Class-Agnostic Object Counting with Text-to-Image Diffusion Model

Xiaofei Hui[1,2], Qian Wu[2], Hossein Rahmani[1], and Jun Liu[1,2(✉)]

[1] Lancaster University, Lancaster, UK
{h.rahmani,j.liu81}@lancaster.ac.uk
[2] Singapore University of Technology and Design (SUTD), Singapore, Singapore

Abstract. Class-agnostic object counting aims to count objects of arbitrary classes with limited information (*e.g.*, a few exemplars or the class names) provided. It requires the model to effectively acquire the characteristics of the target objects and accurately perform counting, which can be challenging. In this work, inspired by that text-to-image diffusion models hold rich knowledge and comprehensive understanding of real-world objects, we propose to leverage the pre-trained text-to-image diffusion model to facilitate class-agnostic object counting. Specifically, we propose a novel framework named CountDiff with careful designs, leveraging the pre-trained diffusion model's comprehensive understanding of image contents to perform class-agnostic object counting. The experiments show the effectiveness of CountDiff on both few-shot setting with exemplars provided and zero-shot setting with class names provided.

Keywords: Class-agnostic object counting · Text-to-image diffusion model · Few-shot and zero-shot

1 Introduction

The task of object counting aims to accurately estimate the number of instances of a specified object in an image. It is crucial for various real-life scenarios, such as traffic surveillance [25], crowd monitoring [42], and wildlife conservation [28]. Many existing counting methods focus on specific classes such as crowd [20,41], cars [15,26], and animals [28]. However, such methods usually rely on laborious data annotations and the trained model often cannot be easily adapted to count novel object classes. To alleviate these limitations, class-agnostic counting aims to count objects of novel target classes with very few annotations required (*e.g.*, a few exemplars or class names) [35,45]. Under this setting, a single model can be employed to count a wide range of object categories at a low cost of annotation, offering enhanced flexibility and applicability for diverse real-world scenarios.

X. Hui and Q. Wu—Both authors contribute equally.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72890-7_1.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15127, pp. 1–18, 2025.
https://doi.org/10.1007/978-3-031-72890-7_1

Specifically, some methods [24,34,35] count objects of arbitrary classes with the information in the exemplars or repeating patterns in the image, while some other methods [1,45] explore object counting based on the provided class names.

Despite lots of research efforts, achieving class-agnostic object counting is still challenging. In particular, to adapt to counting objects of arbitrary classes, the model needs to first effectively learn the characteristics of an arbitrary object class from only a few exemplars (or the provided class names). Then, to accurately find and count the specific target objects in the image, the model is further required to have a thorough understanding of the detailed information in the given image (*e.g.*, textures, shapes, and spatial locations of the objects). This can be difficult, especially when the image contains numerous objects [47], where the features of each object can be subtle.

On the other hand, recently, pre-trained text-to-image diffusion models (*e.g.*, Stable Diffusion [36]) have shown unprecedented power in generating images with rich details and reasonable spatial structures guided by user prompts [9,38,49]. The success of this line of works implies that, being able to accurately generate images with large amounts of details, the pre-trained text-to-image diffusion model has a comprehensive and detailed understanding of images, covering from pixel-level contents to overall layouts (*e.g.*, object textures, object shapes, and spatial structures and locations in the image) [17]. Importantly, having such knowledge and understanding can be very beneficial for class-agnostic object counting.

Inspired by the knowledge and understanding of text-to-image diffusion models toward image contents, in this paper, we aim to investigate *leveraging such understanding to boost the performance of class-agnostic object counting*. However, using pre-trained text-to-image diffusion models to perform counting is not straightforward, as such diffusion models are generally built for text-to-image generation and cannot be directly applied to object counting in the image, *i.e.*, though they hold rich knowledge, the knowledge is implicitly embedded in the model. Thus, it can be challenging to effectively extract and leverage the desired knowledge to facilitate counting. To cope with this issue, we delve deep into the characteristics of text-to-image diffusion model and investigate strategies for its effective utilization to facilitate class-agnostic counting.

A key advantage of text-to-image diffusion model (*e.g.*, Stable Diffusion) lies in its ability to link the semantic information in text to image content. As observed in [17], such ability can be explicitly reflected in the *cross-attention maps*, which can highlight specific regions in the image based on text descriptions. This inspired us that, leveraging this property of the diffusion model, we can obtain cross-attention maps that can explicitly point out the useful regions in the image to aid in performing class-agnostic counting. However, obtaining such cross-attention maps can be difficult, because the image regions are activated in cross-attention maps based on the text inputs. This means that, to obtain the cross-attention maps, we need proper text inputs to describe such useful regions for counting the target objects. Yet, finding such proper text inputs can be challenging: 1) As we need to perform counting for the target

object class, the text input should contain specific information about the target objects. However, no such readily available text information is provided in few-shot class-agnostic counting, and it can be difficult to precisely describe the objects or object parts of interest in real-world scenarios. 2) There can also be some general information that can facilitate counting objects of various classes, which can be hard to describe in text. To tackle this challenge, instead of finding such text descriptions, we propose to specifically learn *embeddings* that encodes the instruction in the embedding space to obtain cross-attention maps that can facilitate class-agnostic counting.

More specifically, in this paper, we aim to boost the performance of class-agnostic counting by effectively extracting the knowledge in the pre-trained diffusion model. To achieve this, we propose a novel framework named **CountDiff** that can flexibly perform counting for both few-shot (with a few exemplars provided) and zero-shot (with class names provided) settings. The key design of our CountDiff lies in activating the ability of text-to-image diffusion models to facilitate class-agnostic counting. This is achieved by extracting knowledge from the pre-trained diffusion model using embeddings that specifically encode useful information. To better tackle this task, we propose to learn two embeddings to properly describe such useful information. In the few-shot setting, first, to encode the specific information of the target object in each image, CountDiff learns an *object-specific embedding* by mapping the image information in the exemplars to the text embedding space. In this way, the diffusion model can understand the visual information of the target object in the exemplars. Then, as class-agnostic object counting requires the model to have good generalization ability for diverse object classes, we further propose to learn an *object-agnostic embedding* to extract knowledge that can be shared for counting objects of different classes. On the other hand, CountDiff can be flexibly switched to the zero-shot setting by obtaining the object-specific embeddings using the provided class names. Combining the power of both object-agnostic and object-specific embeddings, CountDiff can effectively extract knowledge from the cross-attention layers that explicitly reveal useful information for class-agnostic counting.

Besides leveraging the cross-attention maps to point out image regions that are useful for class-agnostic counting, we also take advantage of the rich semantic grouping information in self-attention maps of the pre-trained diffusion model, which can also benefit object counting. Moreover, in the few-shot setting, to better extract knowledge for the target objects in the testing image, we apply a lightweight test-time adaptation to further adapt (fine-tune) the object-specific embedding to better represent the novel objects leveraging the provided exemplars. Leveraging the pre-trained diffusion model and our careful designs, CountDiff can effectively perform class-agnostic counting.

Inspired by the powerful capability of the text-to-image diffusion model, from a new perspective, we investigate leveraging its comprehensive understanding of image contents to explicitly highlight useful information for class-agnostic counting. To achieve this, we propose a novel framework, CountDiff, with careful designs that enable the pre-trained diffusion model to perform counting, taking

advantage of the cross-attention and self-attention maps as well as the specifically learned embeddings. Our CountDiff achieves state-of-the-art performance on both zero-shot and few-shot settings.

2 Related Work

Class-Specific Counting. Class-specific counting aims to count the objects of a specific class in an image (*e.g.*, crowd counting [3,19,33,42], car counting [15,26], and animal counting [2,28]). Most existing methods can be roughly categorized into two groups: detection-based methods that rely on object detectors to localize and count the objects [7,18], and regression-based methods that predict density maps and obtain the final results by summing the pixel values in the density maps [5,22,33]. Different from class-specific counting methods, in this paper, we focus on class-agnostic counting.

Class-Agnostic Counting. Recently, class-agnostic counting has attracted much research attention due to its flexibility. Specifically, in the few-shot setting [6,11,21,24,27,35,40,43,46,47], a few exemplars of the target class are provided in the image to specify the target object. To further reduce the annotations required, exemplar-free class-agnostic counting was proposed [34], whereby no exemplar is provided during inference. These methods usually identify target objects automatically based on the general information in the image (*e.g.*, repetitive object patterns) and perform counting [14,44]. However, we cannot specify the target objects for exemplar-free methods to count, which potentially limits their practical use [45]. To address this limitation and meanwhile have low requirements on annotations, more recently, zero-shot class-agnostic counting that specifies the target of interest by its class name was proposed [45]. In this work, we investigate few-shot and zero-shot class-agnostic object counting tasks and develop a novel framework CountDiff that handles both tasks.

To facilitate the model to count an arbitrary class, existing few-shot approaches usually extract the image and exemplar features using pre-trained networks (*e.g.*, SwAV [4]) and match the exemplar features with the image features to predict a density map. Gong *et al.* [11] proposed to incorporate a pre-trained edge detection module to enhance class-agnostic feature learning. Liu *et al.* [6] proposed a transformer-based counting method with a two-stage training scheme to boost performance. Dukić *et al.* [43] proposed an iterative prototype adaptation method with enhancement on the shape information of the target objects. On the other hand, most zero-shot methods adapt to counting objects of arbitrary classes leveraging semantic information extracted from the class name with the help of the pre-trained CLIP model [30]. Xu *et al.* [45] first extracted semantic information of the target object using CLIP and then generated exemplar prototypes with a variational autoencoder. Jiang *et al.* [16] proposed to align class names and image patches with patch-text contrastive loss. Different from these methods, in this paper, we take advantage of

the detailed image understanding ability of pre-trained diffusion model (e.g., spatial layout and locations of the objects), and propose specific designs to extract useful knowledge for class-agnostic counting. To the best of our knowledge, we are the first to leverage the cross-attention maps as well as the self-attention maps in the pre-trained diffusion model with the help of the specifically learned embeddings, to effectively perform class-agnostic counting.

Text-to-Image Diffusion Models. Recently, text-to-image diffusion models have shown their strong ability and have been studied in various vision domains [17,29,38,49,50]. Specifically, with the advent of pre-trained text-to-image diffusion models (*e.g.*, Stable Diffusion [36], DALL-E [31,32], and Imagen [39]), research studies have proposed to leverage such pre-trained models to further improve specific image generation ability. For example, ControlNet [49] flexibly manipulates the image generation process with a set of conditional controls (*e.g.*, pose and depth). DreamBooth [38] fine-tunes a pre-trained diffusion model to inject novel user-defined concepts to generate images of novel subjects. The strong ability to generate photorealistic images implies that pre-trained diffusion models contain rich and detailed knowledge, and can understand how real-world objects should look like [8,17]. Inspired by this, in this work, we propose to leverage the pre-trained diffusion model to benefit class-agnostic counting.

3 Method

3.1 Preliminaries: Text-to-Image Diffusion Models

Text-to-image diffusion models learn to reconstruct a data distribution z_0 from normally distributed noise z_T, by gradually denoising z_T in the reverse diffusion process conditioned on the text input. In our framework, we adopt the widely-used Stable Diffusion [36], which has shown its power in generating photorealistic images of diverse contents [10,12,38,49]. Specifically, during training, given an image input I and a text input y, Stable Diffusion first maps I into the latent space with a pre-trained image encoder $\mathcal{E}$, and encodes y into the text embedding space with a text encoder τ_θ. The conditional diffusion process is learned by a diffusion model ϵ_θ that is a denoising UNet [37]. At time step t, given the noisy image latent z_t and the text embedding $\tau_\theta(y)$, the objective is to accurately remove the noise ϵ added to the image latent $\mathcal{E}(I)$:

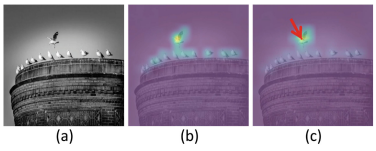

Fig. 1. Visualization of cross-attention and self-attention maps in Stable Diffusion. (a) shows the input image, (b) shows the cross-attention map obtained using the text input "seagulls", and (c) shows the 2D attention map of the marked pixel location (pointed by the red arrow) obtained from the self-attention map.

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(I),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y)\|_2^2\right] \quad (1)$$

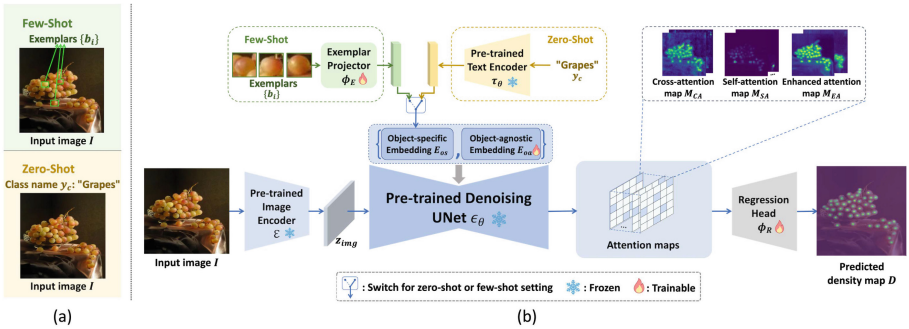

Fig. 2. Best viewed in color. (a) Illustration of the class-agnostic object counting task with few-shot (providing exemplars) and zero-shot (providing class name) settings. (b) Illustration of the proposed CountDiff framework. For the few-shot setting, we use the exemplar projector ϕ_E to obtain the object-specific embedding E_{os} from the exemplars (following the green arrows), and construct the object-agnostic embedding E_{oa} as a learnable vector. Then, given the input image and the embeddings, we extract the attention maps (cross-attention map M_{CA}, self-attention map M_{SA}, and enhanced attention map M_{EA}) to obtain useful information for counting the target objects. The attention maps are then sent to the regression head ϕ_R to predict the density map D. CountDiff can also be switched to perform zero-shot class-agnostic object counting, by obtaining the object-specific embedding leveraging the pre-trained text encoder τ_θ (following the yellow arrows). In CountDiff, parameters of the pre-trained text-to-image diffusion model are frozen. During training, the learning for both object-specific and object-agnostic embeddings as well as the regression head are optimized such that the embeddings can well encode the semantic information that is beneficial for class-agnostic object counting and ϕ_R can generate high-quality density maps.

In this process, the interaction between the image content and the text content is modeled in the *cross-attention layers* employed in the denoising UNet. Specifically, in the l-th cross-attention layer, the query Q_c^l is derived from the noisy image latent z_t, while the key K_c^l and value V_c^l are obtained from the text embedding $\tau_\theta(y)$. The output of this layer is then computed as $\text{SoftMax}(\frac{Q_c^l {K_c^l}^T}{\sqrt{d_l}}) \cdot V_c^l$, where d_l denotes the projection dimension of Q_c^l, K_c^l and V_c^l. Consequently, we can obtain the cross-attention map M_{CA}^l, computed as:

$$M_{CA}^l = \text{SoftMax}(\frac{Q_c^l {K_c^l}^T}{\sqrt{d_l}}) \tag{2}$$

As the cross-attention map is obtained via computing the correlation between the image content in Q_c^l and the semantic text content in K_c^l, it can reflect the correlation between the image and text [8,17]. As shown in Fig. 1 (b), the cross-attention map explicitly depicts the semantically correlated image regions corresponding to the text input.

Besides cross-attention layers, Stable Diffusion also has *self-attention layers* in its UNet structure that can model the semantic association between image

pixels. In the m-th self-attention layer, the query Q_s^m, key K_s^m, and value V_s^m are all derived from the noisy image latent z_t. The self-attention map M_{SA}^m in this layer can be obtained via:

$$M_{SA}^m = \text{SoftMax}\left(\frac{Q_s^m K_s^{mT}}{\sqrt{d_m}}\right), \tag{3}$$

where d_m denotes the projection dimension of Q_s^m and K_s^m. As the self-attention map is computed with the correlation between Q_s^m and K_s^m which are both derived from the image contents, it can contain information about the semantic association between each pixel and other pixels [17]. As shown in Fig. 1 (c), pixels belonging to the same object tend to have higher correlations. Such semantic grouping information can provide insight for class-agnostic counting.

3.2 Proposed CountDiff Framework

In class-agnostic object counting, generally the model is trained on a set of known object classes C_{tr}, supervised by ground-truth density maps. Then, the model is required to generalize to counting objects of novel object classes $c \in C_{te}$. Specifically, as shown in Fig. 2 (a), in the few-shot setting, the target class is specified by n exemplars, annotated by n bounding boxes $\{b_i\}_{i=1}^n$ in the image. Meanwhile, in the zero-shot setting, the target class is indicated by the class name y_c. In this paper, our proposed CountDiff framework can be flexibly switched to perform few-shot or zero-shot object counting as shown in Fig. 2 (b).

The key insight in CountDiff is that the cross-attention maps in the diffusion model can explicitly reveal the semantic association between the image content and the text input. Thus, given any image, we can leverage the cross-attention maps to conveniently highlight useful information to count the target objects with proper descriptions. However, it can be non-trivial to get such textual descriptions during testing, such as small objects shown in the final two rows in Fig. 4. To this end, we seek to exploit the text embedding space of the pre-trained diffusion model and learn to encode such information in the embeddings.

As illustrated in Fig. 2 (b), CountDiff comprises the following processes. 1) To extract useful information from the pre-trained diffusion model, we propose to learn object-specific embedding and object-agnostic embedding in the text embedding space. 2) Leveraging the learned embeddings, we extract relevant knowledge for class-agnostic counting, such as cross-attention maps that highlight useful information in the given image. 3) With the obtained attention maps, we predict the density map via a regression head. Below, we first introduce these processes in few-shot class-agnostic counting, and subsequently explore how CountDiff can be flexibly switched for the zero-shot setting.

Embedding Set Construction. To facilitate few-shot class-agnostic counting, we propose to learn two types of embeddings to extract knowledge: 1) to count the target objects, we propose to learn an *object-specific embedding* E_{os} to extract the specific knowledge of target objects; and 2) to enhance the generalization

ability, we propose to further learn an *object-agnostic embedding* E_{oa} to extract the general knowledge that can be shared in counting objects of different classes. Below we introduce these two embeddings in detail.

As shown in Fig. 2 (a), in the few-shot setting, no text information is given and only a few exemplars are provided to represent the target class. To enable the diffusion model to comprehend the information of the target object represented by the exemplars, we propose to project the visual information in the exemplars into the text-embedding space.

Specifically, given an input image I and n exemplars (*i.e.*, n bounding boxes $\{b_i\}_{i=1}^{n}$ in this image), we first extract the image feature f_{img} and obtain the exemplar features $\{f_i\}_{i=1}^{n}$ by RoI pooling. These exemplar features are then passed to the exemplar projector ϕ_E as shown in Fig. 2 (b). The exemplar projector maps the visual features to text embedding space and obtains E_{os} as:

$$f_i = \text{RoIPooling}(f_{img}, b_i),\ E_{os} = \frac{1}{n}\sum_{i=1}^{n}\phi_E(f_i), E_{os} \in \mathbb{R}^{d_e} \tag{4}$$

where d_e denotes the dimension of the text embedding in the diffusion model. Via this manner, the object-specific embedding E_{os} is obtained, which links the information in the exemplars to the text embedding space and represents the semantic information of the target objects in this image.

Moreover, in class-agnostic object counting, there could exist some common counting knowledge that can be shared for counting various object classes (*e.g.*, repetitive object patterns in the image can be exploited to assist counting). Ideally, during training, the model should acquire some common counting ability that can be shared for counting novel classes during inference. Thus, to enhance the learning of such general ability, we further incorporate an object-agnostic embedding $E_{oa} \in \mathbb{R}^{d_e}$, as shown in Fig. 2 (b).

Particularly, different from the object-specific embedding that is obtained as an output of the projector ϕ_E for a specific input, the object-agnostic embedding E_{oa} is designed as a *learnable embedding*, which is updated to apply to every image as the model parameters. Thus, during training, it is pushed to capture the general understanding that can contribute to counting for every object class.

After obtaining both the object-specific embedding E_{os} and the object-agnostic embedding E_{oa}, we formulate an embedding set as $E = \{E_{os}, E_{oa}\}$.

Knowledge Extraction Using Object-Specific and Object-Agnostic Embeddings. Now that we have constructed the embeddings, we then aim to leverage them to extract knowledge from the pre-trained diffusion model. Considering that the *cross-attention layers* in the diffusion model can explicitly highlight the relevant image regions given the text input, we propose to extract knowledge that explicitly highlights the relevant information for class-agnostic object counting.

Specifically, as shown in Fig. 2 (b), given an input image I, we first encode it into latent $z_{img} = \mathcal{E}(I)$ using the encoder $\mathcal{E}$ of the diffusion model and add noise

to obtain the noisy latent z. The noisy latent z is sent into the denoising UNet ϵ_θ, along with the embedding set $E = \{E_{os}, E_{oa}\}$ containing the object-specific and object-agnostic embeddings. Then, conditioning on E, we can extract cross-attention map M_{CA}^l from the l-th cross-attention layer as:

$$M_{CA}^l = \text{SoftMax}\left(\frac{Q_c^l {K_P^l}^T}{\sqrt{d_l}}\right), M_{CA}^l \in \mathbb{R}^{h^l \times w^l \times 2}, \tag{5}$$

where query Q_c^l is projected from the image latent, and key K_P^l is derived from the embeddings $\{E_{os}, E_{oa}\}$, and h^l and w^l represent the scale of the attention map. The obtained M_{CA} contains two 2D cross-attention maps corresponding to E_{os} and E_{oa} respectively. As the diffusion model (e.g., Stable Diffusion) has multiple cross-attention layers, following Eq. (5), we obtain cross-attention maps for the embeddings over all the layers. Then, we resize and average these cross-attention maps from different layers, and take the averaged result as the final cross-attention map $M_{CA} \in \mathbb{R}^{h \times w \times 2}$.

Besides cross-attention maps, we also leverage the self-attention maps in the pre-trained diffusion model. As introduced in Sect. 3.1, self-attention maps can encode the semantic correlation between each pixel and other pixels, which can reveal semantic grouping information to facilitate class-agnostic counting. Via Eq. (3), we can obtain the self-attention map M_{SA}^m for the m-th self-attention layer. We also average self-attention maps from different layers to obtain the final self-attention map $M_{SA} \in \mathbb{R}^{h \times w \times h \times w}$.

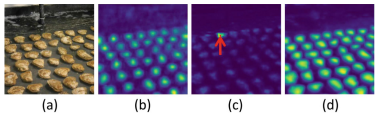

Fig. 3. Visualizations of attention maps. (a) shows the input image, (b) shows the cross-attention map M_{CA} obtained using the embeddings, (c) shows self-attention map M_{SA} at the pixel locations pointed by the red arrows, and (d) shows the enhanced attention map M_{EA}.

Moreover, as self-attention map reflects the correlation between each pixel and other pixels, it inherently contains grouping information for all image contents, including the background. Thus, to incorporate the semantic information of the specific image contents encoded in the embeddings, we apply cross-attention map M_{CA} to self-attention map M_{SA} and obtain the enhanced attention map $M_{EA} \in \mathbb{R}^{h \times w \times 2}$ via:

$$M_{EA}(:,:,p) = \sum_{k=1}^{h} \sum_{j=1}^{w} M_{CA}(k,j,p) M_{SA}(k,j,:,:), \tag{6}$$

where $M_{CA}(k,j,p)$ is the attention value of pixel location (k,j) for the p-th embedding, $p \in \{1,2\}$, and $M_{SA}(k,j,:,:) \in \mathbb{R}^{h \times w}$ is the 2D attention map for pixel location (k,j). Intuitively, this operation gathers the grouping information of the pixels that are highlighted in the cross-attention map in a weighted sum manner, and returns grouping information relevant to the semantic information in the embeddings as the enhanced attention map [17].

As shown in Fig. 3, leveraging the embedding set, the attention maps can explicitly point out the useful information in the image for class-agnostic count-

ing. In the following, we describe how we leverage such information to effectively perform class-agnostic object counting.

Obtaining Density Map Guided by Extracted Knowledge. We obtain the density map by constructing a regression head ϕ_R and leveraging the knowledge extracted with the pre-trained diffusion model, *i.e.*, the cross-attention map M_{CA}, self-attention map M_{SA}, and enhanced attention map M_{EA}. In specific, we rearrange M_{SA} to $M'_{SA} \in \mathbb{R}^{(hw)\times h\times w}$, concatenate it with M_{CA} and M_{EA} and pass the concatenated representations into the regression head. The predicted density map D is obtained via: $D = \phi_R([M_{CA}, M'_{SA}, M_{EA}])$. Finally, the predicted number (n) of the target objects is calculated by summing the pixel values in the predicted density map: $n = \sum_{k,j} D(k,j)$.

To train our model, we calculate the normalized MSE loss between the ground-truth density map D_{gt} and the predicted density map D following [43]. In addition, we incorporate the original loss of the diffusion model L_{LDM} in Eq. (1). Instead of updating the text encoder and diffusion model as in text-to-image diffusion models, here we aim to leverage L_{LDM} to guide the learning of the embeddings. Intuitively, this can help to restrict the learned embeddings to be within the text embedding space of the diffusion model. Overall, the objective is formulated as:

$$L_{CountDiff} = \frac{1}{N}\|d_{gt} - d\|_2^2 + \lambda L_{LDM}, \tag{7}$$

where N is the number of the target objects, and λ is the weight of L_{LDM}.

Test-Time Adaptation for E_{os}. Previous method [35] adopts test-time adaptation to improve the performance. Our CountDiff can also support a light-weight test-time adaptation scheme to further adapt (fine-tune) the object-specific embedding E_{os} such that it can better represent the novel objects for few-shot class-agnostic counting.

In particular, during testing, after obtaining E_{os} using the exemplar projector ϕ_E given the exemplars, we can then regard E_{os} as a learnable embedding with initialized value, and conduct adaptation of E_{os}. As shown in Fig. 2 (a), in few-shot setting, the few (n) exemplars indicated by bounding boxes in the image are provided. Thus, though no ground-truth density map of the whole image is available, we can generate the density map values within these n exemplar boxes based on the center locations and sizes of the exemplar boxes, which can be considered as a guidance signal. We denote the generated density map as D_g. Also, we can obtain the masked predicted density map D_m containing predicted density map values in the exemplar boxes areas by masking the predicted density map D with the exemplar boxes. Then, we compare D_g and D_m to compute the loss L only within the exemplar box areas. The object-specific embedding E_{os} is fine-tuned via:

$$\begin{aligned} L &= \frac{1}{n}\|D_g - D_m\|_2^2, \\ E_{os} &\leftarrow E_{os} - \alpha\nabla_{E_{os}}(L), \end{aligned} \tag{8}$$

where α is the learning rate of this adaptation. Note that as E_{os} is initialized by the output of the exemplar projector, it can have a good starting point and thus it only needs very lightweight fine-tuning. More details are in Supplementary.

CountDiff for Zero-Shot Setting. CountDiff can also be flexibly switched to perform zero-shot class-agnostic counting (providing class name y_c) as shown in Fig. 2 (b). This switch is conveniently achieved by changing the object-specific embedding E_{os}. In the zero-shot setting, E_{os} is obtained using the provided class name y_c. As y_c is in text form, we can leverage it to extract relevant knowledge from the diffusion model. As shown in Fig. 2 (b), we take advantage of the pre-trained text encoder τ_θ in the text-to-image diffusion model to obtain the object-specific embedding: $E_{os} = \tau_\theta(y_c)$. Then, using the obtained object-specific embedding E_{os} along with the object-agnostic embedding E_{oa}, we can extract the attention maps via Eqs. (3), (5) and (6) and predict the density map with the regression head ϕ_R.

3.3 Training and Testing

Few-Shot Setting. During training, each input image is provided with several exemplars and a ground-truth density map. Using the ground-truth density map, we can obtain loss for each image via Eq. (7). The loss is used to update the exemplar projector ϕ_E, the learnable object-agnostic embedding E_{oa}, and the regression head ϕ_R. During testing, we directly adopt the learned object-agnostic embedding E_{oa}, and meanwhile use the trained ϕ_E to obtain the object-specific embedding E_{os}. Then we extract the attention maps to predict the density map. Moreover, to better represent the target objects with the object-specific embedding, we apply test-time adaptation as formulated in Eq. (8) to fine-tune E_{os} for a small number (r) of iterations. Using the updated E_{os}, we can obtain the final predicted density map.

Zero-Shot Setting. During training, each input image is provided with the class name of the target object in this image and the ground-truth density map. The loss is computed as in Eq. (7) and is used to optimize the learnable object-agnostic embedding E_{oa} and the regression head ϕ_R. During testing, given an image of new objects, we obtain the object-specific embedding E_{os} using the pre-trained text encoder of the diffusion model. Together with the object-agnostic embedding E_{oa} learned at the training stage, we can extract knowledge from the diffusion model and leverage such knowledge to predict the density map via the regression head ϕ_R.

In summary, the differences between zero-shot setting and few-shot setting for training and testing are: 1) in zero-shot setting, the object-specific embedding is obtained by the pre-trained text encoder, while in few-shot setting, it is obtained using the trained exemplar projector; 2) in zero-shot setting, as no exemplars are provided, we do not apply test-time adaptation to further adapt E_{os}.

4 Experiments

To evaluate the effectiveness of our proposed CountDiff, we conduct experiments on the common class-agnostic object counting dataset FSC-147 [35] for both few-shot and zero-shot settings. In addition, to further test its generalization ability, we also evaluate CountDiff on CARPK [15].

4.1 Datasets and Evaluation Metrics

The FSC-147 dataset [35] is a large-scale dataset for class-agnostic object counting. It contains 6135 images of 147 object classes. For each image, three bounding boxes indicating the target objects are given. The class name of the target object class in each image is also provided. There is no overlap between the target object classes in the training, validation, and testing sets.

The CARPK dataset [15] is proposed for class-specific counting for cars, and is used to evaluate the model's ability for cross-dataset generalization. It contains 1448 images of different parking lots collected by drones.

Following previous works [35,40,43], we adopt the standard metrics for class-agnostic object counting: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to evaluate the performance of our method.

4.2 Implementation Details

Our experiments are conducted on RTX 3090 GPUs. In our experiments, we adopt Stable Diffusion [36] (version 1-4) as the pre-trained text-to-image diffusion model. To obtain the noisy latent z, we add the noise to the image latent using the DDPM scheduler [13]. We set h and w to 64. For both few-shot and zero-shot experiments, the embedding dimension d_e is set to the text embedding dimension of Stable Diffusion. During training, we adopt AdamW optimizer [23] with a learning rate of 5e-4. We freeze the parameters of Stable Diffusion, and train CountDiff for 100 epochs. We set λ in Eq. (7) to 0.005. The test-time adaptation for the few-shot setting is performed for $r = 3$ iterations with learning rate $\alpha =$ 5e-5. See Supplementary for more implementation details.

4.3 Experiment Results on Few-Shot Setting

We evaluate CountDiff for the few-shot object counting on the common class-agnostic counting dataset FSC-147. Following [35,40,43], during training and inference, the model is provided with 3 exemplars for each input image. We report the MAE and RMSE scores in Table 1. As shown, our CountDiff outperforms previous methods on the few-shot setting, demonstrating its effectiveness.

To further evaluate the generalization ability, we also conduct cross-dataset experiments on the CARPK dataset following the evaluation protocol in [43]. The results are shown in Table 2. Our CountDiff also achieves state-of-the-art performances, demonstrating its generalization ability for different settings.

Table 1. Few-shot results on FSC-147.

Method	Validation set		Test set	
	MAE	RMSE	MAE	RMSE
FamNet [35]	23.75	69.07	22.08	99.54
CFOCNet [46]	21.19	61.41	22.10	112.71
RCAC [11]	20.54	60.78	20.21	81.86
BMNet+ [40]	15.74	58.53	14.62	91.83
SAFECount [47]	15.28	47.20	14.32	85.54
SPDCN [21]	14.59	49.97	13.51	96.80
CounTR [6]	13.13	49.83	11.95	91.23
LOCA [43]	10.24	32.56	10.79	56.97
Ours	**8.43**	**31.03**	**9.24**	**53.41**

Table 2. Few-shot cross-dataset results on CARPK.

Method	MAE	RMSE
FamNet [35]	28.84	44.47
BMNet+ [40]	10.44	13.77
LOCA [43]	9.97	12.51
Ours	**8.36**	**10.84**

4.4 Experiment Results on Zero-Shot Setting

We evaluate CountDiff with the zero-shot setting (providing class names) on FSC-147. As shown in Table 3, our CountDiff achieves the best performance compared to the existing methods, demonstrating the effectiveness of our design.

We also observe that CountDiff achieves better performance in few-shot setting compared to zero-shot setting (*e.g.*, on validation set, we achieve 8.43 on MAE for few-shot setting as shown in Table 1 and achieve 15.50 on MAE for zero-shot setting as shown in Table 3). A possible explanation can be, in the few-shot setting, the model is provided with exemplars in the input image, which can contain more detailed and specific information about the target objects in the image than the general class name provided in the zero-shot setting, and thus the learned embeddings can better represent the objects in this image.

Also, following [16,43], we conduct cross-dataset experiment for the zero-shot setting on CARPK. As shown in Table 4, our method also achieves outstanding performance on cross-dataset experiment, showing that CountDiff has good generalization ability.

Table 3. Zero-shot results on FSC-147.

Method	Validation set		Test set	
	MAE	RMSE	MAE	RMSE
Xu *et al.* [45]	26.93	88.63	22.09	115.17
Shi *et al.* [48]	–	–	24.79	137.15
Jiang *et al.* [16]	18.79	61.18	17.78	106.62
Amini-Naieni *et al.* [1]	17.70	63.61	15.73	106.88
Ours	**15.50**	**54.33**	**14.83**	**103.15**

Table 4. Zero-shot cross-dataset results on CARPK.

Method	MAE	RMSE
Jiang *et al.* [16]	11.96	16.11
Ours	**10.32**	**12.92**

4.5 Ablation Study

We conduct experiments on FSC-147 in the few-shot setting to evaluate the design of CountDiff. More experiments and visualizations on both few-shot and zero-shot settings are in Supplementary.

Evaluation on the Embeddings. CountDiff utilizes two embeddings, *i.e.*, object-specific embedding E_{os} and object-agnostic embedding E_{oa} to extract knowledge. To evaluate the design for E_{os}, we train the following variants from scratch and evaluate them: **1) w/o** E_{os} that uses only E_{oa}, **2) using text for** E_{os} that uses text embedding of ground-truth class name as E_{os}, and **3) using CLIP for** E_{os} that uses pre-trained CLIP image encoder to map the exemplars to embedding space for E_{os}. As shown in Table 5, CountDiff outperforms all these variants, showing the effectiveness of the object-specific embedding. In specific, using text for E_{os} shows a decrease in performance, which can be because the general class name can provide less specific information compared with exemplars cropped from the input image. Also, CountDiff outperforms the variant with E_{os} naively obtained using the off-the-shelf CLIP image encoder, implying that E_{os} learned in our framework contains information that can help counting.

Table 5. Evaluation on the embeddings.

Method	Validation set		Test set	
	MAE	RMSE	MAE	RMSE
w/o E_{os}	18.10	59.68	17.45	108.31
using text for E_{os}	15.23	54.08	14.22	103.01
using CLIP for E_{os}	12.85	36.64	13.67	58.74
w/o E_{oa}	13.18	36.58	14.34	60.87
using text for E_{oa}	12.67	35.47	13.85	59.77
CountDiff	**8.43**	**31.03**	**9.24**	**53.41**

Also, to evaluate the design for E_{oa}, we test the variants: **1) w/o** E_{oa} that uses only E_{os}, and **2) using text for** E_{oa} that uses the text embedding of the word "*object*" for E_{oa}. As shown in Table 5, CountDiff outperforms both variants. This may be because, without E_{oa}, the model may not be able to leverage general knowledge (*e.g.*, repetitive object patterns) to facilitate counting. Also, the general knowledge may not be easily represented by text.

Qualitative Results. We visualize the cross-attention maps and enhanced attention maps obtained using the embeddings, as well as the predicted density maps for few-shot settings in Fig. 4. As shown, attention maps for E_{os} can provide precise and targeted information about the target object, implying that E_{os} can encode semantic information for the specific target objects in each image. Also, we observe that the attention maps for E_{oa} can show explicit information about the objects, regardless of the object classes (as shown in the second row), implying that E_{oa} can implicitly contain instructions to extract general knowledge for class-agnostic counting.

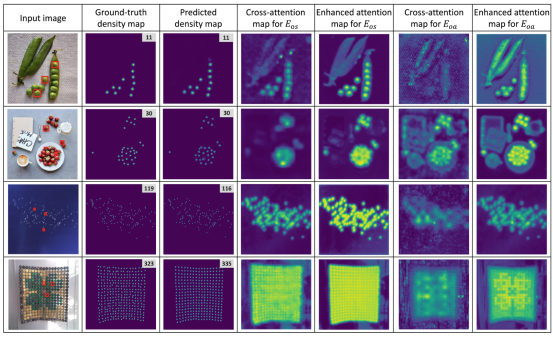

Fig. 4. Qualitative results with few-shot setting on FSC-147 dataset. In each row, we show the input image with exemplar boxes (red), ground-truth density map, predicted density map, and the attention maps obtained using the embeddings. The ground-truth and predicted counting results are shown at the top right corner of the density maps.

5 Conclusion

In this paper, we have proposed a novel framework named CountDiff to leverage the powerful capability of text-to-image diffusion model to perform class-agnostic object counting. We design an object-specific embedding that encodes specific information about the target object and an object-agnostic embedding that contains general information that can be useful for counting different classes, which are used to extract knowledge from the pre-trained diffusion model to facilitate class-agnostic counting. Along with other designs, our CountDiff can achieve state-of-the-art performance in both few-shot and zero-shot settings.

Acknowledgements. This work is supported by Lam Research (ND-00102-E0901-E0901-000).

References

1. Amini-Naieni, N., Amini-Naieni, K., Han, T., Zisserman, A.: Open-world text-specified object counting. In: British Machine Vision Conference (2023)
2. Arteta, C., Lempitsky, V., Zisserman, A.: Counting in the Wild. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 483–498. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_30

3. Babu Sam, D., Agarwalla, A., Joseph, J., Sindagi, V.A., Babu, R.V., Patel, V.M.: Completely self-supervised crowd counting via distribution matching. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) European Conference on Computer Vision, pp. 186–204. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19821-2_11
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9912–9924 (2020)
5. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: counting people without people models or tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2008). https://doi.org/10.1109/CVPR.2008.4587569
6. Chang, L., Yujie, Z., Andrew, Z., Weidi, X.: CounTR: transformer-based generalised visual counting. In: British Machine Vision Conference (BMVC) (2022)
7. Chattopadhyay, P., Vedantam, R., Selvaraju, R.R., Batra, D., Parikh, D.: Counting everyday objects in everyday scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1135–1144 (2017)
8. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: attention-based semantic guidance for text-to-image diffusion models. ACM Trans. Graph. (TOG) **42**(4), 1–10 (2023)
9. Foo, L.G., Rahmani, H., Liu, J.: AIGC for various data modalities: a survey. arXiv preprint arXiv:2308.14177 (2023)
10. Gal, R., et al.: An image is worth one word: personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=NAQvF08TcyG
11. Gong, S., Zhang, S., Yang, J., Dai, D., Schiele, B.: Class-agnostic object counting robust to intraclass diversity. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) European Conference on Computer Vision, pp. 388–403. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19827-4_23
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=_CDixzkzeyb
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)
14. Hobley, M., Prisacariu, V.: Learning to count anything: reference-less class-agnostic counting with weak supervision. arXiv preprint arXiv:2205.10203 (2022)
15. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal networks. In: The IEEE International Conference on Computer Vision (ICCV). IEEE (2017)
16. Jiang, R., Liu, L., Chen, C.: CLIP-count: towards text-guided zero-shot object counting. In: MM '23, Proceedings of the 31st ACM International Conference on Multimedia, pp. 4535–4545. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3581783.3611789
17. Khani, A., Asgari, S., Sanghi, A., Amiri, A.M., Hamarneh, G.: SLime: segment like me. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=7FeIRqCedv
18. Laradji, I.H., Rostamzadeh, N., Pinheiro, P.O., Vazquez, D., Schmidt, M.: Where are the blobs: counting by localization with point supervision. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 547–562 (2018)

19. Liang, D., Xie, J., Zou, Z., Ye, X., Xu, W., Bai, X.: CrowdCLIP: unsupervised crowd counting via vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2893–2903 (2023)
20. Lin, W., Chan, A.B.: 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), pp. 21663–21673 (2023)
21. Lin, W., et al.: Scale-prior deformable convolution for exemplar-guided class-agnostic counting. In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21–24, 2022. BMVA Press (2022). https://bmvc2022.mpi-inf.mpg.de/0313.pdf
22. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5099–5108 (2019)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11363, pp. 669–684. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20893-6_42
25. Michel, A., Gross, W., Schenkel, F., Middelmann, W.: Class-aware object counting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 469–478 (2022)
26. 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: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 785–800. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_48
27. Nguyen, T., Pham, C., Nguyen, K., Hoai, M.: Few-shot object counting and detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) European Conference on Computer Vision, pp. 348–365. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20044-1_20
28. Norouzzadeh, M.S., et al.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. **115**(25), E5716–E5725 (2018)
29. Peng, D., Zhang, Z., Hu, P., Ke, Q., Yau, D., Liu, J.: Harnessing text-to-image diffusion models for category-agnostic pose estimation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) European Conference on Computer Vision. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72624-8_20
30. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
31. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
32. Ramesh, A., et al.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831. PMLR (2021)
33. Ranasinghe, Y., Nair, N.G., Bandara, W.G.C., Patel, V.M.: Diffuse-denoise-count: accurate crowd-counting with diffusion models. arXiv preprint arXiv:2303.12790 (2023)
34. Ranjan, V., Nguyen, M.H.: Exemplar free class agnostic counting. In: Proceedings of the Asian Conference on Computer Vision, pp. 3121–3137 (2022)

35. Ranjan, V., Sharma, U., Nguyen, T., Hoai, M.: Learning to count everything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3394–3403 (2021)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
37. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
38. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22500–22510 (2023)
39. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems, vol. 35, pp. 36479–36494 (2022)
40. Shi, M., Lu, H., Feng, C., Liu, C., Cao, Z.: Represent, compare, and learn: a similarity-aware framework for class-agnostic counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9529–9538 (2022)
41. Song, Q., et al.: Rethinking counting and localization in crowds: a purely point-based framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
42. Sundararaman, R., De Almeida Braga, C., Marchand, E., Pettre, J.: Tracking pedestrian heads in dense crowd. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3865–3875 (2021)
43. Đukić, N., Lukežič, A., Zavrtanik, V., Kristan, M.: A low-shot object counting network with iterative prototype adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 18872–18881 (2023)
44. Wang, M., Li, Y., Zhou, J., Taylor, G.W., Gong, M.: GCNet: probing self-similarity learning for generalized counting network. arXiv preprint arXiv:2302.05132 (2023)
45. Xu, J., Le, H., Nguyen, V., Ranjan, V., Samaras, D.: Zero-shot object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15548–15557 (2023)
46. Yang, S.D., Su, H.T., Hsu, W.H., Chen, W.C.: Class-agnostic few-shot object counting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 870–878 (2021)
47. You, Z., Yang, K., Luo, W., Lu, X., Cui, L., Le, X.: Few-shot object counting with similarity-aware feature enhancement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6315–6324 (2023)
48. Zenglin Shi, Ying Sun, M.Z.: Training-free object counting with prompts. In: WACV (2024)
49. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847 (2023)
50. Zhang, Z., Xu, L., Peng, D., Rahmani, H., Liu, J.: Diff-tracker: text-to-image diffusion models are unsupervised trackers. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) European Conference on Computer Vision. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73390-1_19