Granularity-Aware Class-Agnostic Object Counting

Anonymous authors
Paper under double-blind review

Abstract

Exemplar-based class-agnostic counters can count novel objects from a few examples, but they often ignore the granularity specified by those examples. Given exemplars of red apples, should the model count only red apples, or all visually related apples? Existing methods typically lack this control and tend to collapse fine-grained object groups into a parent category. We introduce granularity-aware class-agnostic counting and propose S-Seg, an exemplar-conditioned segmentation model controlled by a sensitivity parameter α. Small α enforces strict exemplar matching, while large α expands the target to semantically related objects. The predicted mask then restricts existing counters to the intended target regions.

We train S-Seg from synthesized FSC-147 samples using only point supervision, without dense multi-granularity annotations. We also introduce FSC-FG, a benchmark with multiple fine-grained object groups per image. Experiments show that S-Seg consistently improves state-of-the-art counters on FSC-FG, reducing MAE by up to 54%, with further gains when negative exemplars are used to resolve ambiguous target definitions.

1 Introduction

Exemplar-based class-agnostic object counting (COC) Ranjan et al. (2021) enables objects to be counted at test time without re-training. In this setting, users specify the target by selecting a few exemplar boxes, and the model counts objects that match them. However, the intended counting target is not always fixed by the exemplars alone. As illustrated in Fig. 1, given red-apple exemplars, a user may want to count only red apples in one case, but count all visually related apples in another. Both are valid, depending on the desired granularity. Existing COC methods typically lack such control: they often collapse visually related objects into a single parent category and therefore cannot reliably adjust the counting scope.

This limitation is especially problematic for fine-grained counting. For example, when given exemplars of green grapes, state-of-the-art methods Ranjan et al. (2021); Shi et al. (2022); Liu et al. (2022) tend to count all grapes rather than only the target subtype. This behavior reflects both dataset limitations and the difficulty of modeling intra-class variability without extensive fine-grained training data Welinder et al. (2010); Van Horn et al. (2015); Khosla et al. (2011); Xu et al. (2021). More broadly, the issue is not only whether a model can recognize the parent category, but whether it can follow the user-specified granularity of the exemplars.

Text-based counting methods Dai et al. (2024); Jiang et al. (2023); Kang et al. (2024) can partially address fine-grained target specification, but they are better suited to settings with predefined vocabularies Xu et al. (2023) rather than rapid, image-specific target selection. They are also less effective for objects that lack standardized names, exhibit subtle visual differences, or belong to categories that are too fine-grained to describe reliably with text. For example, in industrial quality control, distinguishing nearly identical machine components, such as screws with slightly different threading, can be difficult to specify textually but straightforward with visual exemplars. In medical imaging, counting abnormal cells or rare anomalies, where precise terminology may not exist, naturally calls for an exemplar-based approach. Similarly, in wildlife monitoring, differentiating visually similar species or subspecies can be easier with a few selected examples than with ambiguous text descriptions. Exemplar-based counting therefore provides a flexible way to specify the target object group on the fly, but it still requires a mechanism for controlling how broadly the exemplars should generalize.

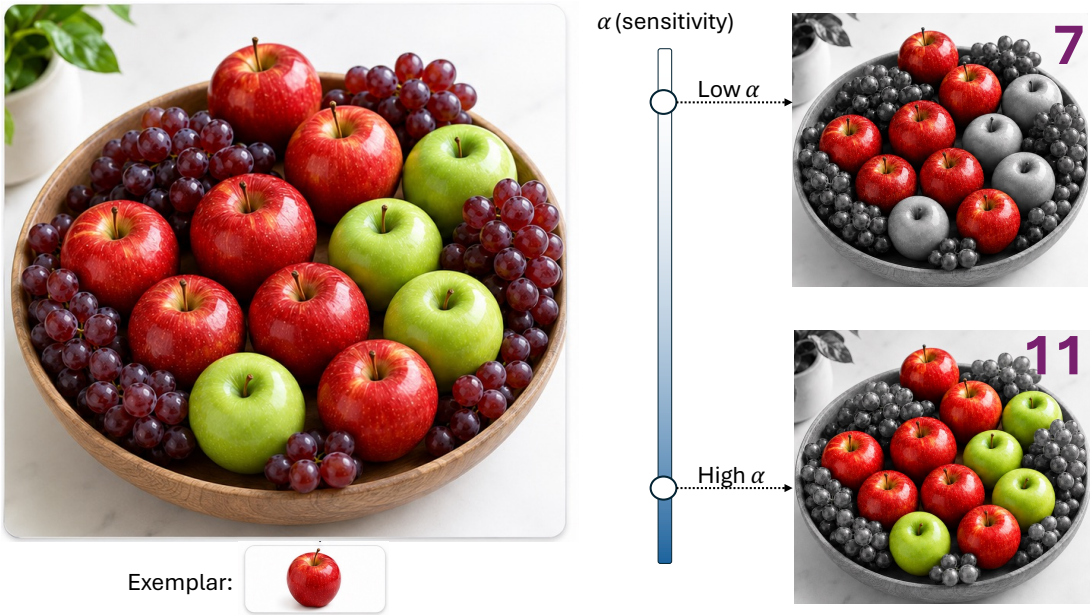


Figure 1: **One knob changes what counts.** Given target exemplars, the sensitivity parameter α controls how broadly the model generalizes from them. Low α counts only objects that closely match the exemplars, while high α expands the target to include related objects, with unrelated objects remaining excluded.

In this paper, we propose an exemplar-based COC method that makes this granularity controllable. Specifically, we introduce a segmentation network that dynamically segments objects at different granularity levels through a sensitivity parameter α. A low α produces a strict segmentation that focuses on objects closely matching the given exemplars, while a high α broadens the segmentation to include semantically related objects within the target category. The resulting mask is then used to restrict counting to the selected target scope. This design directly matches the interactive behavior shown in Fig. 1: moving α changes what counts, from fine-grained exemplar matching to broader category-level counting, while unrelated objects remain excluded.

To train this network, we propose a training paradigm based on synthesized samples derived from FSC-147 Ranjan et al. (2021). Each training sample is formed by concatenating three images: two images from the same parent category, such as 'green apples" and 'red apples", and a third image from a different category, such as "grapes". Under a high sensitivity setting, the network is trained to localize all regions containing apples. Under a low sensitivity setting, it is trained to localize only the specific type of apple depicted by the given exemplars. In effect, the model learns to distinguish object groups from the same parent category using only readily available coarse-grained data, where images are selected such that the object appearances remain sufficiently different. After training, the network can localize object groups from the same category at varying granularity levels, controlled by the input sensitivity parameter. At test time, we determine the sensitivity level based on the given exemplars and count only the objects within the resulting segmented regions.

To evaluate the proposed model, we collect a dataset with multiple levels of class granularity, including both parent classes and subclasses, for each image. Specifically, from FSC-147 Ranjan et al. (2021), COCO Lin et al. (2014), and Visual Genome Goyal et al. (2017); Antol et al. (2015), we select images that contain objects from multiple subcategories, such as donuts of different types, to form the FSC147-FG test set with 418 images. On this challenging test set, our proposed S-Seg model, when applied to state-of-the-art class-agnostic object-counting methods, produces more accurate object counts and substantially outperforms the original methods.

Furthermore, we observe that in some cases, the commonality among positive exemplars is insufficient to fully characterize the target object group. For example, if all flying birds in an image are white, it may be unclear whether the intended target is 'flying birds" or 'white birds". To address such ambiguities and facilitate future research, we also annotate three negative box exemplars for each image. These negative exemplars indicate visually similar objects that should not be counted. Our method can be easily extended to this setting and achieves further improvements in object counting.

Our method is compatible with various existing COC methods Ranjan et al. (2021); Shi et al. (2022); Liu et al. (2022); You et al. (2023), significantly improving their performance in fine-grained scenarios. For example, it reduces LOCA's MAE by 53.88% on the proposed FSC147-FG dataset, from 46.03 to 21.23. On the traditional FSC-147 dataset, it also yields minor improvements in some cases by removing distractors and background noise.

In summary, our contributions are:

- We propose a granularity-aware class-agnostic object-counting method with a controllable sensitivity input.
- We introduce a training paradigm that does not require multi-granularity annotated training data.
- We contribute synthetic and real-image test sets for evaluating counting methods under different granularity levels. When applied to state-of-the-art counters, our proposed segmentation method achieves promising counting results, reducing Mean Absolute Error by up to 54%.

2 Related Work

Object counting Pelhan et al. (2024); Farjon et al. (2023); Djukic et al. (2022); Amini-Naieni et al. (2023); Hsieh et al. (2017b); Lin et al. (2022); Yang et al. (2021b); Wang et al. (2024b); Zenglin Shi (2024); Zhu et al. (2024); Wan et al. (2021); Zhizhong et al. (2024); Zhang & Chan (2022); Xu et al. (2024); Paiss et al. (2023); Kong et al. (2006); Hui et al. (2025); Nguyen et al. (2024); Qiao et al. (2020); Su et al. (2024a); Wang et al. (2024a); Hobley & Prisacariu (2024); Su et al. (2024b) is a classical problem in computer vision. In this work, we focus on granularity-aware class-agnostic counting. The most relevant prior work falls into three categories: single-category counting, class-agnostic counting, and text-based counting. We briefly review them below.

Single-Category Counting. Many recent works study visual counting in category-specific settings, such as crowd counting Ma et al. (2019); Ranjan et al. (2018); Babu Sam et al. (2018); Sam et al. (2017); Li et al. (2018); Liu et al. (2018); Cao et al. (2018); Ranjan et al. (2019); Shi et al. (2019); Liu et al. (2019a); Wang et al. (2019); Zhang et al. (2019); Wan & Chan (2019); Ranjan et al. (2020); Abousamra et al. (2021), cell counting Arteta et al. (2016); Xie et al. (2018); Khan et al. (2016), and fruit counting Rahnemoonfar & Sheppard (2017). Existing methods typically follow either a detection-based or regression-based strategy. Detection-based methods Chattopadhyay et al. (2017); Hsieh et al. (2017a); Laradji et al. (2018) count objects by detecting individual instances. Regression-based methods Zhang et al. (2015); Cholakkal et al. (2019; 2022); Wang et al. (2020); Zhang et al. (2016); Chan et al. (2008); Liu et al. (2019b); Sun et al. (2023) predict the count directly or indirectly through a density map generated from dot annotations. These methods are effective for predefined categories, but they cannot directly count categories unseen during training.

Class-Agnostic Counting. Most related to our work are class-agnostic object counting methods Lu et al. (2018); Ranjan et al. (2021); Yang et al. (2021a); Shi et al. (2022); Gong et al. (2022); Nguyen et al. (2022); Liu et al. (2022); You et al. (2023); Arteta et al. (2014); You et al. (2022). Class-agnostic counting aims to adapt to novel object categories at test time without additional data collection, annotation, or training. In this setting, the target class is specified by a few visual exemplars. Lu *et al.* Lu et al. (2018) concatenate the feature maps of the input image and exemplars before feeding them into a matching module to regress the object count. Yang *et al.* Yang et al. (2021a) propose CFOCNet, which convolves exemplar embeddings

over query-image feature maps to produce density maps. FamNet Ranjan et al. (2021) follows a similar strategy and further applies test-time adaptation based on network activations at exemplar locations. Gong *et al.* Gong et al. (2022) improve robustness to intra-class diversity through exemplar feature augmentation and edge matching. More recently, Liu *et al.* Liu et al. (2022) show that transformer-based architectures can effectively capture patch-level similarity between exemplars and query images. In this paper, we build our segmentation architecture on the similarity-aware design of Shi *et al.* Shi et al. (2022), which jointly learns representations and similarity metrics for class-agnostic counting.

Text-Based Counting. Recent advances in large language models and vision-language models have led to increasing interest in text-based counting Xu et al. (2023). Representative methods include GroundingREC Dai et al. (2024), CLIP-Count Jiang et al. (2023), and VLCounter Kang et al. (2024). Text-based counting is effective when target classes are predefined, well defined, and easy to describe using natural language. However, fine-grained counting targets can be difficult to specify textually when the relevant distinction is subtle, image-specific, or lacks a standard name. In contrast, exemplar-based counting allows users to specify such targets visually, but existing methods do not explicitly control how broadly the model should generalize from the exemplars. Our work addresses this limitation by introducing a sensitivity-controlled segmentation module for granularity-aware counting.

3 Method

Overview. The teaser illustrates the central behavior we want: given the same image and the same exemplars, the desired count can change depending on how strictly the model should follow the exemplars. For example, red-apple exemplars may indicate only red apples under a strict setting, but may also include visually related green apples under a broader setting. Our goal is therefore not only to count objects specified by exemplars, but to make the counting granularity controllable.

We formulate this control through an exemplar-conditioned segmentation model, **S-Seg**. Given an image I, exemplar boxes E, and a sensitivity parameter α, S-Seg predicts a segmentation mask

$$P = \text{S-Seg}(I, E, \alpha), \qquad P \in [0,1]^{H \times W}, \tag{1}$$

where high values indicate image regions that should be counted. The parameter α controls how broadly the model generalizes from the exemplars. A small α corresponds to strict exemplar matching and selects only objects that closely resemble the exemplars. A large α expands the selected region to include semantically related objects from the same parent category. Unrelated objects should remain excluded under both settings.

Once the mask P is predicted, it can be used as a spatial filter for existing class-agnostic counting models: the counter is applied only to the regions selected by S-Seg. In this way, S-Seg converts a standard exemplar-based counter into a granularity-aware counter whose target scope can be adjusted by α.

The main training challenge is that counting datasets do not provide dense masks or ground-truth sensitivity values. In principle, S-Seg could be trained from tuples of the form [**image, exemplars, sensitivity parameter, segmentation mask**], but such multi-granularity segmentation annotations are unavailable. We therefore synthesize pseudo multi-granularity training samples from FSC-147 Ranjan et al. (2021), which provides exemplar boxes and point annotations. The key idea is to stitch together images from the same and different parent categories, then use α to determine which stitched regions should be treated as foreground during training.

3.1 Sensitivity-Based Training

The main challenge is to train S-Seg to respond differently to different values of α without having ground-truth sensitivity labels. We construct pseudo multi-granularity samples from FSC-147 by exploiting a simple observation: objects within the same image usually share more visual details than objects from another image, even if both images belong to the same parent category. Thus, two images from the same FSC-147 class can be treated as different fine-grained groups under the same parent category.

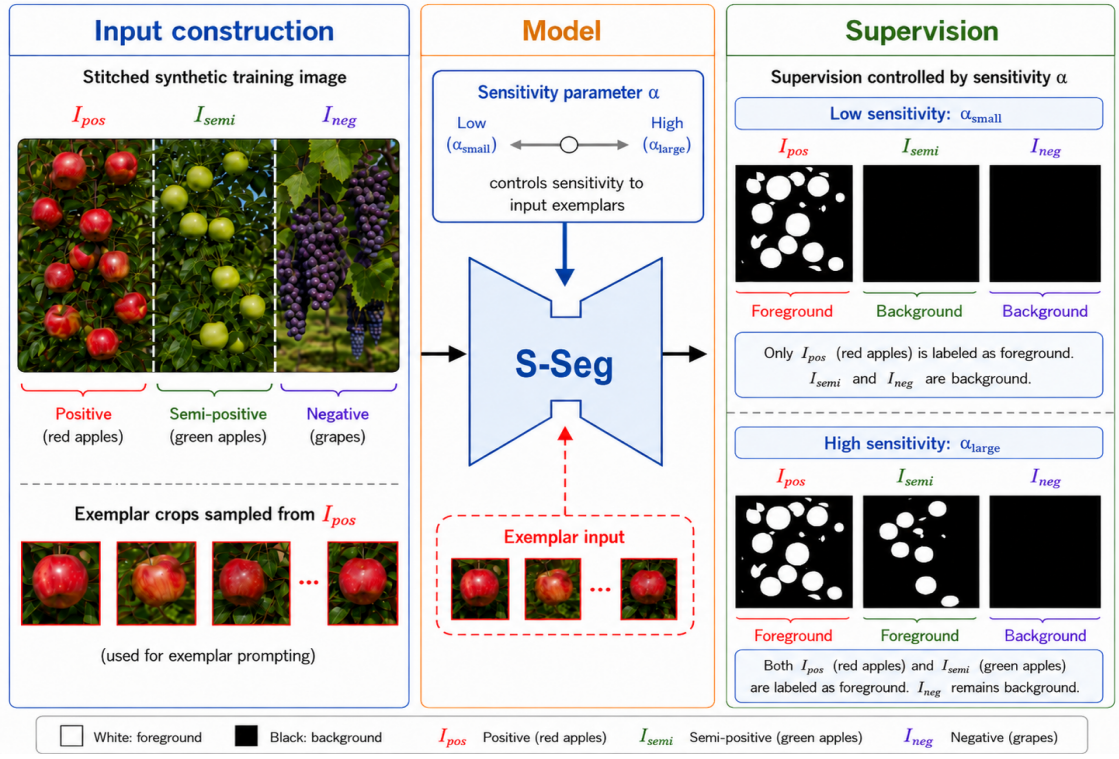


Figure 2: **Training pipeline for S-Seg.** We synthesize a training image by stitching together three images: I_{pos}, which provides the positive exemplars; I_{semi}, which contains objects from the same parent category but with different appearance; and I_{neg}, which contains objects from a different category. With a small sensitivity value α_{small}, S-Seg is trained to segment only the region corresponding to I_{pos}. With a large sensitivity value α_{large}, it is trained to segment both I_{pos} and I_{semi}. The region corresponding to I_{neg} is always treated as background.

Training Data. For each training sample, we draw three images:

- I_{pos}: an image from category c_i, together with its exemplar boxes E;
- I_{semi}: another image from the same category c_i;
- I_{neg}: an image from a different category c_j, where $c_j \neq c_i$.

We stitch these images into a single synthetic image $I = \text{Stitch}(I_{\text{pos}}, I_{\text{semi}}, I_{\text{neg}})$. Note that not every triplet $(I_{\text{pos}}, I_{\text{semi}}, I_{\text{neg}})$ provides a useful granularity signal. We therefore select triplets so that I_{semi} is visually different from I_{pos} but still belongs to the same parent category, while I_{neg} is even farther away and belongs to a different category. We measure the visual distance between two images using their exemplar embeddings. Let $E_a^{(i)}$ denote the i-th exemplar crop from image I_a, and let $G(\cdot)$ be a pretrained embedding model. We define

$$d(I_a, I_b) = \frac{1}{|E_a||E_b|} \sum_{i=1}^{|E_a|} \sum_{j=1}^{|E_b|} d_{\cos}\left(G(E_a^{(i)}), G(E_b^{(j)})\right), \tag{2}$$

where $d_{\cos}(\cdot,\cdot)$ is cosine distance. In our experiments, G is MoCo-v3 Chen et al. (2021). We keep a sampled triplet only if

$$d(I_{\text{pos}}, I_{\text{neg}}) > d(I_{\text{pos}}, I_{\text{semi}}) > d(I_{\text{pos}}, I_{\text{pos}}). \tag{3}$$

This ordering encourages I_{pos} and I_{semi} to represent two visually distinct fine-grained groups within the same parent category, while ensuring that I_{neg} remains a clear background category. The resulting triplets provide pseudo multi-granularity supervision: small α selects only I_{pos}, while large α selects both I_{pos} and I_{semi}.

Pseudo Mask Supervision and Training. Let Ω_{pos}, Ω_{semi}, and Ω_{neg} denote the spatial regions in I corresponding to the three stitched images. We train S-Seg under two sensitivity settings:

$$P_{\text{small}} = \text{S-Seg}(I, E, \alpha_{\text{small}}), \tag{4}$$

$$P_{\text{large}} = \text{S-Seg}(I, E, \alpha_{\text{large}}). \tag{5}$$

The pseudo-supervision is defined as follows (see Figure 2):

- For α_{small}, only Ω_{pos} is treated as foreground. Both Ω_{semi} and Ω_{neg} are treated as background.
- For α_{large}, both Ω_{pos} and Ω_{semi} are treated as foreground. Ω_{neg} is treated as background.

This forces S-Seg to associate small α with strict exemplar matching and large α with broader category-level generalization. After training, α acts as a continuous control over the segmentation scope.

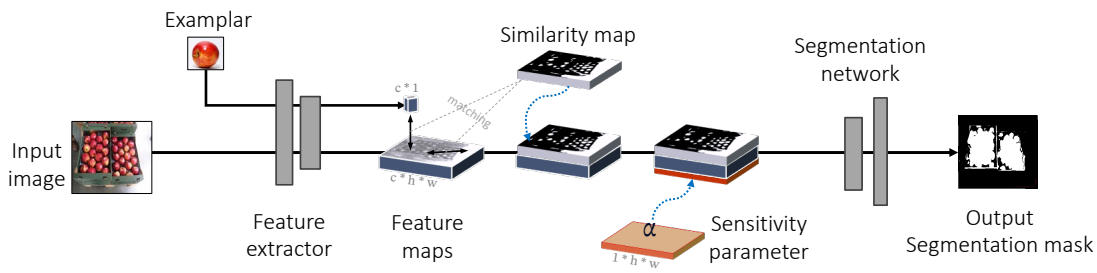


Figure 3: **S-Seg Architecture.** Our exemplar-based segmentation model, S-Seg, consists of a feature extractor and a segmentation network. The input image and the exemplar are input into the feature extractor to obtain a feature map and a feature vector (via average pooling), respectively. Based on this feature map and the feature vector, we compute a similarity map based on the bilinear similarity technique of Shi et al. (2022). The similarity map, feature map, and sensitivity parameter (enlarged to the size of the feature map) are concatenated and input into the segmentation network to output the final segmentation map.

S-Seg architecture. S-Seg is an exemplar-conditioned segmentation network that extends the similarity-based design used in exemplar-based counting models. Given an input image I and exemplar crops E, we first extract an image feature map $F(I)$ and exemplar features $F(E)$ using a shared feature extractor F. The exemplar features are average-pooled into a feature vector, which is compared with the image feature map to produce a similarity map following the bilinear similarity operation of Shi et al. (2022). To make the segmentation controllable, we replicate the sensitivity parameter α spatially to match the feature-map resolution and concatenate it with the image feature map and the similarity map. The concatenated representation is then passed to a segmentation head S to predict the final mask:

$$P = S\left(\left[F(I),\ \text{Sim}\big(F(I), F(E)\big),\ \alpha\mathbf{1}\right]\right), \qquad P \in [0,1]^{H\times W}. \tag{6}$$

where $\text{Sim}(\cdot,\cdot)$ denotes the bilinear similarity map, $[\cdot]$ denotes channel-wise concatenation, and $\alpha\mathbf{1}$ is the spatially replicated sensitivity map. In our implementation, S-Seg follows the architecture of BMNet Shi et al. (2022), but replaces the density-regression head with a segmentation head and explicitly conditions the prediction on α. Figure 3 depicts S-Seg architecture.

3.2 Point Annotation Supervision for Segmentation

The above training paradigm specifies which stitched regions should be foreground or background, but FSC-147 provides only point annotations rather than dense segmentation masks. We therefore use a weakly supervised segmentation loss based on point annotations.

For a given sensitivity value α, let $\Omega_{\text{fg}}(\alpha)$ denote the region or regions that should be segmented as foreground, and let $\Omega_{\text{bg}}(\alpha)$ denote the regions that should be segmented as background. Specifically,

$$\Omega_{\text{fg}}(\alpha_{\text{small}}) = \Omega_{\text{pos}}, \qquad \Omega_{\text{bg}}(\alpha_{\text{small}}) = \Omega_{\text{semi}} \cup \Omega_{\text{neg}}, \tag{7}$$

and

$$\Omega_{\text{fg}}(\alpha_{\text{large}}) = \Omega_{\text{pos}} \cup \Omega_{\text{semi}}, \qquad \Omega_{\text{bg}}(\alpha_{\text{large}}) = \Omega_{\text{neg}}. \tag{8}$$

Since dense foreground masks are unavailable, we construct a soft certainty map $W \in \mathbb{R}^{H \times W}$ from the point annotations. For each annotated object center in the foreground region, we place a 2D Gaussian centered at that point. The resulting map assigns high confidence near annotated object centers and low confidence elsewhere. This allows the model to receive positive supervision where the foreground evidence is reliable.

Let $P = \text{S-Seg}(I, E, \alpha)$ be the predicted segmentation mask. We define the foreground loss as

$$L_{\text{fg}} = \|(P - \mathbf{1}) \odot W \odot M_{\text{fg}}\|_2^2 . \tag{9}$$

where $\mathbf{1} \in \mathbb{R}^{H \times W}$ is an all-one mask, M_{fg} is the binary mask corresponding to $\Omega_{\text{fg}}(\alpha)$, and $\odot$ denotes element-wise multiplication.

For background regions, the supervision is more direct: all pixels in $\Omega_{\text{bg}}(\alpha)$ should be predicted as background. We define

$$L_{\text{bg}} = \|P \odot M_{\text{bg}}\|_2^2 . \tag{10}$$

where M_{bg} is the binary mask corresponding to $\Omega_{\text{bg}}(\alpha)$.

The final training loss is

$$L_{\text{seg}} = L_{\text{fg}} + \lambda_{\text{bg}} L_{\text{bg}}, \tag{11}$$

where λ_{bg} balances foreground and background supervision. In practice, this loss trains S-Seg to activate near annotated foreground objects while suppressing predictions in regions that should not be counted.

3.3 Sensitivity Parameter Selection

We consider two strategies for choosing α_{small} and α_{large} during training.

Fixed sensitivity training. The simplest strategy is to use two fixed values: $\alpha_{\text{small}} = 0.1, \alpha_{\text{large}} = 0.8$. This explicitly teaches the model that a small value corresponds to strict exemplar matching, while a large value corresponds to broader inclusion of semantically related objects.

Adaptive sensitivity training. The second strategy computes α from the visual diversity of the exemplars. Let $\phi(\cdot)$ be an embedding function applied to exemplar crops. For a set of positive exemplars E_{pos} from I_{pos}, we define α_{small} as the average pairwise distance among their embeddings:

$$\alpha_{\text{small}} = \frac{1}{|E_{\text{pos}}|(|E_{\text{pos}}| - 1)} \sum_{e_a \neq e_b \in E_{\text{pos}}} |\phi(e_a) - \phi(e_b)|_2 . \tag{12}$$

This value reflects the intra-group variation among objects that should be treated as the same fine-grained target.

To compute α_{large}, we compare exemplars from I_{pos} with exemplars from I_{semi}. Let E_{semi} denote exemplar crops from I_{semi}. We define

$$\alpha_{\text{large}} = \frac{1}{|E_{\text{pos}}||E_{\text{semi}}|} \sum_{e_a \in E_{\text{pos}}} \sum_{e_b \in E_{\text{semi}}} |\phi(e_a) - \phi(e_b)|_2 \,. \tag{13}$$

This value captures the distance between two fine-grained groups within the same parent category. In practice, distances are normalized to the range $[0, 1]$ before being used as sensitivity values.

At test time, when only user-provided exemplars are available, we compute α from the average pairwise distance among the exemplar embeddings. If the exemplars are visually similar, the resulting α is small and S-Seg performs strict matching. If the exemplars are visually diverse, the resulting α is larger and S-Seg broadens the segmentation scope.

4 Experiments

4.1 FSC-FG Test Set

The FSC-147 dataset Ranjan et al. (2021) is designed to count all objects from a target category, without testing whether methods can follow the granularity specified by exemplars. To evaluate granularity-aware counting, we construct FSC-FG, a test set built from images in FSC-147 Ranjan et al. (2021), COCO Lin et al. (2014), and Visual Genome Goyal et al. (2017); Antol et al. (2015). In each FSC-FG test case, the goal is not to count all objects of a category, but only the subset matching the properties indicated by the exemplars.

These properties may include color, e.g., green vs. red apples, attributes, e.g., strawberries with or without chocolate, or material, e.g., carton vs. plastic milk containers. For each image, we annotate 3 positive exemplars. Since positive exemplars alone can be ambiguous, e.g., whether to count 'flying birds" or 'white birds" when all flying birds are white, we also annotate 3 negative exemplars. These negatives mark visually similar objects that should not be counted. FSC-FG contains 418 test images, with examples shown in Figure 4.

Figure 4: **FSC-FG dataset.** Each image includes 3 positive exemplars (blue boxes) and 3 negative exemplars (red boxes).

4.2 Results on FSC-FG

Table 1 reports the results on FSC-FG. Under *"With pos exemplars"*, we evaluate five class-agnostic counting methods: FamNet Ranjan et al. (2021), BMNet Shi et al. (2022), CounTR Liu et al. (2022), SAFECount You et al. (2023), and LOCA Đukić et al. (2023). None of these methods achieves reliable counts on FSC-FG, with MAEs of 43.04, 40.27, 37.85, 39.13, and 46.03, respectively. These errors are much higher than their MAEs on the original FSC-147 dataset, i.e., 22.08, 14.62, 12.49, 13.56, and 10.79, respectively.

This gap shows that methods trained on standard class-agnostic counting datasets, where each image usually contains one target class and only positive exemplars, tend to over-count by including all objects from the same category while ignoring exemplar granularity. In contrast, our S-Seg model identifies regions matching the intended fine-grained target, leading to substantially improved counting accuracy within these regions. Table 1 also shows that S-Seg remains effective even with the Fixed-Sensitivity-Parameter strategy, where fixed α values are used across different training images.

	With Pos Exemplars				With Pos+Neg Exemplars	
	FSC-FG		FSC147		FSC-FG	
Method	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
FamNet Ranjan et al. (2021)	43.04	73.01	22.08	**99.54**	43.04	73.01
FamNet + PosNeg-Seg	-	-	-	-	26.65	57.05
FamNet + S-Seg-Fixed	20.63	51.65	**21.35**	100.02	-	-
FamNet + S-Seg-Adaptive	**20.05**	**50.33**	21.66	99.55	**16.09**	**44.82**
BMNet Shi et al. (2022)	40.27	71.31	14.62	**91.83**	40.27	71.31
BMNet + PosNeg-Seg	-	-	-	-	31.75	62.08
BMNet + S-Seg-Fixed	22.17	54.43	**14.46**	92.25	-	-
BMNet + S-Seg-Adaptive	**21.18**	**52.78**	14.60	91.90	**16.75**	**44.88**
CounTR Liu et al. (2022)	37.85	67.89	12.49	**89.84**	37.85	67.89
CounTR + PosNeg-Seg	-	-	-	-	31.08	59.32
CounTR + S-Seg-Fixed	20.99	50.25	12.44	94.55	-	-
CounTR + S-Seg-Adaptive	**20.79**	**49.20**	**12.28**	92.37	**16.90**	**41.23**
SAFECount You et al. (2023)	39.13	69.40	13.56	91.30	39.13	69.40
SAFECount + PosNeg-Seg	-	-	-	-	28.67	59.72
SAFECount + S-Seg-Fixed	20.92	52.45	13.39	91.30	-	-
SAFECount + S-Seg-Adaptive	**20.63**	**51.69**	**13.33**	**91.02**	**15.97**	**45.22**
LOCA Đukić et al. (2023)	46.03	88.31	10.79	56.97	46.03	88.31
LOCA + PosNeg-Seg	-	-	-	-	32.68	72.38
LOCA + S-Seg-Fixed	21.75	53.74	**10.46**	56.82	-	-
LOCA + S-Seg-Adaptive	**21.23**	**52.85**	10.48	**55.97**	**15.99**	**43.47**

Table 1: **Comparison of counting results.** Columns under 'With Pos. Exemplars" report the standard setting where the target class is specified only by positive exemplars. Columns under 'With Pos. + Neg. Exemplars" report the setting where both positive and negative exemplars are available. PosNeg-Seg is only applicable in the positive+negative setting because it explicitly takes negative exemplars as input. S-Seg-Fixed is evaluated only in the positive-exemplar setting; in the positive+negative setting, S-Seg selects α using the negative exemplars. Best performance for each metric within each dataset block is highlighted in bold. "–" indicates that the method is not applicable to the corresponding setting.

4.3 Results on FSC147 and Additional Settings

We also evaluate S-Seg on the original FSC147 dataset Ranjan et al. (2021). Unlike FSC-FG, most FSC147 test images contain objects at a single granularity level, *i.e.*, objects typically share similar colors, types, or contexts, and the images contain relatively few fine-grained distractors. Existing class-agnostic counting methods therefore already perform well on this dataset by counting all objects from the target category. In this setting, we apply S-Seg on top of FamNet Ranjan et al. (2021), BMNet Shi et al. (2022), CounTR Liu et al. (2022), and SAFECount You et al. (2023). We use a fixed large value of α so that S-Seg includes all objects from the category of interest.

The results are reported in Table 1 under the FSC147 columns. S-Seg provides marginal improvements over the counting baselines. This is expected because FSC147 is largely a single-granularity benchmark and is therefore less challenging for existing counting methods. The modest gains mainly come from S-Seg suppressing irrelevant background regions or occasional distractors, rather than from resolving fine-grained ambiguity.

Table 1 also includes an additional setting where both positive and negative exemplars are available. Negative exemplars specify visually similar objects that should not be counted, and are used to resolve cases where positive exemplars alone are ambiguous. During inference, given an image I, positive exemplars E, and negative exemplars E_{neg}, we search for a suitable sensitivity parameter α by running S-Seg from large to small values of α until none of the negative exemplars is included in the predicted mask. This search can be performed efficiently in batches by evaluating multiple sensitivity values simultaneously. As shown

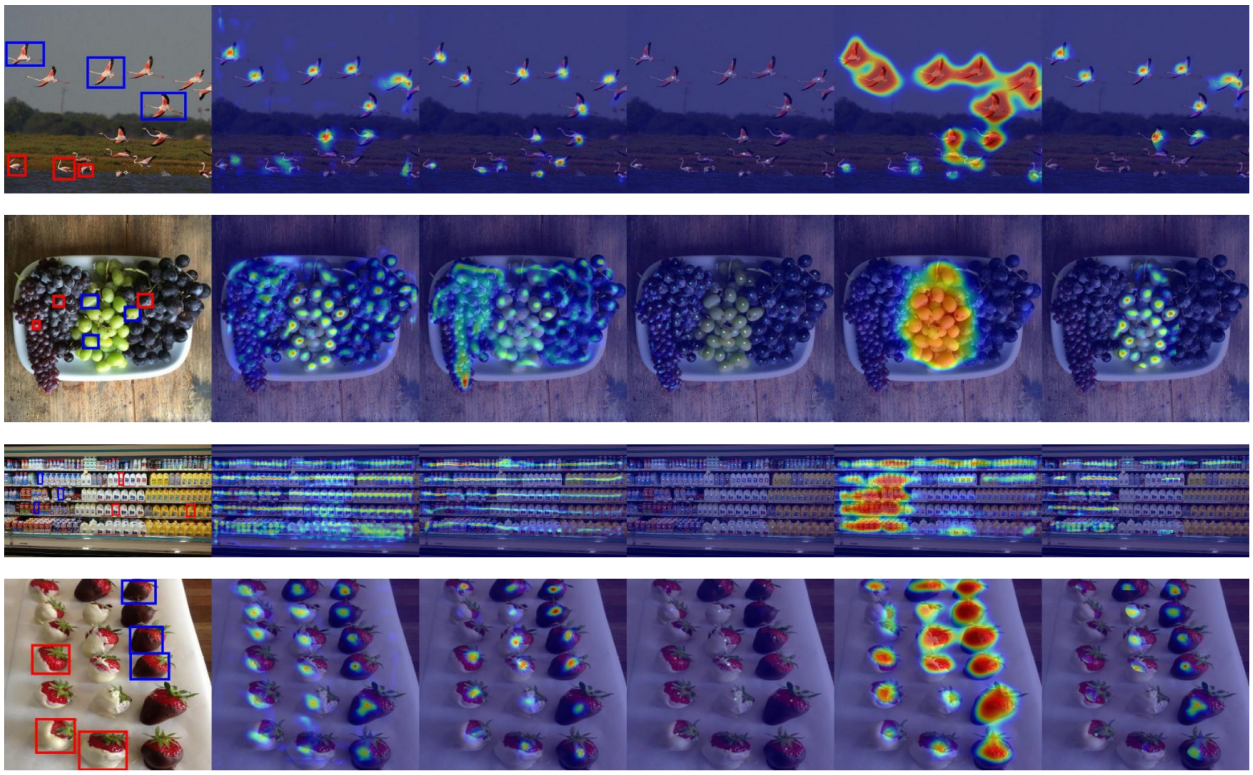


Figure 5: **Qualitative Results.** From left to right: input images with positive exemplars in blue and negative exemplars in red, although only positive exemplars are used for this experiment; output density maps from FamNet Ranjan et al. (2021), BMNet Shi et al. (2022), and CounTR Liu et al. (2022); the S-Seg segmentation mask; and the density map obtained by applying BMNet within our predicted mask.

under *With Pos. + Neg. Exemplars*, negative exemplars further improve counting performance by explicitly defining which visually similar objects should be excluded.

Finally, we compare S-Seg with a direct positive-negative segmentation baseline, denoted as PosNeg-Seg. PosNeg-Seg has an architecture similar to S-Seg, but it does not use the sensitivity parameter α. Instead, it directly takes both positive and negative exemplars as input, computes positive and negative similarity maps, concatenates them with image features, and predicts a segmentation mask. The results in Table 1 show that PosNeg-Seg reduces counting errors compared to the original counting baselines, confirming the usefulness of negative exemplars. However, S-Seg consistently outperforms PosNeg-Seg, indicating that selecting the counting scope through the sensitivity parameter α is more effective than directly conditioning on positive and negative exemplars alone.

We provide qualitative results in Fig. 5. Baseline counting methods tend to count objects across multiple granularities within the same image, whereas S-Seg localizes the intended target group before counting. For example, S-Seg can distinguish between visually similar subgroups such as flying and standing flamingos when the target is specified by exemplars. We further visualize the effect of α in Fig. 6. As α decreases, the segmentation mask becomes more selective and focuses on regions that most closely resemble the exemplars.

4.4 Comparison with an End-to-End Counting Baseline

S-Seg follows a segment-then-count design: it first localizes image regions that match the desired target granularity, and an existing class-agnostic counting model is then applied within the selected regions. To evaluate whether this decomposition is necessary, we compare against an end-to-end counting baseline trained

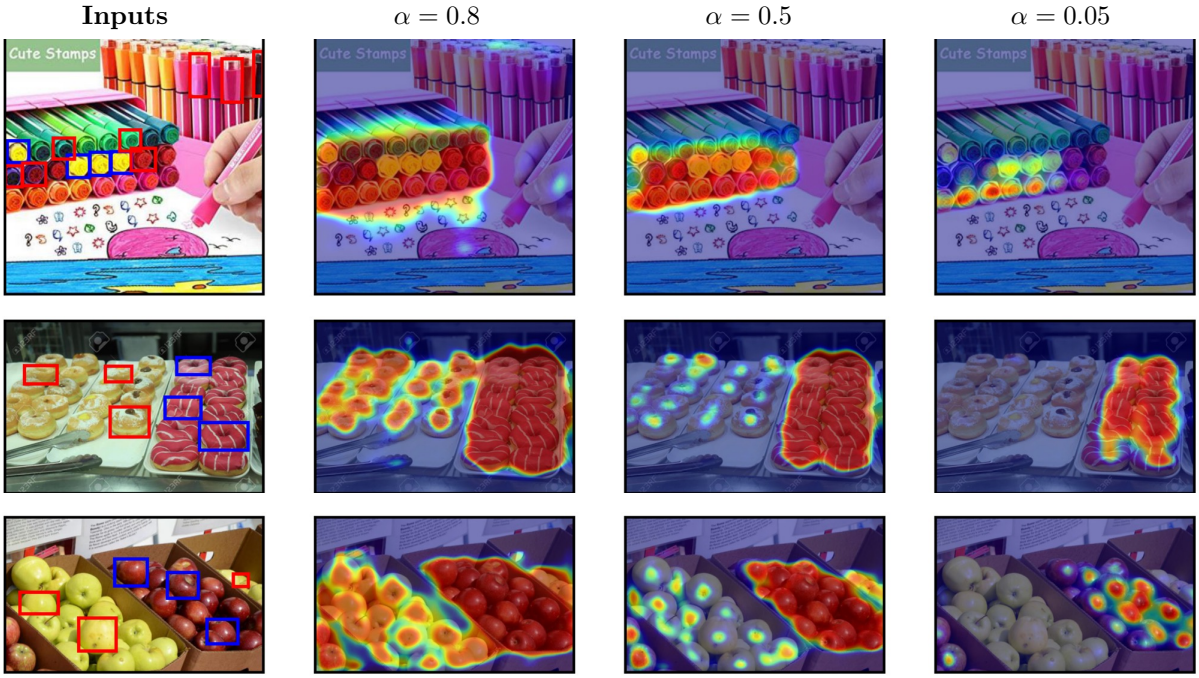


Figure 6: **Impact of the sensitivity parameter α.** As α decreases, the output of the segmentation model covers fewer areas but always prioritizes the image regions that resemble the exemplars.

	FSC-FG		FSC147	
Method	MAE ↓	MSE ↓	MAE ↓	MSE ↓
End-to-End Counting Baseline	29.82	60.06	21.20	121.28
FamNet Ranjan et al. (2021) + S-Seg-Adaptive	20.05	50.33	21.66	99.55
BMNet Shi et al. (2022) + S-Seg-Adaptive	21.18	52.78	14.60	91.90
CounTR Liu et al. (2022) + S-Seg-Adaptive	20.79	49.20	12.28	92.37
SAFECount You et al. (2023) + S-Seg-Adaptive	20.63	51.69	13.33	91.02
LOCA Đukić et al. (2023) + S-Seg-Adaptive	21.23	52.85	10.48	55.97

Table 2: **Comparison with an end-to-end counting baseline.** The end-to-end baseline is trained on the same synthesized data to directly predict counts for granularity-aware targets.

on the same synthesized data. This baseline directly predicts counts for granularity-aware targets, without using an explicit intermediate segmentation mask.

The results are shown in Table 2. The end-to-end baseline improves over several original counting models on FSC-FG, but it remains substantially worse than S-Seg combined with existing counters. This suggests that directly learning counting and fine-grained target selection in a single network is less effective in this setting. In contrast, separating the problem into target localization followed by counting allows S-Seg to handle granularity selection explicitly while preserving the counting ability of strong existing models.

4.5 Comparison with a Vanilla Segmentation Model

We conduct this comparison to evaluate the effectiveness of incorporating the sensitivity threshold α into the segmentation model. Specifically, we train a vanilla segmentation model, denoted as Vanilla-Seg. This model has the same architecture as S-Seg, but does not use the sensitivity parameter. Both models are

Method	FSC-FG		FSC147	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓
FamNet Ranjan et al. (2021)	43.04	73.01	22.08	**99.54**
FamNet Ranjan et al. (2021) + Vanilla Seg.	27.04	59.00	23.01	107.65
FamNet + S-Seg-Adaptive (Ours)	**20.05**	**50.33**	**21.66**	99.55
BMNet Shi et al. (2022)	40.27	71.31	14.62	**91.83**
BMNet Shi et al. (2022) + Vanilla Seg.	28.95	61.10	28.85	113.98
BMNet + S-Seg-Adaptive (Ours)	**21.18**	**52.78**	**14.60**	91.90
CounTR Liu et al. (2022)	37.85	67.89	12.49	**89.84**
CounTR Liu et al. (2022) + Vanilla Seg.	26.79	55.67	22.53	113.60
CounTR + S-Seg-Adaptive (Ours)	**20.79**	**49.20**	**12.28**	92.37
SAFECount You et al. (2023)	39.13	69.40	13.56	91.30
SAFECount You et al. (2023) + Vanilla Seg.	27.95	60.62	23.41	109.77
SAFECount + S-Seg-Adaptive (Ours)	**20.63**	**51.69**	**13.33**	**91.02**
LOCA Đukić et al. (2023)	46.03	88.31	10.79	56.97
LOCA Đukić et al. (2023) + Vanilla Seg.	29.88	64.14	22.32	103.01
LOCA + S-Seg-Adaptive (Ours)	**21.23**	**52.85**	**10.48**	**55.97**

Table 3: **Comparison with a vanilla segmentation model on FSC-FG and FSC147.** Best performance is highlighted in bold.

trained on the same synthetic dataset under identical settings. At test time, we input only the image and exemplars to Vanilla-Seg to obtain a single segmentation mask, which is then used for counting.

The results are summarized in Table 3. We use the predicted mask from Vanilla-Seg to localize the target objects before applying five class-agnostic counting methods: FamNet, BMNet, CounTR, SAFECount, and LOCA. While Vanilla-Seg improves the baseline results by a large margin on FSC-FG, its gains are consistently smaller than those obtained with our sensitivity-based S-Seg. This shows that explicitly conditioning the segmentation model on α is important for handling different counting granularity.

5 Conclusion

In this paper, we introduced granularity-aware class-agnostic object counting, where the target is defined not only by exemplar category but also by the desired visual granularity. We showed that existing exemplar-based counters often collapse fine-grained object groups into a parent category, causing over-counting when only a specific subgroup is intended. To address this, we proposed S-Seg, an exemplar-conditioned segmentation model that uses a sensitivity parameter α to control how broadly the model generalizes from the exemplars. By adjusting α, S-Seg localizes either objects closely matching the exemplars or a broader set of semantically related objects, enabling existing counters to operate within the intended target scope.

We also proposed a sensitivity-based training strategy that learns this behavior from synthesized FSC-147 samples using only point supervision, without requiring dense multi-granularity annotations. Finally, we introduced FSC-FG, a benchmark for evaluating granularity-aware counting on real images with multiple object subgroups. Experiments show that S-Seg consistently improves state-of-the-art class-agnostic counters on FSC-FG, with further gains when negative exemplars are used to resolve ambiguous target definitions.

References

Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. 2021.

Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specified object counting. *ArXiv*, abs/2306.01851, 2023. URL `https://api.semanticscholar.org/CorpusID:259075464`.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Carlos Arteta, Victor S. Lempitsky, Julia Alison Noble, and Andrew Zisserman. Interactive object counting. In *ECCV*, 2014.

Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Detecting overlapping instances in microscopy images using extremal region trees. *Medical image analysis*, 27:3–16, 2016.

Deepak Babu Sam, Neeraj N Sajjan, R Venkatesh Babu, and Mukundhan Srinivasan. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. 2018.

Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. 2018.

Antoni B. Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2008.

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4428–4437, 2017.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629, 2021. URL `https://api.semanticscholar.org/CorpusID:233024948`.

Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12389–12397, 2019.

Hisham Cholakkal, Guolei Sun, Salman Hameed Khan, Fahad Shahbaz Khan, Ling Shao, and Luc Van Gool. Towards partial supervision for generic object counting in natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:1604–1622, 2022.

Siyang Dai, Jun Liu, and Ngai-Man Cheung. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16985–16995, June 2024.

Nikola Djukic, Alan Lukevzivc, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18826–18835, 2022. URL `https://api.semanticscholar.org/CorpusID:253523039`.

Guy Farjon, Huijun Liu, and Yael Edan. Deep-learning-based counting methods, datasets, and applications in agriculture: a review. *Precision Agriculture*, 24:1683–1711, 2023. URL `https://api.semanticscholar.org/CorpusID:257365692`.

Shenjian Gong, Shanshan Zhang, Jian Yang, Dengxin Dai, and Bernt Schiele. Class-agnostic object counting robust to intraclass diversity. *2022 European Conference on Computer Vision(ECCV)*, 2022.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Michael Hobley and Victor Prisacariu. Abc easy as 123: A blind counter for exemplar-free multi-class class-agnostic counting. In *European Conference on Computer Vision*, pp. 304–319. Springer, 2024.

Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4165–4173, 2017a.

Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4165–4173, 2017b. URL https://api.semanticscholar.org/CorpusID:11908837.

Xiaofei Hui, Qian Wu, Hossein Rahmani, and Jun Liu. Class-agnostic object counting with text-to-image diffusion model. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 1–18. Springer Nature Switzerland, 2025.

Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. *arXiv preprint arXiv:2305.07304*, 2023.

Seunggu Kang, WonJun Moon, Euiyeon Kim, and Jae-Pil Heo. Vlcounter: Text-aware visual representation for zero-shot object counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2714–2722, 2024.

Aisha Khan, Stephen Gould, and Mathieu Salzmann. Deep convolutional neural networks for human embryonic cell counting. Springer, 2016.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization(FGVC), IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2011.

D. Kong, D. Gray, and Hai Tao. A viewpoint invariant approach for crowd counting. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pp. 1187–1190, 2006. doi: 10.1109/ICPR.2006.197.

Issam H. Laradji, Negar Rostamzadeh, Pedro H. O. Pinheiro, David Vázquez, and Mark W. Schmidt. Where are the blobs: Counting by localization with point supervision. *ArXiv*, abs/1807.09856, 2018.

Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. 2018.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. URL https://api.semanticscholar.org/CorpusID:14113767.

Wei Lin, Kunlin Yang, Xinzhu Ma, Junyuan Gao, Lingbo Liu, Shinan Liu, Jun Hou, Shuai Yi, and Antoni B. Chan. Scale-prior deformable convolution for exemplar-guided class-agnostic counting. In *British Machine Vision Conference*, 2022. URL https://api.semanticscholar.org/CorpusID:256903360.

Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. 2022.

Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. 2019a.

Weizhe Liu, Mathieu Salzmann, and Pascal V. Fua. Context-aware crowd counting. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5094–5103, 2019b.

Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. 2018.

Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *ACCV*, 2018.

Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. 2019.

E-Ro Nguyen, Hieu Le, Dimitris Samaras, and Michael Ryoo. Instance-aware generalized referring expression segmentation. *arXiv preprint arXiv:2411.15087*, 2024.

Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. *2022 European Conference on Computer Vision(ECCV)*, 2022.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3147–3157, 2023. URL `https://api.semanticscholar.org/CorpusID:257102813`.

Jer Pelhan, Alan Lukevzivc, Vitjan Zavrtanik, and Matej Kristan. Dave – a detect-and-verify paradigm for low-shot counting. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23293–23302, 2024. URL `https://api.semanticscholar.org/CorpusID:269362449`.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020.

Maryam Rahnemoonfar and Clay Sheppard. Deep count: fruit counting based on deep simulated learning. *Sensors*, 17(4):905, 2017.

Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. 2018.

Viresh Ranjan, Mubarak Shah, and Minh Hoai Nguyen. Crowd transformer network. *arXiv preprint arXiv:1904.02774*, 2019.

Viresh Ranjan, Boyu Wang, Mubarak Shah, and Minh Hoai. Uncertainty estimation and sample selection for crowd counting. 2020.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3393–3402, 2021.

Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. 2017.

Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. 2019.

Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Wei Su, Peihan Miao, Huanzhang Dou, and Xi Li. Scanformer: Referring expression comprehension by iteratively scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13449–13458, 2024a.

Yuejiao Su, Yi Wang, Lei Yao, and Lap-Pui Chau. Few-shot class-agnostic counting with occlusion augmentation and localization. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5. IEEE, 2024b.

Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13791–13801, June 2023.

Nikola Đukić, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18872–18881, October 2023.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Jia Wan and Antoni Chan. Adaptive density map generation for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1130–1139, 2019.

Jia Wan, Ziquan Liu, and Antoni B. Chan. A generalized loss function for crowd counting and localization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1974–1983, 2021.

Boyu Wang, Huidong Liu abd Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. In *NeurIPS*, 2020.

Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. 2019.

Yutian Wang, Bin Yang, Xi Wang, Chao Liang, and Jun Chen. Satcount: A scale-aware transformer-based class-agnostic counting framework. *Neural Networks*, 172:106126, 2024a.

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*, 2024b.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Weidi Xie, J Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, 6(3):283–292, 2018.

Jingyi Xu, Hieu Le, Mingzhen Huang, ShahRukh Athar, and Dimitris Samaras. Variational feature disentangling for fine-grained few-shot classification. 2021.

Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *CVPR*, 2023.

Yuanwu Xu, Feifan Song, and Haofeng Zhang. Learning spatial similarity distribution for few-shot object counting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/167. URL `https://doi.org/10.24963/ijcai.2024/167`.

Shuo Yang, Hung-Ting Su, Winston H. Hsu, and Wen-Chin Chen. Class-agnostic few-shot object counting. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 869–877, 2021a.

Shuo-Diao Yang, Hung-Ting Su, Winston H. Hsu, and Wen-Chin Chen. Class-agnostic few-shot object counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 870–878, January 2021b.

Zhiyuan You, Yujun Shen, Kai Yang, Wenhan Luo, X. Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6304–6313, 2022.

Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2023.

Mengmi Zhang Zenglin Shi, Ying Sun. Training-free object counting with prompts. In *WACV*, 2024.

Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Attentional neural fields for crowd counting. 2019.

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

Qi Zhang and Antoni B. Chan. Calibration-free multi-view crowd counting. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pp. 227–244, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20076-2. doi: 10.1007/978-3-031-20077-9_14. URL https://doi.org/10.1007/978-3-031-20077-9_14.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pp. 3754–3762, 2016.

Huang Zhizhong, Dai Mingliang, Zhang Yi, Zhang Junping, and Shan Hongming. Point, segment and count: A generalized framework for object counting. In *CVPR*, 2024.

Huilin Zhu, Jingling Yuan, Zhengwei Yang, Yu Guo, Zheng Wang, Xian Zhong, and Shengfeng He. Zero-shot object counting with good exemplars. In *Proceedings of the European Conference on Computer Vision*, 2024.

A FSC-FG dataset

In this section, we describe in detail how we collect the multi-granularity dataset FSC-FG, with images originating from the datasets FSC-147 Ranjan et al. (2021), COCO Lin et al. (2014), and Visual Genome Goyal et al. (2017); Antol et al. (2015). For granularity-aware counting, we specifically chose images from these datasets that exhibit more than one class granularity such as green and black grapes, or flying and standing birds. In some cases, it is ambiguous to define the class of interest using only positive exemplars. To avoid this ambiguity, we annotate 3 positive exemplars and 3 negative exemplars for each sub-category of interest, on each image. In Figure 7, we show multiple examples of images in our dataset, featuring different types of class-granularity including:

- ***object color***, *e.g.*, green and red apples; black and white horses; green/red/black grapes. (Figure 7a)
- ***object attribute***,*e.g.*, strawberries with and without chocolate; sunglasses and regular glasses; flower pot with green and brown leaves. (Figure 7b)
- ***object status***,*e.g.*, flamingos on land or in water; flying and standing birds. (Figure 7c)
- ***object material***,*e.g.*, carton and plastic milk containers. (Figure 7d)

In figure 8, we demonstrate some examples in FSG-FG in which there exist different levels of granularity in the same images. For example, in figure (8a), the same positive exemplars could define different granularity levels of donut types: (i) Old Fashion Donut, and (ii) Glazed Old Fashion Donut. The user's choices of granularity are simulated by the negative exemplars.

B Numbers of negative exemplars

To evaluate the influence of the number of negative exemplars on performance, we present the results of our method applied to FSC-FG using 1 to 3 negative exemplars. The base counter is SAFECount, and the only varying factor segmentation mask. As shown in Table 4, the performance of S-Seg improves progressively with the addition of more negative exemplars. This trend indicates that an increased number of negative exemplars contributes to a more precise differentiation between categories, thereby enhancing overall accuracy. However, while the inclusion of additional negative exemplars yields better performance, it is important to consider the practical implications. Additional exemplar requires extra labeling effort. Thus, in real-world applications, a balance must be struck between the potential performance gains and the associated labeling labor. This trade-off emphasizes the need to carefully evaluate resource availability and task requirements when determining the optimal number of negative exemplars to use.

We also demonstrate the impact of the sensitivity threshold α in Figure 10. As expected, as α gets smaller, the segmented regions shrink to cover only the sub-classes of interest.

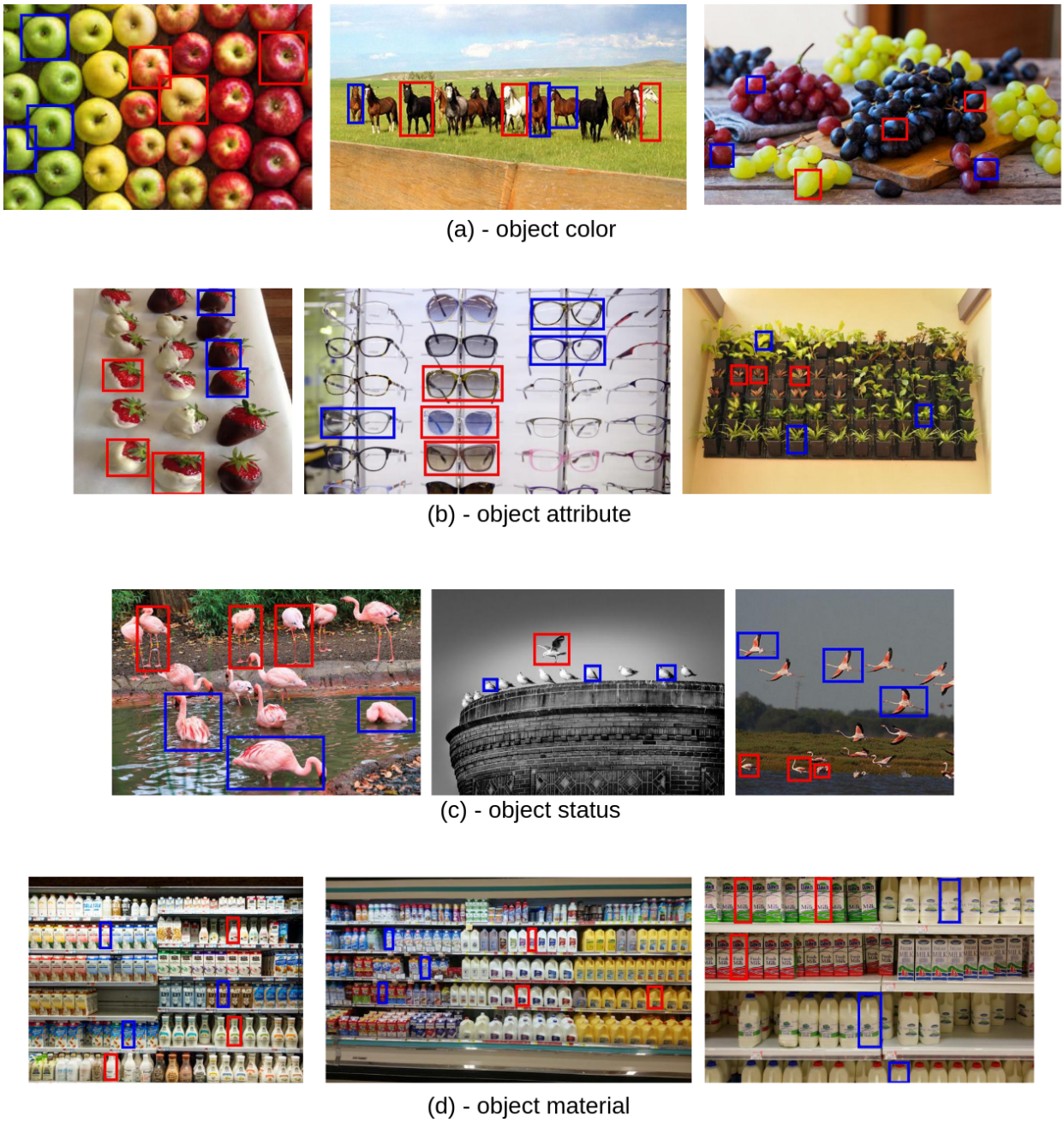


Figure 7: Examples of the FSC-FG dataset. Each row shows a property that defines sub-categories within images. Blue boxes are positive exemplars, red boxes are negative exemplars.

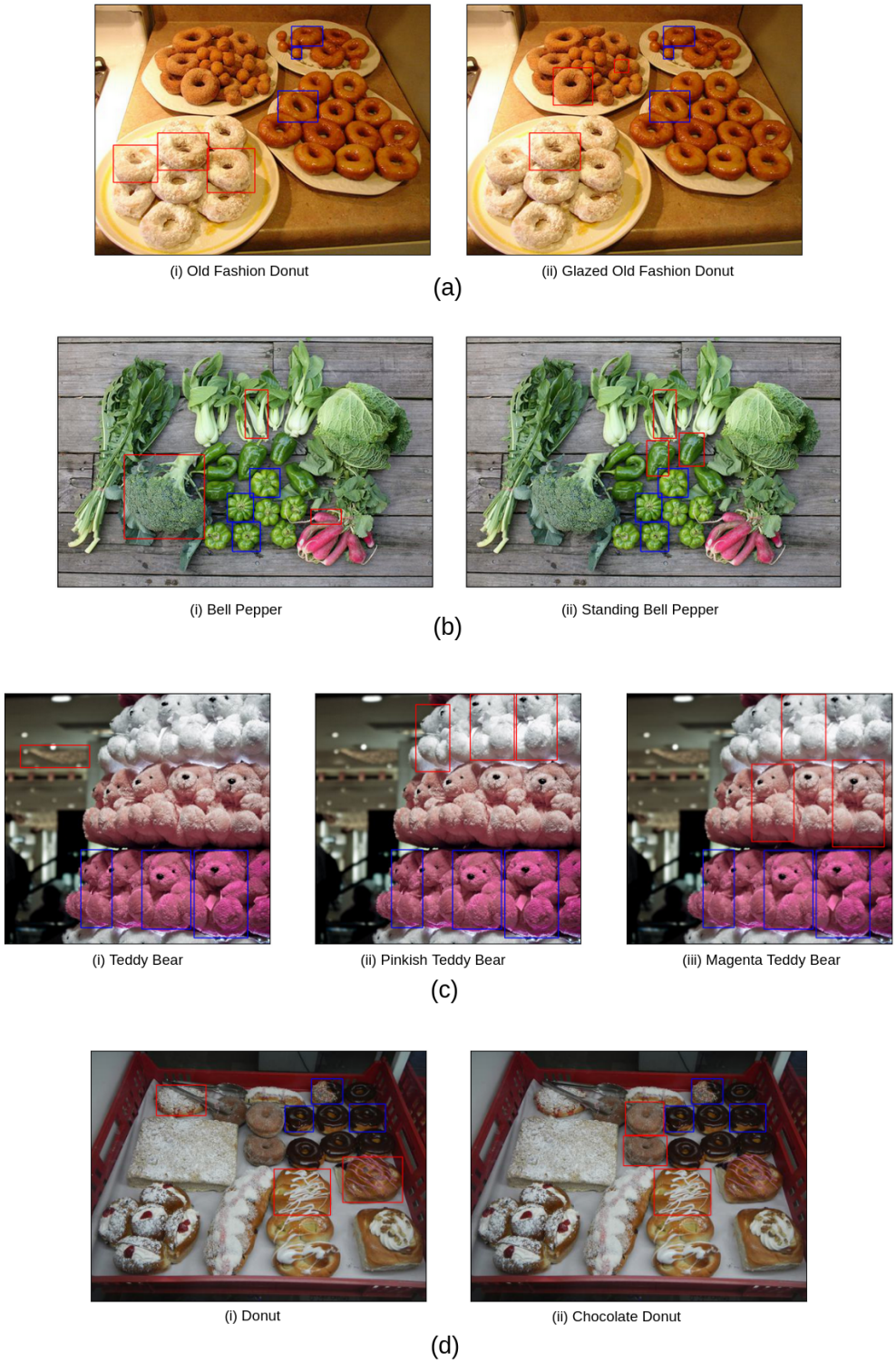
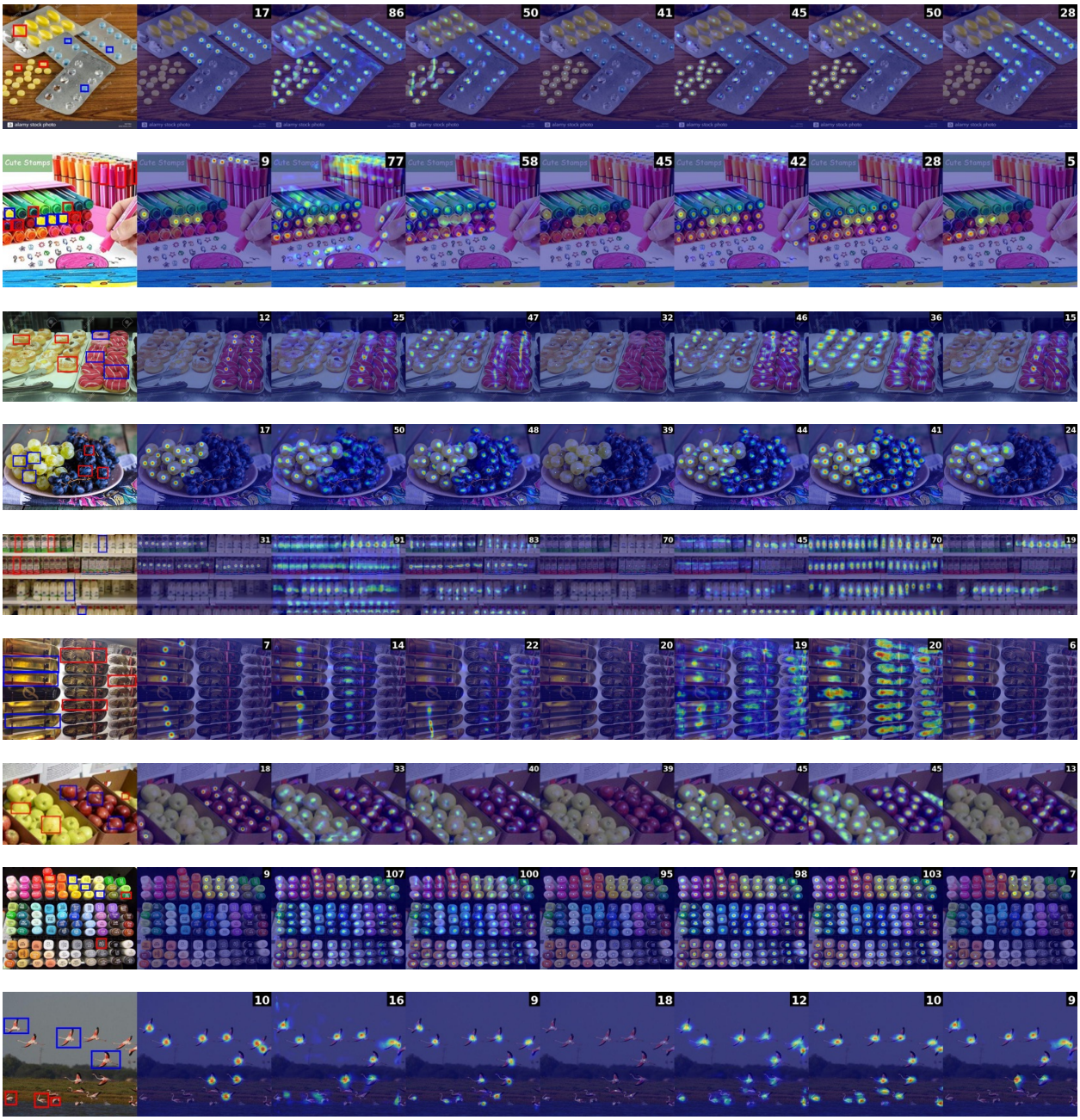
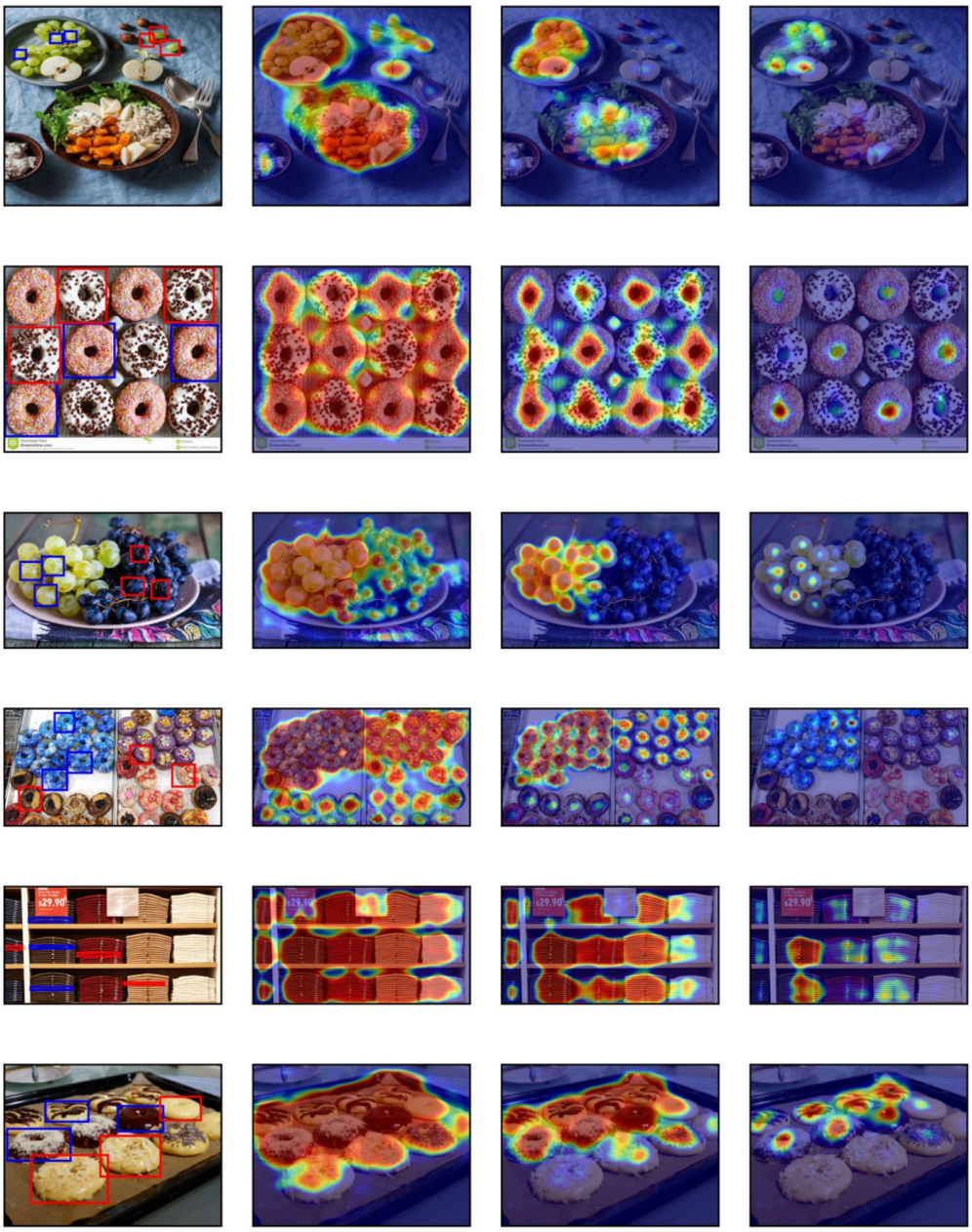


Figure 8: Examples with different levels of granularity in the FSC-FG dataset. Each row shows different levels of the target granularity within the same image. Blue boxes are positive exemplars, red boxes are negative exemplars.



Input GT FamNet BMNet CountTR SafeCount LOCA Ours

Figure 9: **Qualitative Results.** We visualize the density maps of different methods. From left to right are input images with positive and negative exemplars (blue and red boxes respectively), ground truth density map, output density maps obtained from FamNetRanjan et al. (2021), BMNetShi et al. (2022), CounTR (CTR) Liu et al. (2022), SAFECount (SC) You et al. (2023), LOCA Đukić et al. (2023), and the density map obtained by using FamNet together with our segmentation maps (the predicted counts are shown on top-right corner of each method). While FamNet, BMNet, CounTR, SAFECount, and LOCA tend to count all objects, our proposed method only counts objects of the category of interest.



Input $\alpha = 0.8$ $\alpha = 0.5$ $\alpha = 0.05$

Figure 10: **Impact of the sensitivity threshold α.** As α decreases, the output of the segmentation model covers fewer areas but always prioritizes the image regions that resemble the exemplars. We can control α to bias the network to only cover the objects of the category of interest.

# Neg Exemplars	0	1	2	3
MAE	20.63	18.64	16.12	15.97

Table 4: **Counting Accuracy for Different Exemplar Numbers.** Our method does not rely on, but benefits from having more negative examples.