consistency at the instance level can therefore directly improve counting reliability.

Based on this insight, we propose **DualCount**, an instance-aware dual-decoder framework for zero-shot object counting. Our model jointly predicts a density map and a set of instance centers from shared visual–text features. Rather than treating these predictions independently, we interpret them as complementary views of a shared latent object set. The predicted centers induce a soft instance-wise decomposition of the density map, allowing each object to be represented as an individual density component.

On top of this decomposition, we introduce two geometric constraints. First, a per-instance mass conservation constraint ensures that each object contributes

approximately one unit of density mass, directly aligning spatial allocation with the counting objective. Second, a center-of-mass alignment constraint encourages each density component to concentrate around its corresponding predicted center, promoting coherent and compact instance representations. Together, these constraints transform density estimation from pure pixel-wise regression into a structured instance-level modeling problem. By reducing ambiguous mass allocation and preventing part-biased density drift, our approach leads to more accurate mass distribution and consequently reduces counting error.

Extensive experiments on FSC-147, PUCPR+, and CARPK demonstrate that our method consistently reduces counting error and establishes new state-of-the-art performance in zero-shot object counting. The improvements are particularly pronounced in crowded and geometrically complex scenes, where standard density regression tends to produce unstable spatial representations.

The contributions of this paper are threefold:

- **Instance-aware formulation for zero-shot counting.** We reinterpret density estimation as a structured mass allocation problem over a latent set of object instances, revealing the importance of instance-level geometric constraints.
- **Geometric consistency through dual decoders.** We propose an instance-aware dual-decoder framework that couples density and point representations via per-instance mass conservation and center-of-mass alignment constraints.
- **Improved counting performance.** Extensive experiments and ablation studies on FSC-147, PUCPR+, and CARPK demonstrate consistent reductions in counting error and new state-of-the-art results in zero-shot object counting.

2 Related Work

Class-Specific Object Counting. Object counting has long been studied in computer vision, with early research primarily focusing on specific object categories such as cars, cells, pedestrians, or polyps. Detection-based approaches were widely adopted in this stage, where object detectors are trained to localize and count instances of predefined categories. Modern detection frameworks such as YOLO and Faster R-CNN have been successfully applied in vehicle counting and surveillance systems, while segmentation-based methods are often used in biological cell counting [25] and crowd analysis [27]. However, detection-based counting typically struggles in highly crowded scenes due to severe occlusions and object overlaps. To alleviate these limitations, density-based methods were introduced, which estimate object counts by regressing a continuous density map whose integral corresponds to the object count. Density-based formulations are more robust to occlusions and scale variations, and have become the dominant paradigm in crowd counting and dense object counting scenarios. Subsequent

studies further improved density-based supervision beyond standard Gaussian-map regression. Bayesian Loss [14] models probabilistic pixel-to-point contributions, while DM-Count [27] formulates counting as distribution matching with an optimal transport (OT)-based loss. Our density decoder follows this OT-based formulation, but further introduces instance-level geometric constraints to guide density allocation around individual objects.

Class-Agnostic Object Counting. To improve generalization beyond predefined categories, class-agnostic few-shot counting has recently gained significant attention. In this setting, the target object is specified at inference time using a small number of visual exemplars, typically provided as bounding boxes. Early methods adopt Siamese matching networks to learn similarity between exemplars and image regions for density estimation. BMNet+ [22] incorporates non-linear similarity metrics for improved matching, while FamNet [21] enhances backbone design for better density prediction. CounTR [12] combines convolutional feature extraction with transformer-based modeling and cross-attention to improve feature fusion. SAFECount [29] introduces feature refinement modules to enhance generalization, and LOCA [26] proposes object prototype extraction for iterative adaptation to exemplar appearance and shape. Despite their effectiveness, these approaches depend heavily on the quality and representativeness of visual exemplars. Sparse, biased, or noisy exemplars can degrade performance. Moreover, requiring manual exemplar input at inference time limits scalability and full automation in open-world scenarios.

Text-Specified Object Counting. Text-specified or zero-shot object counting eliminates the need for visual exemplars by allowing users to specify the target object using textual descriptions. Leveraging vision–language models, these approaches enable open-world counting across arbitrary object categories. Paiss et al. [15] demonstrated that fine-tuned CLIP models can perform limited counting. Building upon this direction, CLIP-Count [8] and VLCounter [9] condition density regression on text embeddings extracted from CLIP, enabling open-vocabulary density-based counting. TFOC [23] further explores training-free counting by prompting the segmentation model SAM [10] with text, points, or boxes. While these methods demonstrate promising generalization, most existing approaches rely on density regression with limited instance-level spatial guidance. As a result, although the predicted density map is encouraged to match the target density, supervision does not explicitly regulate how density mass should be distributed across object instances. This can lead to background leakage, fragmented instance representations, or part-biased density allocation, particularly in cluttered or geometrically complex scenes. In contrast, our work introduces instance-level geometric consistency into density-based zero-shot counting. By coupling density estimation with predicted instance centers and enforcing per-instance mass conservation and center-of-mass alignment constraints, we transform density regression into a structured mass allocation problem, reducing ambiguous spatial distribution and improving counting reliability.

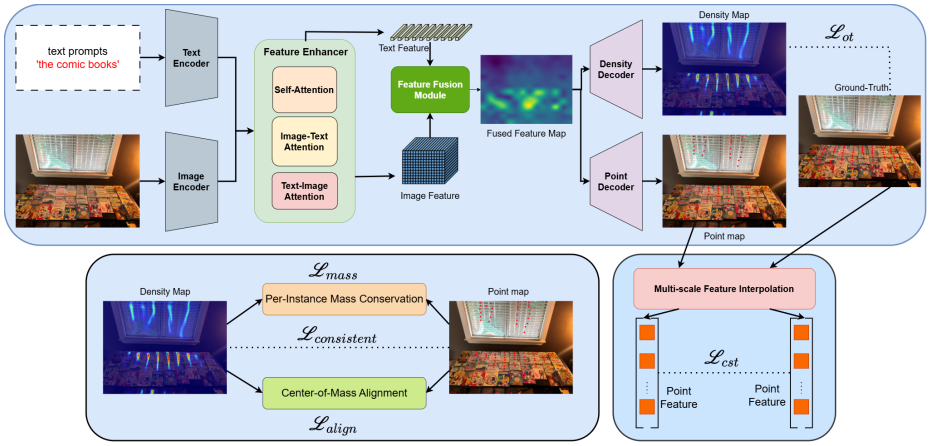

Fig. 2: Overview of DualCount. Visual and textual features are extracted by the image encoder E_v and text encoder E_t, enhanced via cross-modal interaction $\mathcal{E}$, and fused by $\mathcal{F}$ to produce a fused feature map. A dual-decoder head (D_ρ and D_c) predicts the density map and instance centers.

3 Methodology

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a text query t specifying the target object category, zero-shot object counting aims to estimate the number of instances of the queried category in the image. Let N denote the ground-truth object count and $\hat{N}$ the predicted count. In this setting, the object category is defined solely by the textual query t, and the model is required to predict the corresponding object count.

3.1 Architecture Overview.

Figure 2 presents the proposed DualCount framework. Our framework consists of five components: (1) an image encoder E_v, (2) a text encoder E_t, (3) a cross-modal feature enhancement module $\mathcal{E}$, (4) a feature fusion module $\mathcal{F}$, and (5) a dual-decoder head composed of a density decoder D_ρ and a point decoder D_c. The image and text encoders extract unimodal representations, which are subsequently refined and fused into a text-conditioned spatial feature map. The dual decoders then produce complementary outputs, a density map for count estimation and a set of instance centers for instance-level structural modeling.

Encoders. We employ a visual encoder E_v and a text encoder E_t to extract unimodal representations from the input image and text query. The image encoder maps $I \in \mathbb{R}^{H\times W\times 3}$ to a spatial feature map $F_v = E_v(I) \in \mathbb{R}^{C\times H'\times W'}$, where C denotes the feature dimension and (H', W') are the spatial resolutions after downsampling. Each spatial location corresponds to a C-dimensional feature

vector encoding local visual information. The text encoder projects the query t into a semantic embedding $f_t = E_t(t) \in \mathbb{R}^C$. The visual and textual features share the same embedding dimension, enabling effective cross-modal interaction in subsequent modules.

Feature Enhancement via Cross-Modal Attention. Although the backbone feature map $F_v \in \mathbb{R}^{C\times H'\times W'}$ encodes rich visual semantics, it is not explicitly optimized for fine-grained object counting. Counting requires both precise spatial discrimination and strong semantic alignment with the queried object. To strengthen the representation, we design a cross-modal enhancement module $\mathcal{E}(\cdot)$ composed of image self-attention and bidirectional cross-attention. We first reshape F_v into a sequence of spatial tokens $\mathbf{V} \in \mathbb{R}^{(H'W')\times C}$, where each row corresponds to a spatial location.

Image Self-Attention. We apply multi-head self-attention to capture long-range spatial dependencies among image features. Given the visual feature sequence $\mathbf{V}$, the query, key, and value features are computed as $\mathbf{Q}_v = \mathbf{V}W_q^v$, $\mathbf{K}_v = \mathbf{V}W_k^v$, and $\mathbf{V}_v = \mathbf{V}W_v^v$, where $W_q^v, W_k^v, W_v^v \in \mathbb{R}^{C\times C}$ are learnable projection matrices. The self-attended features are then obtained as

$$\mathbf{V}^{\text{sa}} = \text{Softmax}\Big((\mathbf{Q}_v\mathbf{K}_v^\top)/\sqrt{C}\Big)\,\mathbf{V}_v. \tag{1}$$

This operation allows each spatial location to aggregate contextual information from all other regions, producing more robust visual features for counting.

Image-to-Text Cross-Attention. To incorporate semantic guidance from the queried category, we allow visual tokens to attend to the text embedding $f_t \in \mathbb{R}^C$. The text key and value are computed as $\mathbf{k}_t = f_tW_k^t$ and $\mathbf{v}_t = f_tW_v^t$, while the visual queries are computed as $\mathbf{Q}_{vt} = \mathbf{V}^{\text{sa}}W_q^{vt}$. The cross-attended visual features are then obtained by

$$\mathbf{V}^{vt} = \text{Softmax}\Big((\mathbf{Q}_{vt}\mathbf{k}_t^\top)/\sqrt{C}\Big)\,\mathbf{v}_t. \tag{2}$$

This step conditions the visual representation on the queried category and highlights spatial regions that are semantically relevant for counting.

Text-to-Image Cross-Attention. To further strengthen alignment, the text representation attends back to the visual features. We compute $\mathbf{q}_t = f_tW_q^{tv}$, $\mathbf{K}_{tv} = \mathbf{V}^{\text{sa}}W_k^{tv}$, and $\mathbf{V}_{tv} = \mathbf{V}^{\text{sa}}W_v^{tv}$.

The updated text feature is computed as

$$f_t' = \text{Softmax}\Big((\mathbf{q}_t\mathbf{K}_{tv}^\top)/\sqrt{C}\Big)\,\mathbf{V}_{tv}, \tag{3}$$

which encourages tighter cross-modal correspondence.

Feed-Forward Refinement. Finally, the enhanced visual tokens and text embedding are refined via position-wise feed-forward networks, yielding $\mathbf{V}^{enh} = \text{FFN}_v(\mathbf{V}^{vt})$ and $f_t^{enh} = \text{FFN}_t(f_t')$. The visual tokens are reshaped back to spatial form $\tilde{F}_v \in \mathbb{R}^{C\times H'\times W'}$.

The enhanced feature map $\tilde{F}_v$ encodes strengthened spatial context and improved semantic alignment with the text query, providing a robust foundation for both density estimation and center prediction.

Feature Fusion. After cross-modal enhancement, we obtain the refined visual feature map $\tilde{F}_v \in \mathbb{R}^{C\times H'\times W'}$ and the enhanced text embedding $f_t^{enh} \in \mathbb{R}^C$. We fuse these modalities to construct a text-conditioned spatial feature map F that serves as input to the dual decoders.

We adopt feature-wise modulation to inject global semantic information from the text embedding into spatial visual features. Specifically, we generate channel-wise modulation parameters from the text feature:

$$\gamma = W_\gamma f_t^{enh}, \quad \beta = W_\beta f_t^{enh}, \tag{4}$$

where $W_\gamma, W_\beta \in \mathbb{R}^{C\times C}$ are learnable projections, and $\gamma, \beta \in \mathbb{R}^C$.

The fused feature map is computed as

$$F(x) = \gamma \odot \tilde{F}_v(x) + \beta, \tag{5}$$

for each spatial location $x \in \Omega$, where Ω denotes the set of all spatial coordinates in the image domain and $\odot$ denotes channel-wise multiplication.

This modulation mechanism allows the text embedding to dynamically reweight and shift visual feature channels, emphasizing dimensions relevant to the queried object category while suppressing irrelevant responses. As a result, the fused feature map $F \in \mathbb{R}^{C\times H'\times W'}$ encodes spatially grounded representations that are conditioned on the target semantics. This representation is then shared by both the density decoder and the center decoder.

Dual Decoder Given the fused feature map $F \in \mathbb{R}^{C\times H'\times W'}$, we design a dual-decoder architecture that produces (i) a density map for count estimation and (ii) a binary point map for instance-level supervision.

Shared Upsampling and Refinement. Both decoders share a lightweight refinement head composed of bilinear interpolation layers for progressive upsampling and 2×2 convolutional layers for channel reduction. Let F^{up} denote the upsampled and refined feature map at the original image resolution (H, W). This shared structure ensures that both density and point predictions are derived from spatially aligned high-resolution features.

Density Prediction Head. The density prediction head applies a fully connected feed-forward network (FFN) on F^{up} to produce a scalar density value at each spatial location: $\rho = \mathrm{FFN}_\rho(F^{up})$ with $\rho \in \mathbb{R}^{H\times W}$. The predicted count is obtained by integrating the density map: $\hat{N} = \sum_{x\in\Omega} \rho(x)$. Instead of pixel-wise regression, following [4] we supervise the density map using an optimal transport (OT) loss $\mathcal{L}_{ot}$ with the ground-truth point annotations. Let $\mathcal{P}^{gt} = \{p_j\}_{j=1}^N$ denote the set of ground-truth object centers. The OT loss aligns the predicted density distribution ρ with the discrete empirical distribution defined by $\mathcal{P}^{gt}$, encouraging mass to concentrate around true object locations while preserving

global count consistency. This formulation provides spatially meaningful supervision beyond simple ℓ_2 density regression.

Point Prediction Head. In parallel with density estimation, we predict a sparse set of candidate instance centers using a point prediction branch. Concretely, the point head outputs a *logit* map $S = \mathrm{FFN}c(F^{up}) \in \mathbb{R}^{H\times W}$, which is converted to a confidence map by a sigmoid, $h(x) = \sigma(S(x)) \in (0, 1)$ with $x \in \Omega$. At training and inference time, we extract a set of center proposals $\hat{\mathcal{P}} = \{\hat{p}_i\}_{i=1}^{M}$ by first applying non-maximum suppression (NMS) to the confidence map h with a fixed window size, and then selecting all remaining local maxima whose confidence exceeds a threshold τ:

$$\hat{\mathcal{P}} = \{x \in \Omega \mid x \text{ is a local maximum under NMS and } h(x) \geq \tau\}. \tag{6}$$

Each proposal $\hat{p}_i = (u_i, v_i)$ is associated with a confidence score $s_i = h(\hat{p}_i)$, and the number of proposals $M = |\hat{\mathcal{P}}|$ varies per image. To train this, we use Focal loss $\mathcal{L}_{focal}$ following [2]. To further establish a one-to-one correspondence between predicted proposals and ground-truth centers $\mathcal{P}^{gt} = \{p_j\}_{j=1}^{N}$, we solve a bipartite matching problem using the Hungarian algorithm. The matching cost is quadric cost and we compute an optimal assignment $\mathcal{M} \subset \hat{\mathcal{P}} \times \mathcal{P}^{gt}$. Unmatched predictions are treated as negatives (false positives), and unmatched ground-truth points are treated as missed detections. For each predicted point $\hat{p}_i$, we first feed the input image I into a CNN backbone to obtain multi-scale feature maps. We then extract a discriminative point-wise feature for $\hat{p}_i$ by bilinear sampling from these feature maps. Let $\{F^{l_j}\}_{j=1}^{L}$ denote features from L selected layers (e.g., intermediate decoder stages). For each scale l_j, we sample at location $\hat{p}_i$ using bilinear interpolation $f^{l_j}(\hat{p}_i) = \mathrm{Bilinear}\big(F^{l_j}, \hat{p}_i\big)$. We then concatenate features across scales to obtain the final point representation:

$$f(\hat{p}_i) = \mathrm{Concat}\left(f^{l_1}(\hat{p}_i), \ldots, f^{l_L}(\hat{p}_i)\right) \in \mathbb{R}^d. \tag{7}$$

Following [3], we apply a contrastive objective to encourage matched predictions to be similar to their corresponding ground-truth counterparts while pushing away unmatched predictions. Specifically, for each matched pair $(\hat{p}_i, p_j) \in \mathcal{M}$, we treat $f(\hat{p}_i)$ and $f(p_j)$ as a positive pair, while features extracted at unmatched predicted points $\hat{p}_k$ (or randomly sampled background locations) serve as negatives. We use an InfoNCE-style loss:

$$\mathcal{L}_{\mathrm{cst}} = -\frac{1}{|\mathcal{M}|} \sum_{(\hat{p}_i, p_j)\in\mathcal{M}} \log \frac{\exp\big(\mathrm{sim}(f(\hat{p}_i), f(p_j))/\tau\big)}{\sum_{q\in\{p_j\}\cup\mathcal{N}_i} \exp\big(\mathrm{sim}(f(\hat{p}_i), f(q))/\tau\big)}. \tag{8}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature hyperparameter, and $\mathcal{N}_i$ is the negative set (unmatched predictions and/or background samples). This loss improves the discriminability of point features and suppresses false-positive centers by contrasting them against matched instances.

3.2 Soft Instance Decomposition

The density map $\rho \in \mathbb{R}^{H\times W}$ provides a continuous representation for counting but does not explicitly encode instance-level structure. To introduce object-wise

modeling, we decompose ρ into soft instance-specific components guided by the predicted point set $\hat{\mathcal{P}} = \{\hat{p}_i\}_{i=1}^{M}$ obtained from the point decoder, where each $\hat{p}_i = (u_i, v_i)$ denotes a 2D coordinate on the image grid.

For each spatial location $x \in \Omega$, we define a soft assignment weight with respect to each predicted point using a Gaussian kernel:

$$w_i(x) = \frac{\exp\left(-\frac{\|x-\hat{p}_i\|^2}{2\sigma^2}\right)}{\sum_{j=1}^{M} \exp\left(-\frac{\|x-\hat{p}_j\|^2}{2\sigma^2}\right)}, \tag{9}$$

where σ controls the spatial spread of the assignment and $\sum_{i=1}^{M} w_i(x) = 1$ for all $x \in \Omega$. The global density map is then decomposed as

$$\rho_i(x) = w_i(x)\rho(x), \tag{10}$$

which satisfies $\rho(x) = \sum_{i=1}^{M} \rho_i(x)$.

Each component ρ_i represents the soft density contribution associated with the predicted point $\hat{p}_i$. This decomposition introduces differentiable instance-level structure into the density representation.

3.3 Per-Instance Mass Conservation

Although standard density-based counting supervises the predicted density map with pixel-wise regression, it provides limited explicit control over how much density mass is assigned to each individual instance. To address this limitation, we introduce Per-Instance Mass Conservation, which constrains each object instance to receive a consistent unit of density mass.

Given the decomposition $\{\rho_i\}_{i=1}^{M}$, we compute the total mass assigned to instance i as

$$m_i = \sum_{x \in \Omega} \rho_i(x). \tag{11}$$

Ideally, each object contributes approximately one unit of density mass, i.e., $m_i \approx 1$ for valid instances. We therefore introduce a per-instance mass conservation loss:

$$\mathcal{L}_{\text{mass}} = \frac{1}{M} \sum_{i=1}^{M} (m_i - 1)^2. \tag{12}$$

This regularization encourages balanced mass allocation across predicted instances and prevents excessive concentration or dispersion of density mass.

3.4 Center-of-Mass Alignment

Beyond mass conservation, we further enforce geometric consistency between density allocation and predicted points. For each density component ρ_i, we compute its centroid as

$$\bar{p}_i = \frac{1}{m_i} \sum_{x \in \Omega} x\, \rho_i(x), \tag{13}$$

where $x = (u, v)$ denotes spatial coordinates and m_i is defined above.

We then align the centroid $\bar{p}_i$ with the corresponding predicted point $\hat{p}_i$ using

$$\mathcal{L}_{\text{align}} = \frac{1}{M} \sum_{i=1}^{M} \|\bar{p}_i - \hat{p}_i\|_2^2. \tag{14}$$

This alignment encourages each density component to concentrate around its associated predicted point, reducing fragmented or part-biased mass allocation and promoting instance-level geometric coherence.

3.5 Overall Loss Function

In addition to $\mathcal{L}_{ot}$, $\mathcal{L}_{cst}$, $\mathcal{L}_{focal}$, $\mathcal{L}_{mass}$, and $\mathcal{L}_{align}$ introduced in previous sections, we enforce agreement between the two counting mechanisms through a count consistency loss that penalizes discrepancies between the count estimated from the density map and the number of predicted points:

$$\mathcal{L}_{\text{consistent}} = \left(\hat{N} - M\right)^2, \tag{15}$$

where $\hat{N} = \sum_{x \in \Omega} \rho(x)$ denotes the total mass of the predicted density map and M is the number of predicted points.

The final training objective is

$$\mathcal{L} = \lambda_{\text{ot}}\mathcal{L}_{\text{ot}} + \lambda_{\text{cst}}\mathcal{L}_{\text{cst}} + \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} + \lambda_{\text{consistent}}\mathcal{L}_{\text{consistent}} + \lambda_{\text{mass}}\mathcal{L}_{\text{mass}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}, \tag{16}$$

where $\lambda_{\text{ot}}, \lambda_{\text{cst}}, \lambda_{\text{focal}}, \lambda_{\text{consistent}}, \lambda_{\text{mass}}$, and λ_{align} balance the contributions of the loss terms.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on three widely used object counting benchmarks: FSC-147, CARPK, and PUCPR+.

FSC-147 [21] is a large-scale few-shot counting dataset comprising 6,135 images across diverse object categories. The number of instances per image ranges from 7 to 3,731, with an average of 56. Each image is paired with three exemplar crops, randomly chosen object instances annotated with bounding boxes. The dataset is split into 89 categories for training, and 29 disjoint categories each for validation and testing, making FSC-147 a challenging open-set benchmark.

CARPK [7] focuses on vehicle counting in aerial imagery, captured by drones over parking lots. It provides 989 images for training and 459 for testing, with annotations for more than 89,000 cars.

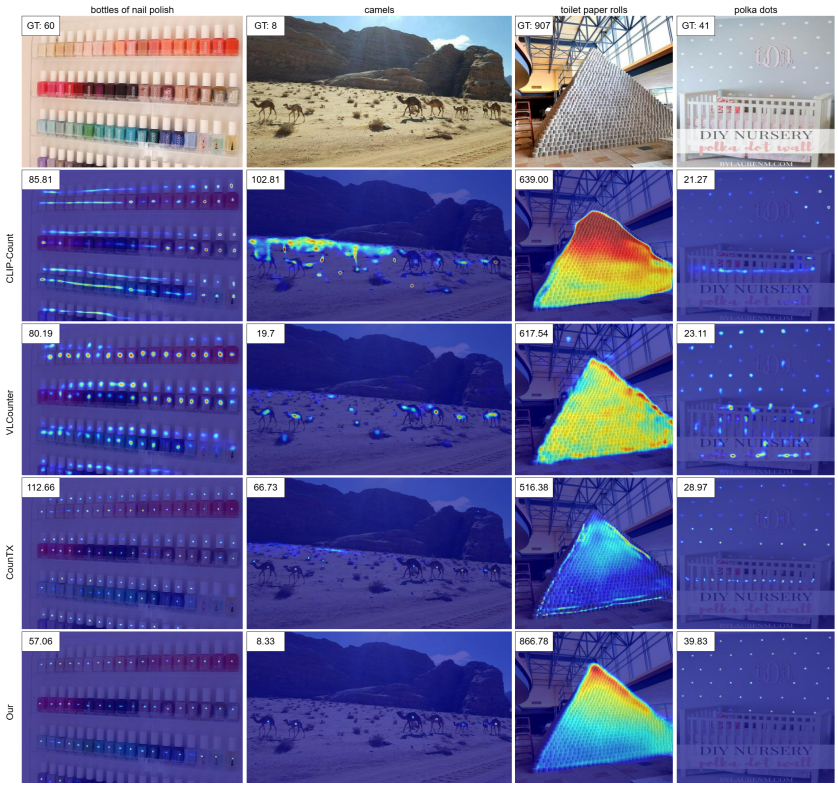

Fig. 3: Qualitative comparison between DualCount and density-based zero-shot counting baselines, including VLCounter [9], CLIP-Count [8], and CounTX [1].

PUCPR+ [7] is another vehicle counting dataset collected from fixed surveillance cameras at the Pontifical Catholic University of Paraná, serving as a complementary fixed-camera benchmark to the drone-based CARPK dataset.

We assess model performance using two standard evaluation metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE):

$$MAE = \frac{1}{N_I}\sum_{i=1}^{N_I}|N_i^{pred} - N_i^{gt}| \qquad RMSE = \sqrt{\frac{1}{N_I}\sum_{i=1}^{N_I}(N_i^{pred} - N_i^{gt})^2} \tag{17}$$

where N_I is the number of test images, N_i^{pred} is the predicted count for image i, and N_i^{gt} is the corresponding ground-truth count.

Table 1: Quantitative comparison on FSC147. We report MAE and RMSE under different settings: few-shot, reference-less, and zero-shot. In each setting, the best results are highlighted in **bold** while the runner-up is highlighted in **underline**.

Scheme	Method	Venue	Shot	Val Set		Test Set	
				MAE	RMSE	MAE	RMSE
Few-shot	FamNet [21]	CVPR'21	3	24.32	70.94	22.56	101.54
	BMNet [22]	CVPR'22	3	15.74	<u>58.53</u>	14.62	<u>91.83</u>
	LOCA [26]	ICCV'23	3	**10.24**	**32.56**	**10.97**	**56.97**
	SAM [23]	WACV'24	3	-	-	19.95	132.16
	PseCo [11]	CVPR'24	3	<u>15.31</u>	68.34	<u>13.05</u>	112.86
	GMN [13]	ACCV'19	1	29.66	89.81	26.52	124.57
	FamNet [21]	CVPR'21	1	24.32	70.94	22.56	101.54
	BMNet [22]	CVPR'22	1	19.06	67.95	16.71	103.31
Reference-less	FamNet [21]	CVPR'21	0	32.15	98.75	32.27	131.46
	LOCA [26]	ICCV'23	0	**17.43**	**54.96**	**16.22**	**103.96**
	RCC [6]	CVPR'23	0	<u>17.49</u>	<u>58.81</u>	<u>17.12</u>	<u>104.53</u>
Zero-shot	Patch-selection [20]	CVPR'23	0	26.93	88.63	22.09	115.17
	CLIP-Count [8]	ACM MM'23	0	18.79	61.18	17.78	106.62
	CounTX [1]	BMVC'23	0	17.10	65.61	15.88	106.29
	VLCounter [9]	AAAI'24	0	18.06	65.13	17.05	106.16
	PseCo [11]	CVPR'24	0	23.90	100.33	16.58	129.77
	DAVE [17]	CVPR'24	0	15.48	52.57	14.90	103.42
	VA-Count [30]	ECCV'24	0	17.87	73.22	17.88	129.31
	GeCo [16]	NeurIPS'24	0	14.81	64.95	13.30	108.72
	CountGD [2]	NeurIPS'24	0	<u>12.14</u>	<u>47.51</u>	12.98	98.35
	T2ICount [19]	CVPR'25	0	13.78	58.78	<u>11.76</u>	<u>97.86</u>
	DualCount (CLIP)	**-**	0	14.12	54.27	12.33	99.78
	DualCount (SwinT + BERT)	**-**	**0**	**11.21**	**42.77**	**11.32**	**97.16**

4.2 Implementation Details

We train the model using the Adam optimizer with a learning rate of 6.25×10^{-6}. Point-level representations are extracted from the ResNet-50 backbone using feature activations from `layer1` to `layer4`. All loss balancing weights in Eq. 16 are selected via grid search. For the encoder, we employ two image–text encoder pairs: CLIP image and text encoders, and a Swin Transformer paired with BERT. More implementation details are provided in the Appendix.

4.3 Comparison with the State-of-the-Arts

We compare DualCount with prior counting methods on the FSC147 benchmark, and the results are summarized in Table 1. The baselines span three categories: *few-shot*, *reference-less*, and *zero-shot* methods. Our approach belongs to the zero-shot category.

In the zero-shot setting, where no exemplar information is provided, existing methods typically suffer from higher errors due to the lack of reference guidance. Among the previous approaches, CountGD [2] achieves the runner-up result on the validation set with 12.14 MAE and 47.51 RMSE, while T2ICount [19] achieves the runner-up result on the test set with 11.76 MAE and 97.86 RMSE. In contrast, DualCount achieves the best performance on both validation and test sets, obtaining 11.21 MAE / 42.77 RMSE on the validation set and 11.32 MAE

Table 2: Performance comparison between DualCount and zero-shot counting baselines on CARPK and PUCPR+.

Dataset	Method	Venue	MAE ↓	RMSE ↓
CARPK	CLIP-Count [8]	ACM MM'23	11.96	16.61
	CounTX [1]	BMVC'23	8.13	10.87
	T2ICount [19]	CVPR'25	8.61	13.47
	CountGD [2]	NeurIPS'24	**3.83**	**5.41**
	DualCount (Ours)	–	<u>7.95</u>	<u>9.78</u>
PUCPR+	CLIP-Count [8]	ACM MM'23	55.82	91.23
	CounTX [1]	BMVC'23	75.24	113.44
	VLCounter [9]	AAAI'24	48.94	69.08
	CountGD [2]	NeurIPS'24	24.31	32.16
	DualCount (Ours)	–	**24.27**	**28.91**

Table 3: Ablation study on the proposed loss components. ✓ indicates the loss is applied during training, while × indicates it is removed. Test MAE and RMSE are reported for each variant.

$\mathcal{L}_{ot}$	$\mathcal{L}_{cst}$	$\mathcal{L}_{consistent}$	$\mathcal{L}_{mass}$	$\mathcal{L}_{align}$	MAE	RMSE
✓	✓	✓	✓	✓	11.32	97.16
✓	✓	✓	✓	×	12.84	100.45
✓	✓	✓	×	✓	13.05	100.98
✓	✓	×	✓	✓	14.25	102.17
✓	×	✓	✓	✓	14.49	103.11

/ 97.16 RMSE on the test set. These improvements highlight the effectiveness of DualCount in capturing object-level information without exemplars through structural consistency between density and point representations.

To evaluate cross-dataset generalization, we further test DualCount on the CARPK and PUCPR+ datasets [7]. The results are summarized in Table 2. Although CountGD reports the lowest error on CARPK, DualCount ranks second overall and outperforms recent zero-shot counting methods including CLIP-Count [8], CounTX [1], and T2ICount [19]. On the PUCPR+ dataset, DualCount achieves an MAE of 24.27 and RMSE of 28.91, surpassing previous methods such as CountGD [2]. These results demonstrate the strong generalization capability of DualCount across different datasets and confirm its robustness in zero-shot object counting scenarios.

4.4 Qualitative Results

Fig. 3 presents qualitative comparisons between DualCount and existing density based zero-shot counting methods, including VLCounter, CLIP-Count, and CounTX. We visualize the predicted density maps to highlight differences in spatial allocation of density mass. Compared with prior approaches, DualCount produces more concentrated and accurate density estimates. This improvement is mainly attributed to the proposed instance-level structural constraints. The Per-Instance Mass Conservation encourages balanced density allocation across predicted instances and prevents excessive concentration or dispersion of mass.

Table 4: Ablation study on architectural components of DualCount. We evaluate the impact of the feature enhancer, decoder design, and backbone choice.

Enhancer	Density Decoder	Point Decoder	Backbone	MAE ↓	RMSE ↓
✓	✓	×	SwinT+BERT	14.12	103.87
✓	×	✓	SwinT+BERT	14.54	105.29
×	✓	✓	SwinT+BERT	13.11	101.08
✓	✓	✓	CLIP	12.33	99.78
✓	✓	✓	SwinT+BERT	**11.32**	**97.16**

In addition, the Center-of-Mass Alignment constraint encourages each density component to align with its corresponding predicted point, reducing fragmented or part-biased density assignments. As a result, DualCount effectively reduces false positives in challenging scenarios, including self-similar objects (e.g., nail polishes), non-convex object layouts (e.g., camels), crowded scenes (e.g., toilet paper rolls), and distracting backgrounds (e.g., polka dots). In contrast, previous methods often spread density mass into irrelevant regions, leading to inaccurate counting predictions.

4.5 Ablation Study

We conduct ablation studies to analyze the contribution of the proposed loss functions and architectural components of DualCount. The results are summarized in Tables 3 and 4.

Effect of loss components. Table 3 evaluates the impact of each proposed loss term. The full model achieves the best performance with an MAE of **11.32** and RMSE of **97.16**. Removing the center-of-mass alignment loss $\mathcal{L}_{align}$ leads to a noticeable degradation (MAE increases from 11.32 to 12.84), indicating its importance in aligning density components with predicted instance centers. Similarly, removing the per-instance mass conservation loss $\mathcal{L}_{mass}$ increases the MAE to 13.05, suggesting that balanced density allocation across predicted instances is critical for accurate counting. Excluding the count consistency loss $\mathcal{L}_{consistent}$ further degrades performance (MAE 14.25), demonstrating the importance of enforcing agreement between density-based and point-based counting estimates. Finally, removing the contrastive loss $\mathcal{L}_{cst}$ leads to the largest performance drop (MAE 14.49), highlighting the role of discriminative point representations in improving localization and counting accuracy.

Effect of architectural components. Table 4 studies the influence of different architectural choices. Using only a single decoder significantly degrades performance. Specifically, the density-only model obtains an MAE of 14.12, while the point-only model results in an MAE of 14.54. This demonstrates that the two decoders provide complementary information for counting. Removing the feature enhancer also leads to worse performance (MAE 13.11), indicating that cross-modal feature refinement helps produce more informative spatial representations. We further compare different backbone configurations and observe

that replacing the CLIP backbone with the SwinT+BERT encoder improves the MAE from 12.33 to **11.32**. This confirms that the proposed architecture benefits from richer multi-scale visual features and stronger language representations.

Overall, these results verify that both the proposed loss functions and architectural components contribute significantly to the final performance of DualCount.

5 Conclusion

In conclusion, we presented DualCount, a zero-shot object counting framework that jointly models density estimation and point-based instance centers. By introducing instance-level structural constraints, per-instance mass conservation and center-of-mass alignment, our method enforces balanced density allocation and geometric consistency, reducing fragmented or biased density predictions. Extensive experiments on FSC147, CARPK, and PUCPR+ demonstrate that DualCount achieves state-of-the-art performance in the zero-shot setting while maintaining strong cross-dataset generalization. Ablation studies confirm the contribution of each component, and qualitative results show that DualCount produces more accurate and concentrated density maps with fewer false positives in challenging scenes. These findings highlight the value of incorporating instance-level structure into density-based counting, and suggest a promising direction for building more robust and generalizable zero-shot counting systems. More broadly, many computer vision tasks rely on point- or density-map prediction, such as human pose estimation and temporal repetition counting in videos. We believe that the proposed instance-level structural modeling may provide a useful direction for improving density-based prediction beyond object counting.

References

1. Amini-Naieni, N., Amini-Naieni, K., Han, T., Zisserman, A.: Open-world text-specified object counting. In: British Machine Vision Conference (2023)
2. Amini-Naieni, N., Han, T., Zisserman, A.: Countgd: Multi-modal open-world counting. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 48810–48837. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/57c56985d9afe89bf78a8264c91071aa-Paper-Conference.pdf
3. Cuong, N.X.: Contrastive point feature matching for open-world object counting. In: 36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025. BMVA (2025), https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_1183/paper.pdf
4. Cuong, N.X., Mai, T.D.: Distribution-guided object counting with optimal transport and dino-based density refinement. In: Buntine, W., Fjeld, M., Tran, T., Tran, M.T., Huynh Thi Thanh, B., Miyoshi, T. (eds.) Information and Communication Technology. pp. 182–192. Springer Nature Singapore, Singapore (2025)

5. Guerrero-Gómez-Olmedo, R., Torre-Jiménez, B., López-Sastre, R., Maldonado-Bascón, S., Onoro-Rubio, D.: Extremely overlapping vehicle counting. In: Iberian conference on pattern recognition and image analysis. pp. 423–431. Springer (2015)
6. Hobley, M., Prisacariu, V.: Learning to count anything: Reference-less class-agnostic counting with weak supervision. arXiv preprint arXiv:2205.10203 (2022)
7. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: Proceedings of the IEEE international conference on computer vision. pp. 4145–4153 (2017)
8. Jiang, R., Liu, L., Chen, C.: Clip-count: Towards text-guided zero-shot object counting. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4535–4545 (2023)
9. Kang, S., Moon, W., Kim, E., Heo, J.P.: Vlcounter: Text-aware visual representation for zero-shot object counting. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2714–2722 (2024)
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
11. Li, G., Li, X., Wang, Y., Wu, Y., Liang, D., Zhang, S.: Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In: European Conference on Computer Vision. pp. 457–472. Springer (2022)
12. Liu, C., Zhong, Y., Zisserman, A., Xie, W.: Countr: Transformer-based generalised visual counting. arXiv preprint arXiv:2208.13721 (2022)
13. Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: Asian conference on computer vision. pp. 669–684. Springer (2018)
14. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6142–6151 (2019)
15. Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., Dekel, T.: Teaching clip to count to ten. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3170–3180 (2023)
16. Pelhan, J., Lukezic, A., Zavrtanik, V., Kristan, M.: A novel unified architecture for low-shot counting by detection and segmentation. Advances in Neural Information Processing Systems **37**, 66260–66282 (2024)
17. Pelhan, J., Zavrtanik, V., Kristan, M., et al.: Dave-a detect-and-verify paradigm for low-shot counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23293–23302 (2024)
18. Qian, Y., Humphries, G., Trathan, P., Lowther, A., Donovan, C.: Counting animals in aerial images with a density map estimation model. ecol. evol. 13 (4), e9903 (2023)
19. Qian, Y., Guo, Z., Deng, B., Lei, C.T., Zhao, S., Lau, C.P., Hong, X., Pound, M.P.: T2icount: Enhancing cross-modal understanding for zero-shot counting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25336–25345 (2025)
20. Ranjan, V., Nguyen, M.H.: Exemplar free class agnostic counting. In: Proceedings of the Asian Conference on Computer Vision. pp. 3121–3137 (2022)
21. Ranjan, V., Sharma, U., Nguyen, T., Hoai, M.: Learning to count everything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3394–3403 (2021)
22. Shi, M., Hao, L., Feng, C., Liu, C., Cao, Z.: Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

23. Shi, Z., Sun, Y., Zhang, M.: Training-free object counting with prompts. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 323–331 (January 2024)
24. Sindagi, V.A., Patel, V.M.: A survey of recent advances in cnn-based single image crowd counting and density estimation. Pattern Recognition Letters **107**, 3–16 (2018)
25. Tyagi, A.K., Mohapatra, C., Das, P., Makharia, G., Mehra, L., AP, P., et al.: Degpr: Deep guided posterior regularization for multi-class cell detection and counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23913–23923 (2023)
26. Đukić, N., Lukežič, A., Zavrtanik, V., Kristan, M.: A low-shot object counting network with iterative prototype adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18872–18881 (2023)
27. Wang, B., Liu, H., Samaras, D., Nguyen, M.H.: Distribution matching for crowd counting. Advances in neural information processing systems **33**, 1595–1607 (2020)
28. Xu, J., Le, H., Nguyen, V., Ranjan, V., Samaras, D.: Zero-shot object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15548–15557 (2023)
29. You, Z., Yang, K., Luo, W., Lu, X., Cui, L., Le, X.: Few-shot object counting with similarity-aware feature enhancement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6315–6324 (2023)
30. Zhu, H., Yuan, J., Yang, Z., Guo, Y., Wang, Z., Zhong, X., He, S.: Zero-shot object counting with good exemplars. In: European Conference on Computer Vision. pp. 368–385. Springer (2024)