Proceedings of Machine Learning Research vol 278:1–7, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

Zero-shot object counting with visual feature extraction and language-guidance

Gaoxin Ma MAGAOXIN@GS.ZZU.EDU.CN and **Zhenfeng Zhu** IEZFZHU@ZZU.EDU.CN
Zhengzhou University, 100 Science Avenue, Zhengzhou, Henan, China

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Zero - Shot Object Counting (ZSC) focuses on counting objects of any class in a query image without the need for user - supplied exemplars. Recently, ZSC has attracted growing interest because of its broad applicability and higher efficiency when contrasted with Few - Shot Object Counting (FSC). Different from FSC, a significant problem in existing ZSC methods is their failure to efficiently recognize high - quality exemplar features. In this paper, we propose a Zero-Shot Object Counting network with Visual Feature Extraction and Language-Guidance (VELG). Through the visual feature extraction module, we progressively fuse the scale and geometric information of the exemplars. Meanwhile, we introduce a language-guidance module that helps the exemplar learn informative image-level visual representations and refine the exemplar features using Contrastive Language-Image Pre-training. Extensive experiments on the FSC147 and CARPK datasets verify the accuracy and strong generalizability of the proposed approach.

Keywords: Zero-shot, Object counting, Clip

1. Introduction

Object counting involves estimating the number of particular objects within a query image. Owing to its practical applications, this area has been attracting increasing interest. For example, crowd counting (Lin et al., 2022) are instrumental in gauging pedestrian density across public spaces. Animal counting is critical for monitoring wildlife populations and managing their impact on wildlife and marine species.

Traditional counting methods are typically limited to specific categories. To address these challenges of previous methods, class-agnostic approaches have been created for situations involving unseen classes, such as Few-shot and Zero-shot object counting. Few-shot counting methods treat the task as a matching problem, utilizing a limited number of annotated bounding boxes to recognize and count objects. In practice, these methods often require users to users manually mark the exemplars of the objects they are interested in. To address the limitations of bounding box annotation, zero-shot counting methods have been developed.

Our method is designed to enumerate objects belonging to specified classes within an image, even without any prior annotations for these classes. We propose a novel Zero-Shot Object Counting network with Visual Extraction and Language-Guidance , which consists of two fundamental components: Visual Feature Extraction module (VE), Language-Guidance module (LG). The VE module separately extracts the exemplar geometric and scales information. The LG model is developed by leveraging Contrastive Language-Image Pre-training . This enables our model to possess the zero-shot image-text alignment capacity. Extensive experiments demonstrate the proposed method can attain a counting performance compared to previous methods.

2. Related Works

Class-specific object counting focuses on objects that belong to pre-defined categories, such as humans (Lin et al., 2022) and vehicles (Hsieh et al., 2017). These class-specific counting models require individual training for each class, involving millions of point annotations. Few-shot counting methods have emerged as an alternative, allowing models to count objects across unseen categories. Few-shot object counting usually consist of three main stages. Firstly, according to the given image and the bounding boxes of its example images, the corresponding query image features and example features are generated. Secondly, obtain a similarity map by matching the query image features with the example image features. Finally, using the similarity map to generate density map and summed up to obtain the number of objects.

During the whole process, two main factors play a crucial role: the general feature representation and similarity matching. As a pioneering work in few-shot object counting, GMN (Lu et al., 2019) firstly proposed a matching network architecture that can calculate the number of objects without specific category. After this work, FamNet (Ranjan et al., 2021) learns to use ROI pooling to predict the density map and further introduces a new dataset specifically for the counting, namely FSC-147. Subsequent advancements in this field can be categorized into two streams. The first direction involves leveraging more sophisticated visual backbones to enhance the quality of the extracted feature representations. The second approach centers on refining the exemplar matching process.

Zero-shot object counting method is a technology that can identify and count the specified objects in an image only by the category name without any labeled samples. The proposal of this method aims to address the limitations of previous few-shot counting methods. RepRPN (Ranjan and Nguyen, 2022) first introduced the concept of zero-shot object counting, aiming to remove the necessity of annotating target exemplars. Subsequently, to enhance the applicability of exemplar-free object counting, ZSC (Xu et al., 2023) proposed a method that incorporates semantic information as guidance. CLIP-Count employs CLIP to encode text and images independently, establishing semantic connections that are essential for intuitive counting. Moreover, PseCo (Huang et al., 2024) introduces a multi-task framework based on SAM, leveraging the powerful capabilities of SAM in segmentation to enhance the performance and functionality of the zero-shot object counting tasks.

3. Our Methods

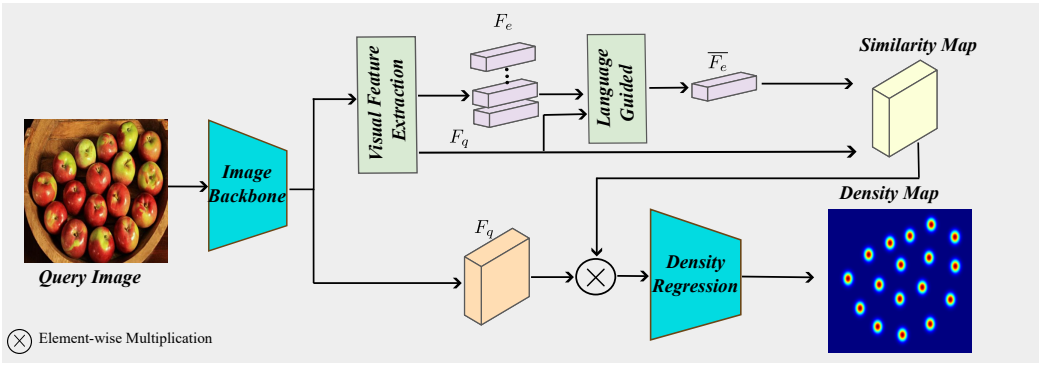

Figure 1: The overall architecture of the VELG.

Figure 1 summarizes the proposed method. Detailed descriptions of the key components are provided in the following sections.

3.1. Visual Feature Extraction

In the zero-shot scenario, the lack of annotations means that scale and geometry queries tailored to annotations cannot be obtained. Therefore, a slight modification is required to compute the exemplar features. We introduce a Visual Feature Extraction module to construct exemplar features for the given class, as shown in Figure 2.

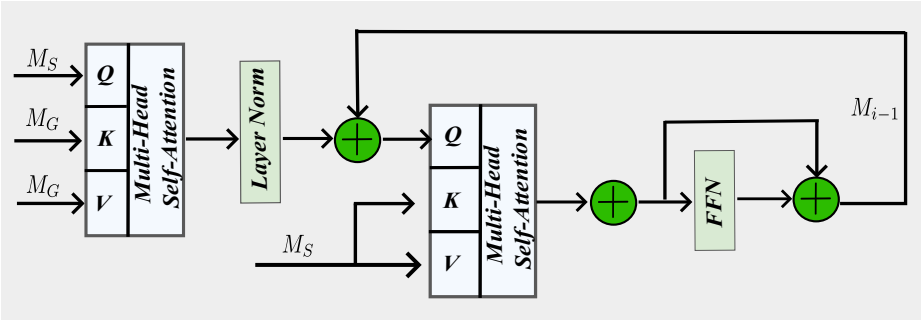

Figure 2: Visual Feature Extraction module

First, the trainable geometry query $F_g \in \mathbb{R}^{c\times s\times s}$ and the scale query $F_s \in \mathbb{R}^{c\times s\times s}$ are initialized. Specifically, the trainable scale queries F_s are reshaped into a matrix $M_S \in \mathbb{R}^{n\times s^2\times c}$ and in the same way, the geometry queries F_g and query image features F_q are reshaped into $M_G \in \mathbb{R}^{n\times s^2\times c}$ and $M_Q \in \mathbb{R}^{h\times w\times c}$, respectively. Then follows the sequence

$$\begin{aligned} M_i^1 =& MHSA\left(M_S, M_G, M_G\right) \\ M_i^2 =& MHSA\left(LN(M_i^1), M_Q, M_Q\right) + M_{i-1} \\ M_i =& FFN\left(M_i^2\right) + M_i^2 \end{aligned} \tag{1}$$

where the MHSA is the standard multi-head self-attention, LN is layer normalization and FFN represents a small feed-forward network. This process is carried out for L iterations, $i \in \{1, ..., L\}$, and the output is M_L. The output $M_L \in \mathbb{R}^{n\times p^2\times c}$is reshaped into a set of object exemplar features $F_e \in \mathbb{R}^{c\times s\times s}$. Distinct from prior approaches, our strategy integrates the geometric and scales feature. In this way, the rich information from geometric and scales can be incorporated into the exemplar features, which addresses the inadequacy of original object counting methods.

3.2. Language-Guidance Module

Although the visual feature extraction module proposed in the previous chapter can effectively capture the geometric information, size information and spatial position information in the image. However, there are still certain challenges when dealing with complex scenes that contain multiple categories of objects.To address this issue, we propose a Language-Guided (LG) module based on CLIP. This module utilizes the powerful image-text alignment capability of CLIP to optimize and correct the example features through semantic information, thereby obtaining high-quality sample image features.

To achieve this goal, we first applies a linear projection to ensure that the input image features and sample features are mapped to the same channel dimension as the text embedding features. For brevity, we call the reshaped $E_p \in \mathbb{R}^{p\times p\times c}$ as image embeddings of the exemplar feature, as our E_p encodes cross-patch information due to the image-level perceptive field within global attention. Denote $\left[E_{p_0}, E_{p_1}, \cdots, E_{p_{i-1}}\right]$ as the parts in which $E_{p_i} \in \mathbb{R}^{p\times p\times c_p}$, $c_p = c/p^2$,$i \in \left\{0, 1, \cdots, p^2\right\}$.

At the same time, in order to accurately locate the target objects corresponding to the category labels, a maximum pooling operation is performed on the ground truth density map in advance to generate an image-level binary mask. This mask clearly divides the image regions into target regions (P) and background regions (N). Then, we introduce the following loss function based on InfoNCE:

$$L = -\log \frac{\sum_{i\in P} \exp\left(Cos\left(E_{pi}, E_t\right)/\tau\right)}{\sum_{i\in P} \exp\left(Cos\left(E_{pi}, E_t\right)/\tau\right) + \sum_{j\in N} \exp\left(Cos\left(E_{pj}, E_t\right)/\tau\right)} \tag{2}$$

where Cos represents the cosine similarity, and τ represents the temperature parameter. In the design of the loss function, the alignment between the image embedding space and the text embedding space is achieved through the contrastive learning mechanism. Specifically, this loss function fine-tunes the text embeddings, pulling the text embeddings closer to the image embeddings of the positive samples, while pushing the image embeddings of the negative samples away. This optimization process is similar to the pre-training objective of CLIP, but it has been adaptively improved for the specific requirements of zero-shot object counting.

4. Experiments

4.1. Datasets and Details

The FSC-147 dataset, proposed by FamNet (Ranjan et al., 2021), is the first dataset specifically customized for few-shot counting tasks. This dataset contains 6,135 images and a total of 147 different categories, including but not limited to food, fruits, vehicles, and animals, among others. In addition to FSC-147, we use the car counting dataset to evaluate its generality.

To ensure efficient training, images with lower resolutions are padded with zeros, and those with higher resolutions are resized to maintain uniform dimensions of 384 × 384 pixels for all query images. We leverage the OpenAI pre-trained CLIP model featuring ViT-B/16 and choose ResNet-50 as our backbone network. When conducting training, we fix the batch size at 16. We adopt the AdamW optimizer with a learning rate of 1e-4. All models are trained for a total of 200 epochs.

4.2. Comparison with State-of-the-arts

4.2.1. FSC-147

To evaluate the performance VELG, this chapter compares it with several few-shot object counting and reference-free object counting methods that have performed well on the FSC-147 dataset, such as FamNet (Ranjan et al., 2021), CounTR (Liu et al., 2022), LOCA (Đukić et al., 2023) and other zero-shot counting methods. All these comparisons are carried out on the FSC-147 dataset, and the results are presented in Table 1.

The analysis of the experimental results shows that the VELG performs well in the zero-shot object counting task mainly due to its unique feature extraction mechanism. Specifically, the visual feature extraction module and the language guidance module generate representative average

Table 1: Performance comparison on FSC-147 dataset.

Methods	shot	Val		Test	
		MAE	RMSE	MAE	RMSE
GMN	3	29.66	89.81	26.52	124.57
FamNet	3	23.75	69.07	22.08	99.54
BMNet+	3	15.74	58.53	14.62	91.83
SAFECount	3	15.28	47.20	14.32	85.54
CounTR	3	13.13	49.83	11.95	91.23
Loca	3	10.24	32.56	10.76	56.97
RepRPN	0	29.24	98.11	26.66	129.11
ZSC	0	26.93	88.63	22.09	115.17
Pseco	0	23.90	100.33	**16.58**	129.77
CLIP-Count	0	18.79	61.18	17.78	**106.62**
VELG(ours)	0	**17.32**	**60.84**	16.67	109.87

features by averaging the size and attribute features of the example images. This design has a dual effect in the zero-shot scenario: on the one hand, the average feature strategy effectively avoids the occurrence of extreme counting results, thus reducing the negative impact on the RMSE index; on the other hand, since the count values tend to be distributed around the true values rather than precise counting results, it leads to a slight increase in the MAE index.

4.2.2. CROSS-DATASET GENERALIZATION

We perform a cross-dataset validation on the CARPK dataset to demonstrate the generalization ability of our VELG model for class-specific object counting tasks. In detail, the model is trained using the FSC-147 dataset and then evaluated on the test set of CARPK without any fine-tuning process. The experimental results are presented in Table 2. Among the models in the zero-shot group, our VELG model attains the optimal performance. In general, this experiment showcases the superior cross-dataset generalization capability of the VELG model.

Table 2: Cross-dataset generalization experiments on CARPK.

Methods	Shot	MAE	RMSE
GMN	3	32.92	39.88
FamNet	3	28.84	44.47
SPDCN	3	18.15	21.61
SAFECount	3	16.66	24.08
BMNet+	3	**10.44**	**13.77**
Grounging-Dino	0	29.72	31.60
RCC	0	21.38	26.61
CLIP-Count	0	11.96	**16.61**
VELG(ours)	0	**11.49**	17.26

4.3. Ablation Study

To rigorously evaluate the contributions of the Visual Feature Extraction (VE) module, Language-Guidance (LG) module within our framework, we conduct a series of ablation studies on the FSC-147 benchmark. We implement different versions of the baseline by incrementally adding the VE module and the LG module, respectively. The results are shown in Table 3.

Table 3: Ablation study of individual architectural components.

VE	LG	Val		Test	
		MAE	RMSE	MAE	RMSE
×	×	23.88	81.67	24.71	125.02
√	×	18.86	67.31	17.22	113.56
×	√	19.56	73.11	17.69	118.92
√	√	**17.32**	**60.84**	**16.67**	**109.87**

The VE module demonstrates the most significant impact, improving MAE by 21% and RMSE by 17.5%. Incorporating the LG module further enhances performance, achieving an additional improvement of 18.1% in MAE and 10.5% in RMSE. The combined effect of both modules reduces the MAE and RMSE to 17.32 and 60.84, respectively. These results validate the robustness and efficiency of the our method.

4.4. Qualitative Results

To illustrate the performance of our method, the qualitative results are presented in Figure 3. The comparison results clearly highlight the counting advantage of VELG. VELG can accurately locate the central area of the target objects and generate highly accurate density maps, while the prediction results of other method often appear scattered.

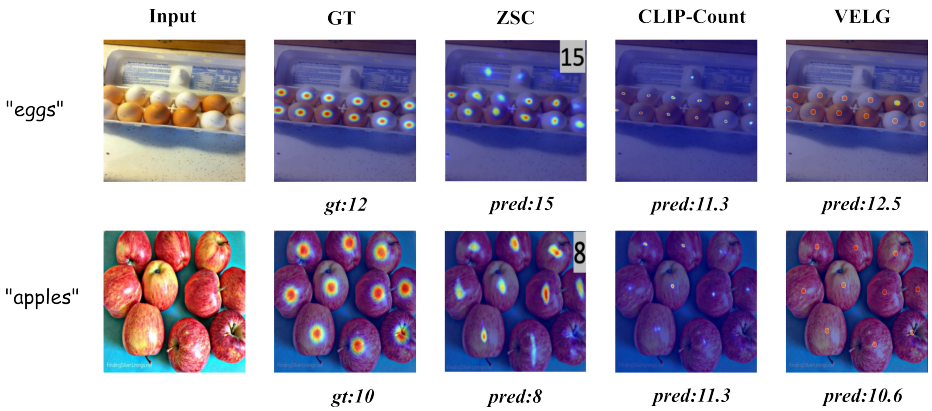

Figure 3: Qualitative results on the FSC147 dataset.

5. Conclusion

In this paper, we present a novel zero-shot object counting method, VELG, which addresses the limitations of previous approaches in feature extraction and similarity map construction. Firstly,

we employ a visual feature extraction module to construct the example features of a given category through an adaptive feature learning mechanism, overcoming the limitation of the lack of labeled data in the zero-shot scenario. Then, by utilizing the image-text alignment ability of the pre-trained CLIP model, the loss between the ground truth density map and the text features is calculated to optimize the representation of the text features. The optimized text features are aligned with the example features to screen out high-quality sample example features. Our method yields promising outcomes in zero-shot counting scenarios and performs effectively across various datasets, providing accurate visual associations and scalability. In future work, we plan to investigate and make better use of advanced vision-language models to further strengthen the capabilities of zero-shot object counting.

References

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

Zhizhong Huang, Mingliang Dai, Yi Zhang, Junping Zhang, and Hongming Shan. Point segment and count: A generalized framework for object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17067–17076, 2024.

Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19596–19605, 2022.

Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. In *33rd British Machine Vision Conference 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *Asian Conference on Computer Vision*, pages 669–684. Springer, 2019.

Viresh Ranjan and Minh Hoai Nguyen. Exemplar free class agnostic counting. In *Proceedings of the Asian Conference on Computer Vision*, pages 3121–3137, 2022.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3393–3402, June 2021.

Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15548–15557, 2023.

Nikola Đukić, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. In *2023 IEEE/CVF International Conference on Computer Vision*, pages 18826–18835, 2023.