CountSE: Soft Exemplar Open-set Object Counting

Shuai Liu, Peng Zhang, Shiwei Zhang, Wei Ke*
School of Software Engineering, Xi'an Jiaotong University
{wei.ke, sh_liu}@mail.xjtu.edu.cn, {pppzhang, zhangshiwei96}@stu.xjtu.edu.cn

Abstract

Open-set counting is garnering increasing attention due to its capability to enumerate objects of arbitrary category. It can be generally categorized into two methodologies: text-guided zero-shot counting methods and exemplar-guided few-shot counting methods. Previous text-guided zero-shot methods only provide limited object information through text, resulting in poor performance. Besides, though exemplar-guided few-shot approaches gain better results, they rely heavily on manually annotated visual exemplars, resulting in low efficiency and high labor intensity. Therefore, we propose CountSE, which simultaneously achieves high efficiency and high performance. CountSE is a new text-guided zero-shot object counting algorithm that generates multiple precise soft exemplars at different scales to enhance counting models driven solely by semantics. Specifically, to obtain richer object information and address the diversity in object scales, we introduce Semantic-guided Exemplar Selection, a module that generates candidate soft exemplars at various scales and selects those with high similarity scores. Then, to ensure accuracy and representativeness, Clustering-based Exemplar Filtering is introduced to refine the candidate exemplars by effectively eliminating inaccurate exemplars through clustering analysis. In the text-guided zero-shot setting, CountSE outperforms all state-of-the-art methods on the FSC-147 benchmark by at least 15%. Additionally, experiments on two other widely used datasets demonstrate that CountSE significantly outperforms all previous text-guided zero-shot counting methods and is competitive with the most advanced exemplar-guided few-shot methods. Codes will be available. Code is available at https://github.com/pppppz22/CountSE.

1. Introduction

The object counting task has experienced substantial advancements with the development of deep learning [7, 14, 34]. Its application is extended from traditional counting of a specific category to methods that support open-set counting. Closed-set [9, 17, 21, 23, 31, 36, 39, 40, 46, 47] object counting methods are limited to predefined categories present in the train set. In contrast, open-set [1, 2, 5, 16, 19, 28, 32, 35, 37, 42, 44, 48, 54] counting methods overcome this restriction by enabling the counting of objects from any category during inference without the need for additional training. Specifically, open-set counting allows the model to identify the target category via visual exemplars (e.g., bounding boxes of target objects in an image) or text prompts, thereby facilitating the counting of objects from any category in an image.

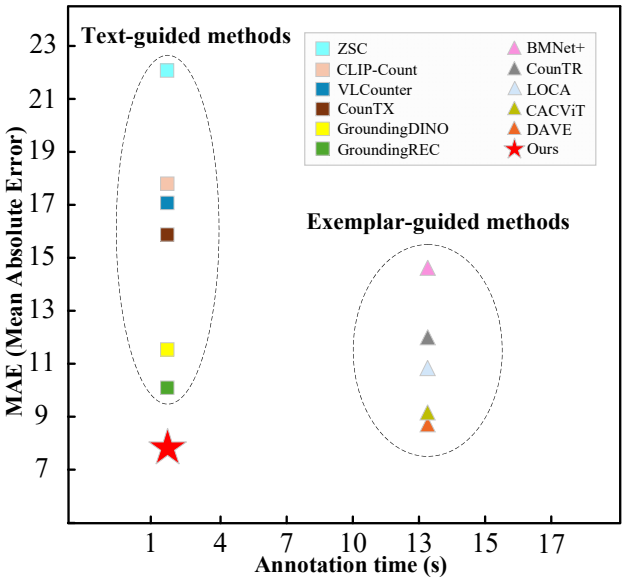

Figure 1. Our method achieves few-shot level performance with zero-shot annotation cost. We measure annotation time as either class name input (e.g., "tomato") or three exemplar box annotations per image, averaged over 100 images. While text-only zero-shot methods have minimal annotation cost but poor performance, and exemplar-based few-shot methods perform better but require costly annotations, our method bridges this gap effectively.

Some open-set counting methods (e.g., [CounTX [1],

*Corresponding author

CLIP-Count [19], VLCounter [20], GroundREC [5]]) rely entirely on text prompts to specify the counting targets, known as text-guided zero-shot object counting (ZSC) methods. However, in current ZSC methods, text prompts often provide limited descriptive information for novel categories and cannot provide object detail characteristics. This means that relying solely on text may not provide enough visual information to accurately identify and count objects, resulting in poor counting performance. Exemplar-guided few-shot counting (FSC) methods (e.g., [CounTR [28], LOCA [42], DAVE [32]]) can offer more comprehensive object information by manually specifying visual exemplars of the target objects. The performance of FSC methods is positively correlated with the accuracy of manually annotated exemplars. Denser images typically require more annotated representative exemplars to ensure satisfactory results, which significantly increases the workload and decreases efficiency in real-world scenarios. Furthermore, manually annotated exemplars face the challenge of high intra-class variance [19], meaning that differences (e.g., scale variations) between objects within the same category can easily cause the selected exemplars to be insufficiently representative. Although methods that combine text and visual exemplars (e.g., [CountGD [2]]) have been proposed for counting, the problems of high annotation costs and visual exemplar selection still exist.

To overcome the limitations mentioned above while combining the advantages of previous methods, we propose CountSE, a novel text-guided object counting method in which soft exemplars are selected only guided by text prompts. We call manually annotated exemplar bounding box is termed as the hard exemplar. Dynamically selecting exemplars in image features without annotation is called soft exemplar. Under the zero-shot setting, CountSE avoids manual annotation of exemplars and improves efficiency in real-world applications. The advantages of our model are shown in Figure 1, where lower Mean Absolute Error (MAE) represents better results.

Specifically, the method consists of two steps: Semantic-guided Exemplar Selection (SES) and Clustering-based Exemplar Filtering (CEF). The SES module addresses the challenge of object scale diversity. In dense scenes, object instances often vary significantly in size (e.g., far and near objects). Consequently, larger receptive fields may blur small object features, while smaller receptive fields fail to capture the complete semantics of large objects. SES extracts candidate soft exemplars from multi-resolution image features at different scales based on semantic information, effectively solving the multi-scale problem. Moreover, it adaptively adjusts the number of soft exemplars for each scale based on their similarity, ensuring that the selected exemplar features are more representative. The CEF module refines candidate soft exemplars output by the SES module using clustering techniques. The goal is to improve the representativeness and accuracy of the selected soft exemplar. Directly using the raw candidate exemplars may introduce noise, such as outliers that are similar in semantic features but have large visual representation differences. CEF performs clustering analysis on the candidate exemplars based on spectral clustering [50], effectively filtering out low-quality soft exemplars. The difference between our method and traditional methods is that, under the condition of using only text descriptions with low annotation cost, we supplement the text with high-quality selected visual exemplars to obtain richer object details. Furthermore, when dealing with dense images containing objects of various scales, we do not need to manually annotate a large number of multi-scale exemplars to avoid expensive annotation costs.

We conducted experiments on three datasets, demonstrating that our method outperforms all existing ZSC methods. On the FSC-147 [35], our method achieves 15.4% and 22.5% lower MAE than the previous state-of-the-art approach on the val and test sets, respectively. Considering the generalization issue, we experimented on CARPK [17] and ShanghaiTech-A [53], and including methods that use visual exemplars, our method achieved the best performance.

In summary, we make the following three contributions:

1. We propose CountSE, the counting method based on soft exemplars, which simultaneously achieves low annotation cost and high counting performance.

2. We propose a simple yet effective soft exemplar selection method. Through semantic-guided selection and clustering filtering, it accurately chooses representative exemplar features, and addresses the multi-scale challenges of objects, while significantly reducing the cost of manually annotated exemplar boxes.

3. We evaluate the model on three standard counting benchmarks: FSC-147 [35], CARPK [17], and Shanghai Tech-A [53]. The results show that CountSE significantly improves the state-of-the-art performance of zero-shot methods that rely solely on text prompts, and it is competitive with few-shot methods that use exemplars.

2. Related Work

2.1. Category-specific Counting

Object counting is one of the core tasks in computer vision, with the primary goal of accurately estimating the number of target instances in images or videos. Early research primarily focused on category-specific counting methods, where the counting categories are predefined as a closed set. The core assumption behind these methods is that the target category space remains completely consistent during both training and testing, and that the training data and test scenarios align strictly in terms of target types, density distribution, and environmental condi-

tions. Typical category of study include human [12, 13, 23–25, 27, 33, 38, 41, 43, 46, 51, 52], animals [4, 36, 40, 47], car [17, 21, 31], and cells [9, 10, 45]. While high accuracy has been achieved in these closed-set scenarios, they fail to meet the evolving needs of practical applications. Category-specific counting methods require the construction of a new dataset and the training of a specialized model for each new category. Although existing methods have been proposed to mitigate the severe performance degradation caused by cross-domain [3, 11, 18, 33] distribution shifts, they still fail to recognize novel categories absent during training. Open-set counting addresses the above issues by leveraging text and exemplar guidance to extend the counting categories, thereby meeting the technical demands of any categories.

2.2. Open-set Counting

Text-guided Zero-shot Counting. The visual language model provides a new solution paradigm for open-set counting. The text-guided category-agnostic counting extends the information of the novel category through the text description, that is, the zero shot setting. ZSC [48] first introduced the zero-shot counting task, which uses pure text descriptions without image exemplars, achieving target matching under zero-shot conditions through constructing class prototypes. CLIP-Count [19] utilizes the cross-modal alignment capabilities of the CLIP model, aligning text embeddings and visual features with contrastive loss to enable CLIP to perform pixel-level density predictions. CounTX [1] leverages pre-trained visual language models and constructs a Transformer decoder to achieve end-to-end text-conditioned counting. GroundingREC [5] uses the pre-trained GroundingDINO [29] model and performs global-local feature fusion to focus the counting model on the text description part. VA-Count [54] finds potential examples by detector to improve the adaptability of new categories.

Exemplar-guided Few-shot Counting. Exemplar-guided open-set counting enables the model to generalize to new categories by offering exemplars of counting categories. Early methods, such as FamNet [35], extend counting to novel categories by matching query images with exemplar templates of the new category. BMNET+ [37] proposes a dual optimization strategy involving joint learning of representations and similarity metrics, effectively improving adaptation to intra-class variations. CACViT [44] simplifies the category-agnostic counting (CAC) pipeline to a single pre-trained ViT [8], where exemplar feature extraction and similarity matching are simultaneously performed in self-attention, effectively shortening training time and improving counting accuracy. SAFECount [49] enhances model attention to areas related to exemplars in the image through similarity comparison and feature enhancement. CounTR [28] introduces a Transformer and uses cross-attention mechanisms to establish fine-grained associations between support samples and query images. LOCA [42] proposes an object prototype extraction module and integrates exemplar information with the query image. DAVE [32] generates a broad range of detection results and removes outliers, effectively addressing the overestimation of counts. CountGD [2] achieves richer object information by utilizing both object exemplars and text descriptions.

3. Method

Unlike visual exemplar-based counting methods, our approach uses semantic to identify object soft exemplars in encoded image features for counting and performance enhancement. The key challenge lies in accurate exemplar feature extraction. We address this with an adaptive method that extracts variable numbers and sizes of exemplars based on object size distribution. Since leveraging pretrained image and text encoders for feature matching, our subsequent modules require no learnable parameters.

3.1. Overview

Counting models based on density map regression [22] often suffer from reduced accuracy when counting occluded or small objects. Recent advancements in object counting have been driven by object detection models, particularly those based on GroundingDINO[29]. GroundingDINO, as a vision-language model, adopts a Transformer architecture and undergoes large-scale pre-training. It fuses text and image inputs at multiple levels, providing strong support for open-set object counting. Given these advantages, our model is developed on the GroundingDINO framework.

The framework of our method is shown in Figure 2. Specifically, given an input image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$, a text description t, we obtain a set of image features $I = \{I_{Stage1}, I_{Stage2}, I_{Stage3}, I_{Stage4}\}$ and text features T from encoder outputs. We use the Semantic-guided Exemplar Selection module to process I and T, selecting relevant features as candidate soft exemplar for each stage. Then, we obtain final soft exemplars through Clustering-based Exemplar Filtering module and concatenate the final soft exemplars with T to obtain the updated text feature T'. The image features I and text features T' are input into the enhancer and decoder to calculate the object count. We will discuss the details of the modules below.

3.2. Semantic-guided Exemplar Selection

To enable the model to adaptively recognize object instances of different sizes in an image, we divide the image features output by the SwinTransformer encoder into four size categories based on resolution. Specifically, the features from Stage 1, with a resolution of $H/8 \times W/8$, correspond to very small objects; the features from Stage 2, with a resolution of $H/16 \times W/16$, correspond to small objects; the features from Stage 3, with a resolution of

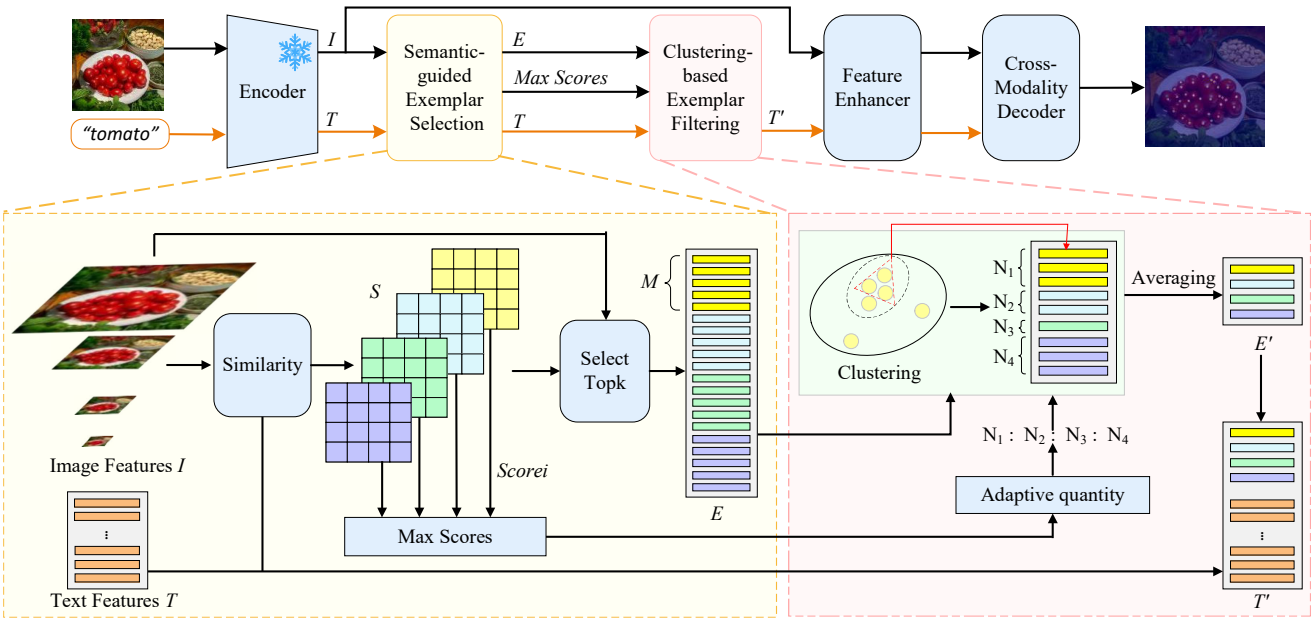

Figure 2. Overview of the CountSE architecture. The model takes both an image and a textual description as inputs. Encoders process these inputs to obtain multi-scale feature maps and text features. The overall flow of our method is illustrated in an enlarged schematic. Firstly, candidate soft exemplars are extracted through a Semantic-guided Exemplar Selection module. Then, the candidate soft exemplars are filtered through Clustering-based Exemplar Filtering. Finally, the generated soft exemplars are concatenated with the text features. The details of the module that has not been enlarged can be found in GroundingDINO[29].

$H/32 \times W/32$, correspond to medium-sized objects; and the features from Stage 4, with a resolution of $H/64 \times W/64$, correspond to large objects. To obtain relevant object information across different scales and enable the model to adapt to varying object sizes, we extract a different number of soft exemplars at each stage.

Assume that the total number of soft exemplars selected from features of all stages is N, with N_i exemplars selected from the Stage i. We use matrix multiplication to calculate the similarity map S_i between the text features T and the image features I_{Stagei} of Stage i:

$$S_i = I_{Stagei} \cdot T \tag{1}$$

Assuming the number of candidate soft exemplars at each stage is M, we extract the top M most relevant object features from the image features at different stages as candidate soft exemplars, denoted as $E_i = \{e_0, e_1, \ldots, e_{M-1}\}$. Meanwhile, we compute the maximum similarity score $Score_i$ between the text features T and the image features I_{Stagei} to determine the number N_i of soft exemplars selected from Stage i. In the detailed image of Figure 2, the maximum similarity scores are represented as Max Scores $= \{Score_i \mid i \in \{1, 2, 3, 4\}$.

$$E_i = \text{SelectTopK}(I_{Stagei}, S_i, M) \tag{2}$$

$$Score_i = \text{Max}(S_i) \tag{3}$$

3.3. Clustering-based Exemplar Filtering

For the candidate soft exemplars selected in 3.2, their semantic accuracy cannot be guaranteed. To minimize the impact of inaccurate candidate soft exemplars, we introduce a filtering stage to remove erroneously selected image features that share superficial feature similarities. Specifically, following DAVE [32], we apply spectral clustering [50] as the clustering technique. First, we compute the similarity among the M candidate soft exemplars selected at each stage to construct an affinity matrix. Based on this similarity, we cluster the candidate soft exemplars E_i.

As shown in Figure 2, we select the top N_i most similar features as the soft exemplar features from the cluster c_i that has the largest number of features in Stage i. Specifically, for the image features at different resolutions, we extract a different number of exemplars based on their similarity scores with the semantics. The Max Scores obtained in 3.2 represent the similarity between different resolution features and the semantics. We first compute the normalized probability corresponding to the scores:

$$P_i = \frac{Score_i}{\sum_{j=1}^{4} Score_j}, i \in \{1, 2, 3, 4\} \tag{4}$$

Using these normalized probabilities, we calculate the number of soft exemplars N_i to be extracted for Stage i:

$$N_i = N \times P_i \tag{5}$$

Method	Prompt Format	Publication Venue	Year	Val set		Test set	
				MAE	RMSE	MAE	RMSE
ZSC [48]	Text	CVPR	2023	26.93	88.63	22.09	115.17
CLIP-Count [19]	Text	ACM MM	2023	18.79	61.18	17.78	106.62
VLCounter [20]	Text	AAAI	2024	18.06	65.13	17.05	106.16
CountTX [1]	Text	BMVC	2023	17.10	65.61	15.88	106.29
GroundingDINO [29]	Text	ECCV	2024	12.07	57.02	11.52	<u>97.37</u>
COUNTGD [2]	Text	NeurIPS	2024	12.14	<u>47.51</u>	12.98	98.35
GroundingREC [5]	Text	CVPR	2024	<u>10.06</u>	58.62	<u>10.12</u>	107.19
FamNet [35]	Visual Exemplars	CVPR	2021	23.75	69.07	22.08	99.54
BMNet+ [37]	Visual Exemplars	CVPR	2022	15.74	58.53	14.62	91.83
CountTR [28]	Visual Exemplars	BMVC	2022	13.13	49.83	11.95	91.23
LOCA [42]	Visual Exemplars	ICCV	2023	10.24	32.56	10.79	56.97
CACViT [44]	Visual Exemplars	AAAI	2024	10.63	37.95	9.13	48.96
DAVE [32]	Visual Exemplars	CVPR	2024	8.91	28.08	8.66	32.36
COUNTGD [2]	Visual Exemplars & Text	NeurIPS	2024	7.10	26.08	5.74	24.09
CountSE (ours)	Text	–	–	**8.51**	**54.93**	**7.84**	**82.99**

Table 1. Comparison with state-of-the-art methods on FSC-147. Bold results are the best performance and the results with an underline are the second best.

After determining the number of soft exemplars N_i to be extracted in each stage, we select the top N_i features from the cluster c_i as the final soft exemplars E_i'. To ensure robustness and limit the length of the text features, we average the feature dimensions of the selected soft exemplars E_i' at each stage to obtain E_{Stagei}. Following CountGD[2], we then concatenate these averaged soft exemplars E_{Stagei} with the text features to update the text features to T':

$$T' = \{T, E_{Stage1}, E_{Stage2}, E_{Stage3}, E_{Stage4}\} \quad (6)$$

3.4. Loss

Following CountGD[2], the final loss $\mathcal{L}$ consists of both the localization loss $\mathcal{L}_{loc}$ and the classification loss $\mathcal{L}_{cls}$.

Localization Loss measures the discrepancy between the predicted object coordinates and the ground truth coordinates. This is calculated using the L_1-norm:

$$\mathcal{L}_{loc} = \sum_{i=1}^{K} |\hat{c}_i - c_i| \quad (7)$$

where $\hat{c}_i$ is the predicted object coordinate for the i-th object, and c_i is the ground truth object coordinate. K is the number of predicted points that match the ground truth.

Classification Loss calculates the discrepancy between the predicted object classes and the corresponding ground truth labels.

$$\mathcal{L}_{cls} = FocalLoss(\hat{L}, Y) \quad (8)$$

where $\hat{L}$ represents the classification probability predictions output by the model, and Y denotes the classification labels obtained through Hungarian matching between the predicted points and ground truth.

The final loss function is a weighted sum of the localization and classification losses. The weight λ balance the contributions of each term:

$$\mathcal{L} = \mathcal{L}_{loc} + \lambda \mathcal{L}_{cls} \quad (9)$$

4. Experiments

4.1. Implementation details

Training. We use the GroundingDINO-B[29] model as the backbone of our method. During training, we freeze the Swin Transformer [30] image encoder and the BERT [6] text encoder. All other trainable network parameters are optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} for 30 epochs. CountSE is trained on an RTX 3090 with a batch size of 4, and the training process takes approximately 16 hours. Additionally, the number of soft exemplars N is set to 18 and we set λ_1 to 1 and λ_2 to 5. For the FSC-147 [35] dataset, we apply the same corrections and text description modifications as COUNTGD [2].

Inference. During inference, the shortest side of each image is resized to 800 pixels while maintaining the aspect ratio. Additionally, we apply the adaptive cropping from CountGD[2] to address the issue of outputting a maximum of 900 counting at a time. For the CARPK [17] and ShanghaiTech-A [53], we use "car" and "people" as the text

Method	Prompt Format	MAE ↓	RMSE ↓
CLIP-Count [19]	Text	11.96	16.61
CounTX [1]	Text	8.13	10.87
VLCounter [20]	Text	6.46	8.68
COUNTGD [2]	Text	3.83	5.41
LOCA [42]	Visual Exemplars	9.97	12.51
CounTR [28]	Visual Exemplars	5.75	7.45
SAFECount [49]	Visual Exemplars	5.33	7.04
COUNTGD [2]	Visual Exemplars & Text	3.68	5.17
CountSE(ours)	Text	**2.79**	**4.20**

Table 2. Comparison with state-of-the-art methods on CARPK.

Method	Category	MAE ↓	RMSE ↓
MCNN [53]	Specific	221.4	357.8
CrwodClip [26]	Specific	217.0	322.7
RCC [15]	Open-set	240.1	366.9
CounTX[1]	Open-set	219.8	351.0
Clip-Count[17]	Open-set	192.6	308.4
COUNTGD[2]	Open-set	141.9	258.0
CountSE(ours)	Open-set	**129.7**	**258.3**

Table 3. Comparison with state-of-the-art methods including specific category counting on ShanghaiTech-A test set.

SES	CEF	Val Set		Test Set	
		MAE	RMSE	MAE	RMSE
-	-	12.07	57.02	11.52	97.37
✓	-	8.56	49.31	8.53	85.10
✓	✓	8.51	54.93	7.84	82.99

Table 4. Ablation experiments for each module.

descriptions respectively.

Metrics. Consistent with previous research, we use the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) to evaluate the performance of the model. These metrics are defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{11}$$

where n represents the number of test images, y_i denotes the actual number of objects, and $\hat{y}_i$ refers to the predicted number of objects for image i.

4.2. Comparison with the state of the art

FSC-147 [35]. In the category-agnostic counting task, we evaluate our method on the FSC-147. This means that we train our model on the train set of FSC-147 and evaluate it on the test and val sets for novel categories. We classify other open-world counting methods based on the type of prompt used: those using only text prompts, those using only visual exemplar prompts, and those using both text and visual exemplar prompts simultaneously.

As shown in Table 1, our method outperforms all other methods in the zero-shot setting relying solely on text prompts. Compared to the previous best method, our method outperforms in terms of MAE by 15.4% and 22.5% on the val and test sets, respectively. Even in the few-shot setting where only visual exemplar prompts are used, our method remains superior, outperforming the best method by 4.4% and 9.4% in MAE on the val and test sets, respectively. Among all methods, CountSE is only slightly inferior to CountGD, which utilizes both text and visual exemplar prompts. This is because, while soft exemplars also provide object information, they do not achieve the same level of accuracy as precisely annotated visual exemplars. Nevertheless, our method significantly improves the performance of pure text-guided methods, and the selected soft exemplars can still provide relatively accurate object features information.

CARPK [17]. We also test generalization ability of CountSE on the CARPK vehicle counting dataset. CountSE was trained solely on the FSC-147 train set, without any additional CARPK images for training. As shown in Table 2, CountSE outperforms all state-of-the-art methods on CARPK, even surpassing previous best-performing method CountGD by 24.1% and 18.7% in terms of MAE and RMSE, respectively.

ShanghaiTech-A [53]. We train the model on the FSC-147 train set and evaluate its generalization on the ShanghaiTech-A. Compared to FSC-147, the Shanghai Tech-A has higher population density and greater scale variation, making it more challenging. In Table 3, we compare the performance of CountSE with other methods. CountSE significantly outperforms the representative specific-category counting methods in the table. Additionally, our method achieves a 32.6% improvement in MAE and a 16.2% improvement in RMSE over the open-set counting method CLIP-Count. In conclusion, the experiments demonstrate the strong generalization ability of our method across different datasets.

4.3. Ablation study

Ablation of each module. We investigate the effectiveness of each module of our proposed method through ablation experiments. The results are shown in Table 4. We first conduct a baseline experiment without any modifications, where only text input is provided and no visual exemplars are used. Introducing the Semantic-guided Exemplar Selection (SES) module reduces the Mean Absolute Error (MAE) by 3.51 and 2.99 on the val and test sets, respectively. Next, we introduce the Clustering-based Exemplar

(a) Accurate soft exemplars.

(b) Inaccurate soft exemplars.

Figure 3. Visualization of soft exemplar. For better viewing, we enlarge the selected soft exemplars.

N	Val Set		Test Set	
	MAE	RMSE	MAE	RMSE
14	9.19	56.01	8.53	96.79
16	8.99	53.37	7.95	83.05
18	8.51	54.93	7.84	82.99
20	9.08	55.86	8.25	90.39
22	9.04	56.18	9.24	106.66

Table 5. Ablation experiment for N

λ	Val Set		Test Set	
	MAE	RMSE	MAE	RMSE
3	9.12	56.49	9.59	106.51
4	8.75	48.23	9.03	98.53
5	8.51	54.93	7.84	82.99
6	8.81	58.92	7.99	90.24
7	9.13	61.97	9.38	89.47

Table 6. Ablation experiment for λ

Filtering (CEF) module, which further reduces MAE by 0.05 and 0.69 on the val and test sets. The smaller improvement on the val set might be because the soft exemplars extracted from the val set are already sufficiently accurate, making the filtering process have a less significant impact on the selected exemplars.

The experimental results show that introducing the SES module, which selects multiple soft exemplars, provides additional object information that helps the model address high intra-class variance and scale variations. When the CEF is removed, performance decreases, demonstrating that CEF effectively filters out inaccurate soft exemplars, thereby improving the richness of object information.

Ablation of N. We also experimented with the impact of the number of soft exemplars N on performance. As shown in Table 5, we set N to 14, 16, 18, 20, and 22. The results clearly show that the model achieves the best performance when N is set to 18. This is because a smaller N doesn't provide enough diversity in the extracted soft exemplars, failing to capture all representative features. On the other hand, a larger N increases the number of extracted features, introducing noise that negatively affects the performance.

Ablation of λ. As shown in Table 6, experimental results indicate that when the λ is set to 5, both the MAE and RMSE on the test and val sets are relatively low, demonstrating that this parameter setting achieves a good balance to simultaneously consider accurate object recognition and essential localization information in an open-set scenario.

In an open-set setting, the model needs to accurately identify objects of various categories. If the λ is too low, the model tends to focus excessively on object localization, attempting to precisely locate objects but failing to classify them correctly, which leads to increasing counting errors. As the model's ability to distinguish object categories weakens, the MAE and RMSE on the test and val sets increase.

Conversely, since the counting task also requires precise localization, setting the λ too high may cause the model being unable to accurately localize the objects. In such a case, the model may only classify whether an object belongs to a certain category without providing accurate posi-

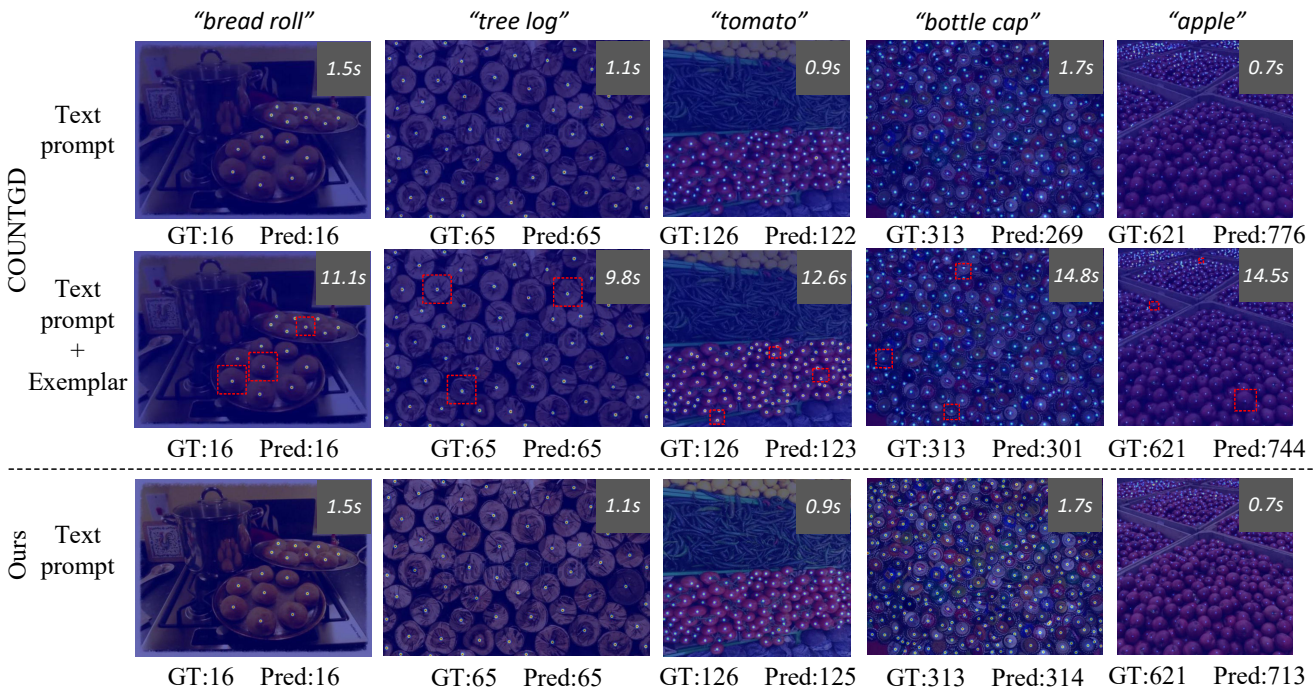

Figure 4. Visualization of counting results. We compare our method with the state-of-the-art COUNTGD in two scenarios: (1) using only text prompts and (2) using both text prompts and exemplars simultaneously. The red dashed box represents the annotated exemplars provided to COUNTGD, while the number in the upper right corner represents the annotation time required.

tional information, resulting in issues like double counting or missed counts.

4.4. Qualitative Result

Visualization of soft exemplar. To validate the effectiveness of semantically guided soft exemplars, we visualize them for representative images (cropped and enlarged for clarity). For simplicity, we visualize one soft exemplar per scale, and different scales are marked with dashed boxes of different colors. As shown in Figure 3a, for images containing multi-scale objects, our method can accurately capture the complete object information, such as a fully visible seagull or the edge features of an object, like the wing of a seagull.

However, as shown in Figure 3b, for complex scenes where objects and backgrounds are highly similar or contain multiple types of objects, our method fails to accurately extract soft exemplars at certain scales. Solving this problem may require more precise pre-training to avoid the association between semantics and irrelevant content.

Visualization of counting results. We compare our method with COUNTGD [2] on images with varying densities and scales (Figure 4). Our approach matches COUNTGD's performance on sparse, uniform-scale images but outperforms it significantly for dense, multi-scale scenes. Although adding visual exemplars reduces the error of COUNTGD, it still does not perform as well as our method. The results demonstrate that our method effectively handles dense object images with large-scale variations while significantly reducing annotation costs.

5. Conclusion

We propose CountSE, a novel zero-shot counting method that eliminates complex manual annotation in few-shot settings and overcomes the limited information from text descriptions in zero-shot settings. CountSE introduces soft exemplars to enrich object representations while handling scale diversity in dense scenes. Extensive experiments on benchmark datasets validate its effectiveness and cross-dataset generalization.

Acknowledgement

This research is supported by the National Key Research and Development Program of China (2023YFB3107400), the Natural Science Basic Research Plan in Shaanxi Province of China (2022JQ-631), the National Natural Science Foundation of China (U24B20185, T2442014, 62161160337, 62132011, 62376210, U20A20177, 62206217, U21B2018), the Shaanxi Province Key Industry Innovation Program (2023-ZDLGY-38). Thanks to the New Cornerstone Science Foundation and the Xplorer Prize.

References

[1] Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specifed object counting. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 1, 3, 5, 6

[2] Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. Countgd: Multi-modal open-world counting. *Advances in Neural Information Processing Systems*, 37:48810–48837, 2025. 1, 2, 3, 5, 6, 8

[3] Binghui Chen, Zhaoyi Yan, Ke Li, Pengyu Li, Biao Wang, Wangmeng Zuo, and Lei Zhang. Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16065–16075, 2021. 3

[4] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12397–12405, 2019. 3

[5] Siyang Dai, Jun Liu, and Ngai-Man Cheung. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16985–16995, 2024. 1, 2, 3, 5

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 5

[7] SongLin Dong, ChengLi Tan, ZhenTao Zuo, YuHang He, YiHong Gong, TianGang Zhou, JunMin Liu, and JiangShe Zhang. Brain-inspired dual-pathway neural network architecture and its generalization analysis. *Science China Technological Sciences*, 67(8):2319–2330, 2024. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019. 1, 3

[10] Giselle Flaccavento, Victor Lempitsky, Iestyn Pope, PR Barber, Andrew Zisserman, J Alison Noble, and Boris Vojnovic. Learning to count cells: applications to lens-free imaging of large fields. *Microscopic Image Analysis with Applications in Biology*, 1:3, 2011. 3

[11] Shenjian Gong, Shanshan Zhang, Jian Yang, Dengxin Dai, and Bernt Schiele. Bi-level alignment for cross-domain crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7542–7550, 2022. 3

[12] Mingyue Guo, Li Yuan, Zhaoyi Yan, Binghui Chen, Yaowei Wang, and Qixiang Ye. Regressor-segmenter mutual prompt learning for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28380–28389, 2024. 3

[13] Tao Han, Lei Bai, Lingbo Liu, and Wanli Ouyang. Steerer: Resolving scale variations for counting and localization via selective inheritance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21848–21859, 2023. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[15] Michael Hobley and Victor Prisacariu. Learning to count anything: Reference-less class-agnostic counting with weak supervision. *arXiv preprint arXiv:2205.10203*, 2022. 6

[16] 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*, pages 304–319. Springer, 2024. 1

[17] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017. 1, 2, 3, 5, 6

[18] Zhi-Kai Huang, Wei-Ting Chen, Yuan-Chun Chiang, Sy-Yen Kuo, and Ming-Hsuan Yang. Counting crowds in bad weather. *arXiv preprint arXiv:2306.01209*, 2023. 3

[19] 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*, pages 4535–4545, 2023. 1, 2, 3, 5, 6

[20] 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*, pages 2714–2722, 2024. 2, 5, 6

[21] Ersin Kilic and Serkan Ozturk. An accurate car counting in aerial images based on convolutional neural networks. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–10, 2023. 1, 3

[22] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. *Advances in neural information processing systems*, 23, 2010. 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 conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 1, 3

[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, 2021.

[25] 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, 2021. 3

[26] Dingkang Liang, Jiahao Xie, Zhikang Zou, Xiaoqing Ye, Wei Xu, and Xiang Bai. Crowdclip: Unsupervised crowd counting via vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2893–2903, 2023. 6

[27] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19628–19637, 2022. 3

[28] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721*, 2022. 1, 2, 3, 5, 6

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 3, 4, 5

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[31] Thomas Moranduzzo and Farid Melgani. Automatic car counting method for unmanned aerial vehicle images. *IEEE Transactions on Geoscience and remote sensing*, 52(3): 1635–1647, 2013. 1, 3

[32] 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*, pages 23293–23302, 2024. 1, 2, 3, 4, 5

[33] Zhuoxuan Peng and S-H Gary Chan. Single domain generalization for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28025–28034, 2024. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[35] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021. 1, 2, 3, 5, 6

[36] Farah Sarwar, Anthony Griffin, Priyadharsini Periasamy, Kurt Portas, and Jim Law. Detecting and counting sheep with a convolutional neural network. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018. 1, 3

[37] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9529–9538, 2022. 1, 3, 5

[38] Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19618–19627, 2022. 3

[39] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3365–3374, 2021. 1

[40] 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*, pages 13791–13801, 2023. 1, 3

[41] Ye Tian, Xiangxiang Chu, and Hongpeng Wang. Cctrans: Simplifying and improving crowd counting with transformer. *arXiv preprint arXiv:2109.14483*, 2021. 3

[42] 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*, pages 18872–18881, 2023. 1, 2, 3, 5, 6

[43] Mingjie Wang, Hao Cai, Yong Dai, and Minglun Gong. Dynamic mixture of counter network for location-agnostic crowd counting. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 167–177, 2023. 3

[44] 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 5832–5840, 2024. 1, 3, 5

[45] 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. 3

[46] Chenfeng Xu, Dingkang Liang, Yongchao Xu, Song Bai, Wei Zhan, Xiang Bai, and Masayoshi Tomizuka. Autoscale: Learning to scale for crowd counting. *International Journal of Computer Vision*, 130(2):405–434, 2022. 1, 3

[47] Jingsong Xu, Litao Yu, Jian Zhang, and Qiang Wu. Automatic sheep counting by multi-object tracking. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 257–257. IEEE, 2020. 1, 3

[48] Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15548–15557, 2023. 1, 3, 5

[49] 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 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6315–6324, 2023. 3, 6

[50] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral

clustering. *Advances in neural information processing systems*, 17, 2004. 2, 4

[51] Chengyang Zhang, Yong Zhang, Bo Li, Xinglin Piao, and Baocai Yin. Crowdgraph: Weakly supervised crowd counting via pure graph neural network. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5):1–23, 2024. 3

[52] Shiwei Zhang, Wei Ke, Shuai Liu, Xiaopeng Hong, and Tong Zhang. Boosting semi-supervised crowd counting with scale-based active learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8681–8690, 2024. 3

[53] 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*, pages 589–597, 2016. 2, 5, 6

[54] Huilin Zhu, Jingling Yuan, Zhengwei Yang, Yu Guo, Zheng Wang, Xian Zhong, and Shengfeng He. Zero-shot object counting with good exemplars. In *European Conference on Computer Vision*, pages 368–385. Springer, 2024. 1, 3