SAM3Count for Zero-Shot Open Vocabulary Counting in Images and Videos

Joana Konadu Owusu and Shivanand Venkanna Sheshappanavar
Geometric Intelligence Research Lab
University of Wyoming
{jowusu1,ssheshap}@uwyo.edu

Abstract

*Open-vocabulary counting (OVC) has garnered significant attention for its ability to count without relying on manual exemplars or class-specific requirements. OVC requires prompt localization, scene understanding, and composition, as well as the ability to distinguish between instances of the same object type. Most methods rely on manually adding exemplars along with text prompts to aid localization and improve performance, but this comes at the cost of ease of use. In the video domain, OVC is equally important for real-time applications, given the challenges posed by occlusions, deformations, and fragmentation. We introduce **SAM3Count**, a zero-shot SAM3-based OVC framework for images and videos. For video counting, SAM3Count builds on SAM3 by designing a lightweight reidentification tracker that maintains an appearance bank to recover lost tracks and curb identity switches. For images, it uses adaptive ROI tiling to improve counting performance across diverse scenes without requiring manual exemplars or priors. SAM3Count achieves impressive results, surpassing the most recent state-of-the-art (SOTA) methods across image (FSCD-147, ShanghaiTech, CARPK) and video (TAO-Count, Penguins) benchmarks. Code is available at https://github.com/Joan947/SAM3Count.*

Figure 1. SAM3Count is a zero-shot counting framework that improves over SAM3 for denser scenes in images and addresses double-counting, identity switches, in video counting.

1. Introduction

Object counting estimates the number of instances of a target object present in an image or video by identifying and aggregating all occurrences in the scene. As a fundamental computer vision task, it has broad real-world applications, including crowd management and urban planning [14, 53], traffic analysis [5, 46], wildlife conservation [33, 36, 47], and scientific imaging [12]. However, obtaining these counts manually is labor-intensive; annotating one hour of drone or surveillance footage requires hours of expert effort. This motivated the emergence of automated counting systems that generalize across scenes, domains, and object types. Earlier methods were class-specific example; people, cells, or animals [4, 20, 41].

Recent works [3, 16, 25, 51] have advanced to class-agnostic OVC in images, with the aim of counting objects from unseen data or categories. Advances [26, 40, 54] in the perception of open vocabulary have made it possible to localize arbitrary textual concepts rather than a set of closed labels. These baselines have provided strong building blocks for image-level counting systems [3, 17, 51, 52]. Video-based OVC remains underexplored. Object counting across video frames poses challenges: counting unique and multiple instances, preventing double-counting, and handling occlusions and appearance variations. Earlier video counting methods [8, 22, 50, 56] were class-specific

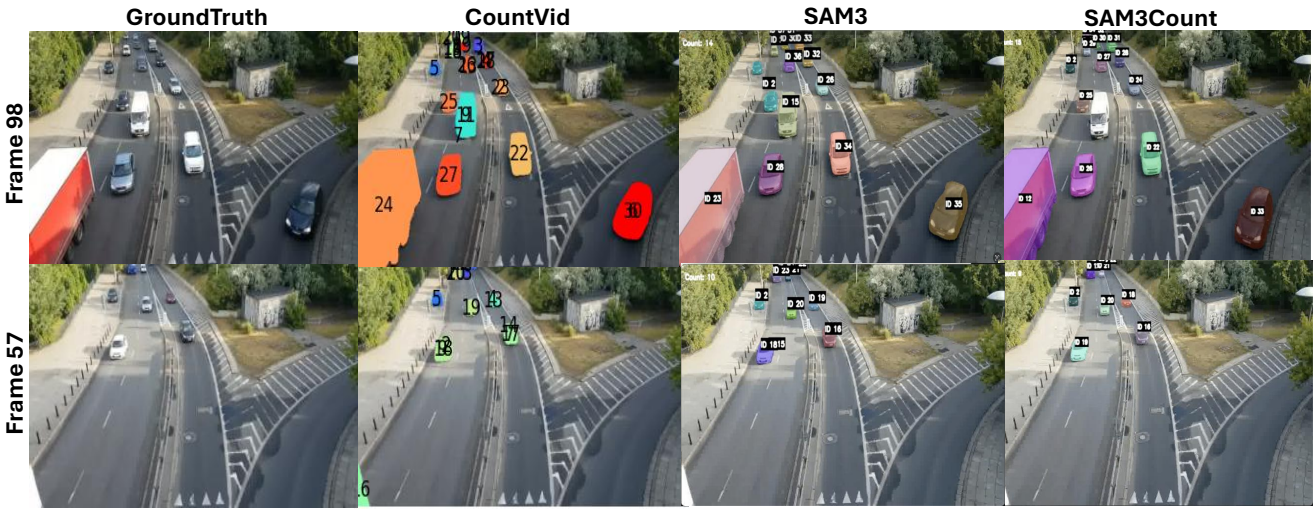

Figure 2. Qualitative comparison illustrating double counting and id inconsistencies across Frame 57 and Frame 98. CountVid[2] exhibits temporal inconsistencies leading to duplicate counting, and SAM3 switches ids for vehicles in the later frames. SAM3Count maintains more consistent identities across frames, reducing double-counting and ID reassignment errors.

and typically followed a detection-track-count pipeline, in which per-frame detections are linked into trajectories using multi-object tracking (MOT) or temporal density regression to predict frame-wise density maps. These pipelines generally assume a closed label set and do not naturally support arbitrary user-specified categories. CountVid [2] introduces the VideoCount benchmark and a prompt-based localization–tracking pipeline for open-vocabulary video counting. However, its results also highlight a central bottleneck: tracking inconsistencies and ID fragmentation during occlusion and reappearance, leading to under- or over-counting.

In this paper, we propose SAM3Count, a text-prompted image and video object-counting framework that improves identity consistency and counting performance in open-vocabulary settings (Figure 1). SAM3Count builds on SAM3 [6] as a frozen backbone for zero-shot segmentation and tracking. We introduce an adaptive multimodal re-identification tracker that fuses appearance, motion, spatial proximity, and temporal continuity to maintain stable IDs. It supports object reappearance, handles occlusions, and deformations in videos. We build on SAM3's image-counting capabilities to perform zero-shot image counting. Figure 2 illustrates the failure modes we address in video counting: double counting and reassignment of one ID to multiple objects during tracking. Our main contributions are fourfold:

- We design SAM3Count for zero-shot, open-vocabulary counting in images and videos.
- We build a multimodal adaptive re-identification tracker on top of SAM3's concept-aware tracker, for SAM3Count to maintain persistent identities under occlusions, deformations, and temporary disappearances, reducing both under-counting and double-counting in videos.
- We design an adaptive ROI tiling module to improve the efficiency and performance of our image-counting model.
- We achieve SOTA results on four image and two video benchmarks.

2. Related Work

2.1. Open Vocabulary Counting in Images

The early object counting methods [4, 20, 21, 55] focused on pedestrian crowd scenes, using detection-based and density-regression paradigms. Detection-based methods [19, 29, 57] estimate counts using a detector head. Advances like CNN-based detectors [13], point-supervised localization [28], RGB-D cues [24], and YOLO-based head detection [15] improved performance in crowded scenes. However, performance dropped under heavy occlusion and with small object sizes, prompting the use of density regression. Context-aware methods such as CSRNet [23] and CAN [27] tackle dense crowds by using dilated convolutions and multi-scale contextual modeling to handle scale variation, perspective effects, and severe crowding. These models are typically category-specific and lack instance identity, limiting temporal or open-category counting.

Open-vocabulary detection (OVD) localize objects from free-form text queries, allowing recognition beyond fixed label sets [26, 29, 57]. Models like [26] enables OVD through image-text pre-training, while SAM-based extensions [18, 40, 54] produce text-guided masks for more precise localization. These models broaden recognition capability, but they are not designed specifically for dense counting and often degrade for small, crowded, or heavily overlapping instances. Building on this direction, recent

work has moved toward open-vocabulary and class-agnostic counting in images.

Class-agnostic counting methods [3, 16, 34, 44, 45] estimate the number of instances of an object in an image without relying on a predefined label set. Wang *et al.* [49] show that vision transformers enable few-shot class-agnostic counting via exemplar–query matching, reducing reliance on fixed taxonomies, but may struggle when exemplars are not representative or when occlusions are severe. [17] uses CLIP features to align text prompts and image regions for zero-shot counting, while other approaches leverage general VLMs to guide counting in a zero-shot or few-shot setting [51]. [16] extends detection-based counting to class-agnostic settings using SAM to generate mark proposals and CLIP for classification, without an explicit design for dense scenarios. Open-vocabulary image counting methods, such as [3], address this issue by using few-shot approach, combining text and visual exemplars to count arbitrary objects in dense scenes. Subsequent work [43, 52] introduced improved prompt queues and loss functions to further improve performance. In our work, we focus on text prompted open vocabulary counting with explicit handling of densely crowded scenes.

2.2. Open Vocabulary Counting in Videos

Earlier methods [8, 9, 50] for video counting followed a detection-plus-tracking paradigm in which detectors localize objects frame by frame and multi-object trackers [56] or universal association models [22]. Open-vocabulary video counting is much less explored. CountVid [2], a method in this setting, decomposes counting into detection, class-agnostic segmentation, and tracking [3, 18, 38]. Although it demonstrates the feasibility of open-vocabulary counting in videos, its tracking pipeline remains sensitive to deformation, occlusion, and distractors, which can produce duplicate identities, fragmented trajectories, and overcounting (see Figure 5). In our approach, we address these limitations using SAM3 [6], together with a re-id tracker to improve identity consistency in and support more reliable open-vocabulary video counting.

3. Open Vocabulary Adaptive Counting

We propose **SAM3Count**, a counting framework with improved identification and tracking consistency for open-vocabulary counting in images and videos. SAM3Count leverages the SAM3 [6] as a backbone for zero-shot detection, tracking and counting. SAM3 uses text, points, boxes and exemplars as its input prompt. We focus on text input only for zero shot counting.

3.1. SAM3Count for Images

We improve SAM3 for zero shot open vocabulary counting based on text inputs only in two stages described below.

3.1.1. Stage 1: Image inference and Density decision

Given an input image I and a prompt q, SAM3 returns a set of candidate instance masks with associated masks, boxes, and scores. We filter them using mask-level non-maximum suppression based on Intersection-over-Minimum (IoM) to produce a set of detections with: $\mathrm{IoM}(\mathbf{m}_a, \mathbf{m}_b) = (|\mathbf{m}_a \cap \mathbf{m}_b|)/(\min(|\mathbf{m}_a|, |\mathbf{m}_b|))$.

Density estimation and tiling trigger: To decide whether the image is a dense scene and proceed to Stage 2, we compute three signals from the clean Stage 1 detections: (i) *coverage area* c, (ii) the number of instances N, and (iii) a *size score* s_{size} that increases as the average box size decreases (objects occupying $> 10\%$ of the image yield $s_{\text{size}} \approx 0$). $s_{\text{dens}} = w_c\, c + w_n \min\left(\frac{N}{C_{\max}}, 1\right) + w_s\, s_{\text{size}}$, where w_c, w_n, and w_s are scalar weights and $\min(N/C_{\max}, 1)$ is the count term normalized and capped to avoid dominance in very crowded scenes.

3.1.2. Stage 2: ROI-guided adaptive tiling (dense scenes)

Stage 2 is activated only for dense scenes to increase recall for small, crowded instances while keeping the computation bounded and efficient. We compute the ROI of the padded union of Stage 1 boxes and generate an adaptive overlapping tile grid based on the boxes' density scores.

Adaptive tiling rule: We map the density outcome to a tiling rule that defines a target grid and an overlap ratio. Given an image of size (W, H), we select a tiling regime r and map it to a target grid (C_r, R_r) (number of tile columns/rows) and the overlap ratio $\gamma_r \in (0, 1)$. We compute the tile size from the grid and clamp it to a valid range by setting $t_r = \mathrm{clip}(\max(W/C_r, H/R_r), t_{\min}, t_{\max})$ and stride $s_r = t_r - \lfloor \gamma_r t_r \rfloor$, then slide a $t_r \times t_r$ window with stride s_r. Each tile is independently processed by SAM3, then mapped to global coordinates, and the detections are merged into a set of unified candidates. For extremely dense scenes, the tiling refinement is applied recursively, such that tiles that remain near detector capacity after initial suppression are subdivided into finer overlapping sub-tiles and reprocessed with a higher confidence threshold. **Cross-tile suppression and Final counting:** To prevent duplicate predictions for the same physical instance, we apply (1) Cross-tile box NMS to remove obvious duplicates across (2) Final mask IoM NMS(same criterion as Stage 1) to handle heavy overlaps; and the filter. Figure 4 provides an overview of the ROI adaptive tiling process.

3.1.3. Image Counting Implementation details.

We report the settings of the image counting pipeline. In the density score, the weights for the coverage, count, and size terms are set to $(0.3, 0.5, 0.2)$, with the count term defined as $\min\left(\frac{N}{C_{\max}}, 1\right)$, where N is the number of clean detections from stage 1 and $C_{\max} = 50$ (the count cap used for normalization). The size score is defined as $s_{\text{size}} =$

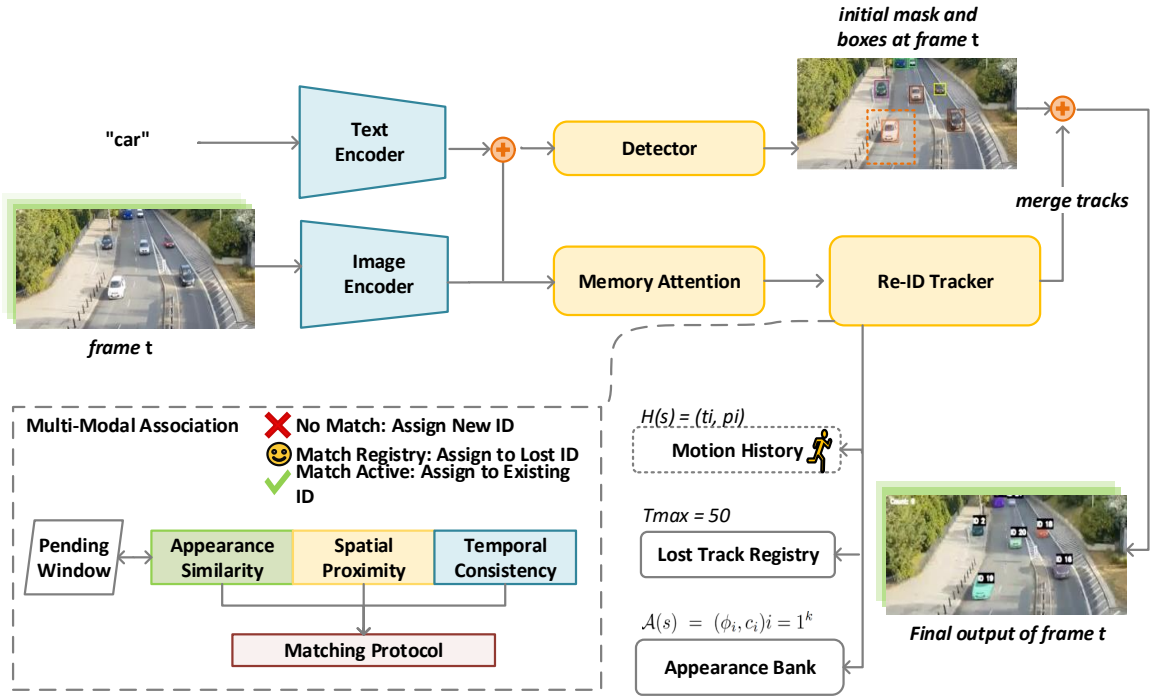

Figure 3. SAM3Count architecture overview. Given a textual query (e.g., "car") and video frame input, joint text–image encoders guide object detection and memory attention, while a Re-ID tracker leverages appearance, spatial, and temporal cues along with motion history and a lost-track registry to maintain consistent identities across frames.

$1 - \min(\bar{a}/0.1, 1)$, where $\bar{a}$ denotes the average box area normalized by the image area; therefore, objects occupying more than roughly 10% of the image contribute a near-zero size score. For adaptive tiling, the density regime is mapped to configurations (C_r, R_r, γ_r), and tile size is constrained to $t_{\min} = 97$ and $t_{\max} = 1024$. The ROI is defined as the padded union of Stage 1 boxes, using padding $p_x = \max(16, 0.04\, b_w)$ and $p_y = \max(16, 0.04\, b_h)$, where b_w and b_h are the ROI width and height.

3.2. SAM3Count in Videos

For video counting, we augment SAM3 with a multi-modal re-identification (Re-ID) tracker that maintains a consistent identity space over time using appearance cues, motion continuity, and recovery of temporarily lost tracks. This reduces duplicate counting caused by split trajectories, identity switches, and reappearing objects.

3.2.1. Input Processing:

Let $V = I_i i = 1^n$ be the video and τ the text prompt. SAM3 encodes each frame as $\Phi_i = \mathrm{PE}(I_i)$ and the prompt as $e_\tau = \mathrm{CLIP}_{\text{text}}(\tau)$, initializes masks on the first frame, and then propagates them through memory-guided tracking. The perception encoder produces high-dimensional feature maps (e.g., 1008×1008 resolution) that we extract for subsequent appearance modeling without any additional training or finetuning.

3.2.2. Multi-Modal Re-Identification Tracker

SAM3 may fragment a single physical object into multiple internal IDs or switch identities when objects are visually similar (see Figure 2). To address this, we introducea multi-modal re-identification (Re-ID) tracker that maintains a consistent identity space on top of SAM3's raw track IDs. For each trajectory, we store (i) an appearance bank $\mathcal{A}(s)$, (ii) a short motion history $\mathcal{H}(s)$, and (iii) a track state indicating whether the identity is currently active, temporarily lost, or terminated.

Appearance bank: We keep up to $K_{\max} = 10$ masked descriptors with their confidence scores, $\mathcal{A}(s) = \{(\phi_i, c_i)\}_{i=1}^{k}$, where $\phi_i \in \mathbb{R}^d$ is L2-normalized and $k \leq K_{\max}$. Given a feature map $\Phi \in \mathbb{R}^{c \times h \times w}$ and a binary mask $M \in \{0,1\}^{h \times w}$, the masked descriptor is

$$\phi = \mathrm{Normalize}\left(\frac{\sum_{i,j} \Phi[:, i, j] \cdot M[i, j]}{\sum_{i,j} M[i, j] + \varepsilon} \right). \quad (1)$$

To preserve diversity, we only add ϕ if it is not redundant with the bank, i.e., $\max_{\phi' \in \mathcal{A}(s)} \cos(\phi, \phi') < 0.95$.

Motion history: We maintain a motion history $\mathcal{H}(s) = (t_i, p_i)$, for each object with $p_i = [c_x, c_y]^\top$, (bounding box

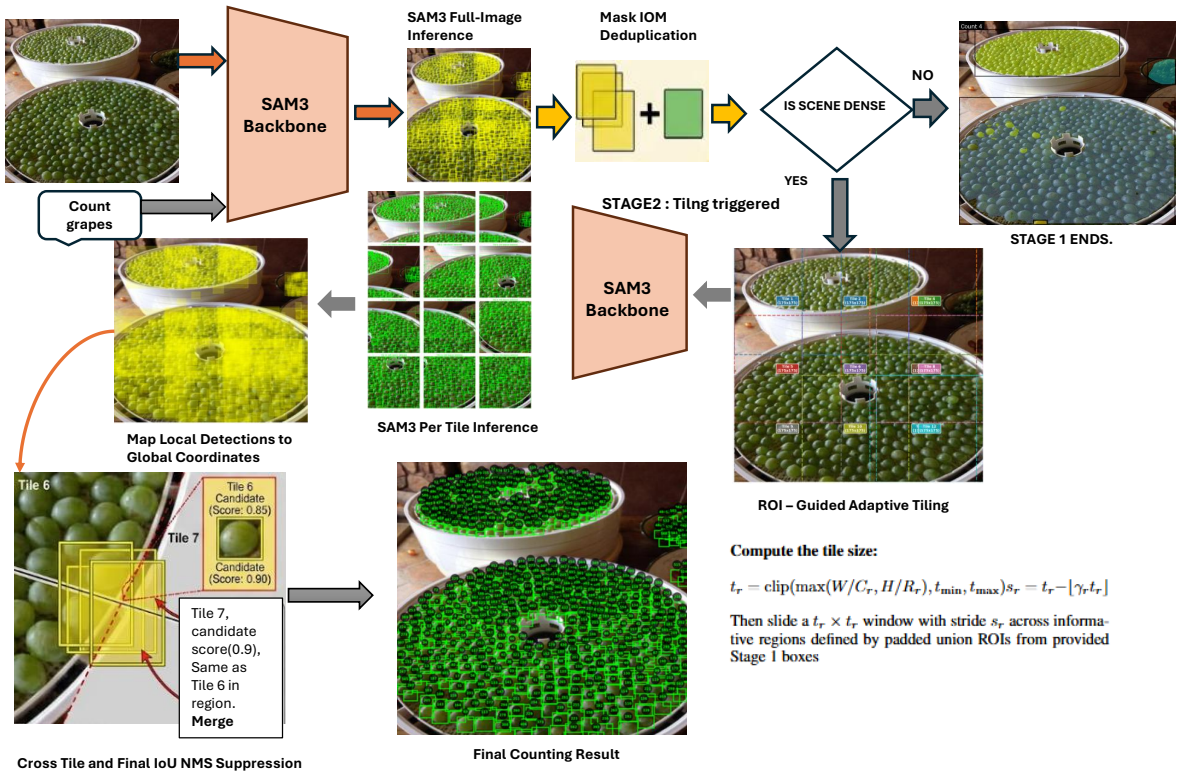

Figure 4. Overview of SAM3Count image counting pipeline. SAM3 initial pass produces initial instance masks, followed by IoM-based deduplication and filtering. Dense scenes activate Stage-2, where SAM3Count performs ROI-guided overlapping tiling, aggregates tile predictions, applies cross-tile suppression, and outputs the final count.

centers over time), and use a constant-velocity extrapolation for short-term gating during association (up to 50 frames).

Lost-track registry: When an object disappears (SAM3 stops tracking), we move it to a lost registry which remain recoverable for up to 50 frames:
$L[\text{id}] = \{\text{id} : s;\ \text{frame} : t;\ \text{feat} : \mathcal{A}(s);\ \text{box} : \text{bbox}(s)\}$

3.3. Multi-Modal Association

SAM3Count adapts and configures its association behavior to the dynamics of each scene. We achieve this by multimodal online association techniques that links SAM3's outputs to consistent identities, prevent false merges, re-identify temporarily lost objects, and merge track splits.

3.3.1. Pending Window for New IDs

To prevent spurious associations, new SAM3 IDs first enter a pending state for $T_{\text{wait}} = 3$ frames, i.e. $Q_{\text{pending}}[s] = \{\text{first frame} : t_0,\ \text{features} : [\phi_1, \phi_2, \phi_3],\ \text{boxes} : [b_1, b_2, b_3],\ \text{masks} : [M_1, M_2, M_3]\}$. This accumulates evidence before committing to ID assignment, reducing the noise from transient false detections and very short-lived segments.

3.3.2. Three-Stage Matching Protocol

After the pending window, we test candidate matches using our complementary modalities: **Appearance similarity:** Let ϕ_{new} be the descriptor of the candidate track and $\mathcal{A}(s')$ the appearance bank of an existing identity s'. We compute $\text{sim}_{\text{app}} = \max_{\phi' \in \mathcal{A}(s')} \cos(\phi_{\text{new}}, \phi')$ and require $\text{sim}_{\text{app}} > 0.92$. **Spatial proximity:** We reject matches that are visually similar but far apart within the frame. Let c_{new} and c_{exist} be the bounding-box centers of the candidate and the existing track. We require $|c_{\text{new}} - c_{\text{exist}}|2 < 50$ pixels. **Temporal consistency:** If the candidate and an existing track overlap in time, they should co-occur and follow similar trajectories. Let F_{pend} and F_{exist} be the sets of frames where the candidate and the existing track are visible. We require an overlap ratio $r_{\text{overlap}} = \frac{|F_{\text{pend}} \cap F_{\text{exist}}|}{|F_{\text{pend}} \cup F_{\text{exist}}|} > 0.6$ with at least two shared frames, and we also enforce that their centers remain close over the shared frames

3.3.3. Re-Identification of Lost Tracks

When a candidate does not match any active identity, SAM3Count attempts re-ID against L (lost registry). We use motion as a gating signal (the candidate must be near the predicted position) and appearance as the main sig-

nal. We keep one combined score (and tighten its spacing): $S_{\text{reid}} = 0.6, s_{\text{app}} + 0.4, s_{\text{mot}}$, where s_{app} is the max cosine similarity to the lost identity's appearance bank, and s_{mot} is a normalized motion-consistency score in $[0, 1]$. We accept re-identification if $s_{\text{app}} > 0.75$, $s_{\text{mot}} > 0.4$, and $S_{\text{reid}} > \theta$.

3.4. Offline Consolidation and Final Counting

As a lightweight post-processing step, we merge trajectories that represent the same object but were split by short gaps. These include merging (i) track pairs (T_1, T_2) with substantial mask overlap over shared frames using mask IoU similarity, and (ii) adjacent tracks T_1 and T_2 separated by a small temporal gap with end or start positions spatially close. Figure 3 demonstrates a visual representation of the SAM3Count model to track and count across video frames.

4. Experiments

We evaluate SAM3Count's zero -shot performance across a diverse set of image and video benchmarks spanning different domains, including UAV, open-world scenes, crowd counting, and web-based images.

4.1. Image benchmarks:

FSC-147 [32] is a standard open-world counting dataset with 147 classes and 6,135 images, covering a wide range of object appearances and densities. FSCD-147 [32] further enhances the validation and test splits of FSC-147 with bounding-box annotations. We use this benchmark to evaluate the counting in cluttered, and small-instance settings. **ShanghaiTech** [55] is a single-class crowd counting benchmark split into two parts. Part A has 182 test images, 66–2256 humans per image, and Part B contains 316 images, 9–539 humans per image (498 images; 9–2,256 people per image). **OmniCount** [31] is a multi-class counting benchmarks with multiple labels of different categories within an image. The Fruit split is used with 300 images of 8 different fruits across images. For web-scale counting, we follow [6] and test SAM3Count on **PixMo-Count** [11], which provides point annotations with a released split that contains 36.9k training examples and human-verified validation/test sets (540 images each). We also test our model on CountBench, which is an LAION-400M-curated evaluation benchmark, containing 540 images with counts between 2 and 10 [35]. **CARPK** is a domain-specific vehicle counting benchmark consisting of 1,448 aerial drone images of parking lots with cars annotated using tight bounding boxes, split into a training set of 989 images and a test set of 459 images. SAM3 is not optimized for tiny, tightly packed instances, so we fine-tune the detection head on **FSCD-147** train split and denote this variant as **SAM3Count (ft)** to better localize tiny objects in dense scenes. The training categories are disjoint from the validation and test categories of FSCD-147 and also differ from those in ShanghaiTech (*human only*). Thus, evaluation remains zero-shot since objects are still specified using only text prompts for unseen categories. We therefore report **SAM3Count (ft)** as a fine-tuned zero-shot open-vocabulary variant, distinct from the frozen training-free setting FSCD-147 and ShanghaiTech to characterize their effect in crowded scenes. For the remaining benchmarks, SAM3Count is evaluated in a zero-shot training-free setting. We report the finetuned results on dense benchmarks (FSCD-147 (val and test split) and ShanghaiTech) to characterize their effect in crowded scenes. For the remaining benchmarks, SAM3Count is evaluated in a training-free setting with **text prompts only**.

4.2. Video benchmarks:

We follow [2] and evaluate SAM3Count on two benchmarks; TAO-Count (357 videos; 139 categories; 8 - 10s lengtgh) built on TAO [10],and Penguins (3 short videos, 28–59 penguins per video) [2].

4.3. Metrics

For image counting, given N images with predicted counts $\hat{y}_i \in \mathbb{Z}_{\geq 0}$ and ground-truth counts y_i, we follow [6] and report exact-match Accuracy and MAE: $\text{Acc}(\%) = 100 \cdot \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = y_i]$ and $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$.

For video counting, given M videos with predicted totals per-video $\hat{Y}_j$ and ground-truth totals Y_j, we follow [2] and report MAE (as above) and RMSE: $\text{RMSE} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(\hat{Y}_j - Y_j)^2}$.

4.4. Quantitative results

FSCD-147: In Table 1, we test SAM3Count on FSCD-147 benchmark images, which range from moderate clutter to highly dense configurations with many small instances. SAM3Count improves SAM3 by mitigating the characteristic dense-scene failure mode. Fine-tuned **SAM3Count (ft)** further strengthens performance in the most crowded cases by adapting the detection head to smaller instances. This leads to SOTA RMSE scores in validation set for zero shot counting and SOTA overall results in the test set.

ShanghaiTech: In Table 2, we test SAM3Count in a denser crowd setting under a fixed prompt (*"human"*) across all images. SAM3Count shows better performance in both Parts A and B, with Part A containing the heaviest images of the crowded-scene. The addition of adaptive tiling techniques has improved the model's performance.

CARPK: In Table 3, we assess the zero-shot training-free approach of our model compared to other training-free methods. Most of these method are few-shot such that they use small number of manually labeled exemplars as prompt. Despite this, SAM3Count attains the SOTA results making it a competitive alternative to exemplar-guided approaches for class-agnostic counting.

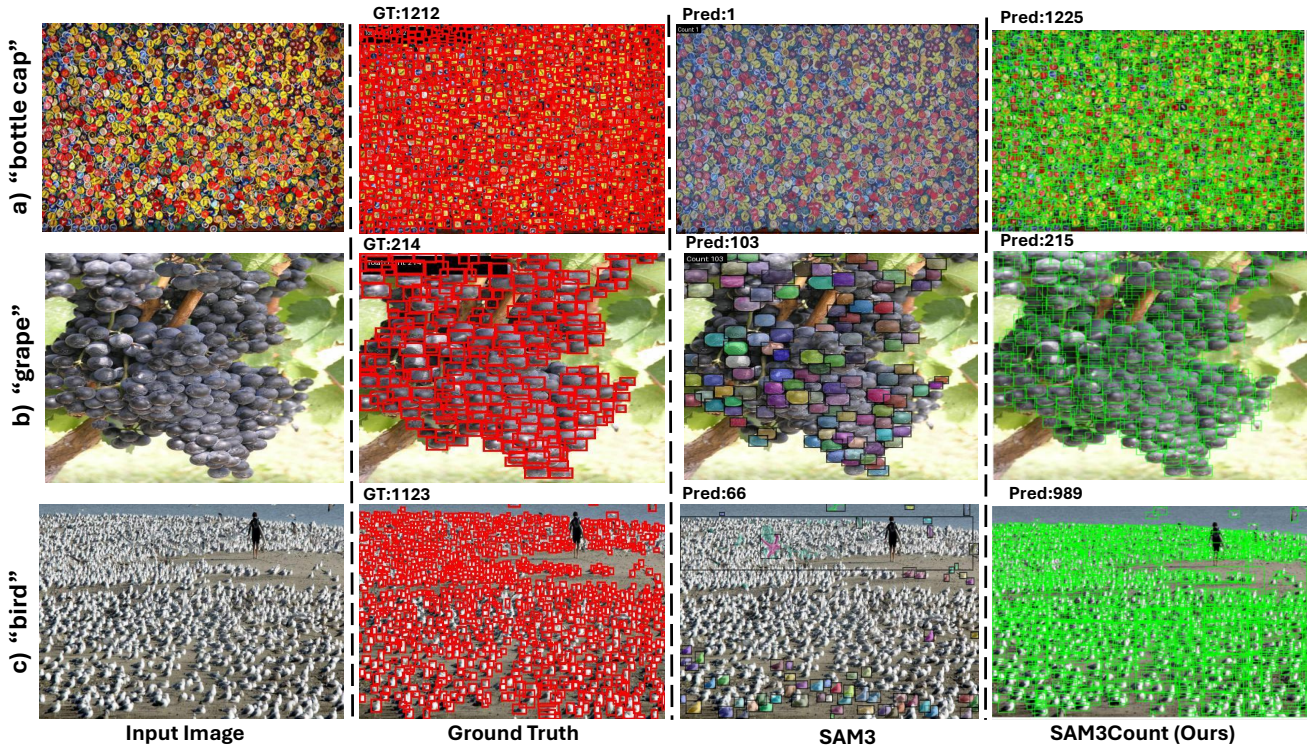

Figure 5. Zero-shot qualitative comparison of SAM3 and SAM3Count on four images from the FSCD-147 benchmark together with the ground truth (GT) count. Input images and their corresponding text prompts (shown on the left). SAM3Count closely matches the GT or slightly overcounts in denser scenes. SAM3 struggles to accurately count in denser crowds, even with lower confidence thresholds.

Table 1. Results on FSCD-147 (Val & Test) for open-world counting based on *text-prompt* only. CountingDINO is exemplar-based. Note: COUNTGD-BOX (CGD-B); Clip-Count(CLP-CT); Grounding DINO (GDINO)

Method	Validation		Test	
	MAE↓	RMSE↓	MAE↓	RMSE↓
OWL_{v2} (NeurIPS '23) [30]	49.02	131.41	41.83	149.82
GDINO (ECCV '24) [26]	54.45	137.12	54.16	157.87
DAVE (CVPR '24) [37]	15.48	152.57	15.52	103.42
PseCo (CVPR '24) [16]	23.90	100.33	16.58	129.77
CLP-CT (MM '23) [17]	18.79	61.18	17.78	106.62
CGD-B (AAAI '25) [2]	**12.24**	66.24	15.01	118.16
SAM3 (ArXiv '25) [6]	23.70	98.72	22.89	144.80
SAM3Count (Ours)	25.02	81.28	19.80	128.59
SAM3Count (ft) (Ours)	14.34	**59.04**	**13.02**	**86.42**

Table 2. Crowd counting results on ShanghaiTech [55] Part A and B text prompt (*"human"*) for all images. Note: COUNTGD-BOX (CGD-B); Clip-Count(CLP-CT); Grounding DINO (GDINO).

Method	Part A		Part B	
	MAE↓	RMSE↓	MAE↓	RMSE↓
OWLv2 (NeurIPS '23) [30]	420.20	553.30	81.50	126.50
G-DINO (ECCV '24) [26]	394.90	537.50	58.30	99.30
CLP-CT (MM '23) [17]	192.60	308.40	45.70	77.40
CGD-B (AAAI '25) [2]	132.20	**253.90**	32.20	57.90
SAM3 (ArXiv '25) [6]	376.73	526.43	52.82	91.94
SAM3Count (Ours)	213.75	376.40	36.43	56.97
SAM3Count (ft) (Ours)	**131.27**	262.82	**28.88**	**44.37**

PixMo & CountBench: Table 4 reports the results of PixMo-Count and CountBench, benchmarks commonly used to evaluate the counting performance of MLMs. SAM3Count improves over SAM3 and remains competitive with strong baselines in vision-language.

OmniCount: Table 5 reports the results for Multi-class counting in images on the Omnicount [31] benchmark which contains multi labels belong to two or more categories within an image. SAM3Count performs better than SOTA CountGD++ which uses text + exemplar prompts.

Video benchmarks: We extend SAM3Count to video counting with our innovative re-id tracker and multimodal matcher and report our results on the Penguins, a subset of the Science-Count dataset from [2], and TAO-Count. Compared with CountVid, SAM3Count performs better across the benchmarks, highlighting the importance of our re-identification and association tracker.

Table 3. Counting results on CARPK test split comparing exemplar-based training-free methods against our text-only training-free approach. CountingDINO is abbreviated as CDINO.

Method	Prompt	MAE↓	RMSE↓
TFOC (WACV '24) [42]	Exemplar	10.97	14.24
OmniCount (AAAI '25) [31]	Exemplar	9.92	12.15
TFCounter (ArXiv '24) [45]	Exemplar	9.71	12.44
$CDINO_{v1}$ (WACV '26) [34]	Exemplar	36.57	47.31
$CDINO_{v2}$ (WACV '26) [34]	Exemplar	21.26	28.20
OCCAM (ArXiv '26) [44]	-	10.06	13.81
SAM3 (ArXiv '25) [6]	Text	3.526	6.571
SAM3Count (Ours)	Text	**3.11**	**5.60**

Table 4. Web-image counting results on PixMo-Test and CountBench. SAM3Count improves over the SAM3 baseline in Pixmo while remaining competitive with other MLM baselines.

Method	PixMo_Test		CountBench	
	MAE↓	Acc↑	MAE↓	Acc↑
DINO-X (ArXiv '24) [39]	0.21	85.0	0.62	82.9
Qwen2 (ArXiv '24) [48]	0.61	63.7	0.28	86.7
Gemini 2.5 (ArXiv '25) [7]	0.38	78.2	0.24	92.4
SAM3 (ArXiv '25) [6]	0.18	87.48	**0.23**	93.23
SAM3Count (Ours)	**0.17**	**89.18**	**0.23**	**93.43**

Table 5. Multi-class counting results on OmniCount [31] (Fruits) split. SAM3Count is text prompt only but the other models are text + exemplars.

Method	MAE↓	RMSE↓
CountGD (NeurIPS '24) [3]	2.76	3.11
CGD-B (AAAI '25) [2]	2.83	3.15
CountGD++ (CVPR '26) [1]	0.56	1.24
SAM3Count (Ours)	**0.43**	**0.93**

Table 6. Video counting results on video benchmarks from [2].

Method	TAO-Count		Penguins	
	MAE↓	RMSE↓	MAE↓	RMSE↓
CountVid [2]	2.60	6.00	4.00	5.30
SAM3Count (Ours)	**0.78**	**1.63**	**2.3**	**3.1**

4.5. Qualitative results

Figure 2 highlights a common failure mode in video counting: inconsistent identity assignment between frames. In this example (Frame 57 and Frame 98), We compared SAM3Count to SAM3 [6] and CountVid [2]. [2] produces temporally unstable instance outputs that can manifest as duplicate counting when objects re-enter the scene or overlap under occlusion. [6] exhibits identity-switching behavior, where existing track IDs are reassigned to different vehicles in later frames (frame 98), indicating a weak long-term association. In contrast, SAM3Count maintains more consistent object identities across the two frames, thus reducing double-counting and preventing track ID reassignment when vehicles move, partially occlude, or undergo scale changes. From figure 5 which illustrates different density levels in images, SAM3Count is noticeably more reliable than [6] once the scene becomes dense. In sparser scenes like "packed fruit" (image, topmost left), both methods behave comparably: localize the target category and can capture meaningful object regions (including whole/part cues). The difference becomes clear in crowded scenes. For example, "bottle cap" or "crowded birds" image. [6] collapses and severely undercounts even under relaxed thresholds. SAM3Count remains substantially closer to the ground truth by recovering instance-level structure through density-triggered ROI tiling.

Table 7. Ablation on the density score threshold.

Density	MAE↓	RMSE↓
6	15.13	59.40
7	14.34	59.04
8	14.43	84.30

Table 8. Ablation on ROI and tiling components.

Setting	MAE↓	RMSE↓
Neither	25.02	81.28
Tiling w/o ROI	16.28	61.71
ROI w/o Tiling	22.40	105.33
Both	14.34	59.04

4.6. Ablation Studies: Density Score, ROI & Tiling

We ablate the density score threshold used to trigger tiling and assess the effects of the ROI and Tiling modules on overall counting performance in SAM3Count on the Val split of the FSCD-147 benchmark. As shown in Table 7, a score threshold of "7" yields the best results, with lower MAE/RMSE. Table 8 also shows the results of the effect of each component on counting performance, with "Both" performing better. "Neither" option in this case is better than "ROI without Tiling", because ROI selects an area to tile and enhance localization when tiling is used.

5. Conclusion

We presented **SAM3Count**, a zero-shot open-vocabulary counting framework for images and videos, built on SAM3, with a re-ID tracker for identity preservation and ROI tiling to handle dense images. Our experiments show improved accuracy (MAE/RMSE) across image and video benchmarks, achieving state-of-the-art results on five image and two video benchmarks. SAM3Count still depends on SAM3 mask quality and may degrade under severe blur, occlusion, heavy overlap, or extreme scale variation. In videos, prolonged occlusions can still cause duplicate counting after the lost track window expires, and the current system relies on fixed hand-crafted mode-selection rules. Future work will explore more robust scene-adaptive calibration and lightweight learned refinements that preserve open-vocabulary zero-shot generality while improving reliability in dense and low-visibility settings.

References

[1] N. Amini-Naieni and A. Zisserman. Countgd++: Generalized prompting for open-world counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 8

[2] N. Amini-Naieni and A. Zisserman. Open-world object counting in videos. In *Association for Advancement of Artificial Intelligence Conference (AAAI)*, 2026. 2, 3, 6, 7, 8

[3] Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. Countgd: Multi-modal open-world counting. *Advances in Neural Information Processing Systems*, 37:48810–48837, 2024. 1, 3, 8

[4] Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Interactive object counting. In *European conference on computer vision*, pages 504–518. Springer, 2014. 1, 2

[5] Jahongir Azimjonov, Ahmet Özmen, and Metin Varan. A vision-based real-time traffic flow monitoring system for road intersections. *Multimedia Tools and Applications*, pages 1–20, 2023. 1

[6] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, François Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. 2, 3, 6, 7, 8

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 8

[8] Meng Cui, Xubo Liu, Haohe Liu, Jinzheng Zhao, Daoliang Li, and Wenwu Wang. Fish tracking, counting, and behaviour analysis in digital aquaculture: a comprehensive survey. *Reviews in Aquaculture*, 17(1):e13001, 2025. 1, 3

[9] Zhe Dai, Huansheng Song, Xuan Wang, Yong Fang, Xu Yun, Zhao Yang Zhang, and Huaiyu Li. Video-based vehicle counting framework. *IEEE Access*, 7:64460–64470, 2019. 3

[10] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *Computer Vision – ECCV 2020*, pages 436–454, 2020. 6

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104, 2025. 6

[12] Zhenwei Dong, Xinyi Liu, Weiyang Shi, Yuheng Lu, Yanyan Liu, Xiaoxiao Hou, Hongji Sun, Ming Song, Zhengyi Yang, and Tianzi Jiang. Neuron counting for macaque mesoscopic brain connectivity research. *IEEE Transactions on Medical Imaging*, 2025. 1

[13] Min Fu, Pei Xu, Xudong Li, Qihe Liu, Mao Ye, and Ce Zhu. Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 43:81–88, 2015. 2

[14] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. A survey of deep learning methods for density estimation and crowd counting. *Vicinagearth*, 2(1):1–37, 2025. 1

[15] Maryam Hassan, Farhan Hussain, Sultan Daud Khan, Mohib Ullah, Mudassar Yamin, and Habib Ullah. Crowd counting using deep learning based head detection. *Electronic Imaging*, 35(9):293–1–293–6, 2023. 2

[16] Zhizhong Huang, Mingliang Dai, Yi Zhang, Junping Zhang, and Hongming Shan. Point segment and count: A generalized framework for object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17067–17076, 2024. 1, 3, 7

[17] Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 4535–4545, 2023. 1, 3, 7

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 3

[19] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *Proceedings of the european conference on computer vision (ECCV)*, pages 547–562, 2018. 2

[20] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. *Advances in neural information processing systems*, 23, 2010. 1, 2

[21] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *2008 19th international conference on pattern recognition*, pages 1–4. IEEE, 2008. 2

[22] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by

segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18973, 2024. 1, 3

[23] 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)*, pages 1091–1100, 2018. 2

[24] Dongze Lian, Xianing Chen, Jing Li, Weixin Luo, and Shenghua Gao. Locating and counting heads in crowds with a depth prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9056–9072, 2022. 2

[25] Yuhao Lin, Haiming Xu, Lingqiao Liu, and Javen Qinfeng Shi. A simple-but-effective baseline for training-free class-agnostic counting. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8155–8164. IEEE, 2025. 1

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pages 38–55. Springer, 2024. 1, 2, 7

[27] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5094–5103, 2019. 2

[28] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6462–6471, 2019. 2

[29] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *Computer Vision – ECCV 2022*, 2022. 2

[30] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023. 7

[31] Anindya Mondal, Sauradip Nag, Xiatian Zhu, and Anjan Dutta. Omnicount: Multi-label object counting with semantic-geometric priors. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 19537–19545, 2025. 6, 7, 8

[32] Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. In *European Conference on Computer Vision (ECCV)*, pages 348–365. Springer, 2022. 6

[33] Naoya Noguchi, Hideaki Nishizawa, Taro Shimizu, Junichi Okuyama, Shohei Kobayashi, Kazuyuki Tokuda, Hideyuki Tanaka, and Satomi Kondo. Efficient wildlife monitoring: Deep learning-based detection and counting of green turtles in coastal areas. *Ecological Informatics*, 86:103009, 2025. 1

[34] Giacomo Pacini, Lorenzo Bianchi, Luca Ciampi, Nicola Messina, Giuseppe Amato, and Fabrizio Falchi. Countingdino: A training-free pipeline for class-agnostic counting using unsupervised backbones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 806–815, 2026. 3, 8

[35] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3170–3180, 2023. 6

[36] Muhammed Patel, Javier Noa Turnes, Jayden Hsiao, Linlin Xu, and David Clausi. Openwildlife: Open-vocabulary multi-species wildlife detector for geographically-diverse aerial imagery. *arXiv preprint arXiv:2506.19204*, 2025. 1

[37] Jer Pelhan, Vitjan Zavrtanik, Matej Kristan, 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 (CVPR)*, pages 23293–23302, 2024. 7

[38] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[39] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024. 8

[40] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 1, 2

[41] Sami Abdulla Mohsen Saleh, Shahrel Azmin Suandi, and Haidi Ibrahim. Recent survey on crowd density estimation and counting for visual surveillance. *Engineering Applications of Artificial Intelligence*, 41:103–114, 2015. 1

[42] Zenglin Shi, Ying Sun, and Mengmi Zhang. Training-free object counting with prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 323–331, 2024. 8

[43] Ziqiang Shi, Rujie Liu, Jun Takahashi, and Shan Jiang. Truecount: Improving open-world object counting with visual-language models and dynamic multi-modal inputs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 1764–1773, New York, NY, USA, 2025. Association for Computing Machinery. 3

[44] Michail Spanakis, Iason Oikonomidis, and Antonis Argyros. Occam: Class-agnostic, training-free, prior-free and multi-class object counting. *arXiv preprint arXiv:2601.13871*, 2026. 3, 8

[45] Pan Ting, Jianfeng Lin, Wenhao Yu, Wenlong Zhang, Xiaoying Chen, Jinlu Zhang, and Binqiang Huang. Tfcounter: Pol-

ishing gems for training-free object counting. *arXiv preprint arXiv:2405.02301*, 2024. 3, 8

[46] Tam Vu, Hong Nam Thai, Viet Ngoc Pham, Huy Tuan Vu, Anh Tuan Luong, and Thien Van Luong. Counting mixed traffic volumes at motorcycle-dominated intersections by using computer vision. *International Journal of Intelligent Transportation Systems Research*, 23(1):146–164, 2025. 1

[47] Guangxu Wang, Jiaxuan Yu, Wenkai Xu, Akhter Muhammad, and Daoliang Li. Automated fish counting system based on instance segmentation in aquaculture. *Expert Systems with Applications*, 259:125318, 2025. 1

[48] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8

[49] Zhicheng Wang, Liwen Xiao, Zhiguo Cao, and Hao Lu. Vision transformer off-the-shelf: A surprising baseline for few-shot class-agnostic counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5802–5810, 2024. 3

[50] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7810–7819, 2021. 1, 3

[51] Jingyi Xu, Hieu Le, and Dimitris Samaras. Zero-shot object counting with language-vision models. *arXiv preprint arXiv:2309.13097*, 2023. 1, 3

[52] Fanfan Ye, Yiqi Fan, Qiaoyong Zhong, Shicai Yang, Di Xie, Jie Song, and Mingli Song. VQCounter: Designing visual prompt queue for accurate open-world counting. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pages 2260–2268. IJCAI, 2025. 1, 3

[53] Ying Yu, Feng Zhu, Jin Qian, Hamido Fujita, Jiamao Yu, Kangli Zeng, and Enhong Chen. Crowdfpn: crowd counting via scale-enhanced and location-aware feature pyramid network. *Applied Intelligence*, 55(5):359, 2025. 1

[54] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. In *European Conference on Computer Vision*, pages 419–437. Springer, 2024. 1, 2

[55] 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 (CVPR)*, pages 589–597, 2016. 2, 6, 7

[56] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Computer Vision – ECCV 2022*, pages 1–21. Springer, 2022. 1, 3

[57] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision – ECCV 2022*, 2022. 2