A Class-Agnostic Object Counting Model Based on GeCo

Kun Qian, Qiran Li, Guang Yang
The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
{kqianae, qlidl, gyangaw}@connect.ust.hk

Abstract

This project introduces a method of Class-Agnostic Counting (CAC), based on a novel model of low-shot counting GeCo, which uses detection and segmentation. In this paper, we first experiment the baseline followed by a refined one with adding 2 designed loss functions. Our model outperforms the baseline in validation MAE and have higher precisions AP/AP50. And the sensitivity of RMSE to outliers is analyzed.

1. Introduction

Class-Agnostic Counting (CAC) is a computer vision task that counts instances from unseen classes via few-shot exemplars. First proposed in 2019 by Lu et al. [9], CAC is a vital for practical scenarios requiring adaptability, such as inventory or wildlife monitoring. CAC is a shift from class-specific counting tasks to generalized object counting in computer vision, handling the limitations of supervised methods requiring extensive labeled data per category.

Yet difficulties also arise from diverse object appearances, causing overgeneralization, occlusions or scales. These lead to false detection, and mixed-class interference inflating counts, where traditional prototypes struggle without direct optimization.

2. Related works

Most popular works emphasize the two primary paradigms in few-shot CAC tasks: density map-based and detection-based methods. The former approach predicts a per-pixel density distribution over the image and then sums values to estimate global counts, offering robustness but lacking object conceptual awareness. The latter, detection-based methods, locates individual objects via bounding boxes or segmentation masks. It gets counts from detection while providing explainable outputs. But often fails in count accuracy due to variations.

Despite progress, CAC faces ongoing hurdles. Low-shot methods arose with the FSC-147 dataset [15], predicting global counts by summing density maps through adaptations like tracking backbones for regression. However, FSC-147's limited 6,135 images across 147 categories, mostly single-class, fail to capture real-world multi-category complexity [1]. Methods underperform on occlusions, scale variations, and backgrounds; Gong et al. [3] highlight intra-image differences in color, shape, and scale as key impediments. Density-based techniques like LOCA [2] use exemplar-guided similarity or prototypes for generalization but over-activate in mixed scenes, inflating counts via false positives. CounTR [7] uses a vision transformer to extract the feature info from images, and a convolutional network to encode the exemplar features.

Detection-based CAC, as in DAVE [13], employs prototypes from exemplars for correlation and localization. It estimates counts from detection. In our project, we adopt GeCo [14], a state-of-the-art extension outperforming DAVE through segmentation integration, dense object queries for robust generalization, and a novel counting loss directly optimizing detection to avoid surrogate issues.

3. Project plan

Our project aims to develop a robust few-shot counting system, focusing on improving the accuracy and robustness of existing methods. Building upon research of advanced methods, we designed a counting model based on GeCo, reconstructing a sub-model and the loss function to make it more capable of effectively handling object occlusion, scale variations, and complex backgrounds. Then evaluation on standard benchmarks is conducted to contribute to the practical application.

3.1. Responsibilities of team members

In our project, we approach correlated tasks. Our name and corresponding tasks are:

Qiran Li – Model deployment, Finetuning, Model architecture redesigning, Loss function designing, Report modifying, Project presentation Q&A
Kun Qian – Experiment comparison, Output analysis, Report drafting, Project presentation
Guang Yang – Example visualizing, Slide deck make, Project presentation

Table 1. Timeline with milestones

Date	Milestone
Oct 19	Project selection and proposal
Nov 2	Baseline deployment and reproduction
Nov 16	Modified model finetuning & test
Nov 21	A full analysis of outputs & Technical report
Nov 26	Slide deck making
Nov 29	Presentation

3.2. Key modules

1. Unified Architecture: GeCo [14] uses integrated object detection, segmentation, and counting in a single framework, based on Segment Anything Model (SAM) [4] backbone.

2. Prototype Generalization: It uses a dense object query approach with iterative cross-attention to adapt prototypes across diverse object appearances.

3. Novel Loss Function: A new dense detection loss directly optimizes for reliable detections by classifying predictions as true positives, false positives, and false negatives via Hungarian matching, avoiding surrogate losses like unit Gaussians that are sensitive to annotations and hyperparameters.

4. Redesign and Retrain: We reconstruct a sub-model, introduce 2 loss functions, and retrain it to achieve a better performance.

3.3. Timeline with milestones

We have set up a detailed pipeline and achieved the milestones of our task, as shown in Table 1.

4. Methodology and technical approach

We choose CeCo [14] as our baseline model. This model already has a strong ability to detect objects in low-shot scenarios by integrated detection and segmentation. It alleviates key problems in CAC, such as incorrect partial detections [13] or failure to locate objects in dense situation [11] (see Figure 1, the second row). GeCo uses the backbone of SAM (Segment Anything Model) [4] to extract features on input image. We fine-tune this model to enhance robustness against cover, scaling, and complex backgrounds.

Based on this model, we have 2 candidate redesigns: One is reconstructing a sub-model with prototype iteration formula, and the other is adding 2 loss functions to the total.

4.1. Baseline: GeCo architecture

The first step is to process an input image $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$ and a small set of exemplar bounding boxes $B^E = \{b_i\}_{i=1:k}$ (typically $k = 1, 2, 3$) to predict object locations, counts, and segmentation masks. In the zero-shot scenario, no typical

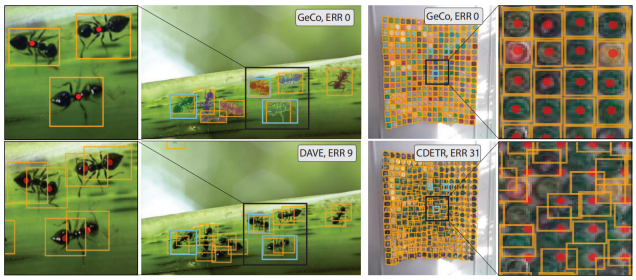


Figure 1. DAVE's [13] predictions have incorrect partial detections of ants (bottom left), while CDETR [11] fails in densely populated regions (bottom right). GeCo addresses these problems.

training data are needed. Instead, this CAC task can be easily down by its upstream pretraining of object detection tasks. In the Fine-tuning stage, the bounding boxes just tell the model to count the objects appearing like these. For example, if exemplar boxes include some colorful balls, given an image containing a pool of all colorful balls and toys. Then the ground truth is exactly the amount of these balls altogether, while fully neglecting the differences among these balls (e.g. color or scale).

Image features are first extracted using the SAM [4] backbone, then yield a feature map $f^I \in \mathbb{R}^{h \times w \times d}$ where $h = H_0/16$, $w = W_0/16$, and $d = 256$. Then generate some prototype to represent the target categories (exactly the positive "labels"). In few-shot, this appearance prototypes p^A are generated through Region of Interest (RoI) pooling from f^I within the exemplars, and shape prototypes p^S are calculated using a Multi Layer Perceptron (MLP) on exemplar dimensions (width and height). Then the prototypes are concatenated to $p \in \mathbb{R}^{2k \times d}$. In zero-shot, a single prototype p is generated by combining a learned objectness prototype p^Z to f^I: $p = \text{CA}(p^Z, f^I, f^I)$ using Cross Attention (CA).

Then, they employ a Dense Query Encoder (DQE). This sub-model is to generalize prototypes across different appearances, but will not be affected by distracting objects. Following the initialization formula $P_0 = f^I$, we adapt the prototypes through CA: $P_i = \text{CA}(P_{i-1}, p, p)$ for $i = 1$ to 3 iterations. This creates image-wide, non-parametric generalizations. Dense queries Q are then formed by further refining with self-attention on f^I: $Q_j = \text{CA}(\text{SA}(f^I), Q_{j-1}, Q_{j-1})$ for $j = 1$ to 2 with $Q_0 = P_3$. Then add a positional encoding to improve the spatial sense. This produces location-specific queries that restrict false positives in dense or multi-class environments.

The Dense Query Decoder (DQD) translates these queries into high-resolution predictions. We upsample the queries $Q^{HR} \in \mathbb{R}^{H \times W \times d}$ ($H = H_0/2$, $W = W_0/2$) through three convolutional blocks (3x3 conv, Leaky ReLU, bilinear upsampling). In the middle we incorporate SAM-HQ features to perform better on small objects. Objectness scores are

predicted as $y^O = \text{LeakyReLU}(W_O \cdot Q^{HR})$, and bounding boxes as $y^{BB} = \sigma(\text{MLP}(Q^{HR})) \in \mathbb{R}^{H \times W \times 4}$.

The loss function of this model is also strong, especially in dense detection. It calculates the detection by using non-maximum suppression (NMS) on y^O maxima, reading corresponding boxes from y^{BB}, and refining with SAM decoder prompts. The final count is the number of refined detections after IoU-based NMS (threshold 0.5). This baseline outperforms density-based methods like LOCA [2] and detection-based like DAVE [13] by providing explainable outputs (locations and masks). It still has low mean absolute error (MAE) in counts.

The base loss $\mathcal{L}_{\text{base}}$ optimizes via Hungarian matching [5]. It predicts maxima to ground-truth centers post forward NMS, and keep the points above median objectness to avoid redundancy:

$$\mathcal{L}_{\text{base}} = -\sum_{i \in \text{TP}} \text{gIoU}(y_i^{BB}, B^{GT}_{\text{HUN}(i)}) + \sum_{i \in \text{TP} \cup \text{FN}} (y_i^O - 1)^2 + \sum_{i \in \text{FP}} (y_i^O - 0)^2, \quad (1)$$

where y_i^{BB} represents the i-th bounding box and $B^{GT}_{\text{HUN}(i)}$ is the ground truth one using Hungarian NMS. y_i^O is the i-th objectness score. TP, FP, and FN denote true positives, false positives, and false negatives, and gIoU is the generalized IoU.

4.2. Improvements and redesigns

Initially we have 2 candidate improvements for GeCo to enhance its performance. We introduce a model-level and a training-level modifications inspired by ablation insights from the original work and related few-shot detection findings [1]. These changes aim to boost precision in densely populated, mixed-class images, and covered or darkened samples.

Multiple Scale Fusion

Firstly, we intended to enhance the prototype generalization in DQE by incorporating a multi-scale attention mechanism in the beginning. We tried to extract features from multiple SAM resolutions using $r = 8$, 16, and 32. We reconstructed the iterative formula during CA step: $P_i = \text{CA}(P_{i-1} + \text{MSF}(f^I), p, p)$, where MSF (Multiple Scale Fusion) block is an added architecture, which concatenates the ROI pooling results of feature maps $f^i \in \{f_r^I\}_{r=8,16,32}$ under different resolutions from SAM. Then it's passed to 1x1 conv to dimension d) and transformed into MSF(f^I). However, after thinking twice and researching for more information, the model's current architecture is robust enough. Eventually this would fiercely undermine the model score, so we aborted this design while finetuning.

Loss function

So we consider the second, refining the loss function to better handle some special scenarios such as partial cover or darkening. Although our existing loss functions are robust enough, it is a little bit conservative in recalling a few true positives.

To improve the detection quality in CAC, we refine the training objective by integrating two advanced loss functions targeting the objectness scores y^O. We use IoU Aware Loss (IAL) and Focal Centering Loss (FCL) inspired by VarifocalNet [20] and FLDOD [6].

Our first loss function IAL correlates objectness predictions with bounding box quality to reduce preference for poorly localized detections. For matched pairs from Hungarian indices, it computes pairwise IoUs between predicted boxes y^{BB} and targets, extracting the diagonal IoU values. Then the method aligns predicted objectness (centers) at corresponding locations, and then pass them to these IoUs through mean squared error (MSE):

$$\text{IAL} = \mathcal{L}_{\text{iou-aware}} = \frac{1}{N} \sum (\hat{c} - \text{IoU}(y^{BB}, B^{GT}))^2, \quad (2)$$

where $\hat{c}$ is the predicted centers for matched indices, and N is the number of matches. If no matches exist, the loss returns zero. This ensures higher objectness scores for accurate localizations, suppressing noise in diverse CAC scenes like those with scale variations or occlusions.

Complementing this, our second loss function FCL addresses imbalance in objectness regression, where positives (centers) are sparse. Applied to masked predictions (where mask is larger than 0), it adapts a binary focal loss to focus less on easy negatives:

$$\text{FCL} = \mathcal{L}_{\text{centerness-focal}} = -\alpha(1 - p_t)^\gamma \log(p_t), \quad (3)$$

with binary cross-entropy (BCE) adjusted by $p_t = e^{(-\text{BCE}(\hat{c}, c_{\text{gt}}))}$, $\alpha = 0.25$, and $\gamma = 2$. Here, $\hat{c}$ and c_{gt} are predicted and ground-truth centers on valid masks. If no valid predictions, it returns zero. This focuses optimization on hard examples, such as small or varying-scale objects, enhancing robustness without surrogate Gaussians.

These losses we have introduced are added to the total $\mathcal{L} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{iou-aware}} + \mathcal{L}_{\text{centerness-focal}}$ and balanced empirically during fine-tuning.

5. Experiments

5.1. Dataset and metrics

We evaluate our baseline on the FSCD-147 dataset [11]. It provides bounding box annotations for all objects across 6,135 images spanning 147 diverse categories. This dataset enables comprehensive assessment of low-shot counting methods, with splits into training (3,659 images), validation (1,286 images), and test (1,190 images) sets. We use 3 exemplars per target category to predict object counts and locations. The global count estimation metrics are:

CVPR #99999

Mean Absolute Error (MAE) quantifies the average miscount in predicted object counts from actual annotations across dataset:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|C_i - \hat{C}_i|, \tag{4}$$

where C_i is ground truth counting result and $\hat{C}_i$ is predicted.

Root Mean Square Error (RMSE) complements MAE for comprehensive error analysis:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i - \hat{C}_i)^2}. \tag{5}$$

Average Precision (AP) integrates precision over recall, providing a single score that balances detection completeness and accuracy across confidence thresholds:

$$\text{AP} = \sum_{k}(r_k - r_{k-1})p_{\text{interp}}(r_k), \tag{6}$$

where k represents the index of the set of different IoU thresholds, r_k means the recall under k-th threshold and is ranked increasingly. $p_{\text{interp}}(r_k)$ is the interpolated precision at recall r_k. This calculates a smoothed curve area, averaged over IoU thresholds in COCO-style [16].

AP50 is a variant of AP computed solely for detections with $IoU \geq 0.5$. It stresses practical localization where moderate overlap suffices.

5.2. Train strategies

The initial training for this model follows a two-phase approach: pretraining with Gaussian loss for initialization, then training for 200 epochs using the dense detection loss, with AdamW optimizer. Augmentations include padding images to a fixed size . Due to computer limitation, our replication of the model undergoes merely the evaluating stage though validation and test set, using original trained GeCo. And our redesign contains finetuning 30 epochs with a batch size of 4 on the training set before reevaluating. We run these tasks on similar two A100 GPUs for about 6 hours per tuning round.

6. Result and analysis

6.1. Comparison and analysis

Two experiments are conducted, replication from GeCo [14] as the baseline, and the redesigned model, with two more losses.

Baseline replication

We reproduced the original GeCo on the same dataset. The results align closely with the original metrics, with only marginal differences such as RMSE attributable to random seeds.

Table 2 shows comparisons (partially from GeCo [14]). Our results confirm the replication without undermining the overall performance. The baseline GeCo-base outperforms CDETR [11] and SAMC [10] outstandingly in MAE and RMSE, and exceeds detection-based DAVE [13] by nearly 21% and density-based LOCA [2] by nearly 16% in AP, with explainable outputs.

Redesigned model

Based on upon the baseline, we introduced the IAL and FCL into the criterion. The same training parameters are used as GeCo-base. The result is displayed in the last row (GeCo-LF) of Table 2. On the validation set, our refined model has a novel performance of AP/AP50 (33.74%/63.13%), outperforming the baseline and origin a bit. And the MAE (9.50) beats the origin (9.52) by 0.02. While on the test set, MAE/RMSE underperforms the baseline and origin obviously, especially RMSE, from nearly 55% to 66.29%. AP50 gets improved while AP decreases marginally.

Our refined model, GeCo-LF has nuanced metric variations compared to the baseline (as shown in Table 2).

On validation, MAE improves marginally to 9.50 (1.96% over baseline, near original), since the IAL aligns objectness scores with bounding box quality using MSE on matched IoUs, reducing the average count deviations by suppressing low-quality false positives, while FCL mitigates imbalance by downweighting easy negatives. It focuses on hard centers for precise low-shot generalization. However, RMSE rises slightly to 44.98 (2.04% over rerun). It is probably because the hard-example emphasis of focal squares larger errors in diverse appearances, while GeCo's direct optimization can balance uniformity. AP and AP50 advance to 33.74 (0.36% gain) and 63.13 (0.96%), benefiting from localization refinement of IoU-aware.

On test, MAE reaches up to 7.97 (0.89% worse than rerun). RMSE surges to 66.29 (15.5% increase). This is because, focal may ignore borderline detection in occluded or scaled test images. Then the squared errors increase hugely, unlike GeCo's robust prototype generalization. AP dips minimally to 43.57 (0.14% drop), as IoU-balanced losses can trade robustness for accuracy in single-stage setups [17], but AP50 improves to 75.62 (0.73%). This tells that, stronger 0.5 IoU detection has a relationship with score-IoU correlation, but the recall is sacrificed in extreme context.

6.2. Investigation of Outliers

Different from MAE taking every error as equal, RMSE amplifies the influence of the samples with high errors, so that RMSE is sensitive to outliers. In our experiments, the outliers are samples with extremely high error. As shown in Figure 2, while MAE equals to 7.9, GeCo-base's error on this sample is 1670.

Table 2. Our replication (GeCo-base) and redesigned model (GeCo-LF, loss-finetuned) compared with GeCo (original paper) and other SOTA methods referred above. Best metrics are bold.

Method	Validation set				Test set			
	MAE↓	RMSE↓	AP↑	AP_{50} ↑	MAE↓	RMSE↓	AP↑	AP_{50} ↑
CDETR [11]	20.38	82.45	17.27	41.90	16.79	123.56	22.66	50.57
SAMC [10]	31.20	100.83	20.08	39.02	27.97	131.24	27.99	49.17
DAVE [13]	9.75	**40.30**	24.20	61.08	10.45	74.51	26.81	62.82
GeCo [14]	9.52	43.00	33.51	62.51	7.91	**54.28**	43.42	75.06
GeCo-base	9.69	44.08	33.62	62.53	**7.90**	57.38	**43.63**	75.08
GeCo-LF	**9.50**	44.98	**33.74**	**63.13**	7.97	66.29	43.57	**75.62**

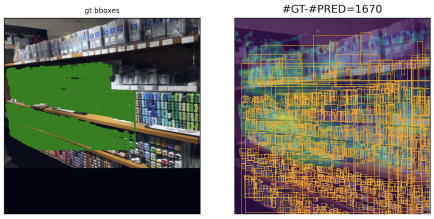


Figure 2. A sample in test dataset with extremely high error.

In order to investigate the influence of the outliers to the metrics, we recompute the MAE and RMSE based on 99% of the datasets by removing the top 1% samples with the highest absolute errors. The comparison of the MAE and RMSE before and after removing are shown in Table 3. Removing 1% of the data causes about 30-40% drops in MAE and about 60-70% drops in RMSE, especially a 85.07% drop in RMSE of our model on test dataset. This results indicates that in some circumstances only 1% of data (about 12 samples) can contributes almost half of the metrics. After the modification, our model outperform the baseline model by both MAE and RMSE on validation and test datasets. So the main weakness of our model lies in handling extreme corner cases. This experiment also exposes the drawbacks of the main metrics in CAC: MAE and RMSE are both sensitive of high error value.

6.3. Samples with visualization

This part shows some visualization of the counting results of the baseline and the redesigned model, as 6 occasions shown in Figure 3.

Color variation and object overlap. On this occasion (see in Figure 3(a)), the baseline fails to count purple sheets and misses overlapping red ones (GT-PRED=4). Because the surrogate loss predicts Gaussians for training [14], indirect for detection task in the baseline. It would cause false negatives in compact structures and eventually undervalue overlapped or colorful objects. Our refined model almost detects all (GT-PRED=1). It benefits from IAL, which aligns centers of colorful objects to localization quality and reduces low-IoU fake positives.

Angle rotation. In Figure 3(b), baseline fails to count the chairs with their angles extremely rotated (GT-PRED=22). It yields fake negatives from suboptimal centering. Our model captures more (GT-PRED=17) chairs, due to FCL wise detection amid fake negatives and boost recal. While IAL ensures scores reflect accurate boxes together with dense queries for prototype generalizing.

Darkening and blurring. In Figure 3(c), baseline missed all blurred and darkened pills under plastic (GT-PRED=11). Because Its surrogate Gaussian loss failed to placed at this pill's center during training, and then optimized local maxima in correlation maps indirectly rather than the detection task itself, making it highly sensitive to annotation errors. Plus, hyper parameters like Gaussian sigma, and low-contrast variations can disrupt precise center predictions. Our model detects a blurred pill (GT-PRED=10). Because FCL downweights easy backgrounds to focus gradients on blurred centers and improve the recall.

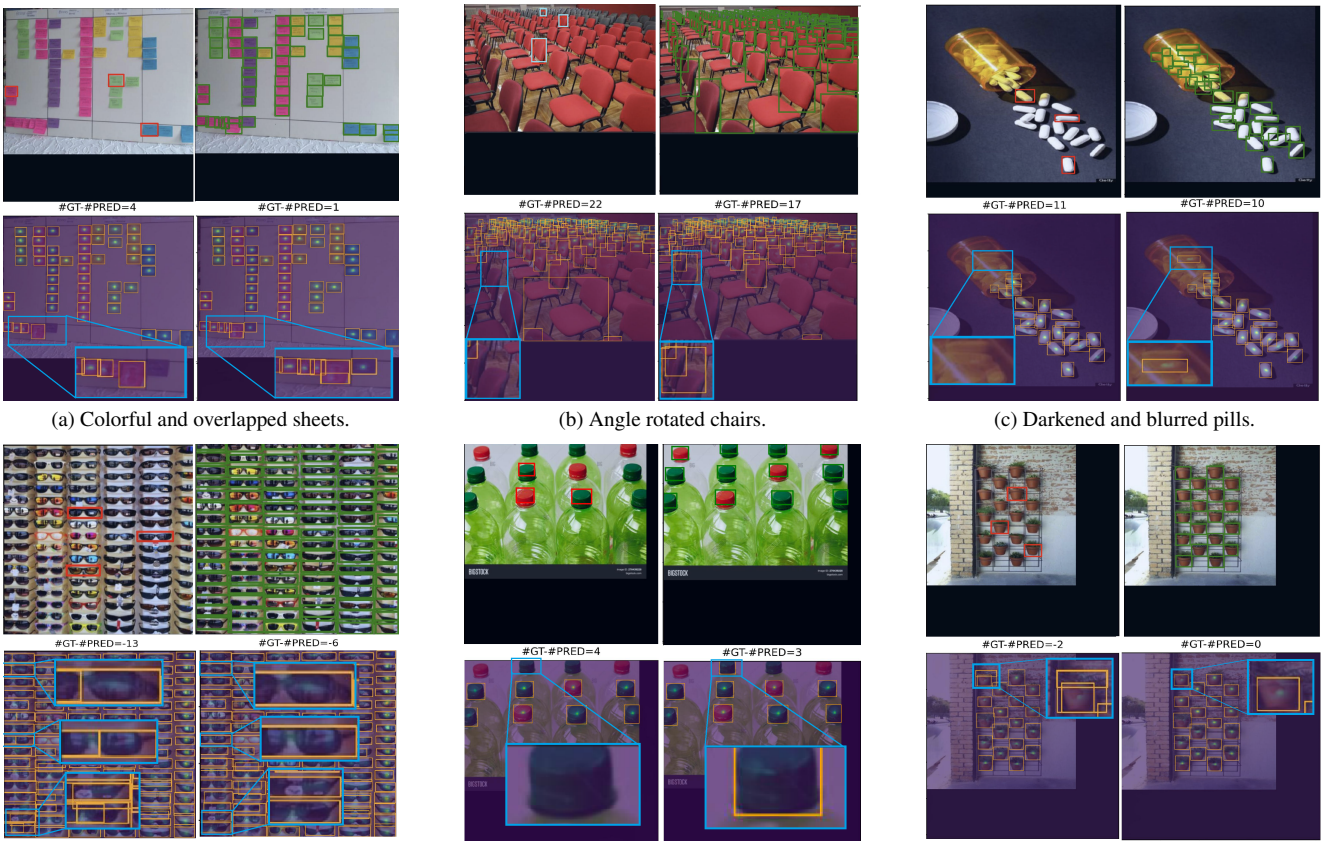
Mis-segmentation. In Figure 3(d), baseline overcounts by segmenting single pairs of sunglasses as twice (GT-PRED=-13), Our model minimizes this (GT-PRED=-6), because IAL suppresses low-IoU duplicates via quality-aligned scores. This loss enhances GeCo's NMS detection and performance in validation set by inhibiting overgeneralization in organized scenes.

Partially cover or incomplete. In Figure 3(e), baseline neglects all incomplete lids for the image boundary (GT-PRED=4), because GeCo is sensitive to partial views with its prototype formulation. This leads to detection failure. Our model identifies one more (GT-PRED=3), Because FCL emphasizes hard and occluded examples for better center regression.

Misinterpretation. In Figure 3(f), baseline misinterprets wall patterns as pots (GT-PRED=-2), but our model accurately counts all without error (GT-PRED=0). This is because unsupervised aggregation of GeCo confuses backgrounds with flower pots, causing fake positives. While in

Table 3. Comparison of MAE and RMSE before (Orig.) and after (Mod.) removing the top1% data.

Model	Validation set						Test set					
	MAE↓			RMSE↓			MAE↓			RMSE↓		
	Orig.	Mod.	Drop (%)	Orig.	Mod.	Drop (%)	Orig.	Mod.	Drop (%)	Orig.	Mod.	Drop (%)
GeCo-base	9.69	6.58	32.09	**44.08**	15.39	65.09	**7.90**	5.00	36.71	**57.38**	10.39	81.89
GeCo-LF	**9.50**	**6.41**	32.53	44.98	**15.30**	65.98	7.97	**4.80**	39.77	66.29	**9.90**	85.07



(a) Colorful and overlapped sheets. (b) Angle rotated chairs. (c) Darkened and blurred pills.

(d) Densely arranged sunglasses. (e) Partially covered bottle lids. (f) Flower pots with a back wall.

Figure 3. For each occasion, the upper left is the input image with exemplar boxes, the upper right is the ground truth counting result with green boxes, the bottom left is the baseline model's counting result with yellow boxes, and the bottom right is our refined model's counting result with yellow boxes. The text is the deviation count from prediction to the ground truth. Positive value is undercounting and negative value is overcounting.

our model, IAL penalize low-quality matches to distinguish the exact pattern. FCL mitigates background imbalance and sharpens focuses on true centers in segmentation.

7. Conclusion and future work

We proposed a redesigned Class-Agnostic Counting (CAC) system based on GeCo and two losses to optimize counting predictions. On FSCD-147 validation set, our model makes marginal improvements in the metrics of MAE and AP, though test RMSE rises due to outlier sensitivity. Our quantified analyses across diverse scenarios show reduced miscounts, using our visualization for explainable outputs.

In the future, researchers of CAC would probably focus on training-free paradigms [12], dynamic example network [8] to address the appearance rarity in CAC task, and structured attention mechanism of modern vision models [19]. Also hybrid reference-less or text-guided methods fused with semantic segmentation, are proposed for complex, crowded environments [18].

References

[1] Luca Ciampi, Ali Azmoudeh, Elif Ecem Akbaba, Erdi Sarıtaş, Ziya Ata Yazıcı, Hazım Kemal Ekenel, Giuseppe Amato1, and Fabrizio Falchi. A survey on class-agnostic counting: advancements from reference-based to open-world text-guided ppproaches. *arXiv preprint arXiv:2501.19184*, 2025. 1, 3

[2] Nikola Djukic, Alan Lukezic, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. In *ICCV*, 2023. 1, 3, 4

[3] Shenjian Gong, Shanshan Zhang, Jian Yang, Dengxin Dai, and Bernt Schiele. Class-agnostic object counting robust to intraclass diversity supplementary material. *European Computer Vision Association*, 2022. 1

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. A novel unified architecture for low-shot counting by detection and segmentation. *arXiv preprint arXiv:2409.18686*, 2024. 2

[5] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 3

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arxiv preprint arxiv:1708.02002*, 2017. 3

[7] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. In *BMVC*. BMVA Press, 2022. 1

[8] Xinyan Liu, Guorong Li, Yuankai Qi, Ziheng Yan, Weigang Zhang, Laiyun Qing, and Qingming Huang. Dynamic example network for class-agnostic object counting. *Pattern Recognition*, 170:111998, 2026. 6

[9] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. *arXiv preprint arXiv:1811.00472*, 2018. 1

[10] Zhiheng Ma, Xiaopeng Hong, , and Qinnan Shangguan. Can sam count anything? an empirical study on sam counting. *arxiv preprint arxiv:2304.10817*, 2023. 4, 5

[11] Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. In *ECCV*, pages 348–365. Springer, 2022. 2, 3, 4, 5

[12] 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, 2025. 6

[13] Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. Dave – a detect-and-verify paradigm for low-shot counting, . *arXiv preprint arXiv:2404.16622*, 2024. 1, 2, 3, 4, 5

[14] Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A novel unified architecture for low-shot counting by detection and segmentation, . *arXiv preprint arXiv:2409.18686*, 2024. 1, 2, 4, 5

[15] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *CVPR*, pages 3394–3403, 2021. 1

[16] Luke Wood and François Chollet. Efficient graph-friendly coco metric computation for train-time model evaluation. *arxiv preprint arxiv:2207.12120*, 2022. 4

[17] Shengkai Wu, Xiaoping Li, and Xinggang Wang. Iou-aware single-stage object detector for accurate localization. *Image and Vision Computing*, 97:103911, 2020. 4

[18] Vathslya Yedidi. Inside the latest computer vision models in 2025. *imagevision.ai*, 2025. 6

[19] Ahmed Zgaren, Wassim Bouachir, and Nizar Bouguila. Save: self-attention on visual embedding for zero-shot generic object counting. *Journal of Imaging*, 11(2)(52):103911, 2025. 6

[20] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Varifocalnet: an iou-aware dense object detector. *arxiv preprint arxiv:2008.13367*, 2020. 3