PBECount: Prompt-Before-Extract Paradigm for Class-Agnostic Counting

Canchen Yang[1], Tianyu Geng[1]*, Jian Peng[1], Chun Xu[2]

[1] College of Computer Science, Sichuan University, Chengdu, 610065, China
[2] Information Construction and Management Office, Sichuan University, 610065, China
volcano.ycc@gmail.com, {tygeng, jianpeng, xchun}@scu.edu.cn

Abstract

In the field of class-agnostic counting (CAC), counting only objects of interest that are similar to exemplars in multi-class scenarios has been a challenging task. To address this challenge, recent research has proposed the extract-and-match paradigm based on the vision transformer (ViT) architecture. However, although this paradigm can improve the accuracy of exemplar-similar object identification, it overly emphasizes the role of the ViT structure. To address this shortcoming, this work introduces a more generalized prompt-before-extract paradigm on top of the extract-and-match paradigm and designs a pure convolutional neural network (CNN) model named PBECount. In addition, an innovative loss function, a post-processing strategy, and a dynamic threshold method are proposed to enhance the detection performance of the proposed model when the probability maps are used as ground truth during model training. The experimental results on the FSC-147 and CARPK datasets demonstrate that the proposed PBECount can identify whether unknown class objects are similar to exemplars and outperform the state-of-the-art CAC methods in terms of accuracy and generalization.

Introduction

The main goal of the class-agnostic counting (CAC) (Gong et al. 2022) is to count objects of interest in images containing arbitrary unknown class objects based on a small number of exemplars. Compared to the class-specific counting (CSC) (Wei et al. 2024; Tyagi et al. 2023), which focuses on counting specific class objects, the CAC has broader application scenarios and greater potential value, making it a subject of increased research attention in recent years.

The CAC, which was first introduced in (Lu, Xie, and Zisserman 2019), is based on the extract-then-match paradigm. This paradigm first extracts features from both a query image and exemplars and then matches their feature similarity, as illustrated in Figure 1(a). Previous CAC studies have mostly followed the extract-then-match paradigm. However, the existing extract-then-match methods tend to count all objects in a query image without effectively assessing whether the objects are truly similar to the exemplars or not, as presented in Figure 1(c).

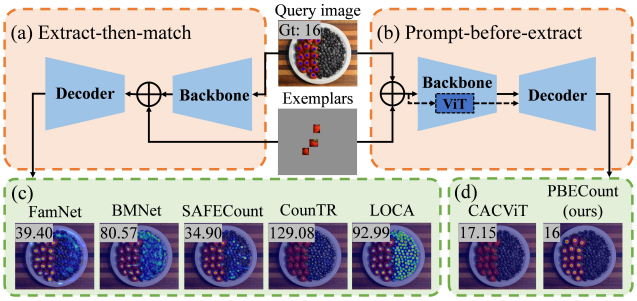

Figure 1: Comparison of the detection process and counting results of the models based on the extract-then-match and prompt-before-extract paradigms: (a) the detection process based on the extract-then-match paradigm; (b) the detection process based on the prompt-before-extract paradigm, and the dashed arrows represent the detection process based on the extract-and-match paradigm, which is a specific case of the prompt-before-extract paradigm; (c) the detection results of the models based on the extract-then-match paradigm; (d) the detection results of the models based on the prompt-before-extract paradigm.

Recently, an innovative extract-and-match CAC paradigm based on the vision transformer (ViT) has been proposed (Wang et al. 2024) . In this paradigm, exemplars and a query image are first cropped and concatenated and then together fed to the ViT. The first two self-attention layers of the ViT structure are used to extract features from both the query image and exemplars, whereas the subsequent two cross-attention layers perform the matching operation between the query image and exemplars. The experimental results in Figure 1(d) demonstrate that the CACViT, based on the extract-and-match paradigm, can filter out objects dissimilar to exemplars. However, this paradigm overly emphasizes the role of the ViT structure and cannot reveal the essence of the CACViT model to assess the similarity between the exemplars and objects.

This work reveals that the main reason the extract-and-match paradigm can help models recognize exemplar-similar objects does not lay in decoupling the ViT structure into feature extraction and feature matching modules but rather refers to the detection process of simultaneously

*Corresponding author

inputting the exemplars and a query image into a model. Particularly, as long as a query image is provided to the backbone together with exemplars in the feature extraction process, the model can accurately assess the similarity between the objects and exemplars. Based on this idea, this study proposes a more generalized CAC paradigm named the prompt-before-extract on top of the extract-and-match paradigm, as shown in Figure 1(b). Compared to the extract-and-match paradigm, the proposed prompt-before-extract removes the requirement for the ViT structure and separate feature-matching module in a model, which makes it applicable to a wider range of network architectures. Based on the proposed prompt-before-extract paradigm, an innovative pure convolutional neural network (CNN) CAC model named PBECount is designed. As presented in Figure 1(d), similar to the CACViT model, the PBECount model can accurately count exemplar-similar objects in a query image, verifying the prompt-before-extract paradigm's ability to assess the similarity between objects and exemplars.

Since the CAC task both emphasizes a model's counting performance and requires the model to be more sensitive to the similarity between objects and exemplars when identifying objects of the same class as exemplars, this study uses the probability maps that can express the detected object confidence better as ground truth during the PBECount training. Furthermore, this study proposes a dynamic threshold method to address the need to manually set thresholds for candidate objects obtained after performing non-maxima suppression on the probability maps in detection. An improved loss function named peak-aware MSE loss and an advanced post-processing method named loop optimization are used to optimize the network's counting performance.

The verification results indicated that the proposed PBECount can achieve excellent counting performance on the FSC-147 dataset (Ranjan et al. 2021), with MAE values of 8.88 and 7.71 on the validation and test sets, respectively. Moreover, it demonstrates superior cross-dataset generalization performance on the CARPK dataset (Hsieh, Lin, and Hsu 2017).

In summary, the main contributions of this work are as follows:

- An innovative CAC paradigm named prompt-before-extract, which counts only the exemplar-similar objects while completing the CAC task, is proposed; also, the reasons for which the proposed paradigm is effective are analyzed;
- A simple but powerful pure CNN model named PBECount, which can achieve the state-of-the-art counting performance on the FSC-147 dataset, is developed;
- A dynamic threshold method is introduced to address the need to manually set thresholds when probability maps are used as ground truth. In addition, peak-aware MSE loss and loop optimization are proposed to enhance the counting performance of PBECount.

Related Works

Class-Agnostic Counting. Compared to CSC (Yang et al. 2024), CAC tackles the counting problem by transforming it into a matching task between the exemplars and objects within query images containing unknown class objects. This approach helps to overcome the reliance on large-scale class-specific datasets during model training. The concept of CAC was initially proposed in (Lu, Xie, and Zisserman 2019), and the publication (Ranjan et al. 2021) introduced the CAC dataset FSC-147, which has become a benchmark for subsequent CAC studies. Previous CAC studies have mostly followed the extract-then-match paradigm. Some of the proposed methods extract features from a query image and then crop exemplar features from the image feature, and finally match the similarity between the query image feature and the exemplar features (Ranjan et al. 2021; Gong et al. 2022; Lin et al. 2022; ukić et al. 2023). Other methods first crop exemplars from a query image, extract features separately from the query image and exemplars and then perform the matching operation between them (Yang et al. 2021; Shi et al. 2022; Liu et al. 2022). Recent studies based on the extract-then-match paradigm has consistently improved CAC performance. However, many proposed methods with high counting accuracy prioritize counting all objects present in the query image rather than focusing on objects similar to the exemplars. This tendency leads to decreased performance in multi-class scenarios (Xu, Le, and Samaras 2023). Designing CAC methods that simultaneously achieve high counting accuracy and reliable classification capabilities has emerged as a new research challenge.

Class-Agnostic Counting in Multi-Class Scenes. To address the aforementioned limitation of classification performance decline in CAC methods with high counting accuracy, (Xu, Le, and Samaras 2023) proposed a two-stage method to refine the CAC model predictions. An additional segmentation network is employed to predict a segmentation mask for exemplar-similar regions, thereby filtering out non-exemplar-similar regions in the density map predicted by CAC model. Similarly, (Pelhan et al. 2024) uses non-maxima suppression on the density map predicted by CAC model to locate all predicted objects. Then, by computing the cosine similarity between exemplars and predicted objects, it filters out non-exemplar-similar objects, and this detection paradigm is named as detect-and-verify. However, both of these methods cannot solve this problem end-to-end and are susceptible to the influence of manual expertise and exemplar background information. (Wang et al. 2024) introduced a CAC paradigm named extract-and-match based on the ViT structure. It crops and concatenates a query image and exemplars, feeding them into ViT together, allowing the attention layers within ViT to simultaneously perform feature extraction and matching. This method can filter out objects dissimilar to exemplars, but it overly emphasizes the role of the ViT structure, and its filtering capability diminishes when non-exemplar-similar classes in the query image are present in the training data. On top of the extract-and-match paradigm, this work introduces an enhanced CAC paradigm named prompt-before-extract, which better identifies object-exemplar similarity and demonstrates stronger generalization ability in cross dataset scenarios.

PBECount for Class-Agnostic Counting

Prompt-Before-Extract Paradigm

This study defines the prompt-before-extract paradigm as follows: "Provide exemplars before inputting a query image into the backbone network and extract features jointly from merged query image and exemplars." The extract-then-match, extract-and-match, and prompt-before-extract paradigms can be respectively represented by $C = D(B(i), e)$, $C = D(V(i, e))$, and $C = D(B(i, e))$, where i and e represent the query image and exemplars. B and D denote the backbone and decoder, respectively; V signifies the ViT structure with $V \in B$, and C represents the counting result. Given that $V \in B$, it can be inferred that the extract-and-match paradigm is essentially a specific case of the prompt-before-extract paradigm.

Proposed Approach Overview

The detection pipeline of the PBECount is illustrated in Figure 2(a). The exemplars are first transformed into an exemplar probability map, which is then concatenated with a query image and input into a pure CNN ResNet34 U-Net model. Subsequently, the probability map predicted by the model's map head is processed using non-maxima suppression (NMS) to obtain the candidate objects. Finally, an optimal threshold obtained from the score head is applied to filter the candidate objects, providing the final count.

Input Data Preprocessing

The PBECount jointly inputs the exemplars and a query image into the model for feature extraction. Inspired by the approach presented in (Zhou, Koltun, and Krähenbühl 2020), in this study, the probability map is used to provide exemplar information to the model. Particularly, for a query image with dimensions $(3, H, W)$, first, an all-zero matrix with a size of $(1, H, W)$ is obtained. Next, in this matrix, a Gaussian kernel is created for each exemplar with a peak value of one, maintaining the same position and size, which yields the exemplar probability map. Finally, the three-channel query image is concatenated with the one-channel exemplar probability map to obtain the four-channel input data for the model. This method naturally preserves the position and shape information of exemplars and can realize CAC tasks with a highly concise model.

Network Structure

In the PBECount model, a U-Net model with the ResNet34 network as a backbone is used to extract features from query images. The detailed model architecture is presented in Figure 2(b). To enhance the model's recognition capability for objects of different scales, the proposed model adopts an additional feature fusion module. After the input data with a size of $(4, H, W)$ are fed to the U-Net model, an image feature with a size of $(64, H, W)$ is generated. This image feature is then fed to the map head and score head separately. In the map head, the image feature passes through a series of convolutional layers in the map conv block, yielding the predicted probability map with a size of $(1, H, W)$. Meanwhile, in the score head, the image feature undergoes another set of convolutional layers in the score conv block, yielding a score feature with a size of $(8, H/6, W/6)$. Finally, the score feature is scaled to $(8, 64, 64)$ and input into the score linear block, where two fully connected layers produce the predicted threshold score.

Dynamic Threshold

The PBECount model uses the probability maps as ground-truth data for model training. Typically, probability maps predicted by the model undergo non-maxima suppression to identify candidate objects. The values at the candidate objects' positions in the probability map represent their confidence scores. Next, a manually set threshold is applied to filter out objects with confidence scores below the threshold, resulting in the final count. However, manually setting the threshold often fails to provide optimal counting results, and an ideal threshold can vary across detection scenarios. To address the need to manually set the probability map thresholds, this study introduces a dynamic threshold method that allows neural networks to predict the threshold automatically based on the input data.

Particularly, the proposed dynamic threshold method introduces an additional score head to the PBECount model. This score head leverages image features extracted by the model to predict an optimal threshold score for each query image, as illustrated in Figure 2. The score head first reduces the feature dimensions through convolutional layers and then estimates the threshold score using fully connected layers. During the model training process, a ground truth threshold score S_{gt} is dynamically computed for each image based on a candidate object count C_{can} and a ground truth object count C_{gt}. If C_{can} is equal to or less than C_{gt}, S_{gt} is set to half of the confidence score of the lowest-confidence object among the candidate objects. Conversely, if C_{can} exceeds C_{gt}, all candidate objects are arranged in descending order of confidence scores, and S_{gt} is set to the average of the confidence scores of the C_{gt}th and $(C_{gt} + 1)$th objects. Setting S_{gt} in this manner ensures that the filtered final count closely matches C_{gt}, thereby enabling the model to predict threshold scores that yield more accurate final counts based on input data after training. During inference, the optimal threshold score predicted by the score head is used after applying a sigmoid function as a probability map threshold to filter candidate objects and obtain the final count results.

Loss Functions

The PBECount model uses the BCE with logits loss calculated between the predicted threshold score S_{pred} and S_{gt} for training the score head, which can be expressed as follows:

$$L_{BEC} = -(S_{gt} \times \ln \sigma(S_{pred}) + (1 - S_{gt}) \times \ln(1 - \sigma(S_{pred}))) \tag{1}$$

where σ is the sigmoid operation.

For the map head training, an innovative peak-aware MSE loss function is derived as follows:

$$L_{paMSE} = \| (M_{pred} - M_{gt}) \times \sqrt{M_w} \|_2^2 \tag{2}$$

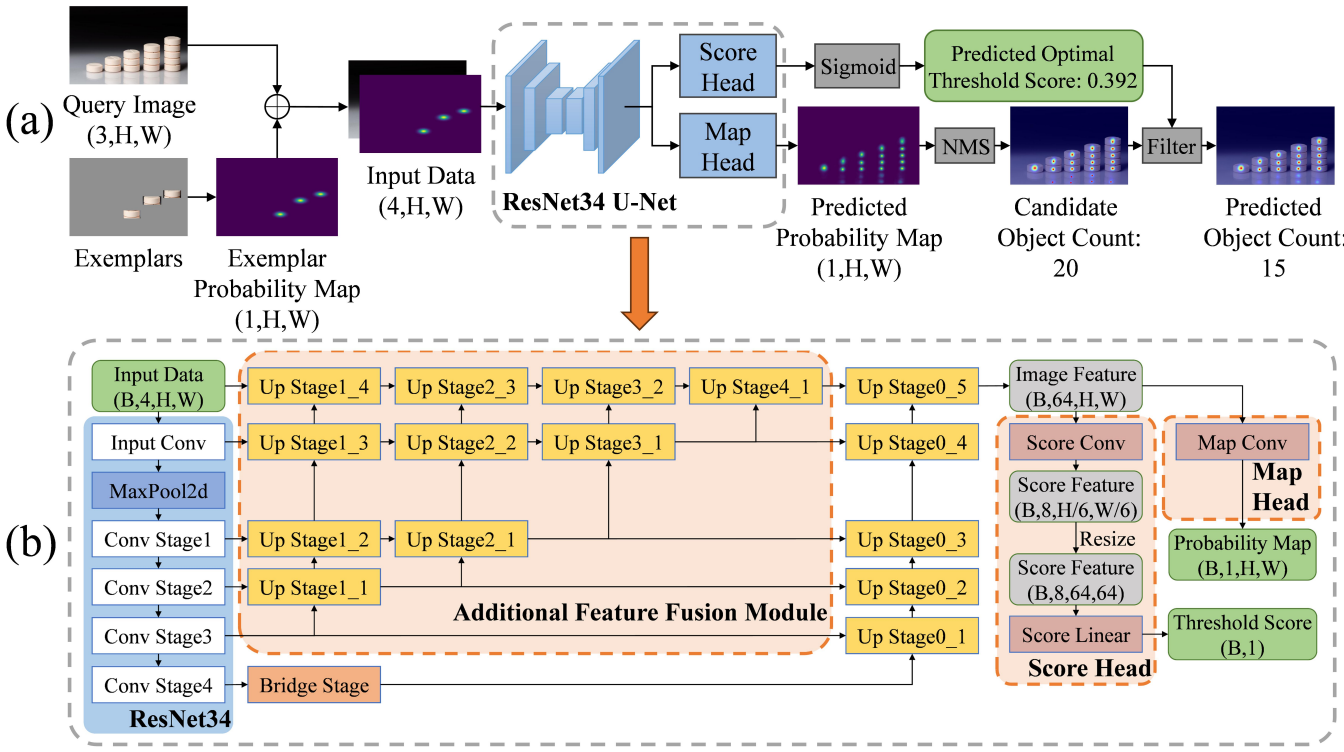

Figure 2: The PBECount framework: (a) the detection pipeline of the PBECount; (b) the architecture of the ResNet34 U-Net model employed in the PBECount.

where, M_{pred} and M_{gt} represent the predicted and ground truth probability maps, respectively; M_w is the weight map obtained by creating an all-one matrix with the same dimensions as M_{gt} first and then dividing it into patches with a size of 32×32. For each patch, the ground-truth object number within that patch is added to all matrix elements in that patch. The MSE loss with an additional weight emphasizes errors adjacent to objects, which have a greater impact on the count result than the errors in non-object regions when the probability maps are used as ground-truth data. The final loss is calculated by $L = L_{BEC} + L_{paMSE}$.

Evaluation Details

Exemplar Size Normalization Following (Pelhan et al. 2024) and (Liu et al. 2022), the size of exemplars is normalized during the inference phase of the proposed model. Particularly, all query images are initially scaled up by a factor of 1.5 and then resized again to ensure that the mean of the exemplars' width and height fall within the range of 80 and 15 pixels.

Loop Optimization In this study, a loop optimization method is proposed to aid the PBECount model in identifying exemplar-similar objects with certain feature differences from the exemplars. In the model testing phase, the counting results are recorded for each image at a fixed threshold of 0.1 and when using the predicted dynamic threshold. If the difference between the two counting results for an image exceeds p, it is considered that the image contains numerous objects with some similarity to the exemplars but with certain feature variations. For these images, the loop optimization is performed to enhance counting performance. Particularly, after model prediction, the distance from each object to its nearest exemplar is calculated. Then, the object farthest from the nearest exemplar is added to the exemplar set, and the model is re-run. Side lengths of the new exemplar on the x-axis and y-axis are respectively expressed as follows:

$$W_x = \min(W_{xm}, W_{xm} \times \sqrt{\frac{A_n}{A_m} \times \frac{D_m}{D_n}}) \tag{3}$$

$$W_y = \min(W_{ym}, W_{ym} \times \sqrt{\frac{A_n}{A_m} \times \frac{D_m}{D_n}}) \tag{4}$$

where, W_{xm} and W_{ym} represent the average lengths of the initial exemplars along the x-axis and y-axis, respectively; A_n is the area of the initial exemplar closest to the new exemplar, and A_m is the average area of the initial exemplars; D_m is the distance from the center point of all initial exemplars to the new exemplar, and D_n indicates the distance between the nearest initial exemplar and the new exemplar.

When the calculation results of q consecutive epochs do not exceed the maximum count result of the previous epochs, the iteration stops, and the detection result from the epoch with the highest predicted object count is considered the final count result. In the proposed method, p and q are set to 300 and seven, respectively.

Count Result Normalisation According to (Liu et al. 2022), in certain detection scenarios, the CAC methods tend to detect the smallest self-similarity unit of an object for counting rather than considering the entire object. To address this issue, this study performs the same count result normalization method in (Liu et al. 2022) during the testing phase. Particularly, the predicted objects are counted within the regions of each exemplar's Gaussian kernel in the exemplar probability map where its value exceeds 0.1. If the detected object count in an exemplar's region exceeds one, the image's counting result is changed to $Sum_p \div (Sum_e \div N_e)$, where Sum_p represents the sum of the predicted probability maps, Sum_e is the sum of the predicted probability maps in regions where the exemplar probability map exceeds 0.1, and N_e is the number of exemplars. In addition, although the probability maps can assign more accurate values to each detected object, thus achieving more precise counting results in most relatively sparse scenes, they might fail in extremely dense scenarios due to insufficient space for Gaussian kernel rendering. To address this issue, this study computes an average distance between the 25 nearest pairs of the predicted objects. When this average distance is less than 7.5, the count result normalization is also applied to the counting result.

Experiments

Implementation Details

To demonstrate the superiority of the PBECount model, as well as the ability of the proposed prompt-before-extract paradigm to assess the exemplar-object similarity, this study conducted experiments on the FSC-147 dataset (Ranjan et al. 2021), and adopted the mean absolute error (MAE) and root mean squared error (RMSE) as evaluation metrics.

Training Data Processing The PBECount model used the probability maps as ground-truth data for model training. In the probability maps, each object was represented by a square Gaussian kernel with a peak value of one. Given that the FSC-147 dataset provided only point-level localization information for objects, the side length of each object's Gaussian kernel was set to $W_g = 2\times\text{int}(4\times\sqrt{D_{no}}+0.5)+1$, where D_{no} represents the distance between the object and its nearest neighbor.

In the training process, data augmentation methods, including random flipping, rotation, scaling, cropping, and HSV enhancement, were employed. In addition, this study combined two customized data augmentation methods. First, an image was randomly concatenated with itself or another image containing objects from a different category. Then, one to four rectangular regions were randomly selected within the image, and the HSV enhancement, blurring, or grayscale operations were performed on these regions.

Training Details The proposed model was trained for 250 epochs using the AdamW optimizer with a weight decay of 10^{-4}. During the first 200 training epochs, the cosine annealing strategy was used to adjust the learning rate, starting from an initial value of 10^{-4} and gradually decreasing it to 10^{-5}. Subsequently, the model weight that achieved an optimal counting performance was fine-tuned using an additional fixed learning rate of 10^{-5} for 50 epochs. The training and testing processes of the proposed model were conducted on an NVIDIA GeForce RTX 4090D GPU with a batch size of one for approximately 25 hours.

Comparison with State-of-the-Art Methods

Quantitative Results The counting performance of the proposed PBECount was compared with other CAC methods on the FSC-147 dataset, as presented in Table 1. All methods were evaluated using the entire set of exemplars provided by the FSC-147 dataset during the testing process. The results indicated that the proposed PBECount achieved state-of-the-art counting performance, surpassing the other methods in terms of the MAE metric on both the validation and test sets. Particularly, the proposed PBECount demonstrated a 10.97% reduction in the MAE value on the test set compared to the state-of-the-art DAVE method, which indicated the reliability of the proposed method's counting performance.

Next, 66 images were selected from the validation set, and 33 images were selected from the test set of the FSC-147 dataset. These images all contained two or more object categories with a number larger than three, forming the FSC-147 Mul dataset. This dataset was used to verify the counting capability of the methods in multi-class scenes. The PBECount achieved excellent performance on the FSC-147 Mul and consistently outperformed all the methods, highlighting the proposed method's accuracy in distinguishing different object categories. Notably, the CounTR and LOCA methods exhibited significant performance degradation on the FSC-147 Mul dataset, even having worse performances than the SAFECount and BMNet+ methods on certain metrics. This result revealed the limitations of these two methods in multi-class scenes.

Performance Comparison for Different Object Densities To further investigate the counting performance of the proposed PBECount for images of different density levels, this study partitioned the validation and test sets of the FSC-147 dataset into eight subsets based on the number of objects in each image. The results are shown in Figure 3, where it can be seen that the PBECount achieved excellent counting performance across various object density scenarios, which confirmed the robustness of the proposed method.

Qualitative Analysis To validate the classification capability of the proposed PBCount, this study examined the detection results of the proposed PBECount and previous CAC methods on multi-class images, as shown in Figure 4. All images used in this analysis were selected from the validation and test sets of the FSC-147 dataset. The exemplars were selected to include both annotated categories (rows 3, 5, 6, and 7) and unannotated categories (rows 1, 2, and 4).

The PBECount method achieved high counting accuracy while surpassing the previous methods in distinguishing different object categories. In complex detection scenarios with multiple object categories, the proposed approach could accurately localize and count objects with varying densities,

Method	FSC-147 Val		FSC-147 Test		FSC-147 Mul Val		FSC-147 Mul Test	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
FamNet (Ranjan et al. 2021)	23.75	69.07	22.08	99.54	19.89	32.43	16.58	24.55
BMNet+ (Shi et al. 2022)	15.74	58.53	14.62	91.83	13.37	17.83	11.66	16.51
SAFECount (You et al. 2023)	15.28	47.20	14.32	85.54	8.66	12.33	7.93	11.27
CounTR (Liu et al. 2022)	13.13	49.83	11.95	91.23	15.45	27.09	9.98	17.76
LOCA (ukić et al. 2023)	10.24	32.56	10.79	56.97	11.09	17.39	6.82	12.85
CACViT (Wang et al. 2024)	10.63	37.95	9.13	48.96	6.00	8.22	5.74	9.90
DAVE (Pelhan et al. 2024)	8.91	**28.08**	8.66	**32.36**	-	-	-	-
PBECount (ours)	**8.88**	30.24	**7.71**	44.92	**4.39**	**5.72**	**4.86**	**9.16**

Table 1: Comparison with the state-of-the-art CAC methods on the FSC-147 dataset. The left part of the table presents the results on the FSC-147 dataset, and the right part of the table presents the FSC-147 Mul results.

Method	FSC-147 Val		FSC-147 Test	
	MAE	RMSE	MAE	RMSE
PBECount$_{\text{no CRN}}$	9.54	33.69	8.78	53.30
PBECount$_{\text{no LO}}$	8.93	**29.91**	9.39	77.72
PBECount$_{\text{no ESN}}$	12.72	56.77	11.07	97.33
PBECount$_{\text{no PA}}$	9.49	39.00	9.79	65.98
PBECount$_{\text{no DT}}$	9.93	32.25	9.89	63.14
PBECount	**8.88**	30.24	**7.71**	**44.92**

Table 2: The results of an ablation study of individual architectural components of the proposed model.

Method	MAE	RMSE
FamNet (Ranjan et al. 2021)	28.84	44.47
BMNet+ (Shi et al. 2022)	10.44	13.77
SAFECount (You et al. 2023)	16.66	24.08
LOCA (ukić et al. 2023)	9.97	12.51
CACViT (Wang et al. 2024)	8.30	11.18
PBECount	**6.27**	**9.08**

Table 3: Generalization performance of the CAC methods on the CARPK dataset.

regardless of whether their categories were included in the training data or not. Notably, the PBECount could also generate probability maps with high-fidelity object localization in multi-class scenes. The experimental results demonstrated that the proposed prompt-before-extract paradigm could enable the model to recognize exemplar-object similarity accurately, providing a significant advantage in solving exemplar-based CAC tasks.

Ablation Study

The ablation experiments were conducted on the FSC-147 dataset, analyzing the impacts of individual architectural components and peak-aware MSE loss on the counting performance of the PBECount. Using the complete version of the proposed method (PBECount) as a baseline, this study retrained the PBECount after removing different architectural components, and the obtained performances are presented in Table 2. The performance was reported by averaging a certain measure on the validation and test sets.

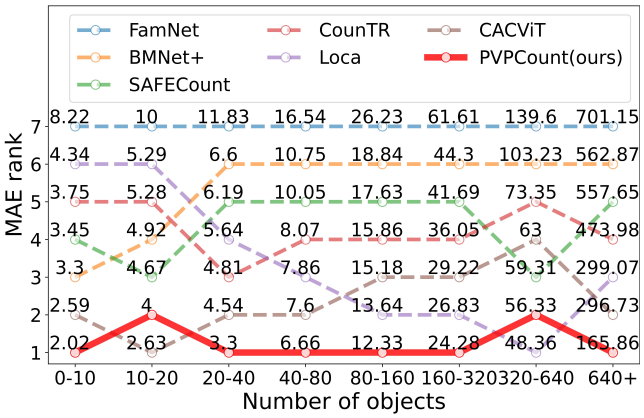

Figure 3: Counting performance of the CAC methods for images of different density levels.

To examine the importance of the dynamic threshold method for models trained with the probability maps as ground truth, this study removed the score head and manually set a fixed threshold that achieved the best counting performance (PBECount$_{\text{no DT}}$). This resulted in a 16.31% MAE performance decrease, which indicated that the dynamic threshold method could successfully predict the optimal threshold for different detection scenarios, enhancing the counting performance of models trained with probability maps. This study also evaluated the effect of the peak-aware MSE loss by replacing it with the regular MSE loss without additional weights (PBECount$_{\text{no PA}}$). This change led to a 13.84% MAE performance decrease, which confirmed the importance of assigning greater attention to errors adjacent to objects when calculating prediction losses during model training with probability maps.

Subsequently, the importance of the three testing strategies used in the testing phase was investigated. For this purpose, three variants, PBECount$_{\text{no ESN}}$, PBECount$_{\text{no LO}}$, and PBECount$_{\text{no CRN}}$, which respectively removed exemplar size normalization, loop optimization, and count result normal-

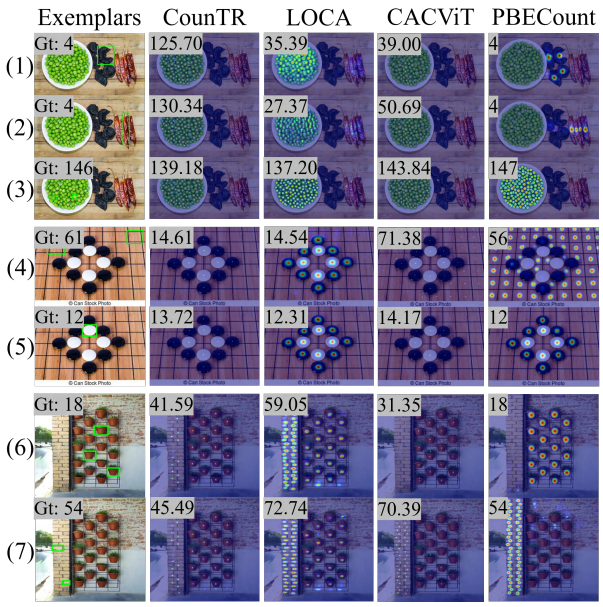

Figure 4: Qualitative results of different methods for three multi-class images from the FSC-147 dataset when objects of different categories were used as exemplars. The green bounding boxes indicate the provided exemplars. The proposed method outperformed the existing approaches in both object counting and localization performance for different object categories in the same image.

ization during the evaluation process, were constructed. The results of the three variants indicated a relative MAE performance decrease of 30.27%, 9.23%, and 9.55% compared to the complete version of the PBECount model. This demonstrated the effectiveness of the testing strategies adopted by the PBECount model in enhancing its counting performance. Based on the results, disabling exemplar size normalization had the most significant impact on the method performance, which highlighted the importance of standardizing exemplar sizes for achieving robust class-agnostic counting performance.

Cross-Dataset Generalization Analysis

Following the evaluation protocol established in (Ranjan et al. 2021), this study evaluates the cross-dataset generalization capability of the PBECount on the CARPK. The "cars" category was excluded from the FSC-147 dataset to avoid object class overlap between the training and test datasets. For counting purposes, three objects were randomly sampled from each image in the CARPK test set as exemplars. The obtained results are presented in Table 3, where it can be seen that the PBECount achieved significantly better cross-dataset generalization performance, achieving a 24.46% reduction in MAE and an 18.78% reduction in RMSE compared to the state-of-the-art method CACViT. These results highlighted the strong generality of the proposed method.

Dynamic Threshold Method Analysis

The mean optimal thresholds predicted by the PBECount model for different object categories in the validation and test sets of the FSC-147 dataset are presented in Figure 5. Among all considered categories, "marbles," "bottle caps," and "tree logs" had the highest mean optimal thresholds. This suggested that the proposed model tended to assign higher confidence scores when predicting these categories, indicating that they were more easily predictable by the proposed model. These categories shared certain characteristics, such as approximately circular shapes and smaller aspect ratios. Conversely, "skis," "shirts," and "books" achieved the lowest mean optimal thresholds among all categories, indicating that they were the most challenging classes for the proposed model to predict. These categories exhibited irregular object shapes characterized by significant aspect ratios and inconsistent textures. Analyzing the information obtained from the dynamic threshold method's predictions provided a better understanding of both the strengths and weaknesses of the PBECount in various detection scenarios. This valuable insight could guide future research in the CAC field and further optimize the proposed approach.

Prompt-Before-Extract Paradigm Effectiveness Analysis

For the prompt-before-extract paradigm, which exhibits outstanding performance in assessing the exemplar-object similarity, this study proposes several possible hypotheses for the reasons behind this.

Many existing CAC methods based on the extract-then-match paradigm employ the strategy of relying only on a decoder for downstream matching tasks. Although this strategy is suboptimal in the CAC tasks, it has shown notable effectiveness in unknown category object instance segmentation tasks (Kirillov et al. 2023; Songa et al. 2024; Ke et al. 2024), indicating that this paradigm is not inherently problematic. Compared to the instance segmentation of individual objects, discriminating the similarity between all objects within a query image and exemplars evidently represents a highly complex and challenging computer vision task. This study posits that this task might exceed the capabilities of lightweight decoders commonly used today.

Further, in the context of CAC tasks, where the same ground-truth data are used to represent different object categories, the feature differences between different category objects might gradually diminish as a model's layers deepen during the image feature extraction process. Consequently, the extract-then-match paradigm, which relies on deeper image features for exemplar-object matching, might struggle to distinguish similarities between distinct objects.

In conclusion, this study deduces that employing larger and more intricate network modules to conduct similarity assessments between objects and exemplars and initiating feature matching at shallow layers of the model contributes to the superior performance of the prompt-before-extract paradigm in the context of CAC tasks.

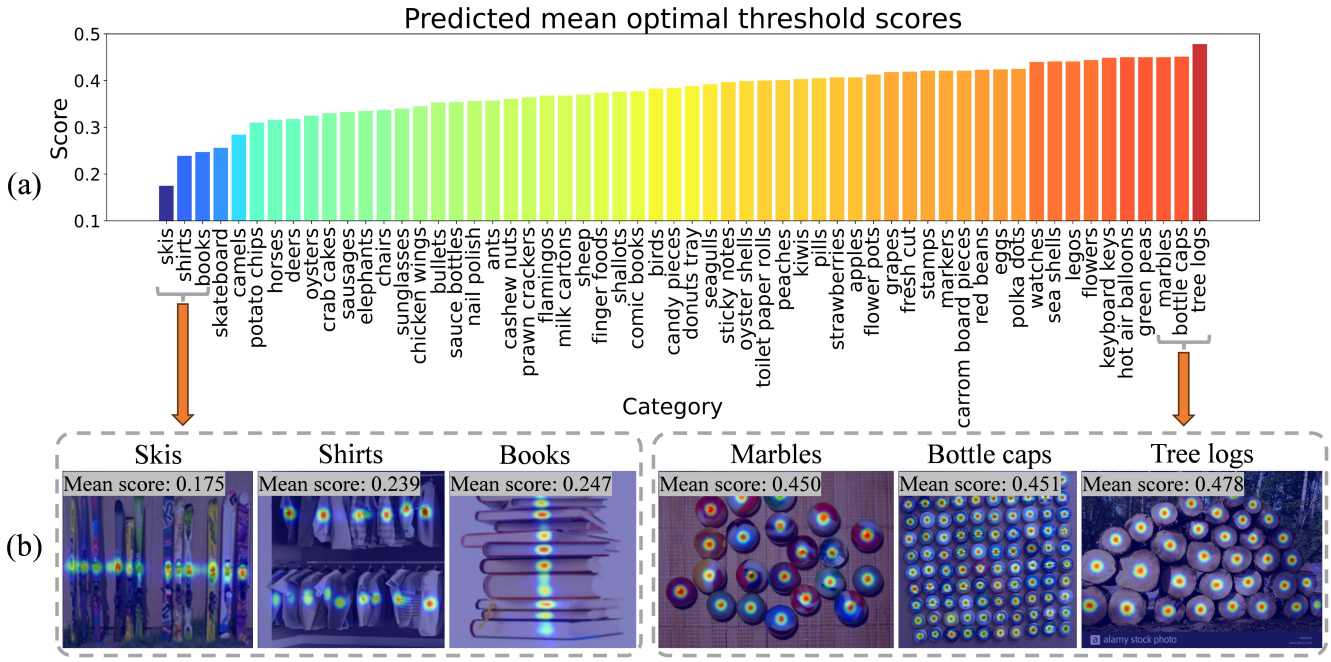

Figure 5: The predicted mean optimal threshold scores for objects of different categories: (a) the mean optimal threshold scores for objects across all categories in the validation and test sets of the FSC-147 dataset; (b) the categories with the maximum and minimum mean optimal threshold scores.

Conclusions

This study proposes an innovative prompt-before-extract paradigm for class-agnostic counting. This paradigm enables better discrimination between exemplars and objects in query images. Building upon this paradigm, this study introduces an advanced and simple but efficient pure CNN method named PBECount. In addition, the dynamic threshold method is proposed to address the requirement for manually setting the probability map detection thresholds. Further, a peak-aware MSE loss and loop optimization are designed to enhance the proposed method's counting accuracy. The verification results show that the proposed PBECount can achieve the state-of-the-art counting performance on the FSC-147 dataset, surpassing the existing methods in recognizing objects similar to exemplars. Finally, the effect of the dynamic threshold method for future CAC research is analyzed, and the effectiveness of the prompt-before-extract paradigm is discussed.

Acknowledgments

This research was funded by the National Natural Science Foundation of China grant No.61906126. Simultaneously, this research was also funded by Key R&D Program of Sichuan Province, China (2023YFG0115, 2023YFG0112), basic research projects of The Sichuan Provincial Industrial Development Fund: "Container Cloud Computing Infrastructure Platform", "Next-generation Big Data Visualization Based Decision-making and Analysis Platform" (Grant. 2023JB06), Cooperative Program of Sichuan University and Zigong, PR China (Grant. 2022CDZG-6).

References

Gong, S.; Zhang, S.; Yang, J.; Dai, D.; and Schiele, B. 2022. Class-agnostic object counting robust to intraclass diversity. In *European Conference on Computer Vision*, 388–403. Springer.

Hsieh, M.-R.; Lin, Y.-L.; and Hsu, W. H. 2017. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, 4145–4153.

Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2024. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lin, W.; Yang, K.; Ma, X.; Gao, J.; Liu, L.; Liu, S.; Hou, J.; Yi, S.; and Chan, A. B. 2022. Scale-Prior Deformable Convolution for Exemplar-Guided Class-Agnostic Counting. In *BMVC*, 313.

Liu, C.; Zhong, Y.; Zisserman, A.; and Xie, W. 2022. Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721*.

Lu, E.; Xie, W.; and Zisserman, A. 2019. Class-agnostic counting. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, 669–684. Springer.

Pelhan, J.; Zavrtanik, V.; Kristan, M.; et al. 2024. DAVE-A Detect-and-Verify Paradigm for Low-Shot Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23293–23302.

Ranjan, V.; Sharma, U.; Nguyen, T.; and Hoai, M. 2021. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3394–3403.

Shi, M.; Lu, H.; Feng, C.; Liu, C.; and Cao, Z. 2022. 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*, 9529–9538.

Songa, Y.; Pua, B.; Wanga, P.; Jiang, H.; Donga, D.; and Shen, Y. 2024. SAM-Lightening: A Lightweight Segment Anything Model with Dilated Flash Attention to Achieve 30 times Acceleration. *arXiv preprint arXiv:2403.09195*.

Tyagi, A. K.; Mohapatra, C.; Das, P.; Makharia, G.; Mehra, L.; AP, P.; et al. 2023. Degpr: Deep guided posterior regularization for multi-class cell detection and counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23913–23923.

ukić, N.; Lukežič, A.; Zavrtanik, V.; and Kristan, M. 2023. A low-shot object counting network with iterative prototype adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18872–18881.

Wang, Z.; Xiao, L.; Cao, Z.; and Lu, H. 2024. Vision transformer off-the-shelf: A surprising baseline for few-shot class-agnostic counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5832–5840.

Wei, X.-Y.; Zhang, L.; Ma, H.-Y.; and Zhang, X.-F. 2024. SCU-Counting: A large-scale benchmark dataset for multi-class object counting. *Transportation Research Part C: Emerging Technologies*, 163: 104608.

Xu, J.; Le, H.; and Samaras, D. 2023. Learning from Pseudo-labeled Segmentation for Multi-Class Object Counting. *arXiv preprint arXiv:2307.07677*.

Yang, C.; Geng, T.; Peng, J.; and Song, Z. 2024. Probability map-based grape detection and counting. *Computers and Electronics in Agriculture*, 224: 109175.

Yang, S.-D.; Su, H.-T.; Hsu, W. H.; and Chen, W.-C. 2021. Class-agnostic few-shot object counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 870–878.

You, Z.; Yang, K.; Luo, W.; Lu, X.; Cui, L.; and Le, X. 2023. Few-shot object counting with similarity-aware feature enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6315–6324.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2020. Tracking objects as points. In *European conference on computer vision*, 474–490. Springer.