# A Fixed-Point Approach to Unified Prompt-Based Counting

**Wei Lin, Antoni B. Chan**

Department of Computer Science, City University of Hong Kong
Tat Chee 83, Kowloon Tong, Hong Kong SAR, China
elonlin24@gmail.com, abchan@cityu.edu.hk

## Abstract

Existing class-agnostic counting models typically rely on a single type of prompt, e.g., box annotations. This paper aims to establish a comprehensive prompt-based counting framework capable of generating density maps for concerned objects indicated by various prompt types, such as box, point, and text. To achieve this goal, we begin by converting prompts from different modalities into prompt masks without requiring training. These masks are then integrated into a class-agnostic counting methodology for predicting density maps. Furthermore, we introduce a fixed-point inference along with an associated loss function to improve counting accuracy, all without introducing new parameters. The effectiveness of this method is substantiated both theoretically and experimentally. Additionally, a contrastive training scheme is implemented to mitigate dataset bias inherent in current class-agnostic counting datasets, a strategy whose effectiveness is confirmed by our ablation study. Our model excels in prominent class-agnostic datasets and exhibits superior performance in cross-dataset adaptation tasks.

## Introduction

Visual counting has been a longstanding topic of interest, driven by the demands of industrial intelligence. However, the majority of existing methods concentrate on counting specific object categories, such as pedestrians (Wang et al. 2020; Lin and Chan 2023; Shu, Wan, and Chan 2023), animals (Arteta, Lempitsky, and Zisserman 2016a), or cars (Wang, Wan, and Li 2018). This limitation hampers their applicability to unseen categories and weakens their transferability. In contrast, numerous commercial and agricultural scenarios necessitate the counting of diverse objects, such as goods on a shelf (Goldman et al. 2019), various crops in farmland (Kitano et al. 2019; Zabawa et al. 2020), or buildings in remote sensing images (Christophe and Inglada 2009; Agarwal and Rajan 2015). To tackle these scenarios, there exists a requirement for class-agnostic models capable of counting objects of any type.

Recent class-agnostic counting models have primarily focused on a box-guided pipeline (Lu, Xie, and Zisserman 2018; Ranjan et al. 2021; Shi et al. 2022; Liu et al. 2022; Lin et al. 2022). These methods utilize three box annotations to indicate the object of interest and subsequently predict a density map to illustrate its distribution and count. Moreover, there are also works that develop models for counting based on text prompts (Xu et al. 2023; Jiang, Liu, and Chen 2023). To establish a connection between vision and language, a language-image pre-training model, CLIP (Radford et al. 2021), is employed to align text and visual embeddings within the same space. Typically, both text and visual encoders remain frozen to retain the zero-shot capability in CLIP, while the density predictor is trained within few-shot counting datasets for task adaptation. However, there are no related works combining these prompts to establish a unified prompt-based counting framework, which can use various types of prompts to predict density maps within only one model. To address this issue, we propose building a unified prompt-based counting framework, considering three types of prompts: box, point, and text, as displayed in Figure 1.

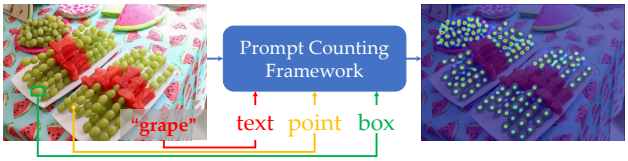



Figure 1: Overview of prompt-based counting. This model takes prompts in various modalities, *e.g.*, box, point, or text annotation to indicate the object of interest, and then predicts the distribution and count accordingly.

Normally, the prompts are represented as a tokens, *i.e.*, feature vectors, in class-agnostic counting (Shi et al. 2022; Liu et al. 2022; Lin et al. 2022). The first problem we need to address is how to translate these prompts from different modalities into the same representation. For boxes and points, we can directly aggregate information from the corresponding regions. However, since text prompts are not visual cues, it cannot be straightforwardly mapped into the visual feature space. Inspired by MaskCLIP (Zhou, Loy, and Dai 2022), we utilize the frozen CLIP to generate a text prompt mask that is consistent with box and point prompts. The key principle of this transformation is that the value embeddings in the last layer of the visual encoder in CLIP contain local semantic features that can be mapped into the language space. Thus, we can directly measure the similarity

between the text feature and local value embeddings to obtain a prompt mask similar to box and point prompts. This approach eliminates the need for a training step to establish the connection between text and image, allowing the training to focus on density prediction, improving performance. This contrasts with previous text-guided zero-shot counting methods (Xu et al. 2023; Jiang, Liu, and Chen 2023), which needs to randomly crop numerous patches from the input images and then proceed to compare them with the text feature in order to select appropriate tokens.

After prompt masks are obtained freely, we employ cross attention to generate density features and utilize a convolutional-based module to make the final predictions. When applying the aforementioned framework to tackle class-agnostic counting, two issues emerge that need to be addressed: (1) The model's robustness is lacking, leading to poor performance during inference; (2) The training dataset FSC-147 (Ranjan et al. 2021) exhibits a bias where most images contain only one type of object. This bias causes the model to count the *salient* object instead of the one indicated by the prompt, especially when the prompt is noisy.

To address the robustness, we introduce a fixed-point inference scheme and its corresponding loss function into our model to enhance its performance without introducing new parameters. Specifically, we observe that the predicted density map can also be considered as a prompt mask to compute the prompt token, establishing a fixed point solution of the prompt-based counting function that converges to a consistent prediction. The density predictor possesses a natural recurrent structure for refining the density map. However, recurrent training is challenging due to the continuous accumulation of gradients. To overcome this issue, we design a fixed-point loss based on implicit differentiation (Chang, Griffiths, and Levine 2022) and bi-level optimization (Jia, Liu, and Huang 2023) to optimize parameters effectively. Experimental results demonstrate that this design can converge the model to a favorable parameter space and result in lower estimation errors compared to a model without it.

To address the dataset bias, we employ a contrastive training scheme to train the counting model. For each training sample (positive), we randomly select an image from the dataset as its contrasting sample (negative). While a token is aggregated according to prompt masks from the image features, we compute two density maps: one for the positive sample and another for the negative sample. Ideally, the density map for the positive sample should closely align with its ground truth, while the density map for the negative sample should resemble an all-zero density map. This approach enables the trained model to make accurate predictions even when multiple objects are present within a single image.

In summary, this paper makes three contributions:

1. We propose a unified prompt-based class-agnostic counting framework to count objects indicated by boxes, points, and texts. These prompts from different modalities are transformed into semantic masks and then counted accordingly.
2. To improve robustness, we propose a fixed-point inference and loss function to train a recurrent structure in the counting framework. This is based on the finding that the predicted density map could also be regarded as a prompt to generate a token for counting objects, creating a fixed point and a recurrent scheme to refine the density map.
3. Addressing dataset bias, we advocate contrastive training of the prompt counting model. The model should predict a density map close to the ground truth for positive samples and an all-zero map for the negative samples.

## Related Works

Visual object counting has been extensively studied in the literature, often concentrating on counting specific objects like pedestrians (Chan, Liang, and Vasconcelos 2008; Zhang et al. 2016; Zhang, Lin, and Chan 2021), vehicles (Wang, Wan, and Li 2018; Hsieh, Lin, and Hsu 2017), cells (Guo et al. 2021; Huang, McKay, and Durr 2021), crops (Akiva et al. 2020; Tong et al. 2021), and animals (Arteta, Lempitsky, and Zisserman 2016b; Xu et al. 2020; Sun et al. 2023). While these models exhibit impressive performance, their applicability is constrained by their inability to generalize to unfamiliar object types on which they were not trained. This limitation has prompted the need for class-agnostic counting. Recent advancements in this field have delved into both box-guided and text-guided approaches.

### Box-Guided Object Counting

Class-agnostic counting with exemplars is commonly referred to as few-shot counting (Lu, Xie, and Zisserman 2018; Yang et al. 2021; Ranjan et al. 2021). In GMN (Lu, Xie, and Zisserman 2018), self-similarity computations are utilized to identify similar patches to an exemplar patch that indicates what should be counted, highlighting them in the predicted heat map. CFOCNet (Yang et al. 2021) employs several reference images to denote the object of interest and then performs counting on a query image. Additionally, Ranjan et al. introduced a class-agnostic counting dataset, FSC-147, along with a baseline model, FamNet, where given exemplars function as convolution kernels to locate objects of a specified type within the image (Ranjan et al. 2021).

Following FSC-147, several box-guided class-agnostic counting approaches have been proposed for general object counting (Shi et al. 2022; Nguyen et al. 2022; Gong et al. 2022; Lin et al. 2022; Liu et al. 2022). Most of these approaches conceratrate on refining the matching strategy between the given exemplars and query images. For instance, BMNet (Shi et al. 2022) employs a dynamic similarity metric and a scale embedding to enhance the matching score between exemplars and ground truth. RCAC (Gong et al. 2022) introduces feature augmentation and an edge matching module to handle intra-class diversity. CounTR (Liu et al. 2022) corporates a transformer-based architecture, leveraging the cross-attention capabilities of transformers for matching tasks. In contrast, SPDCN (Lin et al. 2022) focuses on extracting robust features for matching by integrating exemplar information into the network backbone. Counting-DETR (Nguyen et al. 2022) explores the amalgamation of few-shot counting and detection, treating objects as points and producing not only their locations but also their sizes. One limitation of box-guided approaches is that they

require additional effort to annotate what should be counted using box label for each image.

### Object Counting With Text Prompts

Counting objects according to given text involves taking an image and a text as input, and then predicting the count of the object indicated by the text within the image. ZSOC (Xu et al. 2023) proposes a method using a conditional VAE to transform text cues into visual features, and subsequently utilizes a clustering method to aggregate corresponding object features in the given image. CLIP-C (Jiang, Liu, and Chen 2023) employs the visual-language pre-trained model CLIP (Radford et al. 2021) to establish a connection between text and image. An intra-image contrastive and a hierarchical text-patch interaction module are designed to generate final density maps. Both models achieve outstanding performance on the class-agnostic counting dataset.

In contrast to the above works that focus solely on either box or text as prompts, our proposed unified prompt-based counting framework can count objects with three types of prompts using just one model. A concurrent work, TFPOC (Shi, Sun, and Zhang 2023), presents a training-free counting approach based on both CLIP (Radford et al. 2021) and SAM (Kirillov et al. 2023), also addressing this task. However, it falls short in handling the challenges poses by occlusion and blur in dense scenes, as its counting ability is rooted in instance segmentation rather than density prediction. Thus, as demonstrated in the experiments, our proposed method has lower counting error than TFPOC.

## Method

In this section, we present our token-based prompt counting framework. The structure of the proposed prompt counting framework is illustrated in Figure 2. It takes an image and a prompt mask as input, and then computes the similarity between the prompt and image features, which helps filter unrelated parts in the image. The final density map is predicted by a CNN decoder to represent the distribution and count of the concerned object.

### Prompt Mask

Within our framework, the prompt could take the form of a box or point, indicating the location of an instance, or a piece of text describing what should be counted, as depicted in Figure 1. However, the text prompt differs from the former two, as it cannot be directly transformed into visual cues. Hence, the initial challenge we must address is how to establish a uniform representation for these prompts, even when they take various modalities, *i.e.*, box, point, and text.

The object of interest in an image can typically represented in two ways: a) by an attention mask $\boldsymbol{m}$ that highlights regions occupied by the object; b) by a token $\boldsymbol{t}$ that represents the object's semantic feature. Given an input image feature map $\boldsymbol{F}$, the relationships among these are:

$$\boldsymbol{t} = \frac{\boldsymbol{F}^\top \boldsymbol{m}}{\|\boldsymbol{m}\|_1}, \quad \boldsymbol{m} \in \mathbb{R}_+^N, \quad \boldsymbol{F} \in \mathbb{R}^{N \times C}, \tag{1}$$

where $N = h \times w$ is the size of the flattened feature map, $C$ is the dimension of features, and $\|\cdot\|_1$ donates the $\texttt{sum}$ operation, as the elements in $\boldsymbol{m}$ are non-negative. In the following parts, we detail the process of generating a prompt mask within our unified framework without training.

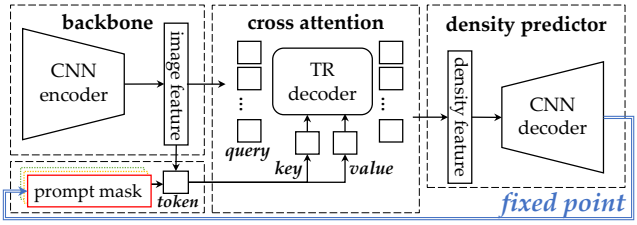



Figure 2: Our unified prompt-based counting framework. A CNN encoder generates image features, and a token is aggregated based on the provided prompt mask. Next, cross-attention is applied to generate density features, which are then decoded to produce the density map. Importantly, the density map can also be viewed as a prompt mask, implying the existence of a fixed point solution. This fixed point enables the utilization of a loop to enhance the output.

For box and point prompts, $\boldsymbol{m}$ can be formulated as a mask in which the labeled region or pixel is set to 1 and the other pixels are set to 0.

For text prompts, we utilize CLIP (Radford et al. 2021) to generate an prompt mask inspired by MaskCLIP (Zhou, Loy, and Dai 2022). Specifically, the matching score between a pre-defined text feature $\boldsymbol{f}_T \in \mathbb{R}^C$ and a given visual feature $\boldsymbol{f}_I \in \mathbb{R}^C$ in CLIP is computed using cosine distance:

$$\mathcal{S}_T(\boldsymbol{f}_I) = \texttt{cosine}(\boldsymbol{f}_T, \boldsymbol{f}_I) = \frac{\boldsymbol{f}_T^\top \boldsymbol{f}_I}{\|\boldsymbol{f}_T\| \|\boldsymbol{f}_I\|}. \tag{2}$$

Looking closely at the last layer of the visual encoder in CLIP, the visual cue can be formulated as:

$$\boldsymbol{f}_I = \mathcal{H}\left(\sum\nolimits_i w_i \boldsymbol{v}_i\right) = \sum\nolimits_i w_i \mathcal{H}(\boldsymbol{v}_i). \tag{3}$$

Here $w_i = \texttt{softmax}\left(\boldsymbol{q}^\top \boldsymbol{k}_i / \sqrt{C}\right)$ measures the saliency level of the $i$-th region. $\boldsymbol{q}$ is the *query* of the class embedding, while $\boldsymbol{k}_i$ and $\boldsymbol{v}_i$ represent the *key* and *value* embeddings at spatial location $i$. The equality in (3) holds because $\mathcal{H}$ is a linear projection. We can then form the approximation:

$$\mathcal{S}_T(\boldsymbol{f}_I) = \mathcal{S}_T(\sum\nolimits_i w_i \mathcal{H}(\boldsymbol{v}_i)) \approx \sum\nolimits_i w_i \mathcal{S}_T(\mathcal{H}(\boldsymbol{v}_i)). \tag{4}$$

Specifically, by taking the first-order Taylor approximation of $\mathcal{S}_T(\mathcal{H}(\boldsymbol{v}))$ around $\boldsymbol{f}_I$:

$$\mathcal{S}_T(\mathcal{H}(\boldsymbol{v})) \approx \mathcal{S}_T(\boldsymbol{f}_I) - \nabla \mathcal{S}_T(\boldsymbol{f}_I)^\top (\mathcal{H}(\boldsymbol{v}) - \boldsymbol{f}_I). \tag{5}$$

Substituting (5) into the RHS of (4), we have

$$\begin{aligned} &\sum\nolimits_i w_i \mathcal{S}_T(\mathcal{H}(\boldsymbol{v}_i)) \\ &\quad \approx \sum\nolimits_i w_i \left[\mathcal{S}_T(\boldsymbol{f}_I) - \nabla \mathcal{S}_T(\boldsymbol{f}_I)^\top (\mathcal{H}(\boldsymbol{v}_i) - \boldsymbol{f}_I)\right] \\ &\quad = \mathbf{S}_T(\boldsymbol{f}_I). \end{aligned} \tag{6}$$

Thus (4) is proved. With this approximation in mind, it become reasonable to interpret $\mathcal{S}_T(\mathcal{H}(\boldsymbol{v}_i))$ as measuring the matching score between the text feature $\boldsymbol{f}_T$ and local visual

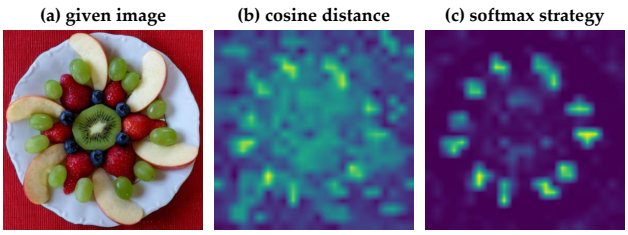



Figure 3: Comparison between `cosine` distance in (7) and `softmax` strategy in (8) on text-prompt mask generation. The list shows the extracted concept dictionary via LLaMA-Adapter V2 and Spacy. The green text represents the user-provided text prompt, whose index is $k$ in (8).

feature at the $i$-th position. This interpretation holds true because $w_i$ functions as a weight for aggregating salient information from the entire image.

Based on the analysis above, we derive a text prompt mask using the representation of `cosine` distance:

$$\boldsymbol{m} = [\texttt{cosine}\left(\boldsymbol{f}_T, \mathcal{H}(\boldsymbol{v}_i)\right)]_i, i \in \{1, \dots, N\}. \quad (7)$$

However, we observe that the mask generated via (7) contains much noise caused by objects close to the target object in the CLIP feature space, as depicted in Figure 3(b). To address this issue, we require a *concept dictionary* that lists object categories present in the input image. This allows us to apply `softmax` to reduce the influence of background objects, similar to the approach adopted by CLIP in zero-shot learning (Radford et al. 2021).

To obtain the concept dictionary, we employ an image caption model, LLaMA-Adapter V2 (Gao et al. 2023), to generate a detailed description of the input image. We then utilize Spacy (Honnibal et al. 2020) to identify all nouns and construct the concept dictionary. Note that each image has its own concept dictionary. Next, we can generate the prompt mask using (4) and the `softmax` strategy from CLIP:

$$\boldsymbol{m} = \left[\frac{\exp\left(\tau \mathcal{S}_{T_k}(\mathcal{H}(\boldsymbol{v}_i))\right)}{\sum_{j=1}^{L} \exp\left(\tau \mathcal{S}_{T_j}(\mathcal{H}(\boldsymbol{v}_i))\right)}\right]_i, i \in \{1, \dots, N\}, \quad (8)$$

where $\tau$ is the temperature, $T_j$ denotes the $j$-th text in the concept dictionary with a length of $L$, and $T_k$ represents the user-provided text in the concept dictionary. An example is in Figure 3.

Finally, we note that our prompt framework is flexible and can also handle other types of prompts, such as instance-masks, as well as multi-prompts (e.g., both text and box are provided). More details are provided in the supplementary.

## Fixed-Point Inference and Loss Function

Given the image feature $\boldsymbol{F}$ and prompt mask $\boldsymbol{m}$, a prompt token $\boldsymbol{t}$ is computed according to (1). The final density map $\boldsymbol{d}$ can be predicted by a function $\mathcal{D}$ with parameters $\theta$:

$$\boldsymbol{d} = \mathcal{D}_\theta\left(\boldsymbol{F},\ \boldsymbol{t}\right), \quad (9)$$

where $\mathcal{D}_\theta$ comprises the cross-attention module and density predictor in Figure 2. With the predicted $\boldsymbol{d}$ and ground truth $\boldsymbol{d}'$, the counting model can be optimized through the L2 loss:

$$\mathcal{L}(\boldsymbol{d},\ \boldsymbol{d}') = \|\boldsymbol{d} - \boldsymbol{d}'\|^2, \quad (10)$$

which is the conventional design for class-agnostic counting.

Here we design a more effective method without altering the network structure by leveraging fixed-point theories (Burden, Faires, and Burden 2015; Jeon, Lee, and Choi 2021). Specifically, we identify that the predicted density map $\boldsymbol{d}$ can also function as a prompt mask to generate the semantic token, thus establishing a fixed-point iteration:

$$\boldsymbol{d}^* = \mathcal{F}(\boldsymbol{d}^*, \Theta) = \mathcal{D}_\theta\left(\boldsymbol{F},\ \frac{\boldsymbol{F}^\top \boldsymbol{d}^*}{\|\boldsymbol{d}^*\|_1}\right), \quad (11)$$

where $\mathcal{F}(\boldsymbol{d}^*, \Theta)$ represents the fixed-point function with respect to the fixed-point $\boldsymbol{d}^*$, and $\Theta = [\boldsymbol{F}, \theta]$ are the weights of $\mathcal{F}$. (11) inspires us to perform (12) to refine the density map $\boldsymbol{d}$ in a loop for improved results, initialized with $\boldsymbol{d}^{(0)} = \boldsymbol{m}$:

$$\boldsymbol{d}^{(t+1)} = \mathcal{F}(\boldsymbol{d}^{(t)}, \Theta), \quad t \in \{0, 1, \dots, T-1\}, \quad (12)$$

where $T$ represents the number of refinement iterations.

The challenge lies in the fact is that the recurrent algorithm can lead to training instability. To address this, we introduce a loss function using implicit differentiation (Chang, Griffiths, and Levine 2022) and bi-level optimization (Jia, Liu, and Huang 2023). We compute the gradient of $\boldsymbol{d}^*$ with respect to $\Theta$ in (11) utilizing the implicit function theorem (Jeon, Lee, and Choi 2021):

$$\frac{\partial \boldsymbol{d}^*}{\partial \Theta} = \frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \Theta} + \frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \boldsymbol{d}^*} \frac{\partial \boldsymbol{d}^*}{\partial \Theta}, \quad (13)$$

$$\Rightarrow \frac{\partial \boldsymbol{d}^*}{\partial \Theta} = \left[\boldsymbol{I} - \frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \boldsymbol{d}^*}\right]^{-1} \frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \Theta} \quad (14)$$

The inverse item of (14) can be expressed using Neumann series:

$$\left[\boldsymbol{I} - \frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \boldsymbol{d}^*}\right]^{-1} = \sum\nolimits_{k=0}^{K=\infty} \left[\frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \boldsymbol{d}^*}\right]^k. \quad (15)$$

Thus, the gradient of $\mathcal{L}$ with respect to $\Theta$ is:

$$\frac{\partial \mathcal{L}}{\partial \Theta} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{d}^*} \left[\boldsymbol{I} - \frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \boldsymbol{d}^*}\right]^{-1} \frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \Theta} \quad (16)$$

$$= \sum\nolimits_{k=0}^{K=\infty} \frac{\partial \mathcal{L}}{\partial \boldsymbol{d}^*} \left[\frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \boldsymbol{d}^*}\right]^k \frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \Theta}. \quad (17)$$

Subsequently, we take the first-order approximation for practical computation of (17):

$$\frac{\partial \mathcal{L}}{\partial \Theta} \approx \frac{\partial \mathcal{L}}{\partial \boldsymbol{d}^*} \frac{\partial \mathcal{F}(\boldsymbol{d}^*, \boldsymbol{F})}{\partial \Theta} = \frac{\partial \mathcal{L}(\mathcal{F}(\boldsymbol{d}^*, \Theta), \boldsymbol{d}')}{\partial \Theta}, \quad (18)$$

which implies that the gradient of $\Theta$ after iterations can be approximated by only computing the last iteration on $\boldsymbol{d}^*$ instead of employing detailedly backpropagation within the recurrent structure.

In the ideal scenario, the last prediction $\boldsymbol{d}^{(T)}$ in (12), the result of infinite iteration $\boldsymbol{d}^{(\infty)}$, the fixed point $\boldsymbol{d}^*$, and the ground truth $\boldsymbol{d}'$ are extremely close to each other:

$$\boldsymbol{d}^{(T)} \approx \boldsymbol{d}^{(\infty)} = \boldsymbol{d}^* \simeq \boldsymbol{d}'. \quad (19)$$

As a result, we replace $\boldsymbol{d}^*$ with $\boldsymbol{d}'$ in (18) and formulate the following loss to incorporate (18) into training,

$$\mathcal{L}_\infty = \mathcal{L}(\mathcal{F}(\boldsymbol{d}', \Theta),\ \boldsymbol{d}'). \quad (20)$$

Furthermore, to minimize the error between $\boldsymbol{d}^{(T)}$ and $\boldsymbol{d}^{(\infty)}$, we also utilize the following loss during training:

$$\mathcal{L}_T = \mathcal{L}(\boldsymbol{d}^{(T)},\ \mathcal{F}(\boldsymbol{d}', \Theta)). \tag{21}$$

By assuming $\boldsymbol{d}^{(\infty)} = \mathcal{F}(\boldsymbol{d}', \Theta)$ and combining (20) and (21), we formulate the proposed fixed-point loss function as:

$$\hat{\mathcal{L}} = \mathcal{L}_\infty + \mathcal{L}_T = \mathcal{L}(\boldsymbol{d}^{(T)}, \boldsymbol{d}^{(\infty)}) + \mathcal{L}(\boldsymbol{d}^{(\infty)}, \boldsymbol{d}'). \tag{22}$$

Experimental results demonstrate that the optimal performance is achieved when $T$ is set to 2.

## Contrastive Training Strategy

FSC-147 (Ranjan et al. 2021) serves as the main dataset for training a class-agnostic counting model. However, a dataset bias exists: for most training images, only one type of object exists in the training image. This bias may cause the model training to take a shortcut and count the salient object rather than the objects indicated by the prompt, especially in case where the provided prompts contain much noise. To address this issue, we employ a contrastive training strategy. The key principle is that a prompt token in one image should correspond to an all-zero density map in another image if they do not contain the same type of objects.

A training sample is represented as $(\boldsymbol{F}, \boldsymbol{m}, \boldsymbol{d}')$, where $\boldsymbol{F}$ denotes image features, $\boldsymbol{m}$ and $\boldsymbol{d}'$ represent the prompt mask and the ground truth, respectively. To generate samples for contrastive training, we concatenate the $i$-th image with the $j$-th one in a batch, creating a new training sample, denoted as $\boldsymbol{F}_{ij} = [\boldsymbol{F}_i, \boldsymbol{F}_j]$. Furthermore, we concatenate $\boldsymbol{m}_i$ and $\boldsymbol{d}'_i$ with two all-zero tensors, $\mathbf{0}_{\boldsymbol{m}_j}$ and $\mathbf{0}_{\boldsymbol{d}'_j}$ respectively, whose shape comes from the subscript, to formulate the new prompt mask and learning target, $\boldsymbol{m}_{ij} = [\boldsymbol{m}_i, \mathbf{0}_{\boldsymbol{m}_j}]$ and $\boldsymbol{d}'_{ij} = [\boldsymbol{d}'_i, \mathbf{0}_{\boldsymbol{d}'_j}]$. This approach allows us to define the fixed-point loss within contrastive training as follows:

$$\hat{\mathcal{L}}_{ij} = \mathcal{L}(\boldsymbol{d}^{(T)}_{ij}, \boldsymbol{d}^{(\infty)}_{ij}) + \mathcal{L}(\boldsymbol{d}^{(\infty)}_{ij}, \boldsymbol{d}'_{ij}). \tag{23}$$

In this context, the definition of $\boldsymbol{d}^{(t)}_{ij}$ is consistent with (12):

$$\boldsymbol{d}^{(t+1)}_{ij} = \mathcal{F}(\boldsymbol{d}^{(t)}_{ij}, \Theta) \quad \text{and} \quad \boldsymbol{d}^{(0)}_{ij} = \boldsymbol{m}_{ij}. \tag{24}$$

Through this approach, we enforce the given prompt to predict a density map similar to ground truth in the positive part (*i.e.*, $i$-th sample), while predicting an all-zero density map in the negative part (*i.e.*, $j$-th sample). Note that we do not introduce category information during training since the probability of concatenating images with the same object type is negligible. The effectiveness of contrastive training without considering category information has been shown in various unsupervised and self-supervised learning methods (van den Oord, Li, and Vinyals 2018; He et al. 2020).

# Experiments

**Datasets.** We employ the *FSC-147* (Ranjan et al. 2021) for training and evaluating the proposed prompt counting model. It encompasses 147 distinct categories. Additionally, *CARPK* (Hsieh, Lin, and Hsu 2017) car counting dataset is used to assess the model's capability for cross-dataset adaptation.

| Prompt | Method | | validation | | test | |
|---|---|---|---|---|---|---|
| | | | MAE | MSE | MAE | MSE |
| box | GMN | *accv'18* | 29.66 | 89.81 | 26.52 | 124.57 |
| | FamNet | *cvpr'21* | 26.55 | 77.01 | 26.76 | 110.95 |
| | BMNet | *cvpr'22* | 17.89 | 61.12 | 16.89 | 96.65 |
| | SPDCN | *bmvc'22* | 21.60 | 71.83 | 19.41 | 128.26 |
| | CounTR | *bmvc'22* | 17.40 | 70.33 | **14.12** | 108.81 |
| | TFPOC | *arxiv'23* | – | – | 19.95 | 132.16 |
| | ours | | **16.87** | **59.45** | 16.68 | **105.08** |
| | SPDCN† | *bmvc'22* | 16.36 | 53.94 | 14.66 | 101.89 |
| | CounTR† | *bmvc'22* | 13.15 | 49.72 | 12.06 | 90.01 |
| | ours† | | **12.80** | **48.65** | **11.86** | **89.40** |
| text | ZSOC | *cvpr'23* | 26.93 | 88.63 | 22.09 | 115.17 |
| | CLIP-C | *mm'23* | 18.79 | 61.18 | 17.78 | 106.62 |
| | TFPOC | *arxiv'23* | – | – | 24.79 | 137.15 |
| | ours | | **16.92** | **58.92** | **16.81** | **105.83** |
| point | TFPOC | *arxiv'23* | – | – | 20.10 | 132.83 |
| | ours | | **17.16** | **59.38** | **15.86** | **103.27** |

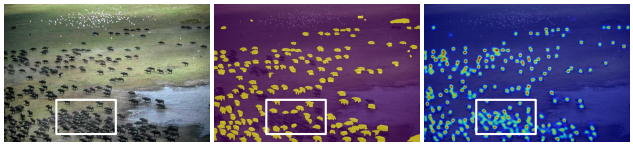
Table 1: Comparison with other prompt counting methods. Both box and point employ one annotation as the prompt. Withing box prompt, † means scale prior.

**Evaluation metrics.** We assess performance using the metrics of Mean Absolute Error (MAE, $\texttt{MAE} = \frac{1}{N}\sum_{i=1}^{N}|C_i - C'_i|$) and root Mean Squared Error (MSE, $\texttt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i - C'_i)^2}$). In the formulation we use $N$ to represent the number of samples in either the validation or test set, $C_i$ and $C'_i$ correspond to the count of the $i$-th prediction and its corresponding ground truth. In ablation studies, we employ the average MAE/MSE across the three prompt types (box/point/text) to conserve space. Comprehensive results can be found in the supplementary material.

## Prompt Counting Results

As shown in Table 1, we conduct a comparative analysis across different prompt types, considering that fewer works have focused on combining these prompts together.

**Box prompts:** In the box-guided counting, our model demonstrates significantly lower MAE and MSE compared to similar simple-structured models, such as GMN (Lu, Xie, and Zisserman 2018) and FamNet (Ranjan et al. 2021). BMNet (Shi et al. 2022) is a more complex network that incorporates a bi-linear matching module, and our model achieves results similar to it (MAE: 16.89 *vs.* 16.68). For SPDCN (Lin et al. 2022) and CounTR (Liu et al. 2022), we compare our model with them in two tracks, vanilla ones and models with scale prior(†). The scale prior, *i.e.*, the width and height of box, is a specific property within box prompts. Specifically, SPDCN use it to adjust the receptive field, and CounTR design a test-time normalization (TT-norm) to refiene the prediction. With scale prior, SPDCN† and CounTR† achieves MAE of 16.36 and 13.15 on the validation set respectively. However, removing the scale prior from these models results in their MAEs increasing to 21.60 and 17.40 on the validation set (without † in Table 1). To create a better comparison by accounting for the scale prior, we have implemented TT-norm, following CounTR's methodol-



(a) input image (b) TFPOC (c) ours

Figure 4: Visualization between TFPOC and our model. TFPOC cannot handle extremely dense regions (white box).

| Mask | box (val) | | text (val) | | point (val) | |
|---|---|---|---|---|---|---|
| | MAE | MSE | MAE | MSE | MAE | MSE |
| cosine | 18.55 | 69.32 | 20.59 | 79.30 | 18.46 | 69.70 |
| softmax | **16.87** | **59.45** | **16.92** | **58.92** | **17.16** | **59.38** |

Table 2: Comparison of text prompt mask generation.

ogy. By including the scale prior (ours$^{\dagger}$), our model's MAE significantly reduces to 12.80 and 11.86, which is better than SPDCN and CountTR with scale prior.

Although the scale prior can boost the performance, we do not recommend implementing it since scale prior contradicts the motivation of unified prompt-based counting. It would be better to omits the scale prior and focuses on considering shared properties among prompts from different modalities (termed "prompt masks" in our paper) for prompt counting.

**Text prompts:** When compared with models trained using text prompts, the strengths of our model become evident. ZSOC (Xu et al. 2023) utilizes a conditional VAE to generate class prototypes representing objects of interest, and then selects exemplars within a given image for counting. CLIP-C (Jiang, Liu, and Chen 2023) directly employs the frozen CLIP model to generate density features. Subsequently, it designs a hierarchical text-patch interaction module to obtain density maps. While our model also leverages CLIP to transform text cues into visual features, we avoid fine-tuning CLIP's features. Instead, we use it to generate a prompt mask for consistent representation in prompt counting. In the text-guided counting, our model achieves an MAE/MSE of 16.84/105.16, both of which are superior to ZSOC (22.09/115.17) and CLIP-C (17.78/106.62).

**Comparison with TFPOC:** The concurrent approach TFPOC (Shi, Sun, and Zhang 2023) is a training-free prompt counting model, leveraging SAM (Kirillov et al. 2023) to count objects based on the segmentation results due to its high-quality zero-shot capability. It also incorporates box, text, and point as prompts facilitate object counting in the image. However, as demonstrated in Table 1, our model outperforms TFPOC across all prompt types. This is attributed to our model's framework based on density prediction rather than instance segmentation, as the former excels in handling challenges like occlusion and blur, while the latter may fail on small or densely placed objects. Additionally, it is worth noting that our approach involves a training process, whereas TFPOC is entirely training-free. Figure 4 provides an illustrative example showcasing our model's advantages, particularly in dense regions.

Finally, in the supplemental, we provide results on using instance masks as prompts, as well as inference using multiple prompts at the same time.

| loss function | validation set | | test set | |
|---|---|---|---|---|
| | MAE | MSE | MAE | MSE |
| L2 | 19.09 | 67.32 | 17.22 | 106.49 |
| fixed-point | **16.98** | **59.25** | **16.45** | **104.72** |

Table 3: Comparison between MSE and fixed-point loss.

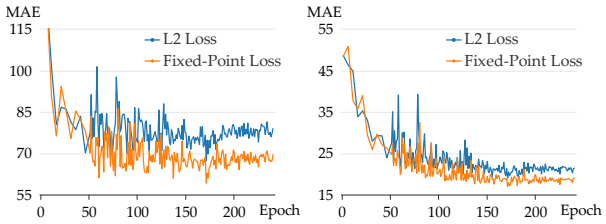



Figure 5: MAE/MSE on the validation set during training.

## Ablation Study on Text Prompt Mask

Figure 3 illustrates that the text prompt mask generated via `cosine` distance method exhibits significant noise, whereas the `softmax` strategy effectively highlights the object of interest. In addition, we compare their respective counting performances on FSC-147, and Table 2 presents the MAE/MSE on the validation set, considering different approaches for generating text prompt masks. The results reveal that a noisy mask from the cosine strategy not only increases estimation errors within its own track (text), but also degrades performance in other tracks (box & point prompts).The model has to contend with noisy input during training, potentially impacting its generalization capability.

## Ablation Study on Fixed-Point Loss

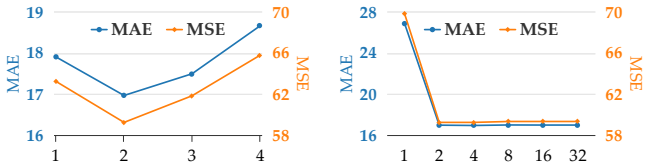
**Overall improvements.** We compare the proposed fixed-point loss in (22) with the vanilla L2 loss in (10) to demonstrate the overall improvements. As shown in Table 3, when L2 loss is applied to prompt counting, the resulting MAE and MSE are 17.22 and 106.49, respectively. While this performance surpasses that of TFPOC (Shi, Sun, and Zhang 2023), the estimation errors are still considerably larger than the text-guided class-agnostic counting model, CLIP-C (Jiang, Liu, and Chen 2023). In contrast, our fixed-point loss reduces the MAE (MSE) from 19.09 (67.32) to 16.98 (59.25) on the validation set, and the MAE (MSE) is lowered to 16.45 (104.72). Note that this improvement is achieved without introducing new network modules; rather, it involves looping the cross-attention and the density predictor while training with the designed fixed-point loss. Consequently, the number of parameters remains unchanged.

In Figure 5, we also demonstrate the variation of MAE/MSE on the validation set during training. The graph illustrates that the proposed fixed-point loss converges to a lower MAE/MSE than the vanilla L2 loss, making it a more effective choice compared to other loss functions.

**Detailed improvements.** We discuss three aspects of the

| infinity | finity | validation set | | test set | |
|---|---|---|---|---|---|
| | | MAE | MSE | MAE | MSE |
| $\mathcal{L}_\infty$ | – | 24.91 | 71.89 | 26.97 | 110.21 |
| – | $\mathcal{L}_{T'}$ | 20.79 | 68.73 | 17.38 | 107.41 |
| $\mathcal{L}_\infty$ | $\mathcal{L}_{T'}$ | 17.14 | 62.14 | **16.29** | 105.91 |
| $\mathcal{L}_\infty$ | $\mathcal{L}_T$ | **16.98** | **59.25** | 16.45 | **104.72** |

Table 4: Ablation study on different combinations of infinity and finity part in fixed-point loss.




(a) Training with different $T$ (b) Inference with different $T$

Figure 6: The influence of $T$ in training and inference.

proposed fixed point: the infinity loss $\mathcal{L}_\infty$ in (20), the finite loss $\mathcal{L}_T$ in (21), and the number of iterations $T$. For the infinity loss, we can only conduct experiments on the model with or without it. Regarding the finite loss, there is an alternative approach which directly computes the L2 loss between $\boldsymbol{d}^{(T)}$ and the ground truth $\boldsymbol{d}'$, resulting in the loss:

$$\mathcal{L}_{T'} = \mathcal{L}(\boldsymbol{d}^{(T)}, \boldsymbol{d}'). \tag{25}$$

Note that this is different from the vanilla L2 loss (10) since the recurrent structure is retained ($T = 2$).

Table 4 presents a comparison of four combinations involving $\mathcal{L}_\infty$, $\mathcal{L}_T$ and $\mathcal{L}_{T'}$. When only loss $\mathcal{L}_\infty$ is used, the performance is poor (MAE 26.97), and even worse than TFPOC (Shi, Sun, and Zhang 2023). This outcome is reasonable since it completely removes the effect of prompts in the computation graph, failing to account for an incomplete prompt mask that differs from the ground truth. In the second experiment, $\mathcal{L}_{T'}$ in (25) serves as the learning objective. The results are significantly better than the previous approach, and its estimation errors on the test set are also lower than the L2 loss. By combining $\mathcal{L}_\infty$ and $\mathcal{L}_{T'}$, better MAE/MSE is achieved on the validation set, and the lowest MAE on the test set is achieved (16.29), but with slightly worse MSE than the proposed method(105.91 *vs.* 104.72).

We next explore the number of iterations $T$ in both training and inference. We trained four models with $T \in \{1, 2, 3, 4\}$ for comparison, as shown in Figure 6(a). The best performance is obtained with $T = 2$. A large $T$ causes the model to lose prompt mask information, while the model cannot reach its full effect without refinement ($T = 1$). Focusing on the model trained with $T = 2$, as shown in Figure 6(b), we investigate whether applying different value of $T$ can reduce estimation errors during inference. It is noticeable that a significantly high MAE/MSE is observed when $T$ is set to 1, but the errors decrease and converge to the lowest value after the second iteration, providing evidence for the existence and validity of the fixed point in prompt counting.

| contrastive training | validation set | | test set | |
|---|---|---|---|---|
| | N-MAE | N-MSE | N-MAE | N-MSE |
| w/o | 38.09 | 73.09 | 48.77 | 78.05 |
| w/ | **1.90** | **17.16** | **3.89** | **15.19** |

Table 5: Ablation study on contrastive training.

| | box | | | text | | |
|---|---|---|---|---|---|---|
| | BMNet | TFPOC | ours | CLIP-C | TFPOC | ours |
| MAE | 10.44 | 10.97 | **7.83** | 11.96 | 11.01 | **7.62** |
| MSE | 13.77 | 14.24 | **9.74** | 16.61 | 14.34 | **9.71** |

Table 6: Cross-dataset adaptation on CARPK dataset.

## Ablation Study on Contrastive Training

To verify the effectiveness of contrastive training, we evaluate the trained model using negative samples. Similar to the training scheme, we extract a prompt token from one test sample, and use it to count on another randomly selected image to measure negative estimation errors, N-MAE and N-MSE, which are MAE and MSE for the negative samples with 0 ground-truth count. Here we use category information to ensure that each pair of test samples contains different types of objects. Table 5 demonstrates that the contrastive training operates as expected, and without it, N-MAE/MSE are significantly higher. The errors are considerably reduced when the model is trained in contrastive training scheme.

## Cross-Dataset Adaptation

Continuing the tradition of previous class-agnostic counting studies (Ranjan et al. 2021; Jiang, Liu, and Chen 2023), we explore cross-dataset adaptation by directly applying the prompt model to count cars in the CARPK dataset (Hsieh, Lin, and Hsu 2017). The results in Table 6 provide a comparison between the proposed method and others on two prompt types: box and text prompts. Our method attains the lowest estimation errors in both tasks, showcasing its strong cross-dataset adaptation capability and the broad generalization of our prompt-based counting approach.

# Conclusion

In this paper, we introduce a unified prompt-based counting model that can accurately count specific objects in an image using box, point, or text prompts. Our proposed method transforms prompts from different modalities into the a consistent representation, the prompt mask, which effectively highlights regions containing the objects of interest. Furthermore, we identify a fixed point within our framework when the predicted density map is treated as a prompt mask. This observation motivates us to implement a recurrent structure for refining the density map, and a fixed-point loss is also derived to make the training process stable and efficient. Addressing the inherent dataset bias, where most samples contain only one type of object in the current class-agnostic counting datasets, we employ a contrastive training shceme to mitigate shortcuts and bolster model robustness. Comprehensive experiments and ablation studies validate the effectiveness of our framework, demonstrating remarkable performance achievements.

## Acknowledgements

This work was supported by a Strategic Research Grant from City University of Hong Kong (Project No. 7005665).

## References

Agarwal, L.; and Rajan, K. S. 2015. Fast ICA based algorithm for building detection from VHR imagery. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, 1889–1892. IEEE.

Akiva, P.; Dana, K.; Oudemans, P.; and Mars, M. 2020. Finding berries: Segmentation and counting of cranberries using point supervision and shape priors. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 50–51.

Arteta, C.; Lempitsky, V.; and Zisserman, A. 2016a. Counting in the wild. In *European conference on computer vision (ECCV)*, 483–498.

Arteta, C.; Lempitsky, V.; and Zisserman, A. 2016b. Counting in the wild. In *European Conference on Computer Vision (ECCV)*, 483–498. Springer.

Burden, R. L.; Faires, J. D.; and Burden, A. M. 2015. *Numerical analysis*. Cengage learning.

Chan, A. B.; Liang, Z.-S. J.; and Vasconcelos, N. 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE conference on computer vision and pattern recognition*, 1–7. IEEE.

Chang, M.; Griffiths, T.; and Levine, S. 2022. Object representations as fixed points: Training iterative refinement algorithms with implicit differentiation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 32694–32708.

Christophe, E.; and Inglada, J. 2009. Object counting in high resolution remote sensing images with OTB. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, volume 4, IV–737–IV–740.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; Li, H.; and Qiao, Y. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*.

Goldman, E.; Herzig, R.; Eisenschtat, A.; Goldberger, J.; and Hassner, T. 2019. Precise detection in densely packed scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 5227–5236.

Gong, S.; Zhang, S.; Yang, J.; Dai, D.; and Schiele, B. 2022. Class-Agnostic Object Counting Robust to Intraclass Diversity. In *European Conference on Computer Vision (ECCV)*, volume 13693, 388–403.

Guo, Y.; Krupa, O.; Stein, J.; Wu, G.; and Krishnamurthy, A. 2021. SAU-Net: A Unified Network for Cell Counting in 2D and 3D Microscopy Images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hsieh, M.; Lin, Y.; and Hsu, W. H. 2017. Drone-Based Object Counting by Spatially Regularized Regional Proposal Network. In *International Conference on Computer Vision (ICCV)*, 4165–4173.

Huang, L.; McKay, G. N.; and Durr, N. J. 2021. A Deep Learning Bidirectional Temporal Tracking Algorithm for Automated Blood Cell Counting from Non-invasive Capillaroscopy Videos. In de Bruijne, M.; Cattin, P. C.; Cotin, S.; Padoy, N.; Speidel, S.; Zheng, Y.; and Essert, C., eds., *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 415–424.

Jeon, Y.; Lee, M.; and Choi, J. Y. 2021. Differentiable Forward and Backward Fixed-Point Iteration Layers. *IEEE Access*, 9: 18383–18392.

Jia, B.; Liu, Y.; and Huang, S. 2023. Improving Object-centric Learning with Query Optimization. In *International Conference on Learning Representations (ICLR)*.

Jiang, R.; Liu, L.; and Chen, C. 2023. CLIP-Count: Towards Text-Guided Zero-Shot Object Counting. *CoRR*, abs/2305.07304.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. In *International Conference on Computer Vision (ICCV)*.

Kitano, B. T.; Mendes, C. C.; Geus, A. R.; Oliveira, H. C.; and Souza, J. R. 2019. Corn plant counting using deep learning and UAV images. *IEEE Geoscience and Remote Sensing Letters (GRSL)*.

Lin, W.; and Chan, A. B. 2023. 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*, 21663–21673.

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

Liu, C.; Zhong, Y.; Zisserman, A.; and Xie, W. 2022. CounTR: Transformer-based Generalised Visual Counting. In *British Machine Vision Conference (BMVC)*, 370.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations (ICLR)*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

Lu, E.; Xie, W.; and Zisserman, A. 2018. Class-agnostic counting. In *Asian conference on computer vision (ACCV)*, 669–684. Springer.

Nguyen, T.; Pham, C.; Nguyen, K.; and Hoai, M. 2022. Few-Shot Object Counting and Detection. In *European Conference on Computer Vision (ECCV)*, 348–365. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, volume 139, 8748–8763.

Ranjan, V.; Sharma, U.; Nguyen, T.; and Hoai, M. 2021. Learning To Count Everything. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 3394–3403.

Shi, M.; Lu, H.; Feng, C.; Liu, C.; and Cao, Z. 2022. Represent, Compare, and Learn: A Similarity-Aware Framework for Class-Agnostic Counting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 9529–9538.

Shi, Z.; Sun, Y.; and Zhang, M. 2023. Training-free Object Counting with Prompts. *arXiv preprint arXiv:2307.00038*.

Shu, W.; Wan, J.; and Chan, A. B. 2023. Generalized Characteristic Function Loss for Crowd Analysis in the Frequency Domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, G.; An, Z.; Liu, Y.; Liu, C.; Sakaridis, C.; Fan, D.-P.; and Van Gool, L. 2023. Indiscernible Object Counting in Underwater Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13791–13801.

Tong, P.; Han, P.; Li, S.; Li, N.; Bu, S.; Li, Q.; and Li, K. 2021. Counting trees with point-wise supervised segmentation network. *Engineering Applications of Artificial Intelligence*, 100: 104172.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *Arxiv*, abs/1807.03748.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Neural Information Processing Systems (NeurIPS)*, 5998–6008.

Wang, Q.; Gao, J.; Lin, W.; and Li, X. 2020. NWPU-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence (T-PAMI)*, 43(6): 2141–2149.

Wang, Q.; Wan, J.; and Li, X. 2018. Robust hierarchical deep learning for vehicular management. *IEEE Transactions on Vehicular Technology*, 68(5): 4148–4156.

Xu, J.; Le, H.; Nguyen, V.; Ranjan, V.; and Samaras, D. 2023. Zero-Shot Object Counting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 15548–15557.

Xu, J.; Yu, L.; Zhang, J.; and Wu, Q. 2020. Automatic Sheep Counting by Multi-object Tracking. In *Conference on Visual Communications and Image Processing (VCIP)*, 257–257. IEEE.

Yang, S.-D.; Su, H.-T.; Hsu, W. H.; and Chen, W.-C. 2021. Class-agnostiFew-shot Object Counting. In *Winter Conference on Applications of Computer Vision (WACV)*, 870–878.

Zabawa, L.; Kicherer, A.; Klingbeil, L.; Töpfer, R.; Kuhlmann, H.; and Roscher, R. 2020. Counting of grapevine berries in images via semantic segmentation using convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 164: 73–83.

Zhang, Q.; Lin, W.; and Chan, A. B. 2021. Cross-view cross-scene multi-view crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 557–567.

Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; and Ma, Y. 2016. Single-image crowd counting via multi-column convolutional neural network. In *Conference on Computer Vision and Pattern recognition (CVPR)*, 589–597.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract Free Dense Labels from CLIP. In *European Computer Vision Conference (ECCV)*, volume 13688, 696–712.