Zero-shot Object Counting with Good Exemplars

Huilin Zhu[1,2,3,†], Jingling Yuan[1,2,†], Zhengwei Yang[4,†], Yu Guo[3,5], Zheng Wang[4], Xian Zhong[1,2,6(✉)], and Shengfeng He[3(✉)]

[1] Sanya Science and Education Innovation Park, Wuhan University of Technology
[2] Hubei Key Laboratory of Transportation Internet of Things, School of Computer Science and Artificial Intelligence, Wuhan University of Technology
zhongx@whut.edu.cn
[3] School of Computing and Information Systems, Singapore Management University
shengfenghe@smu.edu.sg
[4] School of Computer Science, Wuhan University
[5] School of Navigation, Wuhan University of Technology
[6] ROSE@EEE, Nanyang Technological University
[†] Equal Contribution
https://github.com/HopooLinZ/VA-Count

Abstract. Zero-shot object counting (ZOC) aims to enumerate objects in images using only the names of object classes during testing, without the need for manual annotations. However, a critical challenge in current ZOC methods lies in their inability to identify high-quality exemplars effectively. This deficiency hampers scalability across diverse classes and undermines the development of strong visual associations between the identified classes and image content. To this end, we propose the Visual Association-based Zero-shot Object Counting (VA-Count) framework. VA-Count consists of an Exemplar Enhancement Module (EEM) and a Noise Suppression Module (NSM) that synergistically refine the process of class exemplar identification while minimizing the consequences of incorrect object identification. The EEM utilizes advanced vision-language pre-taining models to discover potential exemplars, ensuring the framework's adaptability to various classes. Meanwhile, the NSM employs contrastive learning to differentiate between optimal and suboptimal exemplar pairs, reducing the negative effects of erroneous exemplars. VA-Count demonstrates its effectiveness and scalability in zero-shot contexts with superior performance on two object counting datasets.

1 Introduction

In visual monitoring applications, object counting plays a critical role in analyzing images or videos. Traditional methods focus on high precision within predefined object categories, such as crowds [4, 23], vehicles, and cells [1, 34, 39, 40, 44]. Yet, these methods are limited to specific categories, lacking the flexibility to adapt to new, unseen classes. To address these challenges, class-agnostic methods have been developed for scenarios with unseen classes. These methods, including few-shot, reference-free, and zero-shot object counting [12, 32, 35, 46, 47], provide varying levels of independence from predefined object classes.

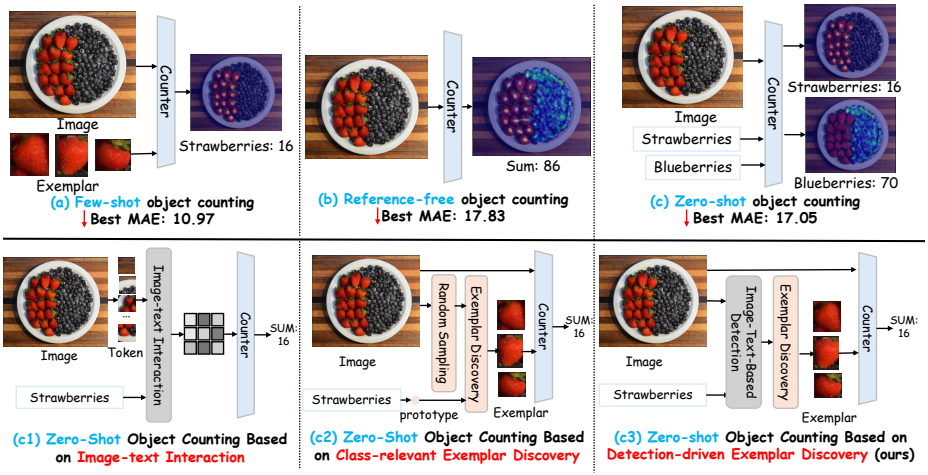

Fig. 1: Illustration of class-agnostic object counting methods. (a) Few-shot uses limited annotations for counting. (b) Reference-free quantifies objects without annotations. (c) Zero-shot counts specific classes without annotations, further divided into: (c1) Image-text association, leveraging direct image-text correlations. (c2) Class-related exemplar search, using prototypes to link classes with images. (c3) Our method introduces a detection-driven exemplar discovery to harmonize text with visual representations, distinguishing it from prior methods.

In this context, different strategies are adopted for object counting under varying constraints, as illustrated in Fig. 1. Few-shot counting methods [29,46,47], depicted in Fig. 1(a), method the task as a matching problem, using a small number of annotated bounding boxes to identify and count objects throughout the image. While effective, this method requires fine-tuning with annotations from novel classes, limiting its scalability in real-world surveillance settings due to the sparse availability of annotated bounding boxes. To circumvent the limitations of bounding box annotations, reference-free counting methods are developed [10,19,32,41], as shown in Fig. 1(b). These methods aim to ascertain the total number of objects in an image without relying on specific cues. Nevertheless, the lack of specificity in counting categories makes these methods prone to errors induced by background noise, as they indiscriminately count all visible objects, leading to a lack of control in the counting process.

In pursuit of more scalable and realistic counting solutions, zero-shot methods [3, 45, 49], illustrated in Fig. 1(c), are introduced. These techniques are designed to count objects from specified classes within an image without prior annotations for those classes, addressing the limitations of both few-shot and reference-free methods by providing enhanced specificity and scalability. These methods can be categorized into two streams. The initial method [13, 14] leans on image-text alignment to comprehend object-related correlations without needing physical exemplars. This method enhances scalability for unidentified classes but

struggles with adequately representing image details for target classes, especially those with atypical shapes, as demonstrated in Fig. 1(c1). Conversely, the second method [45] concentrates on identifying objects through the discovery of class-relevant exemplars. This is achieved by creating pseudo labels that assess the resemblance between image patches and class-generated prototypes. Nevertheless, this method's reliance on arbitrary patch selection hampers its ability to accurately outline entire objects. Additionally, the absence of direct text-image engagement restricts its scalability, tethered to the pre-defined categories present in the training dataset, as illustrated in Fig. 1(c2).

As shown in Fig. 1(c3), we introduce the Visual Association-based Zero-shot Object Counting (VA-Count) framework. VA-Count aims to create a robust link between specific object categories and their corresponding visual representations, ensuring adaptability to various classes. This framework is anchored by three core principles. First, it prioritizes flexibility and scalability, enabling adaptation to novel classes beyond its initial parameters. Second, it enhances precision in identifying exemplary objects, strengthening the connection between visual depictions and their categories. Third, it devises strategies to reduce the effects of localization errors on counting precision. Building on these principles, VA-Count integrates an Exemplar Enhancement Module (EEM) and a Noise Suppression Module (NSM), which are dedicated to refining exemplar identification and mitigating adverse impacts, respectively.

In detail, the EEM expands VA-Count's capacity to handle various classes through the integration of Vision-Language Pretaining (VLP) models, such as Grounding DINO [20]. These VLP models, trained on extensive datasets, excel in identifying a wide range of classes by defining specific categories. In the context of ZOC, it is essential to select exemplars that each contain precisely one object from among the potential bounding boxes that might encompass varying object quantities. To this end, we deploy a binary filter aimed at rigorously refining the set of candidate exemplars, excluding those that fail to comply with the single-object requirement. This filtration step is pivotal for ensuring the precision and consistency necessary for ZOC.

Moreover, even when potential exemplars accurately represent single objects, the unintentional inclusion of exemplars not pertaining to the target category poses a persistent problem. This misalignment introduces uncertainty into the learning process that associates exemplars with images. To counteract this issue, the NSM module operates as a safeguard by identifying negative exemplars, which are unrelated to the intended category. Contrasting with the EEM, which focuses on selecting ideal samples to foster visual connections with images, the NSM employs samples from irrelevant classes to build these associations, utilizing contrastive learning to differentiate between them. This method of contrastive learning acts as a rectifying mechanism, markedly improving the accuracy and efficiency of the associative learning framework.

In summary, our contributions are threefold:

- We introduce a Visual Association-based Zero-shot Object Counting framework, which facilitates high-quality exemplar identification for any class

without needing annotated examples and forges robust visual connections between objects and images.
- We propose an exemplar enhancement model leveraging the universal class-agnostic detection capabilities of the Vision-Language Pretaining model for precise exemplar selection, and a Noise Suppression Module to minimize the adverse effects of incorrect samples in visual associative learning.
- Extensive experiments conducted on two object counting datasets demonstrate the state-of-the-art accuracy and generalizability of VA-Count, underscoring its notable scalability.

2 Related Work

2.1 Class-Specific Object Counting

Object counting plays a crucial role in public safety, public administration, and the liberation of human labor. Currently, class-specific object counting [22, 32, 35, 46, 47] is the predominant method, which entails identifying specific object categories (such as humans [21, 24, 31, 50, 51], vehicles [28, 48], fishes [38], cells [40], *etc.*) leveraging object detection or density estimation and counting accordingly. While these methods show excellence within close-set scenarios with a fixed number of categories, transferring them to arbitrary categories poses challenges. Introducing novel categories necessitates retraining or fine-tuning a counting model with new data, which limits their applicability in real scenarios.

2.2 Class-Agnostic Object Counting

Class-agnostic object counting [8, 26, 29, 36, 42] is proposed for scenarios with less data, which can be divided into few-shot and zero-shot depending on the annotation usage. Specifically, GMN [26] initially frames the class-agnostic counting task as a matching task, leading to FamNet [33], which implements ROI Pooling for broad applicability across FSC-147. As multi-class datasets emerged, the focus shifts towards few-shot methods, where LOCA [41] enhances feature representation and exemplar adaptation; and CounTR [19] utilizes transformers for scalable counting with a two-stage training model. BMNet [**?**] innovates with a bilinear matching network for refined object similarity assessments. In the realm of zero-shot methods, which are categorized into two types, methods like ZSC [45] leverage textual inputs to generate prototypes and filter image patches, thus reducing the need for extensive labeling, albeit with fixed generators that limit scalability. CLIP-Count [13] employs CLIP to encode text and images separately, establishing semantic associations crucial for intuitive counting. VL-Count [14] takes this further by enhancing CLIP's text-image association learning specifically for object counting. Additionally, PseCo [12] introduces a SAM-based multi-task framework that achieves segmentation, dot mapping, and detection on counting data, offering broad application prospects but also necessitating greater computational resources.

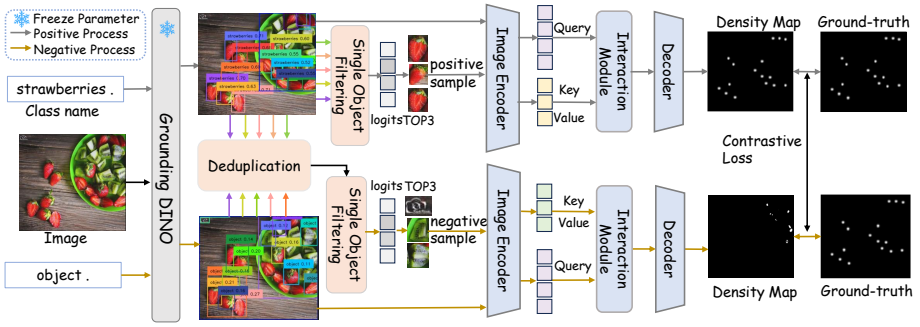

Fig. 2: Overview of the proposed method. Proposed method focuses on two main elements: the Exemplar Enhancement Module (EEM) for improving exemplar quality through a patch selection integrated with Grounding DINO [20], and the Noise Suppression Module (NSM) that distinguishes between positive and negative class samples using density maps. It employs a Contrastive Loss function to refine the precision in identifying target class objects from others in an image.

2.3 Vision-Language Pretaining Model

In recent years, Vision-Language Pretaining (VLP) methods have proven pivotal in enhancing scene understanding and representation learning capabilities. Their adaptability makes them applicable across a wide range of downstream tasks [2, 5–7, 9, 18, 27, 37, 43]. CLIP [30] segregates vision and language features, aligning them through contrastive learning. BLIP [17] introduces a multimodal mixture of encoders and decoders to align different modalities. Building upon this, BLIP2 [16] combines specialized vision and language models to enhance multimodal understanding capabilities through bootstrapping. Grounding DINO [20] incorporates language into close-set detection, improving generalization for open-set detection. The Segment Anything Model (SAM) [15] is based on a prompt-based segmentation task, allowing flexible prompts for zero-shot capabilities across diverse tasks. VLP models, known for their robust multimodal comprehension and scene understanding, significantly advance deep learning and facilitate learning of unknown classes.

3 Proposed Method

3.1 Formula Definition

As shown in Fig. 2, we introduce a Visual Association-based Zero-shot Object Counting framework (VA-Count) focusing on zero-shot, class-agnostic object counting. The categories among the training set C_{train}, validation set C_{val}, and testing set C_{test} are distinguished, ensuring no overlap among them ($C_{\text{train}} \cap C_{\text{val}} \cap C_{\text{test}} = \emptyset$). VA-Count generates density maps D from input images I for

Algorithm 1 Grounding DINO-Guided Exemplar Enhancement Module

I: Input image
T^p: Positive text label ({specific class}), T^n: Negative text label ("object")
B^p: Bounding boxes for positive samples, $\mathcal{S}^p$: Logits for positive samples
B^n: Bounding boxes for negative samples, $\mathcal{S}^n$: Logits for negative samples
τ_l: Logits threshold, τ_{iou}: IoU threshold
$M(\cdot)$: Single Object Classifier
Input: I, T^p, T^n
Output: $\mathcal{O}^p = \{(B^p, \mathcal{S}^p)\}$: Positive outputs, $\mathcal{O}^n = \{(B^n, \mathcal{S}^n)\}$: Negative outputs
Grounding DINO Process:
$F \leftarrow \text{ExtractFeatures}(I)$
$\mathcal{S}^p, B^p \leftarrow \text{Detect}(F, T^p)$, filter by τ_l; *and* $\mathcal{S}^n, B^n \leftarrow \text{Detect}(F, T^n)$, filter by τ_l
Deduplication and Filtering:
Initialize $B^n_{\text{filtered}}, B^p_{\text{new}}, B^n_{\text{new}}$
for b^n in B^n **do** ▷ Remove duplicates
    if b^n is unique in B^p with IoU $< \tau_{\text{iou}}$ **then**
        B^n_{filtered}.append(b^n)
    end if
end for
for all $b \in B^p \cup B^n_{\text{filtered}}$ **do** ▷ Single object filter
    if $M(b)$ is true **then**
        Add b to the appropriate new set
    end if
end for
Update $\mathcal{O}^p, \mathcal{O}^n$ with new sets

any given class C, and counts objects using these density maps. Specifically, VA-Count utilizes pseudo-exemplars E^p to enhance image-text associations, acting as a bridge to establish robust visual correlations between E^p and the images I. To extract exemplars from images, we propose the use of two key modules: the Exemplar Enhancement Module (EEM) (*cf*. Sec. 3.2) and the Noise Suppression Module (NSM) (*cf*. Sec. 3.3).

To alleviate the noise introduced by objects belonging to other classes on the target objects within images, the EEM and NSM are simultaneously used to obtain positive exemplars B^p and negative exemplars B^p. The EEM consists of Grounding DINO $G(\cdot)$ and a filtering module $\Phi(\cdot)$. There are different filtering modules for positive and negative samples $\Phi^p(\cdot)$ and $\Phi^n(\cdot)$ respectively. $\Phi^p(\cdot)$ is a binary classifier, while $\Phi^n(\cdot)$ consists of a binary classifier and a deduplication module. The two kinds of pseudo-exemplars and images are then fed into the Counter $\Gamma(\cdot)$ simultaneously for correlation learning. $\Gamma(\cdot)$ comprises an image encoder, correlation module, and decoder. The optimization goal of this paper is as follows, where $\mu(\cdot)$ denotes the similarity, and D^p, D^n, D^g represent the density maps for positive, negative, and ground truth respectively:

$$D^p = \Gamma\left(\Phi^p\left(G\left(I, T^p\right)\right)\right), \quad D^n = \Gamma\left(\Phi^n\left(G\left(I, T^n\right)\right)\right), \tag{1}$$

$$\text{Objective} = \begin{cases} \max \mu(D^p, D^g), \\ \min \mu(D^n, D^g). \end{cases} \tag{2}$$

3.2 Exemplar Enhancement Module

We introduce an Exemplar Enhancement Module (EEM) for detecting objects within images and refining the detected objects as target exemplars. The workflow of the EEM is outlined in Algorithm 1. The EEM ensures VA-Count's scalability to arbitrary classes by incorporating Vision-Language Pretaining (VLP) models (*e.g.*, Grounding DINO [20]) for potential exemplar discovery, renowned for its efficiency in feature extraction and precision in object localization. Furthermore, the EEM involves meticulously discovering and refining potential exemplars to enhance the quality of positive and negative exemplars for precise object counting.

Grounding DINO-Guided Box Selection. Given the training set input image I_i, accompanied by predefined sets of positive text labels $T_i^p = \{C_i\}$ and negative text labels $T_i^n =$ "object", where C_i represents the specified target class for the input image and T_i^n is fixed as "object". These labels correspond to the target objects and the noise objects, respectively. Taking positive exemplar discovery as an example, Grounding DINO assigns logits value $\mathcal{S}_i^p = \{s_{i,j}\}_{j=0}^m$ to all candidate bounding boxes $B_i^p = \{b_{i,j}\}_{j=0}^m$ based on T_i^p, m denotes the number of candidate boxes within the image. For the j-th box in the i-th image, $s_{i,j}$ represents the likelihood that $b_{i,j}$ belongs to the specified class text C_i. The output of positive candidate boxes $\mathcal{O}^p$ can be formulated as:

$$\mathcal{O}^p = \{G(I_i, T_i^p)\}_{i=0}^k = \{(B_i^p, \mathcal{S}_i^p)\}_{i=0}^k, \tag{3}$$

where k denotes the number of images in the training set.

Negative Samples and Deduplication. To minimize the impact of irrelevant classes on the counting accuracy of the target object, we adopt a filtering method for negative samples. Initially, we obtain all candidate bounding boxes for objects within each image. Similar to Eq. (3), the negative candidate boxes $\mathcal{O}^n$ without filtering can be formulated as:

$$\mathcal{O}^n = \{G\left(I_i, T_i^n\right)\}_{i=0}^k = \{(B_i^n, \mathcal{S}_i^n)\}_{i=0}^k, \tag{4}$$

where for each image I_i, the term $T_i^n =$ "object" is employed to identify and generate all bounding boxes B^n within that image. This method guarantees the detection of bounding boxes for all objects present in the image.

Then, for each image I_i, we assess each bounding box b^n from the negative candidate boxes B^n, and each b^n is evaluated to determine its uniqueness in relation to the boxes within B^p. Specifically, a bounding box is deemed unique if its overlap with any box in B^p is minimal, based on the Intersection over Union (IoU) threshold τ_{iou}, which can be formulated as:

$$\text{IoU}\left(B^p, B^n\right) = \frac{B^p \cap B^n}{B^p \cup B^n}, \tag{5}$$

where $B^p \cap B^n$ and $B^p \cup B^n$ denotes the intersection and union between positive B^p and negative B^n boxes. Unique negative boxes b^n are then included in the final set B^n_{filtered} of negative exemplars.

Single Object Exemplar Filtering. While DINO excels at identifying targets for arbitrary classes, each candidate box does not always contain a single object because boxes encompassing multiple objects may carry higher confidence levels than boxes of single objects. To ensure the integrity of the visual connections established with images, it's imperative to select exemplars that exclusively contain a single object. To achieve this, we treat singular discrimination as a binary classification task, using the binary classifier $\delta(\cdot)$ to refine candidate bounding boxes, ensuring each exemplar contains a single object.

As shown in Fig. 3, $\delta(\cdot)$ leverages a frozen Clip-vit backbone, integrated with a trainable Feed-Forward Network (FFN) for binary classification tasks. Training data is meticulously curated, consisting of samples of single and multiple objects. The labeled single-object samples are the exemplars in the training sets, and the labeled multi-object samples consist of randomly cropped patches and the entire image. To ensure that the class-agnostic counting is maintained, the training data is split for training and evaluation with disjoint samples, ensuring robust exemplar assessment. The classification results for positive candidate boxes $b^p \in B^p$ can be formulated as:

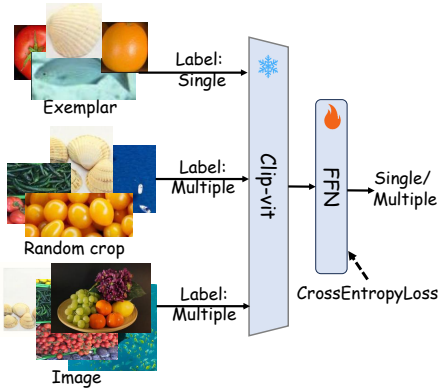

Fig. 3: Illustration of the single object exemplar filtering with a frozen Clip-vit encoder and a trainable FFN to distinguish single from multiple objects.

$$\delta\left(b^p\right) = \text{FFN}\left(\text{Clip-vit}\left(b^p\right)\right), \quad (6)$$

and the filtered set B_{new} contains bounding boxes b^p that are conditioned on the classification results, which can be formulated as:

$$B^p_{\text{new}} \leftarrow B^p_{\text{new}} \cup \left\{b | \delta\left(b^p\right) = 1\right\}, \quad (7)$$

where the symbol $\leftarrow$ signifies the update operation for the set B^p_{new}, and the set builder notation $\{b|\delta(b^p) = 1\}$ represents the collection of bounding boxes for which $\delta(b^p)$ predicts a positive outcome.

3.3 Noise Suppression Module

In the context of the EEM, text-image alignment is redefined as object-image alignment by identifying positive B^p and negative B^n exemplars. We delves

into generating positive and negative density maps and alleviating the noise introduced by the negative exemplars.

Initially, for each image I_i, we select the top three patches with the highest S^p from the positive candidate boxes B^p_{new} as positive exemplars $E^p = \{b^p_i\}^k_{i=1}$, and the top three patches with the highest S^n from the negative candidate boxes B^n_{filtered} as negative exemplars $E^n = \{b^n_i\}^k_{i=1}$. Following CounTR [19], we build the Counter $\Gamma(\cdot)$ with feature interaction to fuse information from both image encoders. Specifically, we merge encoder outputs by using image features as queries and the linear projections of sample features as keys and values, ensuring dimension consistency with image features, in accordance with the self-similarity principle in counting, which can be formulated as:

$$\boldsymbol{F}_{\text{fuse}} = \Gamma_{\text{fuse}}(\boldsymbol{F}_{\text{query}}, \boldsymbol{W}^k \boldsymbol{F}_{\text{key}}, \boldsymbol{W}^v \boldsymbol{F}_{\text{value}}) \in \mathbb{R}^{M \times D}, \tag{8}$$

where $\boldsymbol{F}$ denotes the feature representations, $\boldsymbol{W}^k$ and $\boldsymbol{W}^v$ are the learnable weights for keys and values from $\{E^p, E^n\}$, M denotes the number of tokens, D is the feature dimensionality, and $\mathbb{R}^{M \times D}$ the space of the feature matrix. The decoder outputs the density heatmap after up-sampling the fused features to the input image's dimensions:

$$D^n_i = \Gamma_{\text{decode}}\left(\boldsymbol{F}^n_{\text{fuse}}\right), \quad D^p_i = \Gamma_{\text{decode}}\left(\boldsymbol{F}^p_{\text{fuse}}\right). \tag{9}$$

Contrastive Learning and Loss Functions. The objective of the NSM in VA-Count is to reduce the impact of noise in images on counting performance while ensuring the accuracy of density map predictions. To achieve this, a contrastive loss $\mathcal{L}_C$ is proposed, using specified class density maps as positive samples and non-specified class density maps as negative samples. This involves maximizing the similarity between positive density maps and the ground-truth density maps and minimizing the similarity between negative density maps and the ground-truth density maps, as detailed in Eq. (10). To guide density map generation, we use the loss method from CounTR [19].

The density loss $\mathcal{L}_D$ is calculated as the mean squared error between each pixel of the density map D^p_i generated for positive samples and the ground-truth density map D^g_i, as shown in Eq. (11). H and W respectively denote the height and width of the density map.

$$\mathcal{L}_C(D^p_i, D^g_i, D^n_i) = -\log \frac{\exp \text{sim}\left(D^p, D^g\right)}{\exp \text{sim}\left(D^p, D^g\right) + \exp \text{sim}\left(D^n, D^g\right)}, \tag{10}$$

$$\mathcal{L}_D\left(D^p_i, D^g_i\right) = \frac{1}{HW} \sum \left\| D^p_i - D^g_i \right\|^2_2, \tag{11}$$

$$\mathcal{L}_{\text{total}}\left(D^p_i, D^g_i, D^n_i\right) = \mathcal{L}_C + \mathcal{L}_D. \tag{12}$$

4 Experimental Result

4.1 Datasets and Implementation Details

Datasets. FSC-147 [10] dataset is tailored for class-agnostic counting with 6,135 images and 147 classes. Unique for its non-overlapping class subsets, it

provides class labels and dot annotations for zero-shot counting using textual prompts.

CARPK [11] dataset offers a bird's-eye view of 89,777 cars in 1,448 parking lot images, testing the method's cross-dataset transferability and adaptability.

Evaluation Metrics. Following previous class-agnostic object counting methods [29], the evaluation metrics employed are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). MAE is widely used to assess model accuracy, while RMSE evaluates model robustness.

Exemplar Enhancement Module uses Grounding DINO[7] for bounding box proposals, setting the threshold τ_l to 0.02. For negative sample filtering, the IoU threshold τ_{iou} is set to 0.5. The single object classifier employs CLIP ViT-B/16[8] as its backbone, with an FFN comprising two linear layers, trained over 100 epochs at a learning rate of e-4. The dataset is partitioned in a 7:3 ratio.

Noise Suppression Module follows CounTR's [19] two-stage training: MAE pretraining and AdamW [25]-optimized fine-tuning. It is trained on FSC-147 with a learning rate of 10^{-5}, batch size of 8, on an NVIDIA RTX L40 GPU.

4.2 Comparison with the State-of-the-Arts

For the performance evaluation of our method, it is benchmarked against a variety of state-of-the-art few-shot and zero-shot counting methods on FSC-147. Additionally, we evaluate our method in comparison with class-specific counting models on CARPK.

Quantitative Result on FSC-147. We evaluate the effectiveness of VA-Count on FSC-147, comparing it with state-of-the-art counting methods as detailed in Tab. 1. Our method surpasses the exemplar-discovery method ZSC [45], demonstrating that the exemplars found by VA-Count are of higher quality. VA-Count achieves the best performance in MAE and second in RMSE, validating our method's effectiveness. Despite being second in RMSE, it still outperforms ZSC. In comparison with CLIP-Count [13], VA-Count, due to some noise introduction, has a few inferior samples but, overall, surpasses CLIP-Count in performance.

Quantitative Result on CARPK. In Tab. 2, VA-Count's cross-domain and non-cross-domain performance on CARPK are compared with previous methods. In the zero-shot group, VA-Count achieves the best performance, particularly with its cross-domain performance methoding that of the few-shot group, demonstrating its outstanding transferability. It is worth noting that employing $\Phi(\cdot)$ significantly reduces errors compared to directly using the Grounding DINO [20] method. In the absence of any training data, VA-Count outperforms FamNet [33] in the cross-domain group.

Ablation Study. We conduct both quantitative and qualitative analyses on the contributions of each component in our proposed VA-Count, which includes the Grounding-DINO candidate box extraction and filtering module. The quantitative outcomes are presented in Tab. 3. Using only Grounding DINO method

[7] https://github.com/IDEA-Research/GroundingDINO
[8] https://github.com/openai/CLIP

Table 1: Quantitive results of our VA-Count and other state-of-the-art competitors on FSC-147. F-S, R-F, and Z-S are abbreviated for Few-shot, Reference-free, and Zero-shot settings. Best results for each scheme and the second-best results at the zero-shot setting are highlighted in bold and underline.

Scheme	Method	Venue	Shot	Val Set		Test Set		Avg	
				MAE	RMSE	MAE	RMSE	MAE	RMSE
F-S	FamNet [33]	CVPR'21	3	24.32	70.94	22.56	101.54	23.44	86.24
	CFOCNet [46]	WACV'21	3	21.19	61.41	22.10	112.71	21.65	87.06
	CounTR [19]	BMVC'22	3	13.13	49.83	11.95	91.23	12.54	70.53
	LOCA [41]	ICCV'23	3	**10.24**	**32.56**	10.97	56.97	10.61	44.77
	SAM [36]	WACV'24	3	-	-	19.95	132.16	19.95	132.16
	PseCo [12]	CVPR'24	3	15.31	68.34	13.05	112.86	14.18	90.60
	CACViT [42]	AAAI'24	3	10.63	37.95	**9.13**	**48.96**	**9.88**	**43.46**
	FamNet [33]	CVPR'21	1	26.05	77.01	26.76	110.95	26.41	93.98
R-F	FamNet [33]	CVPR'21	0	32.15	98.75	32.27	131.46	32.21	115.11
	RepRPN-C [32]	ACCV'22	0	29.24	98.11	26.66	129.11	27.95	113.61
	CounTR [19]	BMVC'22	0	18.07	71.84	**14.71**	106.87	**16.39**	89.36
	RCC [10]	CVPR'23	0	17.49	58.81	17.12	104.53	17.31	81.67
	LOCA [41]	ICCV'23	0	**17.43**	**54.96**	16.22	**103.96**	16.83	**79.46**
Z-S	ZSC [45]	CVPR'23	0	26.93	88.63	22.09	<u>115.17</u>	24.51	101.90
	CLIP-Count [13]	MM'23	0	<u>18.79</u>	**61.18**	<u>17.78</u>	**106.62**	18.285	**83.90**
	PseCo [12]	CVPR'24	0	23.90	100.33	**16.58**	129.77	<u>20.24</u>	115.05
	VA-Count	Ours	0	**17.87**	<u>73.22</u>	17.88	129.31	**17.87**	<u>101.26</u>

(first row) achieves an error of 52.82 without training, which, although not as accurate as regression-based methods, ensures the detection of relevant objects. Performance improves slightly after adding a single-object classification filter (second row). With training based on $\mathcal{L}_D$, it already meets counting requirements. In Tab. 2, we compare using Grounding DINO alone and with a single-object classification filter on CARPK (last three rows). Our binary classifier significantly improves performance, reducing MAE and RMSE by about 10.

4.3 Qualitative Analysis

Analysis of the zero-shot performance. To further ensure the effectiveness of the proposed VA-Count framework, we visualize qualitative results in Fig. 4. We provide a side-by-side comparison of the proposed VA-Count against the few-shot counting method [19]. VA-Count achieves a remarkable resemblance to the ground truth, showcasing the method's nuanced understanding of object boundaries and densities and being less affected by the background noise. Specifically, the first row shows there exists a golden egg drowned by white eggs. The few-shot method struggled with this nuanced differentiation, failing to recognize the golden egg distinctly. In the second row, strawberries near flowers also confound the few-shot

Table 2: Quantitative results of our VA-Count and other state-of-the-art competitors on CARPK. $\Phi(\cdot)$ denotes the single-object classification filter. C and F denote CARPK and FSC-147, respectively.

Methods	Venue	Shot	C → C		F → C	
			MAE	RMSE	MAE	RMSE
FamNet [33]	CVPR'21	3	18.19	33.66	28.84	44.47
GMN [26]	CVPR'21	3	7.48	9.90	-	-
BMNet+ [35]	CVPR'22	3	5.76	7.83	**10.44**	**13.77**
CounTR [19]	BMVC'22	3	**5.75**	**7.45**	-	-
RCC [10]	CVPR'23	0	9.21	11.33	21.38	26.61
CLIP-Count [13]	MM'23	0	-	-	11.96	16.61
Grounding DINO [20]	arXiv'24	0	29.72	31.60	29.72	31.60
Grounding DINO + $\Phi(\cdot)$	Ours	0	18.54	21.71	18.54	21.71
VA-Count	Ours	0	**8.75**	**10.30**	**10.63**	**13.20**

Table 3: Ablation study on each component's contribution to the final results on FSC-147. We demonstrate the effectiveness of two parts of our framework and two types of loss: $G(\cdot)$ for Grounding DINO, $\Phi(\cdot)$ for the single-object filtering section, the density loss $\mathcal{L}_D$, and the contrastive loss $\mathcal{L}_C$.

$G(\cdot)$	$\phi(\cdot)$	$\mathcal{L}_D$	$\mathcal{L}_C$	Val Set		Test Set	
				MAE	RMSE	MAE	RMSE
●	○	○	○	52.82	134.49	54.48	159.30
●	●	○	○	52.12	135.29	54.27	159.76
●	●	●	○	19.63	73.94	18.93	**116.65**
●	●	●	●	**17.87**	**73.22**	**17.88**	129.31

method. These examples emphasize VA-Count's superior ability to identify and differentiate between objects with minor differences. The third row presents a challenging scenario with dense keys partially occluded by hands. This situation tests the model's ability to count tiny, closely situated objects under partial occlusion, showcasing VA-Count's advanced capability to accurately identify and count such challenging objects, which is significantly better than the few-shot method. These results highlight the impact of exemplar selection and the incorporation of negative patches in VA-Count, significantly enhancing its object counting and localization capabilities, and showcasing its innovation in zero-shot object counting.

Analysis of Positive and Negative Exemplars. To make our experiment more straightforward, we also conduct a qualitative analysis of the patch selection. As shown in Fig. 5 and Fig. 6, we illustrate selected positive and negative patches for various categories under a zero-shot setting. Taking a closer look at the positive patches for categories such as crab cakes and green peas, the results show a high degree of accuracy in the model's ability to isolate and highlight the regions

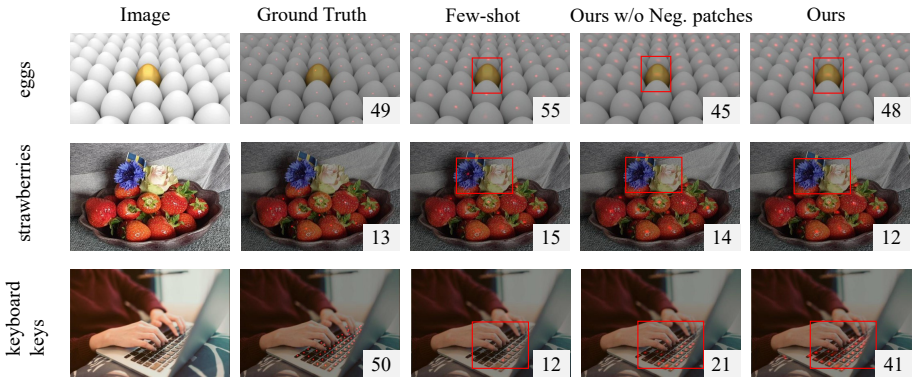

Fig. 4: Illustration of heatmaps compared with few-shot method [19] on FSC-147. Predicted density map is overlaid on the original RGB image. (Best viewed in zoom in)

Fig. 5: Illustration of the positive (Pos.) and negative (Neg.) exemplars on FSC-147.

containing the target objects. This precision underscores the effectiveness of VA-Count framework in discerning relevant features amidst complex backgrounds, affirming its robustness in the exemplar discovery. Negative patches, especially from categories like strawberries and crab cakes, highlight the model's challenges with visually similar or overlapping areas not in the target category, underscoring the need for improved discriminative abilities. This analysis underscores our paper's impact on zero-shot object counting and the importance of refining visual learning and exemplar selection for future advancements.

Effective of the object exemplar filter. The effectiveness of the object exemplar filter is further evaluated by comparing visualization grounding results with and without the filter. Fig. 7 illustrates this comparison for the category of cars on CARPK. Images without the filter show multiple cars within a single

Fig. 6: Illustration of the final positive (Pos.) and negative (Neg.) exemplars for images on CARPK.

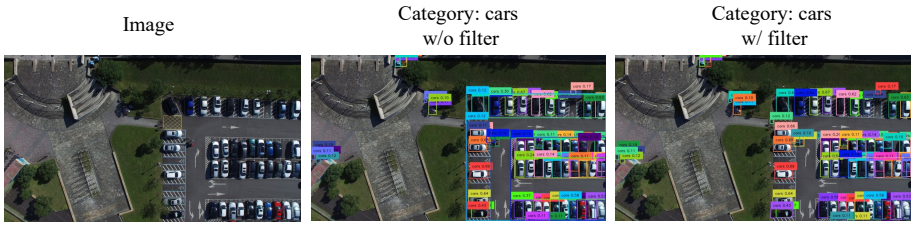

Fig. 7: Illustration of candidate boxes before and after exemplar filter for images on CARPK.

bounding box, indicating Grounding DINO's [20] inability to isolate individual objects effectively. Conversely, images with the filter applied demonstrate a significant improvement, with bounding boxes accurately encompassing single cars. This clear distinction highlights the binary classifier's crucial role in ensuring precise object counting by enforcing the single-object criterion within each exemplar, validating the filter's contribution to enhancing the model's accuracy and reliability in VA-Count framework.

5 Conclusion

This paper addresses the challenges in class-agnostic object counting by introducing the Visual Association-based Zero-shot Object Counting (VA-Count) framework. VA-Count effectively balances the need for scalability across arbitrary classes with the establishment of robust visual connections, overcoming the limitations of existing Zero-shot Object Counting (ZOC) methods. VA-Count comprises an Exemplar Enhancement Module (EEM) and a Noise Suppression Module (NSM), which are dedicated to refining exemplar identification and mitigating adverse impacts, respectively. The EEM utilizes advanced Vision-Language Pretaining models like Grounding DINO for scalable exemplar discovery, while the NSM mitigates the impact of erroneous exemplars through contrastive learning. VA-Count shows promise in zero-shot counting, performing well on three datasets and offering precise visual associations and scalability. In the future, we will explore and better utilize advanced visual language models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62271361, the Sanya Yazhou Bay Science and Technology City Administration scientific research project under Grant 2022KF0021, the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2023B1515020097, and the National Research Foundation Singapore under the AI Singapore Programme under Grant AISG3-GV-2023-011.

References

1. Arteta, C., Lempitsky, V.S., Zisserman, A.: Counting in the wild. In: Proc. Eur. Conf. Comput. Vis. pp. 483–498 (2016)
2. Bai, Y., Cao, M., Gao, D., Cao, Z., Chen, C., Fan, Z., Nie, L., Zhang, M.: RaSa: Relation and sensitivity aware representation learning for text-based person search. In: Proc. Int. Joint Conf. Artif. Intell. pp. 555–563 (2023)
3. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: Proc. Eur. Conf. Comput. Vis. pp. 397–414 (2018)
4. Chai, L., Liu, Y., Liu, W., Han, G., He, S.: CrowdGAN: Identity-free interactive crowd video generation and beyond. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2856–2871 (2022)
5. Chen, C., Ye, M., Jiang, D.: Towards modality-agnostic person re-identification with descriptive query. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 15128–15137 (2023)
6. Dou, Z., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., Gao, J., Wang, L.: Coarse-to-fine vision-language pre-training with fusion in the backbone. In: Adv. Neural Inf. Process. Syst. pp. 32942–32956 (2022)
7. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 14084–14093 (2022)
8. Gong, S., Zhang, S., Yang, J., Dai, D., Schiele, B.: Class-agnostic object counting robust to intraclass diversity. In: Proc. Eur. Conf. Comput. Vis. pp. 388–403 (2022)
9. He, S., Chen, W., Wang, K., Luo, H., Wang, F., Jiang, W., Ding, H.: Region generation and assessment network for occluded person re-identification. IEEE Trans. Inf. Forensics Secur. **19**, 120–132 (2023)
10. Hobley, M., Prisacariu, V.: Learning to count anything: Reference-less class-agnostic counting with weak supervision. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2023)
11. Hsieh, M., Lin, Y., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 4165–4173 (2017)
12. Huang, Z., Dai, M., Zhang, Y., Zhang, J., Shan, H.: Point, segment and count: A generalized framework for object counting. arXiv:2311.12386 (2023)
13. Jiang, R., Liu, L., Chen, C.: CLIP-Count: Towards text-guided zero-shot object counting. In: Proc. ACM Multimedia. pp. 4535–4545 (2023)
14. Kang, S., Moon, W., Kim, E., Heo, J.: VLCounter: Text-aware visual representation for zero-shot object counting. In: Proc. AAAI Conf. Artif. Intell. pp. 2714–2722 (2024)

15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 3992–4003 (2023)
16. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. Int. Conf. Mach. Learn. pp. 19730–19742 (2023)
17. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proc. Int. Conf. Mach. Learn. pp. 12888–12900 (2022)
18. Li, S., Sun, L., Li, Q.: CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. In: Proc. AAAI Conf. Artif. Intell. pp. 1405–1413 (2023)
19. Liu, C., Zhong, Y., Zisserman, A., Xie, W.: CounTR: Transformer-based generalised visual counting. In: Proc. Brit. Mach. Vis. Conf. p. 370 (2022)
20. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv:2303.05499 (2023)
21. Liu, X., Yang, J., Ding, W., Wang, T., Wang, Z., Xiong, J.: Adaptive mixture regression network with local counting map for crowd counting. In: Proc. Eur. Conf. Comput. Vis. pp. 241–257 (2020)
22. Liu, Y., Ren, S., Chai, L., Wu, H., Xu, D., Qin, J., He, S.: Reducing spatial labeling redundancy for active semi-supervised crowd counting. IEEE Trans. Pattern Anal. Mach. Intell. **45**(7), 9248–9255 (2023)
23. Liu, Y., Wen, Q., Chen, H., Liu, W., Qin, J., Han, G., He, S.: Crowd counting via cross-stage refinement networks. IEEE Trans. Image Process. **29**, 6800–6812 (2020)
24. Liu, Y., Xu, D., Ren, S., Wu, H., Cai, H., He, S.: Fine-grained domain adaptive crowd counting via point-derived segmentation. In: Proc. IEEE Int. Conf. Multimedia Expo. pp. 2363–2368 (2023)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. Int. Conf. Learn. Represent. (2019)
26. Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: Proc. Asian Conf. Comput. Vis. pp. 669–684 (2019)
27. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. In: Adv. Neural Inf. Process. Syst. pp. 35087–35102 (2022)
28. Mundhenk, T.N., Konjevod, G., Sakla, W.A., Boakye, K.: A large contextual dataset for classification, detection and counting of cars with deep learning. In: Proc. Eur. Conf. Comput. Vis. pp. 785–800 (2016)
29. Nguyen, T., Pham, C., Nguyen, K., Hoai, M.: Few-shot object counting and detection. In: Proc. Eur. Conf. Comput. Vis. pp. 348–365 (2022)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
31. Ranjan, V., Le, H.M., Hoai, M.: Iterative crowd counting. In: Proc. Eur. Conf. Comput. Vis. pp. 278–293 (2018)
32. Ranjan, V., Nguyen, M.H.: Exemplar free class agnostic counting. In: Proc. Asian Conf. Comput. Vis. pp. 71–87 (2022)
33. Ranjan, V., Sharma, U., Nguyen, T., Hoai, M.: Learning to count everything. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 3394–3403 (2021)

34. Sam, D.B., Agarwalla, A., Joseph, J., Sindagi, V.A., Babu, R.V., Patel, V.M.: Completely self-supervised crowd counting via distribution matching. In: Proc. Eur. Conf. Comput. Vis. pp. 186–204 (2022)
35. Shi, M., Lu, H., Feng, C., Liu, C., Cao, Z.: Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 9529–9538 (2022)
36. Shi, Z., Sun, Y., Zhang, M.: Training-free object counting with prompts. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. pp. 323–331 (2024)
37. Song, S., Wan, J., Yang, Z., Tang, J., Cheng, W., Bai, X., Yao, C.: Vision-language pre-training for boosting scene text detectors. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 15681–15691 (2022)
38. Sun, G., An, Z., Liu, Y., Liu, C., Sakaridis, C., Fan, D., Van Gool, L.: Indiscernible object counting in underwater scenes. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 13791–13801 (2023)
39. Tian, C., Zhang, X., Liang, X., Li, B., Sun, Y., Zhang, S.: Knowledge distillation with fast CNN for license plate detection. IEEE Trans. Intell. Transp. Syst. (2023)
40. Tyagi, A.K., Mohapatra, C., Das, P., Makharia, G., Mehra, L., AP, P., Mausam: DeGPR: Deep guided posterior regularization for multi-class cell detection and counting. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 23913–23923 (2023)
41. Đukic, N., Lukezic, A., Zavrtanik, V., Kristan, M.: A low-shot object counting network with iterative prototype adaptation. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 18872–18881 (2023)
42. Wang, Z., Xiao, L., Cao, Z., Lu, H.: Vision transformer off-the-shelf: A surprising baseline for few-shot class-agnostic counting. In: Proc. AAAI Conf. Artif. Intell. pp. 5832–5840 (2024)
43. Xie, D., Liu, L., Zhang, S., Tian, J.: A unified multi-modal structure for retrieving tracked vehicles through natural language descriptions. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops. pp. 5418–5426 (2023)
44. Xiong, Z., Chai, L., Liu, W., Liu, Y., Ren, S., He, S.: Glance to count: Learning to rank with anchors for weakly-supervised crowd counting. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. pp. 342–351 (2024)
45. Xu, J., Le, H., Nguyen, V., Ranjan, V., Samaras, D.: Zero-shot object counting. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 15548–15557 (2023)
46. Yang, S., Su, H., Hsu, W.H., Chen, W.: Class-agnostic few-shot object counting. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. pp. 869–877 (2021)
47. You, Z., Yang, K., Luo, W., Lu, X., Cui, L., Le, X.: Few-shot object counting with similarity-aware feature enhancement. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. pp. 6304–6313 (2023)
48. Zhang, Z., Liu, K., Gao, F., Li, X., Wang, G.: Vision-based vehicle detecting and counting for traffic flow analysis. In: Proc. IEEE Int. Joint Conf. Neural Networks. pp. 2267–2273 (2016)
49. Zheng, Y., Wu, J., Qin, Y., Zhang, F., Cui, L.: Zero-shot instance segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 2593–2602 (2021)
50. Zhu, H., Yuan, J., Zhong, X., Liao, L., Wang, Z.: Find gold in sand: Fine-grained similarity mining for domain-adaptive crowd counting. IEEE Trans. Multimedia **26**, 3842–3855 (2024)
51. Zhu, H., Yuan, J., Zhong, X., Yang, Z., Wang, Z., He, S.: DAOT: Domain-agnostically aligned optimal transport for domain-adaptive crowd counting. In: Proc. ACM Multimedia. pp. 4319–4329 (2023)