Contrastive Point Feature Matching for Open-world Object Counting

Ngo Xuan Cuong
University of Arkansas
cngo@uark.edu

Abstract

Open-world object counting refers to estimating the number of target objects specified by a text description. Current state-of-the-art density-based methods rely on density map regression trained with counting-supervised losses, which often overlook the spatial and visual correspondence between image regions and object instances. Furthermore, the absence of spatial supervision leads to inconsistent performance in challenging scenarios such as crowding and occlusion. We propose CPMNet (Contrastive Point Matching Network), a novel framework that introduces spatial supervision through contrastive learning with point-level guidance. CPMNet regress object point locations from the predicted density map through Sinkhorn EM (expectation maximization), extracts their corresponding features, and applies contrastive learning to enhance discrimination between target objects and irrelevant regions. Experimental results on a benchmark dataset demonstrate that our method offers enhanced performance in open-world counting, providing a more precise and reliable solution for estimating object counts.

1 Introduction

Object counting traditionally focuses on estimating the number of instances belonging to a specific class within an image, such as crowd counting, vehicle counting, or car counting. Recently, the field has advanced toward class-agnostic few-shot counting, where the target object is specified at inference time using visual exemplars, typically user-provided bounding boxes highlighting a few instances of the object in the image. While class-agnostic few-shot counters are suitable for open-world scenarios, they rely on manual input at test time, limiting automation.

To address this, recent work by Xu et al. [19] proposes specifying the object of interest using a textual description instead of visual exemplars. This opens up the possibility of open-world object counting, where arbitrary object categories can be queried at inference via class names. Existing approaches to this problem typically rely on density-based methods, which estimate object count through a predicted density map. However, these methods are constrained by their reliance solely on textual descriptions, lacking rich object representations and spatial correspondence. As a result, they often generate mismatched or misaligned density regions, leading to poor counting accuracy, especially in cluttered scenes or when the text-to-object match is ambiguous.

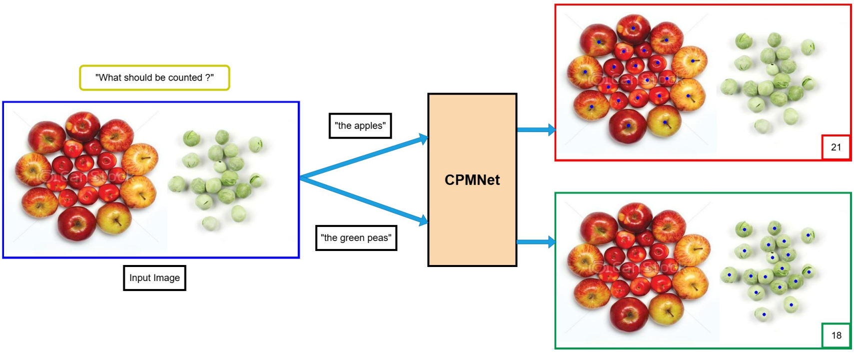

Figure 1: Open-world object counting on multiple object types using our proposed approach, CPMNet.

To overcome these limitations, we propose Contrastive Point Matching Network (CPMNet), which introduces spatial supervision through contrastive learning with point-level annotations. CPMNet first generates a density map using a transformer-based backbone. It then regresses object locations from this density distribution using the Sinkhorn EM algorithm, producing a set of predicted points. These points are matched one-to-one with ground truth points using the Hungarian algorithm, yielding **K** matched pairs. We extract features from the matched points using a CNN backbone. Matched point pairs are treated as positives, while unmatched or false-positive points are considered negatives. Contrastive learning is then applied to these point triplets, guiding the model to distinguish between relevant and irrelevant regions.

This contrastive training encourages the model to produce more informative and spatially aware density maps. It enhances the model's ability to focus attention on the correct object regions and improves its discrimination between target objects and background or distractors. The key idea behind CPMNet is to leverage contrastive learning to inject spatial consistency and better object understanding into the density estimation process. Our approach significantly improves performance on the FSC147 dataset and demonstrates robustness to the challenges inherent in open-world object counting.

The contributions of this paper are threefold:

- First, we propose a novel approach to open-world object counting, Contrastive Point Matching, which introduces spatial supervision into the counting process and enhances object representations within the model.

- Second, we design CPMNet, a model that leverages contrastive learning for point matching. This allows the model to incorporate spatial correspondence and improve density-based counting in open-world settings.

- Third, we conduct extensive experiments on the widely-used FSC147 dataset and perform detailed ablation studies to validate the effectiveness of each component, demonstrating significant improvements in counting accuracy.

2 Related Work

Class-Specific Object Counting. Computer vision researchers have long faced challenges in visual object counting, with much of the early work focusing on specific object categories such as cars, cells, humans, or polyps. Detection-based techniques were commonly used in the early stages of this research, aiming to detect and count instances of known object classes. For example, complex detection algorithms like YOLO are widely employed in vehicle counting systems, while segmentation-based approaches are often used in biological cell counting [16] and human counting [18] to achieve accurate results. However, detection-based methods tend to perform poorly in crowded scenes due to occlusions and object overlaps, which significantly increase counting errors. To address these limitations, density-based methods emerged as a more robust alternative, as they are better suited to handle occlusions by estimating object counts through continuous density maps.

Class-Agnostic Object Counting. More recently, class-agnostic counting has attracted considerable attention due to its ability to generalize across unseen object categories. The key idea is to count the number of object instances in an image, given a small set of visual exemplars at inference time. These exemplars, usually in the form of user-provided bounding boxes—enable the model to adapt to novel categories without retraining. Early approaches utilized Siamese matching networks to predict density maps. More advanced methods, such as BMNet+ [14], incorporate non-linear similarity metrics into the representation learning process. FamNet [12] improves the backbone architecture for enhanced density estimation, while CounTR [8] combines convolutional encoders with vision transformers and cross-attention mechanisms for better feature fusion. SAFECount [20] enhances generalization through a dedicated feature refinement module. LOCA [17] introduces an object prototype extraction module for iterative adaptation based on the exemplar's appearance and shape. Despite their advances, these methods are inherently limited by their reliance on the quality and informativeness of visual exemplars. If the guidance is biased, sparse, or noisy, the performance may degrade. Moreover, these methods do not allow users to specify the class of interest using textual descriptions, limiting their flexibility in open-world scenarios.

Text-Specified Object Counting. Text-specified object counting aims to count instances of an arbitrary object category in an image, given only a textual description of the target class. These methods are inherently zero-shot, eliminating the need for manually annotated visual exemplars or category-specific training. Notably, Paiss et al. [9] fine-tuned CLIP to enable it to count up to ten objects, demonstrating its potential for class-agnostic tasks. Building on this idea, CLIP-Count [3] and VLCounter [4] use CLIP's text encoder to embed class names, which interact with visual features to generate density maps. TFOC [15] further pushes this direction by introducing prompt-based querying (points, boxes, or text) to the segmentation model SAM [5], enabling training-free object counting. However, existing density-based methods have limitations. Trained purely with counting supervision, they lack spatial correspondence and rely on weak object representations from foundation models, leading to ambiguous and inaccurate density maps. To address this, CPMNet introduces a contrastive point matching scheme that encourages the model to produce spatially informative density maps and learn more precise object representations. This enables the model to better distinguish target objects from background noise and generate more focused density predictions, improving counting accuracy.

3 Method

Figure 2 is the overview of the CPMNet framework. The model follows the standard open-world object counting architecture, consisting of three main components: an encoder, a feature interaction module, and a decoder. The encoder extracts visual and textual features using CLIP. These features are fused in the transformer-based feature interaction module to capture fine-grained text-image relationships. The decoder then reconstructs a density map from the interaction features, which is used for point prediction and contrastive learning.

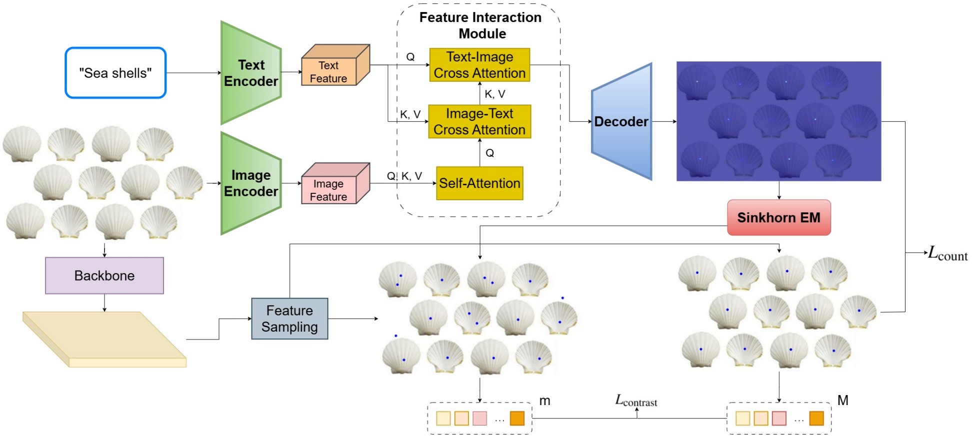

Figure 2: Overall structure of CPMNet, which consists of an encoder, a feature interaction module, and a decoder. The design introduces (i) point regression using Sinkhorn EM, and (ii) Contrastive Point Matching by extracting point-wise features and applying contrastive learning based on positive and negative samples generated via Hungarian Matching.

3.1 Density Regression

Motivated by Amini-Naieni et al. [1], we leverage CLIP to explore the interaction between text and image features. The CLIP encoder takes an input image of spatial resolution $H \times W$ and a user-provided text description, projecting both into a shared embedding space. These features are then passed to a Feature Interaction Module—a transformer-based structure designed to model semantic and spatial interactions. Each block in this module includes three key components: self-attention on image features, image-to-text cross-attention, and text-to-image cross-attention. The goal is to let the model learn fine-grained alignments between textual descriptions and image regions, producing a similarity map that highlights regions of interest aligned with the text.

The similarity map is then fed into a decoder, which serves two purposes: (1) recovering the object distribution and (2) restoring the spatial resolution. The decoder consists of alternating layers—four bilinear interpolation layers for upsampling and four 2×2 convolutional layers for feature refinement. This sequence progressively restores spatial details while enhancing feature quality. Finally, a feedforward network (FFN) outputs the distribution map, which is used by the Sinkhorn EM algorithm to infer object point locations.

3.2 Sinkhorn EM for Point Regression

We use Sinkhorn EM (Expectation Maximization) to compute the position of objects given their density. Unlike traditional point regression methods such as local maxima detection, which rely on hand-crafted thresholds and often fail under dense or ambiguous conditions, Sinkhorn EM leverages the global density distribution to infer object positions. Specifically, Sinkhorn EM uses the total object count implied by the density map as a constraint, making it more faithful to the underlying density.

Given a predicted density map $D = \{(D_i, x_i)\}_{i=1}^{n}$, where $n = H \times W$, $D_i \geq 0$ is the predicted density value at pixel i, and $x_i \in \mathbb{R}^2$ is its spatial coordinate, our goal is to regress a set of object points $K = \{(k_j, \hat{y}_j)\}_{j=1}^{m}$, where $m = \sum_i D_i$, $k_j = 1$ is the weight of the j-th object, and $\hat{y}_j \in \mathbb{R}^2$ is its estimated location.

To align the predicted density map D with the point set K, we adopt the Sinkhorn EM algorithm, which iteratively minimizes the Sinkhorn distance through two alternating steps:

E-step (Expectation): Compute the optimal transport plan $P \in \mathbb{R}^{n \times m}$ between D and K using the Sinkhorn algorithm [10]. The regularized Sinkhorn distance is defined as:

$$\mathcal{L}_{\varepsilon}(D, K) = \min_{P \in U(a,b)} \langle C, P \rangle + \varepsilon \sum_{i,j} P_{ij} \log P_{ij} \tag{1}$$

where $C_{ij} = \|x_i - \hat{y}_j\|^2$ is the quadratic transport cost, and ε is the entropy regularization coefficient.

M-step (Maximization): With the transport plan P fixed, update each point location $\hat{y}_j$ in K by computing the weighted average of pixel locations x_i ([7], [13]):

$$\hat{y}_j = \frac{\sum_{i=1}^{n} P_{ij} x_i}{\sum_{i=1}^{n} P_{ij}} \tag{2}$$

These steps are repeated until convergence, either when the Sinkhorn distance stabilizes or after a fixed number of iterations. The final output is the set of regressed object points $\{\hat{y}_j\}$, which represent the predicted object locations.

3.3 Contrastive Point Matching

To perform contrastive point matching, we first associate the predicted points $\{\hat{y}_i\}_{i=1}^{m}$ with the ground truth points $\{y_j\}_{j=1}^{M}$, where M is the number of ground-truth points, using the Hungarian Matching algorithm in a one-to-one fashion. The matched predicted points form the set of positive samples P, while the unmatched ones are treated as negative samples N.

We extract point-wise features from the image $i-th$ feature map $F_i \in \mathbb{R}^{C \times H_i \times W_i}$, generated by the vision backbone, by bilinearly interpolating at each predicted point location:

$$z_i = \text{Interp}(F, \hat{y}_i), \quad \text{for } \hat{y}_i \in P \cup N, \tag{3}$$

where $z_i \in \mathbb{R}^C$ is the feature vector for point $\hat{y}_i$.

A contrastive loss is then applied to encourage the features of matched (positive) pairs to be close, while pushing apart those of unmatched (negative) pairs. For a positive anchor–target pair (z_i, z_j), and a set of negatives $\{z_k\}_{k \in N}$, we use an InfoNCE-style loss:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\exp(\text{sim}(z_i, z_j)/\tau) + \sum_{k \in N} \exp(\text{sim}(z_i, z_k)/\tau)}, \tag{4}$$

where the cosine similarity is defined as: $\mathrm{sim}(a,b) = \frac{a^\top b}{\|a\|\|b\|}$, and τ is a temperature hyperparameter.

This contrastive objective drives the model to learn spatially discriminative features by aligning predicted points with true object locations and distancing them from background or unrelated regions in the embedding space.

3.4 Training and Inference

The model is trained end-to-end using a combination of two losses: a counting loss and a contrastive loss. The total loss is defined as

$$L = \alpha L_{\text{count}} + \beta L_{\text{contrast}}, \tag{5}$$

where α and β are weights that balance the two objectives. The counting loss L_{count} is implemented as a mean squared error (MSE) between the predicted and ground-truth object counts. Specifically, it compares the total mass of the predicted density map to the number of annotated objects.

During inference, the model predicts a density map for a given input image. The final object count is obtained by summing all pixel values in the density map. Additionally, the point coordinates regressed via the Sinkhorn EM module can be used to visualize or localize the predicted object positions.

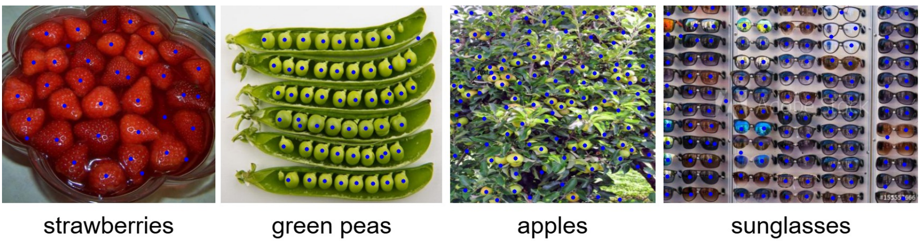

Figure 3: Qualitative results of Sinkhorn EM on the FSC-147 test set. In the first two examples (strawberries and green peas), Sinkhorn EM accurately predicts object positions. However, in the apple image, the textual description is ambiguous, and due to the visual similarity between apples and the background, the model incorrectly places some points in the background. In the sunglasses example, overcounting occurs as a result of strong self-similarity between object instances.

4 Experiment

4.1 Dataset and Metrics

We conduct experiments on FSC-147 [12], a multi-class few-shot object counting dataset containing 6,135 images. The number of objects per image varies significantly, ranging from 7 to 3,731, with an average of 56. Each image is accompanied by three exemplars—randomly selected object instances annotated with bounding boxes. The training set contains 89 object categories, while the validation and test sets include 29 disjoint categories each, making FSC-147 an open-set object counting benchmark.

In addition, we also evaluate the model on the CARPK [2] dataset. CARPK contains aerial images of parking lots captured by drones, with a training set of 989 images and a test set of 459 images.

We evaluate the performance of our model using two standard metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

4.2 Implementation detail

For the encoder, we use CLIP for both the image and text branches. Point-level features are extracted from the ResNet-50 backbone using the output of the final convolutional layer (f5). For the Sinkhorn EM module, we set the entropy regularization parameter ε to 0.9 (scaling factor), limit the number of iterations to 16, and apply early stopping when the average displacement of points falls below 1 pixel.

During training, input images are cropped to a resolution of 224×224 and normalized. We apply standard data augmentation techniques, including Gaussian noise, Gaussian blur, horizontal flipping, and color jittering, to improve generalization.

4.3 State-of-the-Art Comparison

We compare our approach with previous density-based zero-shot object counting methods. The quantitative results are presented in Table 1. **CPMNet** significantly outperforms all prior zero-shot density-based models, including the long-standing state-of-the-art **CounTX**, achieving MAE reduction of 20.59% and 31.46% on the test and validation sets of the FSC-147 dataset, respectively. Although **ZSOC** [19] adopts a two-stage strategy, its error remains notably higher than that of CPMNet, further demonstrating the robustness and effectiveness of our unified approach.

To evaluate cross-dataset generalization, CPMNet was tested on the CARPK [2] dataset. As shown in Table 2, CPMNet achieves the best performance among open-world counting methods, with lower MAE and RMSE compared to CLIP-count [3], CounTX [1], and VL-Counter [4].

Table 1: Performance comparison of different methods on validation and test sets. Pseco [6] is detection-based, while other methods are density-based.

Method	Type	Test set		Validation set	
		MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
Pseco [6]	Detection-based	16.58	129.77	23.90	100.33
ZSOC [19]	Density-based	22.09	115.17	26.93	88.63
RepRPN [11]	Density-based	28.32	128.76	31.69	100.31
CLIP-Count [3]	Density-based	17.78	106.62	18.79	61.18
VLCounter [4]	Density-based	17.05	106.16	18.06	65.13
CounTX [1]	Density-based	15.88	106.29	17.10	65.61
CPMNet (Our)	**Density-based**	**12.61**	**98.76**	**11.72**	**44.68**

Contrastive Point Matching. As illustrated in Figure 4, traditional density-based methods struggle in scenarios where objects are not spatially compact, such as scenes with seagulls,

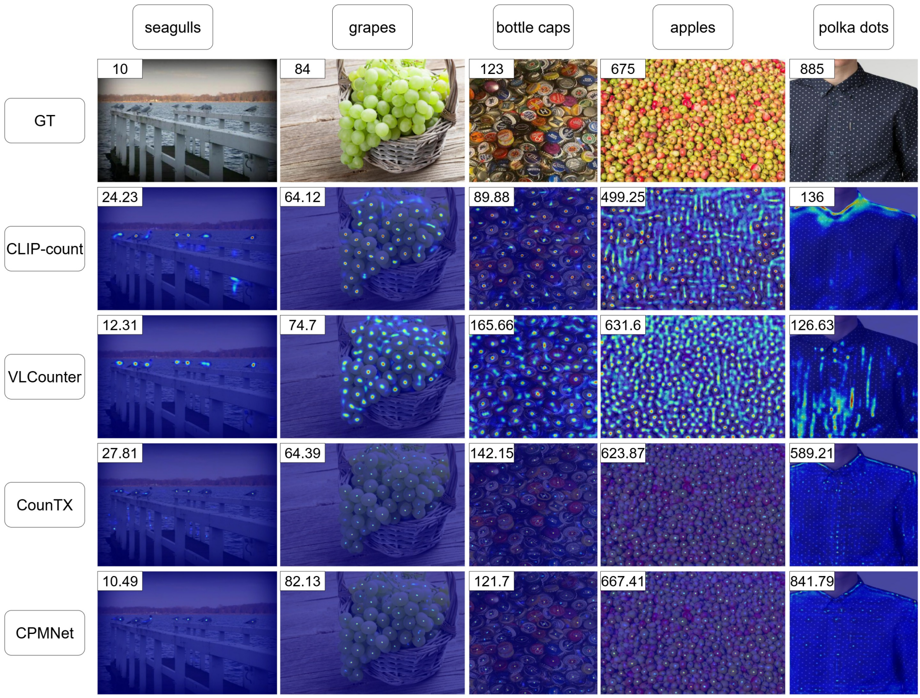

Figure 4: Qualitative result comparison of different density-based model VLCounter [4], CLIP-Count [3], CounTX [1] and CPMNet.

resulting in the predicted density spreading to irrelevant regions. In crowded and occluded scenes, these methods also fail to resolve closely packed instances, often collapsing multiple objects into a single peak. In contrast, **CPMNet** effectively handles these challenging cases by leveraging contrastive learning to distinguish between ambiguous object regions, thereby enhancing both spatial precision and count accuracy.

Table 2: Comparison of different counting methods on the CARPK dataset.

Method	Year	Published	MAE ↓	RMSE ↓
CLIP-count [3]	2023	✓	11.96	16.61
CounTX [1]	2023	✓	8.13	10.87
VLCounter [4]	2023	✓	6.46	8.68
CPMNet	2025	–	**5.79**	**6.68**

4.4 Ablation Studies

We first investigate the impact of the contrastive point matching (CPM) component. To this end, we train two model variants: one with contrastive loss $CPMNet_{const}$ and one without it $CPMNet_{MSE}$. As shown in Table 3, although both models achieve reasonable performance, the model trained with contrastive loss significantly reduces the counting error by 7.75% in MAE and 1.12% in RMSE. This improvement is attributed to the additional spatial supervision introduced by the contrastive objective, which helps the model learn more discrimina-

Table 3: Impact of Contrastive Point Matching on FSC-147 Test Set Performance

	MAE ↓	RMSE ↓
$\text{CPMNet}_{\text{MSE}}$	13.67	99.94
$\text{CPMNet}_{\text{const}}$	**12.61**	**98.76**

Table 4: Performance Comparison of Point Regression Methods on FSC-147 Test Set

	MAE ↓	RMSE ↓
CMNet_{LM}	14.03	101.85
$\text{CPMNet}_{\text{SEM}}$	**12.61**	**98.76**

tive object representations and encourages the generation of more informative and structured density maps.

Point Proposal. We also compare two methods for point prediction: local maxima detection and Sinkhorn EM. The results are presented in Table 4. Local maxima-based prediction performs poorly due to its high sensitivity to kernel size, small kernels are needed in crowded scenes, while sparse scenes require larger ones. This makes it challenging to generalize across different object densities. In contrast, Sinkhorn EM avoids such rigid constraints. Training the model with Sinkhorn EM leads to more than a 10% improvement in MAE compared to local maxima, highlighting the robustness and adaptability of our method for accurate object localization.

Limitation Figure 3 presents qualitative results of predicted points on the FSC-147 test set. When the target object is visually distinct from the background—such as in the cases of strawberries and green peas, the model performs well, accurately localizing object positions. However, in challenging scenarios where the object blends into the background, like the apples image, the model produces false positives due to difficulty distinguishing between the object and its surroundings. In the sunglasses example, the model mistakenly predicts each lens separately, resulting in overcounting.

5 Conclusion

We propose a novel approach to open-world object counting called CPMNet, which incorporates contrastive point matching. By adding spatial information, CPMNet enhances object representation and produces more accurate density maps. Experiments on FSC-147 demonstrate improved performance over previous methods, and ablation studies confirm the effectiveness of contrastive learning and Sinkhorn EM in improving counting accuracy.

References

[1] N. Amini-Naieni, K. Amini-Naieni, T. Han, and A. Zisserman. Open-world text-specified object counting. In *British Machine Vision Conference*, 2023.

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

[3] Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4535–4545, 2023.

[4] Seunggu Kang, WonJun Moon, Euiyeon Kim, and Jae-Pil Heo. Vlcounter: Text-aware visual representation for zero-shot object counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2714–2722, 2024.

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[6] Gang Li, Xiang Li, Yujie Wang, Yichao Wu, Ding Liang, and Shanshan Zhang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *European Conference on Computer Vision*, pages 457–472. Springer, 2022.

[7] Wei Lin and Antoni B. Chan. Optimal transport minimization: Crowd localization on density maps for semi-supervised counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21663–21673, June 2023.

[8] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721*, 2022.

[9] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180, 2023.

[10] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

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

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

[13] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems, 2019. URL https://arxiv.org/abs/1610.06519.

[14] Min Shi, Lu Hao, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[15] Zenglin Shi, Ying Sun, and Mengmi Zhang. Training-free object counting with prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 323–331, January 2024.

[16] Aayush Kumar Tyagi, Chirag Mohapatra, Prasenjit Das, Govind Makharia, Lalita Mehra, Prathosh AP, et al. Degpr: Deep guided posterior regularization for multi-class cell detection and counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23913–23923, 2023.

[17] Nikola Đukić, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18872–18881, 2023.

[18] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *Advances in neural information processing systems*, 33: 1595–1607, 2020.

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

[20] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6315–6324, 2023.