

PPSNet: Perceptual Prior Similarity-guided Network for Class-agnostic Counting

Shengwei Jia

sswwjj@stumail.ysu.edu.cn

Yanshan University

Junhui Liu

Yanshan University

Jiahao Wang

Yanshan University

Yongqiang Cui

Yanshan University

Shihui Zhang

Yanshan University

Research Article

Keywords: class-agnostic counting, vision-guided exemplar counting, perceptual prior, similarity-guided weighting

Posted Date: August 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10614753/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

PPSNet: Perceptual Prior Similarity-guided Network for Class-agnostic Counting

Shengwei Jia¹, Junhui Liu¹, Jiahao Wang¹, Yongqiang Cui^{1,2},
Shihui Zhang^{1,2*}

¹School of Artificial Intelligence (School of Software), Yanshan University,
West of Hebei Street No.438, Qinhuangdao, 066000, Hebei, China.

²Hebei Key Laboratory for Computer Virtual Technology and System
Integration, Qinhuangdao, 066000, Hebei, China.

*Corresponding author(s). E-mail(s): sshhzz@ysu.edu.cn;

Contributing authors: sswwjj@stumail.ysu.edu.cn;

jhliu@stumail.ysu.edu.cn; W1085405029@163.com; hbyd_yqc@sina.com;

Abstract

Class-agnostic counting (CAC) aims to count objects of any class with only a few prompts or even no prompts at all. Existing methods have made remarkable progress by using visual exemplars as prompts, yet this paradigm requires bounding boxes to annotate the target exemplars for counting. The inherent geometric constraints of bounding boxes inevitably make it difficult to precisely fit the actual contours of the exemplars, which leads to the unavoidable inclusion of irrelevant background information in the selected visual exemplars. This background noise may induce pseudo-activations as the network depth increases, which further undermines the matching between visual exemplars and the query image. Furthermore, although existing work adopts the Extract-and-Match paradigm to perform feature extraction and similarity modeling for both query and visual example features, no work analyzes its superiority from a theoretical perspective. Therefore, in this paper, we theoretically analyze the Extract-and-Match paradigm and propose a novel Perceptual Prior Similarity-guided Network (PPSNet) for class-agnostic counting. Specifically, PPSNet incorporates two elaborately designed modules: Perceptual Prior Construction Module (PPCM) and Similarity-guided Weighting Module (SWM). PPCM constructs a perceptual prior by leveraging scale information to emphasize the distribution of visual exemplars within the regions selected via bounding boxes, and it also alleviates the loss of scale information in visual exemplars caused by the resizing operation. SWM leverages the

principle of contrastive learning to perform patch-level supervision on the similarity map between visual exemplars and the query image by constructing a foreground mask of visual exemplars. It then weights the query features with this similarity map to enhance the response of query features to the distribution of visual exemplars. In addition, we extend PPSNet into a general counting model, which can complete class-agnostic counting with visual prompts, text prompts, or no prompts at all. Extensive experiments conducted on the FSC-147, CARPK, and COCO datasets demonstrate that our proposed method achieves superior performance. Code has been released in <https://github.com/Jissaww/PPSNet>.

Keywords: class-agnostic counting, vision-guided exemplar counting, perceptual prior, similarity-guided weighting

1 Introduction

Object counting aims to estimate the number of objects in a given image. Existing mainstream methods typically design models tailored to specific classes, such as crowds [1–3], vehicles [4, 5], or cells [6]. While these class-specific counting models achieve excellent performance in their dedicated scenarios, their architectures and training paradigms are inherently constrained by class-specific prior knowledge. When counting objects of novel classes, researchers must collect new large-scale datasets and train models from scratch. This process consumes substantial human annotation costs and computational resources, while severely limiting model generalization capabilities. Consequently, such methods struggle to adapt to the diverse counting demands of open-world scenarios.

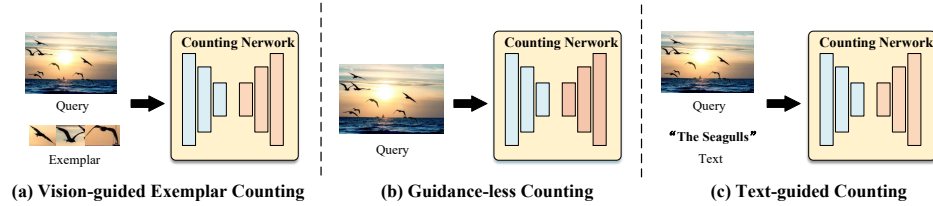


Fig. 1 Three paradigms of class-agnostic counting: (a) Vision-guided Exemplar Counting selects a few visual exemplars using bounding boxes to specify the counting class and achieves higher accuracy than other paradigms; (b) Guidance-less Counting requires no class specification and typically counts dominant objects in images, which presents certain limitations; (c) Text-guided Counting establishes connections between cross-modal features and uses textual descriptions to specify the counting class.

Class-Agnostic Counting (CAC) emerges to address this limitation, enabling counting of arbitrary classes using a single pre-trained model with minimal or no prompts. The advent of CAC overcomes the constraints of class-specific object counting. As shown in Fig. 1, CAC can be divided three paradigms: (a) Vision-guided Exemplar

Counting, (b) Guidance-less Counting, and (c) Text-guided Counting. Specifically: (a) Vision-guided Exemplar Counting [7–10] provides explicit visual prompts by selecting exemplars of counting classes in the query image using a few bounding boxes; (b) Guidance-less Counting [11–13] requires no external prompts and relies solely on direct inference of the pre-trained model on the query image; (c) Text-guided Counting [14–16] leverages formatted textual descriptions as semantic class prompts, where the standard format is “the {class name}”. However, the guidance-less counting and text-guided counting methods underperform the vision-guided exemplar counting method due to their inability to access the appearance details of vision exemplars and the presence of feature misalignment across modalities.

Current vision-guided exemplar methods primarily adopt the Extract-then-Match paradigm [10, 17–19], as illustrated in the Fig. 2. This paradigm performs independent feature extraction on query images and visual exemplars, neglecting the bidirectional interaction mechanism between them during feature extraction. Consequently, the extracted query features lack target awareness, causing subsequent feature matching to suffer from significant background interference and yielding low-discriminability results. Moreover, the Extract-then-Match paradigm requires an additional feature matching stage, introducing redundant design that contradicts the trend toward task-agnostic foundation models [20]. In contrast, as shown in the Fig. 3, the Extract-and-Match paradigm [20] resolves these limitations by jointly encoding query images and visual exemplars at the input stage, enabling synchronous optimization of feature extraction and relational modeling. This design enhances target-aware feature learning and suppresses background interference. In Section 3.1, we theoretically derive the advantages of this paradigm, demonstrating how it overcomes the lack of bidirectional information flow guidance in the Extract-then-Match paradigm, endows query features with target awareness, and effectively improves feature discriminability.

In addition, existing vision-guided exemplar counting methods typically select visual exemplars using bounding boxes. As shown in Fig. 4, due to the inherent geometric constraints of bounding boxes, although they can fully enclose the exemplars, they often cannot precisely fit the irregular contours of visual exemplars. As a result, the visual exemplars input to the model inevitably contain partial background regions. As the network depth increases, the potential background noise attached to the visual exemplars is gradually activated during feature extraction, which eventually leads to mismatches between the query image and background noise in the similarity modeling stage. We argue that this issue of background noise pseudo-activation weakens the model’s ability to perceive target visual exemplars. Similarly, to feed visual exemplars into the model, current methods commonly adopt the resize operation to forcibly adjust visual exemplars of different shapes into fixed-size feature vectors [10, 18, 21, 22]. This causes the loss of the inherent scale information of visual exemplars, thereby affecting the counting accuracy of the model.

To address the aforementioned issues, we propose a novel Perceptual Prior Similarity-guided Network (PPSNet) based on the Extract-and-Match paradigm for class-agnostic counting. The network contains two core modules: Perceptual Prior Construction Module (PPCM) and Similarity-guided Weighting Module (SWM). Specifically, PPCM constructs a gaussian prior using the original dimensions of visual

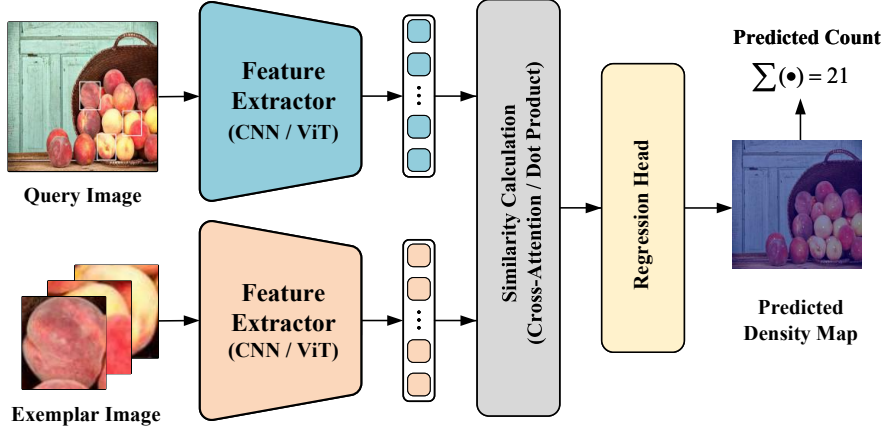


Fig. 2 Overview of the Extract-then-Match paradigm. The Extract-then-Match paradigm follows this workflow: First, it extracts features from the query image and visual exemplars separately; second, it models similarity between the query features and exemplar features; finally, it regresses a density map to obtain the final predicted density map.

exemplars to characterize their central distribution within bounding boxes. This design prevents interference from background noise pseudo-activation during feature extraction and similarity modeling, enhancing the robustness of key target matching. SWM computes the cosine similarity between query image features and visual exemplar features to generate a globally supervised similarity map. It then weights the query features using this similarity map to enhance feature response intensity in regions matching the visual exemplar features. Extensive experiments on the FSC-147, CARPK, and COCO datasets demonstrate that our proposed PPSNet achieves superior counting performance.

In a nutshell, our contributions can be summarized as follows:

- We propose a Perceptual Prior Similarity-guided Network (PPSNet) based on the Extract-and-Match paradigm for class-agnostic counting, and provide theoretical analysis of the optimization mechanism of this paradigm. Our well-designed PPSNet enhances the model’s discriminative capability for visual exemplars in query images by leveraging prior information and similarity map weighting, effectively mitigating background noise interference with counting performance in vision-guided exemplar counting tasks.
- We discuss the background noise pseudo-activation problem arising from using visual exemplars as prompts in vision-guided exemplar counting, and propose Perceptual Prior Construction Module (PPCM) to enhance the model’s perception of visual exemplar distribution within bounding boxes. To the best of

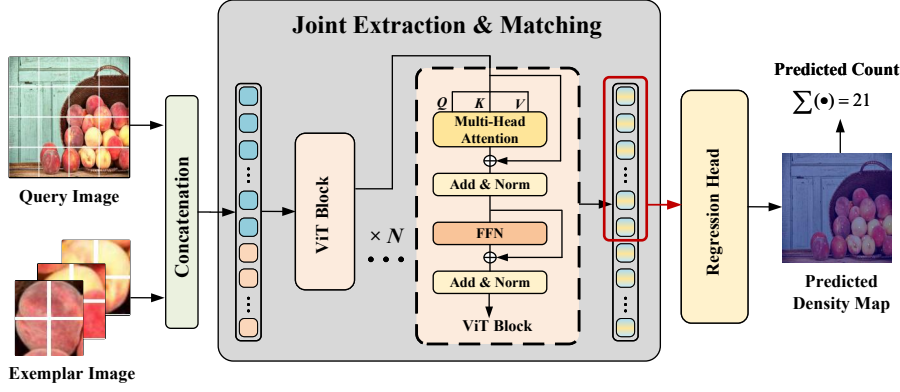


Fig. 3 Overview of the Extract-and-Match paradigm. The Extract-and-Match paradigm simultaneously performs feature extraction and similarity modeling on both the query image and visual exemplars. It then regresses a density map to obtain the final predicted density map.

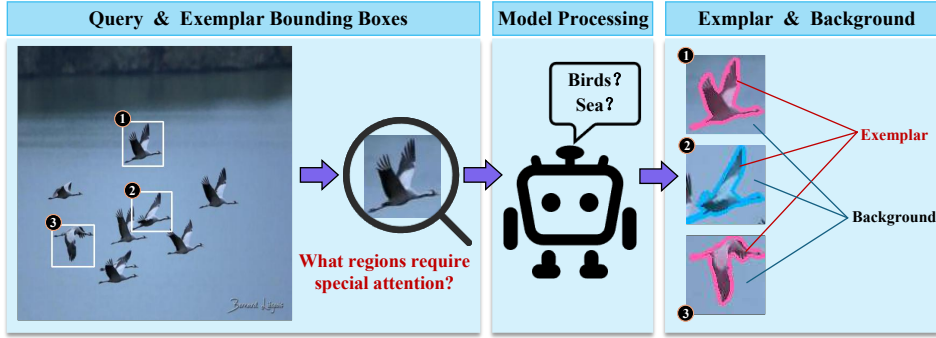


Fig. 4 The background noise pseudo-activation problem caused by bounding box annotations in class-agnostic counting.

our knowledge, our study is the first work to address the non-negligible impact of background noise caused by the inherent geometric constraints of bounding boxes on counting accuracy during feature extraction and similarity modeling.

- We propose Similarity-guided Weighting Module (SWM), a novel weighting module based on cosine similarity. SWM constructs a foreground mask for visual exemplars and leverages contrastive learning principles to supervise the cosine similarity map between query features and visual exemplar features at the patch-level. It then uses this similarity map to weight query features, enhancing feature responses in target regions of the query image.

2 Related Works

In recent decades, class-specific counting [1, 2, 4] has achieved excellent performance. However, this task faces significant limitations in open-world applications. The emergence of class-agnostic counting [7, 23] addresses these limitations and has attracted considerable research attention. Based on the form of prompt information, CAC can be categorized into three paradigms: (a) vision-guided exemplar counting; (b) guidance-less counting; (c) text-guided counting.

2.1 Vision-guided Exemplar Counting

Vision-guided exemplar counting represents the initial paradigm for CAC, using visual exemplars selected by bounding boxes as prompts to perform counting. GMN [23] formulates CAC as a matching problem by leveraging the self-similarity within images to count given exemplars. To address the lack of dedicated datasets for CAC, Ranjan et al. [7] introduce the FSC-147 dataset and propose FamNet. FamNet accomplishes CAC through three sequential stages: Feature Extraction, Feature Correlation, and Density Prediction. However, pre-defined similarity measurement rules often yield unsatisfactory results, e.g., inner product [7]. Consequently, Shi et al. [8] and You et al. [9] optimize similarity map computation methods to produce clearer and more discriminative similarity maps. Unlike previous methods, Liu et al. [10] introduce Vision Transformers (ViT) [24] into CAC to replace convolutional structures for query feature extraction and employ cross-attention for similarity modeling. LOCA [21] iteratively generates exemplar prototypes using both appearance information and shape information of visual exemplars to enhance feature representation capability. DAVE [25] improves counting accuracy by adding a detection stage to LOCA and removing outliers. Unlike existing methods that select visual exemplars using bounding boxes, Huang et al. [26] and Zhai et al. [27] use novel point prompts for counting and reduce the need for annotations. In summary, these methods follow the Extract-then-Match paradigm, which neglects the mutual interaction and guidance of early features. Wang et al. [20] propose the Extract-and-Match paradigm to enhance the model’s early perception of exemplars. In this paper, we also employ the Extract-and-Match paradigm and theoretically analyze its advantages.

2.2 Guidance-less Counting

Vision-guided exemplar counting requires manual annotation of a few bounding boxes, and the quality of visual exemplar prompts heavily depends on these annotations. Consequently, researchers explore performing CAC without any prompts. Ranjan et al. [11] first propose RepRPN-Counter to perform counting by estimating the frequency of target occurrences in images. Hobley et al. [28] formulate guidance-less counting as a repetitive recognition task and employ ViT to directly regress counts from input images. To enhance model generalization, GCNet [12] operates without prompts using a siamese architecture with a carefully designed self-similarity learning strategy for automatic object counting. GCA-SUN [13] employs gating mechanisms to eliminate irrelevant features and significantly improves counting accuracy in the guidance-less counting paradigm. To reduce the dependence on bounding boxes, in this paper, we

also attempt to extend from vision-guided exemplar counting to the guidance-less counting paradigm.

2.3 Text-guided Counting

Existing guidance-less counting methods typically count only dominant objects in images and cannot target arbitrary classes, motivating researchers to leverage text prompts. Xu et al. [14] first employ text prompts for counting. By incorporating text prompts to generate exemplar prototypes, they convert text-guided counting into vision-guided exemplar counting. Unlike this method, CounTX [15] directly computes cross-attention between text features and query features without constructing visual exemplars, regressing the result to a predicted density map. Kang et al. [16] argue that generating exemplar prototypes from text prompts is vulnerable to erroneous information. Consequently, they propose VLCounter for direct text-based counting. However, existing methods typically rely on CLIP [29] to align text and image features. Researchers observe that CLIP lacks sensitivity to textual prompts. Therefore, CountDiff [30] and T2ICount [31] both employ diffusion models to leverage the diffusion models’ rich prior knowledge and fine-grained visual understanding capabilities for text-guided counting. To ensure the generality of our model, we also extend it to the text-guided counting paradigm.

In conclusion, existing class-agnostic counting methods fall into three paradigms. Although guidance-less counting and text-guided counting offer simpler annotation workflows, their counting accuracy remains inferior to vision-guided exemplar counting. Consequently, we focus on advancing the vision-guided exemplar counting paradigm. As previously discussed, while existing methods have introduced the Extract-and-Match paradigm, rigorous theoretical analysis of its advantages remains lacking. Furthermore, the inherent geometric constraints of bounding boxes inevitably introduce background noise into visual exemplars, causing pseudo-activation of background regions during feature extraction and similarity modeling, which degrades counting performance. To tackle these limitations, we provide theoretical analysis of the Extract-and-Match paradigm and propose PPSNet. Our framework enhances vision-guided exemplar counting accuracy by constructing perceptual priors to emphasize exemplar distribution patterns and applying image-exemplar similarity maps to strengthen feature responses in target regions.

3 Method

This Section presents the Perceptual Prior Similarity-guided Network (PPSNet) for class-agnostic counting. We first theoretically analyze the potential advantages of the Extract-and-Match paradigm (Section 3.1). Then, we implement a basic method (EaMethod) to illustrate the fundamental workflow of the Extract-and-Match paradigm (Section 3.2). Finally, we bring two improvements on top of EaMethod to propose PPSNet. Specifically, we utilize perceptual priors to model the distribution of visual exemplars (Section 3.3) and enhance feature responses in target regions through supervised similarity map weighting (Section 3.4). The overall architecture of our proposed PPSNet is illustrated in Fig. 5.

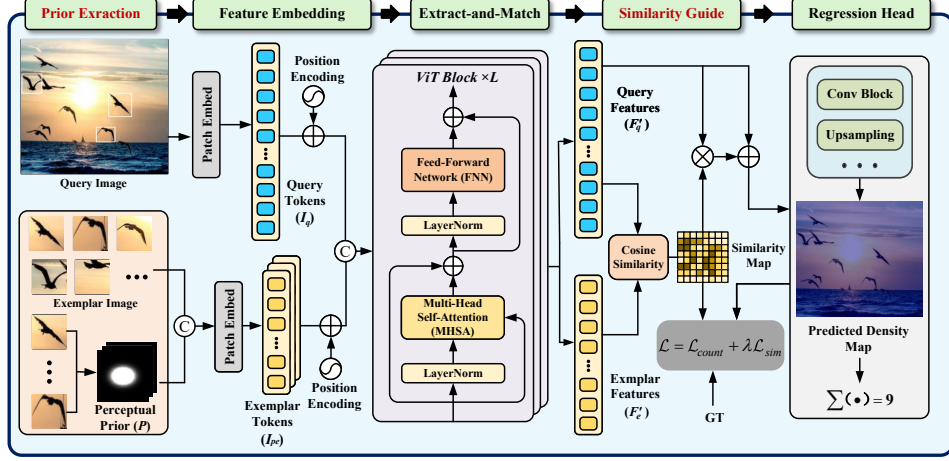


Fig. 5 Overview of Perceptual Prior Similarity-guided Network (PPSNet). PPSNet adopts the Extract-and-Match paradigm and integrates two core modules: Perceptual Prior Construction Module (PPCM) and Similarity-guided Weighting Module (SWM). PPCM constructs perceptual priors to capture the distribution of visual exemplars within bounding boxes. SWM enhances feature responses in target regions through patch-level supervised similarity map weighting. The weighted features then pass through a regression head to generate the final predicted density map.

3.1 Paradigm Analysis

For the Extract-and-Match paradigm, it concatenates the query features $F_q \in \mathbb{R}^{N_q \times d}$ and visual exemplar features $F_e \in \mathbb{R}^{N_e \times d}$ to obtain the concat features $F \in \mathbb{R}^{(N_q+N_e) \times d}$. Applying self-attention to the concat features simultaneously accomplishes feature extraction and similarity modeling. We project the concatenated features into $Q, K, V \in \mathbb{R}^{(N_q+N_e) \times d}$:

$$F = \begin{bmatrix} F_q \\ F_e \end{bmatrix} \quad (1)$$

$$Q = FW_Q = \begin{bmatrix} Q_q \\ Q_e \end{bmatrix}, K = FW_K = \begin{bmatrix} K_q \\ K_e \end{bmatrix}, V = FW_V = \begin{bmatrix} V_q \\ V_e \end{bmatrix}, \quad (2)$$

For the self-attention mechanism, as shown in Eq. (3):

$$Attn(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V = Softmax(S)V = AV \quad (3)$$

Substituting the concatenated features into Eq. (3), we obtain:

$$S = \frac{QK^T}{\sqrt{d}} = \frac{1}{\sqrt{d}} \begin{bmatrix} Q_q \\ Q_e \end{bmatrix} \begin{bmatrix} K_q^T & K_e^T \end{bmatrix} = \frac{1}{\sqrt{d}} \begin{bmatrix} Q_q K_q^T & Q_q K_e^T \\ Q_e K_q^T & Q_e K_e^T \end{bmatrix} \quad (4)$$

$$A = \text{Softmax}(S) = \begin{bmatrix} A_{qq} & A_{qe} \\ A_{eq} & A_{ee} \end{bmatrix} \quad (5)$$

$$\text{Attn}(Q, K, V) = AV = \begin{bmatrix} A_{qq} & A_{qe} \\ A_{eq} & A_{ee} \end{bmatrix} \begin{bmatrix} V_q \\ V_e \end{bmatrix} = \begin{bmatrix} A_{qq}V_q + A_{qe}V_e \\ A_{eq}V_q + A_{ee}V_e \end{bmatrix} \quad (6)$$

As shown in Eq. 6, in the Extract-and-Match paradigm, the concatenated features F simultaneously accomplish feature extraction and similarity modeling. After updating, the query features become $F'_q = A_{qq}V_q + A_{qe}V_e$ and the exemplar features become $F'_e = A_{eq}V_q + A_{ee}V_e$. It is evident that bidirectional information interaction exists between query features and visual exemplar features during each layer's update process. In contrast to the Extract-then-Match paradigm, which typically extracts query and exemplar features independently first and only enables feature interaction in the subsequent similarity modeling stages, the Extract-and-Match paradigm enables true bidirectional mutual feature refinement. Due to inherent limitations of cross-attention in Extract-then-Match, query feature updates contain only the $A_{qe}V_e$ term while exemplar features remain unchanged (lacking the $A_{eq}V_q$ term). Furthermore, the bidirectional information flow during feature extraction allows the model to inject the $A_{qe}V_e$ term into query features from the initial stage, enabling query features to capture target-relevant information from the beginning. This establishes stronger target awareness that progressively intensifies through iterative layer propagation.

3.2 Extract-and-Match Workflow

Overall Workflow. We implement a basic method EaMethod based on the Extract-and-Match paradigm rather than the prevalent Extract-then-Match paradigm. Given a query image $I_q \in \mathbb{R}^{3 \times H_q \times W_q}$ and visual exemplar images $I_e \in \mathbb{R}^{n \times 3 \times H_e \times W_e}$, we flatten them into sequences and project them into the feature space to obtain query features $F_q \in \mathbb{R}^{N_q \times d}$ and visual exemplar features $F_e \in \mathbb{R}^{N_e \times d}$. We concatenate the query features F_q and visual exemplar features F_e to obtain the concatenated features $F \in \mathbb{R}^{(N_q+N_e) \times d}$. The concatenated features F are fed into a feature processor $\psi(\cdot)$ for joint feature extraction and similarity modeling to obtain correlated features $F' \in \mathbb{R}^{(N_q+N_e) \times d}$. We extract the corresponding query feature portion $F'_q \in \mathbb{R}^{N_q \times d}$ from the correlated features and feed it into a regression head $\phi(\cdot)$ to obtain the final predicted density map $D_{pred} \in \mathbb{R}^{H \times W}$. The specific process is shown in the following formula:

$$F' = \psi([F_q; F_e]) \quad (7)$$

$$D_{pred} = \phi_{reg}(F'_q) \quad (8)$$

Counting Loss. We employ the conventional MSE loss to supervise model training.

$$\mathcal{L} = \|D_{pred} - D_{gt}\|^2 \quad (9)$$

where D_{gt} denotes the ground truth density map.

3.3 Perceptual Prior Construction Module

Vision-guided exemplar counting requires bounding boxes to select a few visual exemplars. Due to the inherent geometric constraints of bounding boxes, they can only enclose regions containing target objects but fail to precisely align with object contours. Consequently, the extracted visual exemplars inevitably contain background noise. As network depth increases, this background noise becomes progressively activated during feature extraction and similarity modeling, resulting in erroneous matching between visual exemplars and target objects in query images. Fig. 4 shows that the bounding box region contains both the target bird and background ocean, demonstrating that model inputs are contaminated visual exemplars rather than pure foreground instances. In addition, as shown in Fig. 6, we observe that the manually annotated bounding boxes are of varying sizes. To enable valid input of visual exemplars into the network, mainstream existing methods universally rescale these exemplars to a fixed spatial dimension. When the visual exemplars are elongated objects, this resizing operation distorts their native aspect ratios. Such geometric distortion erodes the intrinsic structural information of the visual exemplars, may mislead the model’s scale perception of the exemplars, and thus degrades the accuracy of the final counting results.

To address the aforementioned issues, we propose a Perceptual Prior Construction Module (PPCM), which constructs a perceptual prior from visual exemplars to characterize their distribution. Specifically, we first compute the standard deviation of each visual exemplar in horizontal and vertical directions, as shown in Eq. 10:

$$\sigma_x^k = w^k R_0, \sigma_y^k = h^k R_0 \quad (10)$$

where h^k, w^k denote the normalized scales of the k -th visual exemplar. $R_0 = H_e/2 = W_e/2$. H_e, W_e denote the scales of the resized visual exemplar.

Then, for each pixel coordinate (i, j) in every visual exemplar, we compute the relative displacement and construct a gaussian prior:

$$\Delta x(i) = i - c_x, \Delta y(j) = j - c_y \quad (11)$$

$$D_k(i, j) = \frac{\Delta x(i)^2}{2(\sigma_x^k)^2} + \frac{\Delta y(j)^2}{2(\sigma_y^k)^2} \quad (12)$$

$$G_k(i, j) = \exp(-D_k(i, j)) \quad (13)$$

where c_x, c_y denote the center coordinates of the resized visual exemplar, $D_k(i, j)$ denotes the normalized squared distance from pixel (i, j) to the center in the k -th visual exemplar (larger $D_k(i, j)$ indicates greater distance from the center), and $G_k(i, j)$ denotes the perceptual prior constructed based on this distance, whose value gradually decreases as (i, j) moves away from the visual exemplar center.

If a visual exemplar has a larger size, fewer exemplars can be accommodated within the same unit area. Based on this prior knowledge, we consider the area of visual exemplars to be approximately inversely proportional to their distribution density per

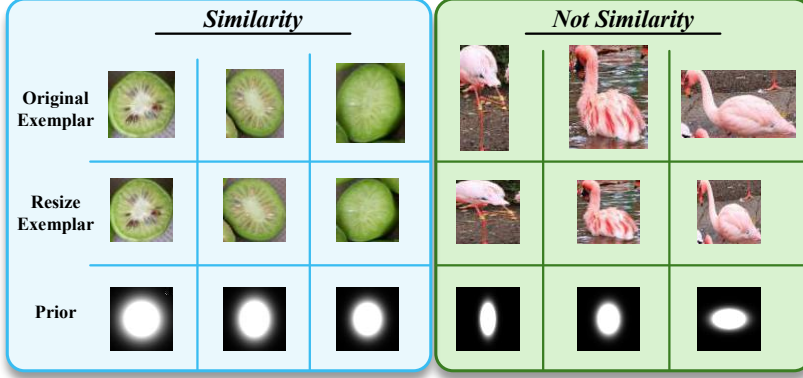


Fig. 6 Visualizing visual exemplar images and perceptual priors before and after resizing. We consider two types of cases: (1) visual exemplars with similar aspect ratios among them, where the resize operation has minimal impact on their aspect ratios. (2) visual exemplars with dissimilar aspect ratios, where the resize operation distorts their aspect ratios causing geometric deformation. Our perceptual prior implicitly encodes the aspect ratios of visual exemplars, thereby compensating for the losses caused by the resize operation. Simultaneously, our perceptual prior reflects the distribution of visual exemplars within bounding boxes, preventing excessive activation of background noise.

unit area. Therefore, we modulate the strength of the perceptual prior through an area coefficient, as shown in Eq. 14:

$$P_k(i, j) = \frac{1}{\sqrt{w^k h^k}} G_k(i, j) \quad (14)$$

where $P_k(i, j)$ denotes the final perceptual prior.

To integrate the perceptual prior constructed by PPCM into EaMethod, we concat the perceptual prior with visual exemplar images to obtain prior-enhanced exemplars $I_{pe} \in \mathbb{R}^{n \times 4 \times H_e \times W_e}$. The perceptual prior encodes the distribution of visual exemplars within bounding boxes, enabling the model to distinguish foreground exemplars from background noise in bounding boxes. This alleviates background noise activation during subsequent feature extraction and similarity modeling, as well as erroneous matching with targets in query images. Additionally, since the original height and width of visual exemplars are explicitly used when constructing the perceptual prior, it implicitly encodes aspect ratio information, mitigating geometric distortion issues. As shown in Fig. 6, for approximately square-shaped visual exemplars, the resize operation has minimal impact on their aspect ratios. However, elongated visual exemplars are often stretched into near-square shapes after resizing. In such cases, our constructed perceptual prior implicitly compensates for their original aspect ratio information, reducing the loss of structural information in visual exemplars.

3.4 Similarity-guided Weighting Module

Existing class-agnostic counting methods typically employ counting loss as supervision during training. However, for vision-guided exemplar counting, we consider its essence as a matching problem between visual exemplars and different targets in query images. The similarity between visual exemplar features and query features can serve as a metric for matching quality. Therefore, we propose Similarity-guided Weighting Module (SWM) to explicitly supervise the similarity map between visual exemplar features and query features, accurately measuring their matching quality. Furthermore, we leverage this similarity map to weight query features, thereby enhancing feature responses in target regions as illustrated in Fig. 5.

We apply SWM to the query features F'_q and visual exemplar features F'_e after the self-attention mechanism in EaMethod. First, we compute the cosine similarity between these two features to obtain the similarity map S , then use this similarity map to weight the query features, as shown in the following formula:

$$S(F'_q, F'_e) = \frac{F_q'^T F_e'}{\|F_q'\|_2 \|F_e'\|_2} \quad (15)$$

$$F'_q = F'_q(1 + S(F'_q, F'_e)) \quad (16)$$

How should we supervise the similarity map S ? We construct patch-level supervision signals from the ground truth (GT) density map. Specifically, the size of the similarity map N_q is $1/r$ of the GT density map size. Thus, each pixel in the similarity map corresponds to an $r \times r$ pixel region in the GT density map. If any object exists within these $r \times r$ pixels, the region is considered a foreground patch; otherwise, it is treated as a background patch. Based on this strategy, we construct visual exemplar masks as supervision labels for the similarity map. We then compute the similarity loss $\mathcal{L}_{\text{sim}}$ using contrastive learning principles to supervise the similarity map, with the loss calculation shown in the following formula:

$$\mathcal{L}_{\text{sim}} = -\log \frac{\sum_{i \in \mathcal{P}_{\text{pos}}} \exp(S_i)}{\sum_{j \in \mathcal{P}_{\text{pos}}} \exp(S_j) + \sum_{j \in \mathcal{P}_{\text{neg}}} \exp(S_j)} \quad (17)$$

where S_i denotes the similarity score of the i -th patch, and $\mathcal{P}_{\text{pos}}$ and $\mathcal{P}_{\text{neg}}$ denote the foreground and background patch indices in the visual exemplar mask, respectively.

Combining Eq. 9 and Eq. 17, the final optimization objective of PPSNet is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{count}} + \lambda \mathcal{L}_{\text{sim}} \quad (18)$$

where λ denotes the balancing factor.

Additionally, we visualize the constructed visual exemplar masks and similarity maps. As shown in Fig. 7, the visual exemplar mask accurately covers the true target regions, demonstrating that the patch-level generation strategy based on ground truth density maps achieves precise foreground separation. This provides reliable supervision signals for subsequent similarity map learning in SWM, thereby laying the foundation for both similarity learning and feature weighting.

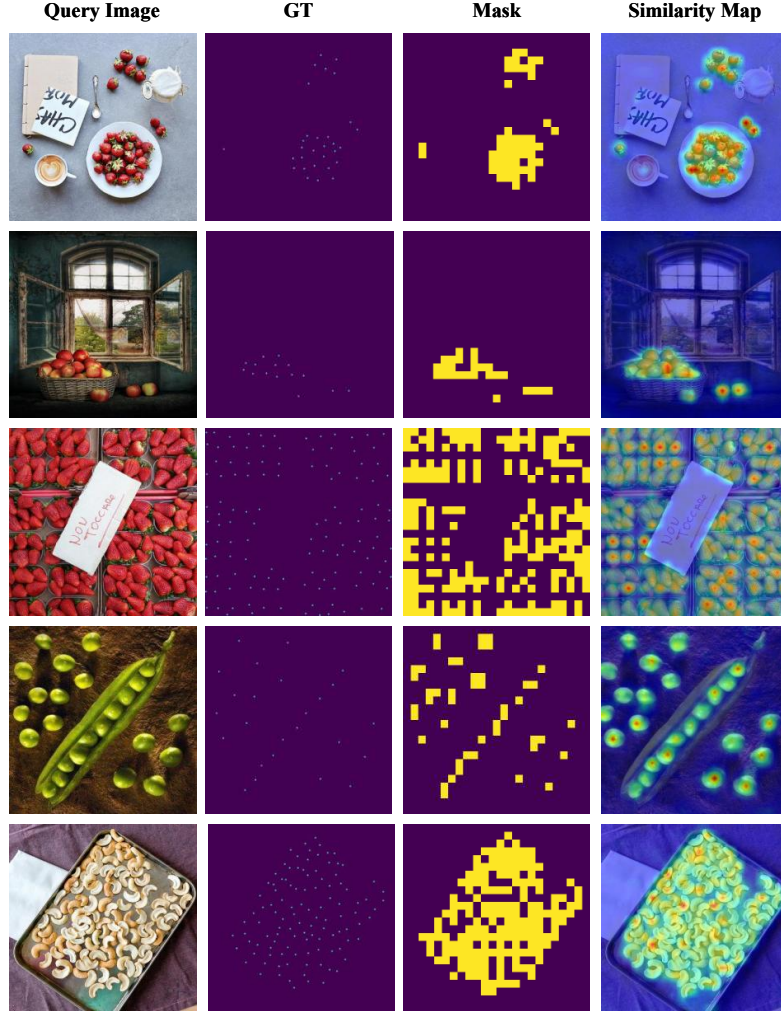


Fig. 7 Visualizing visual exemplar masks and similarity maps. We construct patch-level visual exemplar masks based on ground truth density maps to achieve accurate foreground-background separation. The similarity maps supervised by similarity loss effectively highlight target regions that are semantically consistent with visual exemplars.

4 Experiments

4.1 Datasets and Metrics

FSC-147. To demonstrate the counting capability of our PPSNet, we evaluate it using the FSC-147 dataset [7]. This is the first large-scale dataset designed for class-agnostic counting, containing 6,135 images spanning 147 categories including food, animals, and vehicles. Each image in this dataset includes point annotations, bounding box annotations, and category names. Furthermore, the training, validation, and test sets

have disjoint categories, ensuring accurate evaluation of the model’s generalization capability.

CARPK. CARPK [32] is a dataset specifically designed for vehicle counting. This dataset was captured from drone perspectives at an average height of 40 meters, containing 1,448 parking lot images with nearly 90,000 vehicles.

COCO. COCO dataset is widely used for downstream tasks in computer vision. Ranjan [7] constructed Val-COCO and Test-COCO subsets for evaluating class-agnostic counting capabilities.

Metrics. Following previous works [8, 10], we use the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) to measure the counting performance of our method:

$$MAE = \frac{1}{N} \sum_{i=1}^N |D_i^{pred} - D_i^{gt}| \quad (19)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (D_i^{pred} - D_i^{gt})^2} \quad (20)$$

4.2 Implementation Details

We employ identical data augmentation [10], with input images resized to 384×384 and split into patches of size 16×16 , which are then projected to a 768-dimensional embedding space. Our Extract-and-Match feature processor, pretrained with MAE [33], consists of 12 transformer encoder blocks with a hidden dimension of 768 and 12 attention heads per multi-head attention layer, followed by 3 additional transformer blocks with a hidden dimension of 512 for feature enhancement and dimensionality reduction. Visual exemplars are resized to 64×64 and split into patches of size 16×16 . We employ the AdamW optimizer with a weight decay of 0.05. The model is trained for 200 epochs on the FSC-147 dataset using a learning rate of $1e-4$ and a batch size of 8. All training and testing are conducted on a single NVIDIA GeForce RTX 4090 GPU.

4.3 Comparison with State-of-the-art Methods

4.3.1 Results for FSC-147

Vision-guided Exemplar Counting. We evaluate the counting performance of our PPSNet on the FSC-147 dataset [7]. For the vision-guided exemplar counting paradigm, existing methods typically evaluate counting performance using multiple visual exemplars (three exemplars) and single visual exemplar settings respectively, as shown in Table 1. In the multiple-exemplar setting, our PPSNet outperforms the typical Extract-then-Match paradigm method CounTR [10] with MAE and RMSE improvements of 22.26% and 32.22% on the test set, and 22.24% and 18.98% on the validation set. These results validate the superiority of our Extract-and-Match paradigm. Furthermore, compared to recent methods such as RCFNet [34] and SQLNet [22], our PPSNet achieves consistent and significant improvements across all evaluation metrics, validating the effectiveness of our proposed architecture. Crucially, under the more

challenging single visual exemplar setting, PPSNet maintains superior performance by achieving the best MAE results on both validation and test sets. Collectively, the experimental results in Table 1 confirm that our proposed PPSNet delivers highly competitive performance within the vision-guided exemplar counting paradigm.

Table 1 Comparison results with state-of-the-art methods under the vision-guided exemplar counting paradigm on the FSC-147 dataset. The best results are highlighted in **bold**.

Method	Prompt	Shots	Validation Set		Test Set	
			MAE	RMSE	MAE	RMSE
FamNet [7]	Vision	1	26.55	77.01	26.76	110.95
BMNet+ [8]	Vision	1	17.89	61.12	16.89	96.65
CounTR [10]	Vision	1	13.15	49.72	12.06	90.01
CACL[35]	Vision	1	17.88	68.11	14.71	100.64
CSTrans-B [36]	Vision	1	18.78	65.27	16.93	93.76
FFMP [37]	Vision	1	13.06	47.85	12.68	93.74
DSFNet [38]	Vision	1	16.34	56.42	15.10	96.75
SQLNet [22]	Vision	1	13.78	45.71	12.56	77.59
PPSNet(Ours)	Vision	1	11.48	43.07	11.12	81.17
GMN [23]	Vision	3	29.66	89.81	26.52	124.57
FamNet [7]	Vision	3	23.75	69.07	22.08	99.54
BMNet+ [8]	Vision	3	15.74	58.53	14.62	91.83
SPDCN [39]	Vision	3	14.59	49.97	13.51	96.80
CounTR [10]	Vision	3	13.13	49.83	11.95	91.23
SAFECount [9]	Vision	3	15.28	47.20	14.32	85.54
CACL [35]	Vision	3	14.89	58.48	11.89	79.01
CSTrans-B [36]	Vision	3	18.78	65.27	16.93	93.51
PseCo [26]	Vision	3	15.31	68.34	13.05	112.86
FFMP [37]	Vision	3	11.36	38.84	11.10	73.29
DSFNet [38]	Vision	3	12.94	46.70	11.76	90.71
DFM [40]	Vision	3	12.08	48.63	11.31	75.39
RCFNet [34]	Vision	3	11.18	42.76	11.05	72.53
SQLNet [22]	Vision	3	12.40	42.30	12.49	80.85
PPSNet(Ours)	Vision	3	10.21	40.37	9.29	61.83

Guidance-less Counting. The guidance-less counting paradigm requires no prompts. We adapt our proposed PPSNet to perform class-agnostic counting without visual exemplars or text prompts. Specifically, we replace visual exemplar inputs with learnable tokens, enabling PPSNet to generate predicted density maps from query images alone. As shown in Table 2, we compare against methods specifically designed for guidance-less counting and methods adapted from the vision-guided exemplar counting paradigm. Our PPSNet achieves state-of-the-art performance. It outperforms GCA-SUN [13], a method exclusively designed for guidance-less counting, by 26.03% and 20.43% in MAE on validation and test sets respectively. Compared to adapted methods, PPSNet also demonstrates superior performance.

Text-guided Counting. The text-guided counting paradigm employs text descriptions as prompts instead of visual exemplars. We adapt PPSNet to leverage text prompts for counting. Specifically, we encode text descriptions using CLIP [29] and

Table 2 Comparison results with state-of-the-art methods under the guidance-less counting paradigm on the FSC-147 dataset. The best results are highlighted in **bold**. (gl denotes guidance-less counting, vg $\rightarrow$ gl denotes adaptation from the vision-guided exemplar counting paradigm to guidance-less counting.)

Method	Type	Validation Set		Test Set	
		MAE	RMSE	MAE	RMSE
RepRPN-Counter [11]	gl	30.40	98.73	27.45	129.69
RCC [28]	gl	17.49	58.81	17.12	104.53
GCNet [12]	gl	19.50	63.13	17.83	102.89
GCA-SUN [13]	gl	16.06	53.04	14.00	92.19
CountR [10]	vg $\rightarrow$ gl	18.07	71.84	14.71	106.87
CACL [35]	vg $\rightarrow$ gl	19.23	72.81	15.56	103.44
PseCo [26]	vg $\rightarrow$ gl	23.90	100.33	16.58	129.77
DSFNet [38]	vg $\rightarrow$ gl	22.82	64.51	21.60	106.28
LOCA [21]	vg $\rightarrow$ gl	17.43	54.96	16.22	103.96
PPSNet(Ours)	vg $\rightarrow$ gl	11.88	46.81	11.14	84.21

replace visual exemplars with these text embeddings, which are then concatenated with query tokens to perform text-guided counting. Experimental results in Table 3 demonstrate that our PPSNet achieves state-of-the-art performance. Compared to CountTX [15], a typical Extract-then-Match method for text-guided counting, PPSNet reduces MAE by 29.94% on the validation set and 28.02% on the test set. Against the current state-of-the-art method T2ICount [31], PPSNet achieves 13.06% MAE reduction and 22.32% RMSE reduction on the validation set.

Table 3 Comparison results with state-of-the-art methods under the text-guided counting paradigm on the FSC-147 dataset. The best results are highlighted in **bold**.

Method	Validation Set		Test Set	
	MAE	RMSE	MAE	RMSE
ZSC [14]	26.93	88.63	22.09	115.17
CLIP-Count [41]	18.79	61.18	17.78	106.62
VLCounter [16]	18.06	65.13	17.05	106.16
CountTX [15]	17.10	65.61	15.88	106.29
CountDiff [30]	15.50	54.33	14.83	103.15
T2ICount [31]	13.78	58.78	11.76	97.86
QUANet [42]	14.22	53.39	13.24	97.24
TPECCount [43]	16.37	63.60	14.68	103.88
QICA [44]	13.82	60.24	13.05	104.17
PPSNet(Ours)	11.98	45.66	11.43	82.26

Complexity. We compare the complexity of our proposed PPSNet with current typical Extract-then-Match paradigm methods CountR [10] and CountTX [15], as shown in Table 4. Compared to CountR [10], although our parameter count increases by only 0.31M, we achieve significant MAE and RMSE improvements of 22.26% and 32.22% on the test set, respectively, while reducing inference latency by 5.49ms. When

compared to EaMethod, despite a modest 1.38ms increase in inference time, PPSNet achieves substantially lower MAE (9.29) and RMSE (61.83) with merely a 0.2M parameter increase. This demonstrates our method’s ability to enhance counting accuracy without introducing excessive parameters. Furthermore, against the text-guided counting method CounTX [15], our model maintains nearly identical parameter counts. Although GFLOPs increase to 86.24G, we achieve remarkable MAE and RMSE reductions of 29.94% and 28.02%, respectively. Results across both paradigms consistently validate that PPSNet significantly improves class-agnostic counting accuracy while maintaining parameter efficiency.

Table 4 Computational cost comparison with other counting methods. While PPSNet is not the most computationally efficient approach, it achieves superior counting accuracy with only marginal computational overhead, demonstrating an optimal balance between performance and efficiency.

Method	Prompt	Params(M)	GFLOPs(G)	Time(ms)	MAE	RMSE
CounTR [10]	Vision	99.69	83.23	16.68	11.95	91.23
EaMethod (Ours)	Vision	99.80	88.79	9.81	11.21	91.38
PPSNet(Ours)	Vision	100.00	88.80	11.19	9.29	61.83
CounTX [15]	Text	161.08	43.88	36.29	15.88	106.29
PPSNet(Ours)	Text	163.92	86.24	23.67	11.43	82.26

4.3.2 Cross-dataset generalization

CARPK. To evaluate the cross-dataset generalization capability of our PPSNet, we conduct experiments on the CARPK dataset, with evaluation from two aspects: (i) direct testing without fine-tuning; (ii) fine-tuning followed by testing on CARPK. Detailed results are presented in Table 5. We remove all vehicle images from the FSC-147 training set to preserve generalization to novel objects, then directly test the pretrained model on CARPK. Compared to BMNet+ [8], our method achieves 3.93% and 6.10% lower MAE and RMSE respectively. Against CounTR [10], we obtain a 2.15% MAE reduction. When fine-tuned on CARPK, Table 5 shows our method maintains superior performance, surpassing CounTR [10] by 12.52% in MAE and 10.60% in RMSE. Notably, PPSNet achieves competitive results in both evaluation aspects, demonstrating exceptional cross-dataset generalization capability and consistent performance across diverse scenarios and datasets.

Val-COCO and Test-COCO. Following prior works [7, 36], we also evaluate our method on the Val-COCO and Test-COCO datasets. These subsets serve as dedicated evaluation benchmarks, particularly when comparing against detection-based approaches. As shown in Table 6, our method surpasses all competing methods, including object detectors Faster R-CNN [45], RetinaNet [46], and Mask R-CNN [47] and other class-agnostic counting methods. Compared to the state-of-the-art FFMP [37], we achieve 16.77% and 21.84% improvements in MAE and RMSE on the test set, respectively. This result further confirms our method’s robust generalization capability across diverse scenarios and datasets.

Table 5 Comparison results with state-of-the-art methods on the CARPK dataset. The best results are highlighted in **bold**.

Method	Prompt	Fine-tuned	MAE	RMSE
FamNet+ [7]	Vision	✗	28.84	44.47
SPDCN [39]	Vision	✗	18.15	21.61
SAFECount [9]	Vision	✗	16.66	24.08
BMNet+ [8]	Vision	✗	10.44	13.77
CounTR [10]	Vision	✗	10.25	12.70
RCC [28]	None	✗	12.31	15.40
GCA-SUN [13]	None	✗	10.96	13.95
CounTX [15]	Text	✗	11.64	14.85
CLIP-Count [41]	Text	✗	11.96	16.61
PPSNet(Ours)	Vision	✗	10.03	12.93
FamNet+ [7]	Vision	✓	18.19	33.66
SPDCN [39]	Vision	✓	10.07	14.12
SAFECount [9]	Vision	✓	5.33	7.04
BMNet+ [8]	Vision	✓	5.76	7.83
CounTR [10]	Vision	✓	5.75	7.45
RCC [28]	None	✓	9.21	11.33
CounTX [15]	Text	✓	8.13	10.87
PPSNet(Ours)	Vision	✓	5.03	6.66

Table 6 Comparison results with state-of-the-art methods on the Val-COCO and Test-COCO dataset. The best results are highlighted in **bold**.

Method	Type	Val-COCO		Test-COCO	
		MAE	RMSE	MAE	RMSE
Faster-RCNN [45]	Detection	52.79	172.46	36.20	79.59
RetinaNet [46]	Detection	63.57	174.36	52.67	85.86
Mask-RCNN [47]	Detection	52.51	172.21	35.56	80.00
FamNet [7]	Regression	39.82	108.13	22.76	45.92
CounTR [10]	Regression	24.66	83.84	10.89	31.11
CSTrans-B [36]	Regression	32.27	95.87	15.25	27.08
FFMP [37]	Regression	21.67	76.15	11.27	29.12
PPSNet(Ours)	Regression	21.96	74.17	9.38	22.76

4.4 Qualitative Result

To better analyze our proposed PPSNet, we compare prediction density maps of PPSNet with CounTR [10] and EaMethod to validate method effectiveness. As shown in Fig. 8, PPSNet achieves more accurate counting results and density distributions that closely align with visual exemplar distributions. Specifically, compared to CounTR [10], we observe background regions being misidentified as exemplars (e.g., row 3 and row 5). We attribute this to the Extract-then-Match paradigm’s lack of target-oriented feature extraction, causing background noise to be erroneously matched as visual exemplars during similarity modeling. Additionally, CounTR exhibits undercounting phenomena (e.g., row 1, row 2 and row 4), likely due to insufficient exemplar perception. Compared with the EaMethod, although its counting

results significantly suppress background noise activation due to the Extract-and-Match paradigm’s bidirectional information guidance and target-oriented feature extraction, EaMethod still exhibits inadequate exemplar perception and fails to explicitly capture exemplar distribution patterns. Our PPSNet substantially mitigates these issues (e.g., row 2, row 4 and row 5). These qualitative results demonstrate that PPSNet alleviates pseudo-activation of background noise within bounding boxes, reduces background interference, and enhances feature responses in target regions.

We also visualize the self-attention of the query image (A_{qq}), the attention from the example image to the query image (A_{eq}), and the attention from the query image to the example image (A_{qe}) in the Extract-and-Match paradigm respectively. As observed from Fig. 9, the terms A_{qq} and A_{eq} emphasize foreground features while suppressing background features, whereas A_{qe} focuses more on the background. As stated in Section 3.1, this confirms that the Extract-and-Match paradigm incorporates richer feature information during the update of query features.

Besides, as shown in Fig. 7, we further analyze PPSNet’s similarity maps. These similarity maps precisely highlight target regions semantically consistent with visual exemplars across diverse scenarios while effectively suppressing background responses. This suppression remains significant even in densely packed scenes (e.g., row 3 and row 5), indicating that our intermediate similarity maps comprehensively encode visual exemplar distributions. Consequently, during feature weighting, the model strengthens responses in target regions, thereby improving counting accuracy.

4.5 Ablation Study

We perform ablation studies on the FSC-147 dataset to verify the effectiveness of our proposed method, with the corresponding results presented in Table 7.

Proposed Modules. By integrating PPCM to the EaMethod (B1 and B5 *vs.* B2 and B6), we observe from Table 7 that the MAE and RMSE are reduced of 7.77% and 8.92% on the validation set, and by 3.93% and 4.41% on the test set. This demonstrates that our PPCM mitigates background noise interference and enhances visual exemplar distribution perception through perceptual prior construction. Adding SWM to the EaMethod (B1 and B5 *vs.* B3 and B7) yields MAE values of 12.03 and 11.01 on validation and test sets respectively, confirming that similarity-guided feature weighting strengthens target region responses. When integrating both PPCM and SWM with the EaMethod (B1 and B5 *vs.* B4 and B8), MAE and RMSE improve by 16.45% and 9.57% on the validation set, and 17.13% and 32.34% on the test set. This demonstrates mutual reinforcement between our modules, collectively enhancing counting accuracy.

Similarity Map Supervision. As shown in Table 8, we can observe that supervising the similarity map improves MAE and RMSE on the validation set by 9.16% and 5.83% respectively. This indicates that supervising the similarity map enables it to better distinguish target foreground and background in images, thereby laying a solid foundation for subsequent feature weighting.

Loss Balance Factor. To determine the optimal value of the balancing factor λ for the similarity loss, we conduct ablation studies. As shown in Table 9, when λ decreases from 0.1 to 0.001, the model’s MAE and RMSE on both validation and test sets progressively improve. The model achieves peak performance at $\lambda = 0.001$.

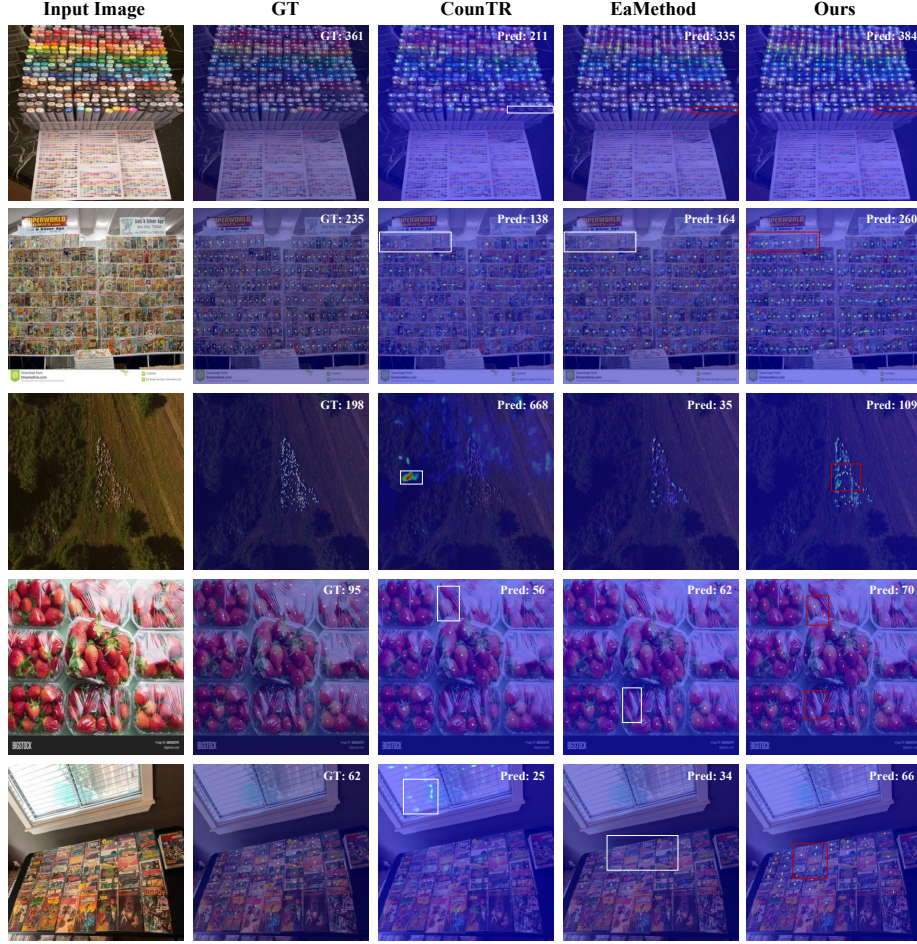


Fig. 8 Qualitative results on FSC-147. Left to right columns show query image, ground truth density map and predictions from CounTR, EaMethod and PPSNet. Predicted counts appear top-right. White boxes mark incorrect counts; red boxes indicate correct counts. Zoomed-in views are provided for clarity.

Further reduction of λ weakens the supervisory signal of the loss, causing counting performance to deteriorate.

Number of Vision Exemplars and Learnable Tokens. The number of visual exemplars directly impacts model counting performance, prompting our ablation study on exemplar quantity as shown in Table 10. Experimental results demonstrate optimal counting performance at $N = 3$ exemplars. We further observe that counting accuracy consistently improves with increasing exemplar numbers, likely because more exemplars provide richer and more diverse visual information to the model. Additionally, as described in Section 4.3.1, we adapt our model to guidance-less counting by replacing

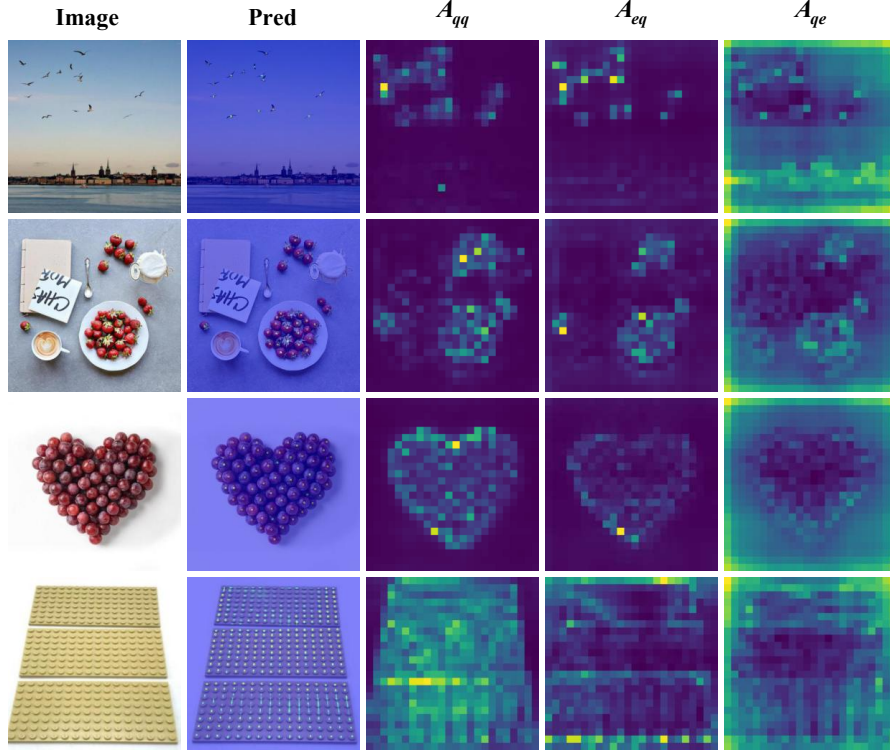


Fig. 9 Visualizations of attention maps. A_{qq} and A_{eq} highlight the foreground feature and suppress the background feature.

Table 7 Ablation study on each component of PPSNet. The best results are highlighted in **bold**.

No.	PPCM	SWM	Val MAE	Val RMSE
B1	✗	✗	12.22	44.64
B2	✓	✗	11.27	40.66
B3	✗	✓	12.03	45.79
B4	✓	✓	10.21	40.37
No.	PPCM	SWM	Test MAE	Test RMSE
B5	✗	✗	11.21	91.38
B6	✓	✗	10.77	87.35
B7	✗	✓	11.01	93.19
B8	✓	✓	9.29	61.83

Table 8 Ablation study of similarity loss. The best results are highlighted in **bold**.

$\mathcal{L}_{\text{count}}$	$\mathcal{L}_{\text{sim}}$	Val MAE	Val RMSE	Test MAE	Test RMSE
✓	✗	11.24	42.87	9.50	72.56
✓	✓	10.21	40.37	9.29	61.83

Table 9 Ablation study of the loss balance factor. The best results are highlighted in **bold**.

λ	Val MAE	Val RMSE	Test MAE	Test RMSE
0.1	13.92	48.86	13.79	105.52
0.01	11.85	46.56	11.34	91.40
0.001	10.21	40.37	9.29	61.83
0.0001	11.28	43.02	10.65	76.90

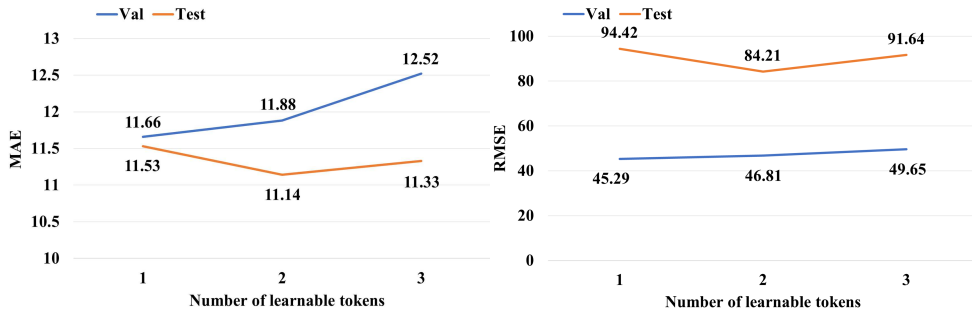


Fig. 10 Influence of varying numbers of learnable tokens on counting performance.

visual exemplars with learnable tokens. We conduct ablation studies on the number of learnable tokens (1, 2, and 3) to maintain correspondence with visual exemplar quantities. As shown in Fig. 10, model performance does not improve monotonically with increasing token numbers. We attribute this to the lack of explicit semantic guidance in learnable tokens compared to visual exemplars. Consequently, we select 2 learnable tokens as the optimal configuration for guidance-less counting.

Text Encoder. In Section 4.3.1, we adapt PPSNet to text-guided counting by integrating CLIP’s text encoder. As shown in Table 11, fine-tuning the text encoder significantly outperforms freezing it in counting performance.

5 Conclusion

In this paper, we propose a novel Perceptual Prior Similarity-guided Network (PPSNet) based on the Extract-and-Match paradigm for class-agnostic counting, alongside a theoretical analysis of its superiority over the traditional Extract-then-Match paradigm. PPSNet comprises two core modules: Perceptual Prior Construction

Table 10 Ablation study of the number of vision exemplar. The best results are highlighted in **bold**.

N	Val MAE	Val RMSE	Test MAE	Test RMSE
1	11.48	43.07	11.12	81.17
2	11.24	45.44	10.48	87.35
3	10.21	40.37	9.29	61.83

Table 11 Ablation study of training strategy of text encoder. The best results are highlighted in **bold**.

Text Encoder	Validation Set		Test Set	
	MAE	RMSE	MAE	RMSE
Freezed	11.84	44.45	12.44	107.79
Fine-tuned	11.98	45.66	11.43	82.26

Module (PPCM) and Similarity-guided Weighting Module (SWM). PPCM captures the distribution of visual exemplars in bounding boxes to alleviate background noise pseudo-activation and scale information loss stemming from the inherent geometric constraints of bounding boxes. SWM adopts the principle of contrastive learning to perform patch-level supervision on the cosine similarity map between query and exemplar features through a foreground mask of visual exemplars, and weights query features to highlight target region distributions. To the best of our knowledge, this is the first work to address the non-negligible background noise pseudo-activation caused by the inherent geometric constraints of bounding boxes. Extensive experiments on the FSC-147, CARPK and COCO datasets verify the effectiveness and generalization capability of our proposed method.

Declarations

- **Funding** The work is partially supported by the National Natural Science Foundation of China No. 62476235; Central Government Guided Local Funds for Science and Technology Development No. 216Z0301G; Hebei Natural Science Foundation No. F2023203012; Science Research Project of Hebei Education Department No. QN2024010 and Innovation Capability Improvement Plan Project of Hebei Province No. 22567626H.
- **Conflicts of Interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- **Data availability** Data will be made available on reasonable request.

References

1. Ling M, Chen J, Liu Y, et al (2025) Dual-branch adjacent connection and channel mixing network for video crowd counting. Pattern Recognition p 111709

2. Mu B, Shao F, Xie Z, et al (2025) Misf-net: Modality-invariant and-specific fusion network for rgb-t crowd counting. *IEEE Transactions on Multimedia*
3. Xiong L, Li Z, Huang X, et al (2025) Csfnet: A novel counting network based on context features and multi-scale information. *Multimedia Systems* 31(1):7
4. Liu Y, Shen H, Wang T, et al (2025) Vehicle counting in drone images: An adaptive method with spatial attention and multiscale receptive fields. *ETRI Journal* 47(1):7–19
5. Wang Y, Xu S, Wang P, et al (2025) Vehicle detection algorithm based on improved rt-detr: Y. wang et al. *The Journal of Supercomputing* 81(1):290
6. Ferreira FRT, do Couto LM (2025) Using deep learning on microscopic images for white blood cell detection and segmentation to assist in leukemia diagnosis. *The Journal of Supercomputing* 81(2):410
7. Ranjan V, Sharma U, Nguyen T, et al (2021) Learning to count everything. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 3394–3403
8. Shi M, Lu H, Feng C, et al (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*, pp 9529–9538
9. You Z, Yang K, Luo W, et al (2023) Few-shot object counting with similarity-aware feature enhancement. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp 6315–6324
10. Liu C, Zhong Y, Zisserman A, et al (2022) Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:220813721*
11. Ranjan V, Nguyen MH (2022) Exemplar free class agnostic counting. In: *Proceedings of the Asian Conference on Computer Vision*, pp 3121–3137
12. Wang M, Li Y, Zhou J, et al (2024) Gcnet: Probing self-similarity learning for generalized counting network. *Pattern Recognition* 153:110513
13. Wu Y, Xu Y, Xu T, et al (2025) Gca-sunet: A gated context-aware swin-unet for exemplar-free counting. URL <https://arxiv.org/abs/2409.12249>, 2409.12249
14. Xu J, Le H, Nguyen V, et al (2023) Zero-shot object counting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 15548–15557
15. Amini-Naieni N, Amini-Naieni K, Han T, et al (2023) Open-world text-specified object counting. In: *BMVC*

16. Kang S, Moon W, Kim E, et al (2024) Vlcounter: Text-aware visual representation for zero-shot object counting. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 2714–2722
17. Zhang Y, Zhang H, Zhao Q (2023) Counting and locating anything: Class-agnostic few-shot object counting and localization. In: 2023 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 2921–2926
18. Wang Y, Yang B, Wang X, et al (2024) Satcount: A scale-aware transformer-based class-agnostic counting framework. *Neural Networks* 172:106126
19. Quan KAC, Nguyen VT, Nguyen TV, et al (2025) Unified vit-cnn for few-shot object counting. *Signal, Image and Video Processing* 19(3):221
20. Wang Z, Xiao L, Cao Z, et al (2024) Vision transformer off-the-shelf: A surprising baseline for few-shot class-agnostic counting. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 5832–5840
21. Đukić N, Lukežič A, Zavrtanik V, et al (2023) A low-shot object counting network with iterative prototype adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 18872–18881
22. Wu H, Chen Y, Liu L, et al (2025) Sqlnet: Scale-modulated query and localization network for few-shot class-agnostic counting. *IEEE Transactions on Image Processing*
23. Lu E, Xie W, Zisserman A (2018) Class-agnostic counting. In: Asian conference on computer vision, Springer, pp 669–684
24. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*
25. 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, pp 23293–23302
26. Huang Z, Dai M, Zhang Y, et al (2024) Point segment and count: A generalized framework for object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 17067–17076
27. Zhai G, Han S, Chen K, et al (2026) Towards a novel few-shot object counting paradigm with point-guided exemplar prompt in industrial internet of things. *Future Generation Computer Systems* 174:107946
28. Hobley M, Prisacariu V (2022) Learning to count anything: Reference-less class-agnostic counting with weak supervision. *arXiv preprint arXiv:220510203*

29. Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PmLR, pp 8748–8763
30. Hui X, Wu Q, Rahmani H, et al (2024) Class-agnostic object counting with text-to-image diffusion model. In: European Conference on Computer Vision, Springer, pp 1–18
31. Qian Y, Guo Z, Deng B, et al (2025) T2icount: Enhancing cross-modal understanding for zero-shot counting. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 25336–25345
32. Hsieh MR, Lin YL, Hsu WH (2017) Drone-based object counting by spatially regularized regional proposal network. In: Proceedings of the IEEE international conference on computer vision, pp 4145–4153
33. He K, Chen X, Xie S, et al (2022) Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16000–16009
34. Xie Z, Wang G (2025) Semantically-guided cross-level feature fusion network for few-shot object counting via dual-stream template discrimination. Knowledge-Based Systems p 115137
35. Zhai C, Du Y, Qu H, et al (2024) Class-agnostic counting and localization with feature augmentation and scale-adaptive aggregation. Knowledge-Based Systems 300:112104
36. Gao BB, Huang Z (2024) Cstrans: Correlation-guided self-activation transformer for counting everything. Pattern Recognition 153:110556
37. Ma G, Zhu X, Tian Z, et al (2025) Few-shot object counting with frequency attention and multi-perception head. Neurocomputing p 130598
38. Liu K, Dou Z, Zou X, et al (2025) Dsfnet: Class-agnostic object counting network with dual enhancement of similarity and feature representation. IEEE Transactions on Industrial Informatics
39. Lin W, Yang K, Ma X, et al (2022) Scale-prior deformable convolution for exemplar-guided class-agnostic counting. In: BMVC, p 313
40. Zhai C, Zhang F, Chen W, et al (2025) Decoupled feature matching for few-shot counting and localization. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 1–5
41. Jiang R, Liu L, Chen C (2023) Clip-count: Towards text-guided zero-shot object counting. In: Proceedings of the 31st ACM International Conference on

Multimedia, pp 4535–4545

- 42. Shi M, Zhang X, Yue Z, et al (2025) Text-promptable object counting via quantity awareness enhancement. arXiv preprint arXiv:250706679
- 43. Zhang S, Liu J, Jia S (2026) Text-guided patch-level exemplar selection for zero-shot counting. Engineering Applications of Artificial Intelligence 167:113986
- 44. Zhang D, Li B, Wang F, et al (2026) Boosting quantitative and spatial awareness for zero-shot object counting. URL <https://arxiv.org/abs/2603.16129>, 2603.16129
- 45. Ren S, He K, Girshick R, et al (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28
- 46. Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
- 47. He K, Gkioxari G, Dollár P, et al (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969