Open-world Text-specified Object Counting

Niki Amini-Naieni[1]
niki.amini-naieni@eng.ox.ac.uk
Kiana Amini-Naieni[2]
kamininaieni@ucdavis.edu
Tengda Han[1]
htd@robots.ox.ac.uk
Andrew Zisserman[1]
az@robots.ox.ac.uk

[1] Visual Geometry Group (VGG), University of Oxford, UK
[2] University of California, Davis, USA

Abstract

Our objective is open-world object counting in images, where the target object class is specified by a text description. To this end, we propose *CounTX*, a class-agnostic, single-stage model using a transformer decoder counting head on top of pre-trained joint text-image representations. CounTX is able to count the number of instances of any class given only an image and a text description of the target object class, and can be trained end-to-end. In addition to this model, we make the following contributions: (i) we compare the performance of CounTX to prior work on open-world object counting, and show that our approach exceeds the state of the art on all measures on the FSC-147 [24] benchmark for methods that use text to specify the task; (ii) we present and release FSC-147-D, an enhanced version of FSC-147 with text descriptions, so that object classes can be described with more detailed language than their simple class names. FSC-147-D and the code are available at https://www.robots.ox.ac.uk/~vgg/research/countx.

1 Introduction

The goal of object counting is to estimate the number of relevant objects in an image. Traditional object counting methods focus on counting objects of a specific class of interest [2, 3, 18, 27]. These techniques are well-suited for solving particular problems, but they cannot operate in *open-world* settings, where the class of interest is not known beforehand and can be arbitrary.

One approach to open-world object counting is class-agnostic few-shot object counting [16, 29]. In this setting, a user specifies the class of interest at inference time with one or more visual exemplars. These exemplars take the form of bounding boxes over different instances of the object in the image. Although class-agnostic few-shot object counters can be deployed in open-world scenarios, they require a human to provide the visual exemplars during inference. An alternative approach is for the object of interest to be specified by text, rather than by visual exemplars. This is the approach proposed recently by Xu *et al.* [28], where the objects to be counted can be specified by an *arbitrary* class name at inference

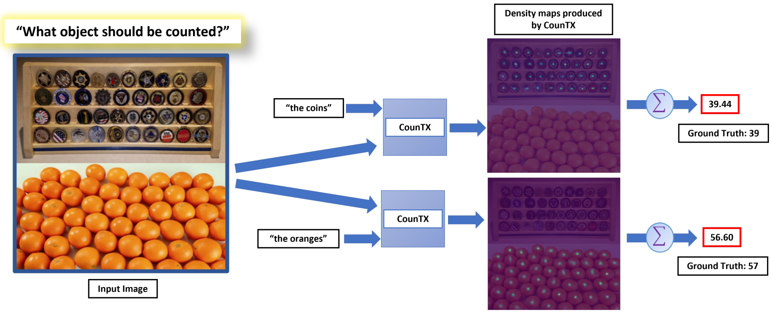

Figure 1: CounTX estimates object counts directly from an image as a response to the question "what object should be counted?". In this example, two text inputs are used to predict the object counts of different objects in the same image. Note, no visual exemplars are required at any stage.

time. This is implemented as a two-stage process where first the text is used to select visual exemplars in the image, and then the counting proceeds using these exemplars as in the standard few-shot setting.

In this paper, we propose *CounTX*, a *single-stage* open-world image counting model, where the objects to be counted are specified by a textual description at inference time. CounTX (pronounced "Count-text") does not use visual exemplars, and can be trained end-to-end. It can count objects of any class, even ones unseen during training. CounTX exceeds the performance of the two-stage approach of [28] on all measures on FSC-147, a standard counting benchmark. Figure 1 shows an example of its use.

The key idea is to benefit from the availability of image encoders that have been pre-trained for a joint text-image encoding using large scale image-text paired data, such as CLIP [22]. Taking inspiration from Chang *et al.* [15], that attention mechanisms can be used to model the similarity between image patches and other inputs, we use a transformer decoder to determine the similarity between the text encoding of the target object description and the spatial map of the image. This generates a density map that is then decoded to the image resolution and used for counting. We also investigate whether it is better to freeze or fine-tune the image or text encoder when training the counting head of the model.

In short, we make three contributions: *First*, we develop CounTX, an open-world counting model that accepts an image and an arbitrary object class description, and directly uses these inputs to predict the object count. We are the first to tackle this open-world counting problem using a single-stage approach, without relying on an exemplar-based counting model; *Second*, we augment the FSC-147 [24] dataset with class descriptions and release the modified dataset, FSC-147-D, for future research; *Third*, we verify the effectiveness of our model and training procedure on the FSC-147 dataset through both quantitative and qualitative results. CounTX significantly improves on both the validation set and the test set performance of [28], the only prior work on open-world text-specified object counting.

2 Related Work

Class-specific Object Counting. Class-specific object counting focuses on counting objects of a specific category [2, 3, 18, 27]. There are two main approaches to this task: detection-based methods [4, 6, 10] and regression-based methods [1, 2, 5, 12, 13, 17, 27].

While detection-based methods rely on an object detector to produce bounding boxes that can be enumerated to predict the object count, regression-based methods directly map the input image to a continuous scalar estimate of the object count.

Class-agnostic Object Counting. The goal of class-agnostic object counting is to count the instances of an arbitrary class in an image given a number of visual exemplars at the time of inference [1, 7, 8, 14, 15, 16, 19, 24, 26, 29, 30]. Class-agnostic object counters require users to provide a positive number of exemplars to specify the class of the object to count. For instance, CounTR uses a two-stream transformer-based architecture to model the similarity between image patches and visual exemplars [15]. While CounTR accepts any number of visual exemplars (zero or more), if the user provides zero exemplars, CounTR will default to counting instances of the *dominant* class. Although CounTX also uses a two-stream transformer-based architecture, it does not require visual exemplars for class specification. There are also general object counting models that do not use exemplars [9, 23]. However, they do not allow the user to specify the class of interest. Instead, like CounTR with zero exemplars, these algorithms count instances of the dominant class in the image.

Text-specified Object Counting. The aim of text-specified object counting is to count objects of an arbitrary class in an image given only the class description. Xu *et al.* [28] are the first to propose a method for this task. Unlike CounTX, the technique presented by Xu *et al.* does not produce object counts directly. Instead, it proposes optimal visual exemplars for use by existing class-agnostic object counting networks such as FamNet [24], BMNet [26], BMNet+ [26], and Xu *et al.*'s own architecture, all already trained with annotated visual exemplars. Xu *et al.* refer to their method as 'zero-shot' as it avoids the user inputting the visual exemplars. In contrast, CounTX eliminates the need for annotated visual exemplars altogether, and also accepts a more detailed specification of the target object to count (rather than simply using a class name).

3 Method

We consider the problem of open-world object counting in images, where the target object class is specified by a text description. In this setting, classes unseen during training may be encountered during inference.

3.1 Overview

Given a training set $\mathcal{D}_{train} = \{(X_1, t_1, Y_1), \ldots, (X_N, t_N, Y_N)\}$, each $X_i \in \mathbb{R}^{H \times W \times 3}$ is a training image with a tokenized class description $t_i \in \mathbb{R}^n$ and a binary map $Y_i \in \mathbb{R}^{H \times W}$ with a one at the center of each object in X_i described by t_i and zeros at all other entries. As shown in Eq. 1, the entries of Y_i can be summed to obtain the count of objects in X_i described by t_i:

$$Count(X_i, t_i) = \sum_{p,q} (Y_i)_{p,q} \tag{1}$$

where p, q specify the pixel index.

Our goal is to develop an open-world object counter and to train it on $\mathcal{D}_{train}$ such that it generalizes well to $\mathcal{D}_{test}$, a held-out test set of images with classes not in D_{train}. To achieve

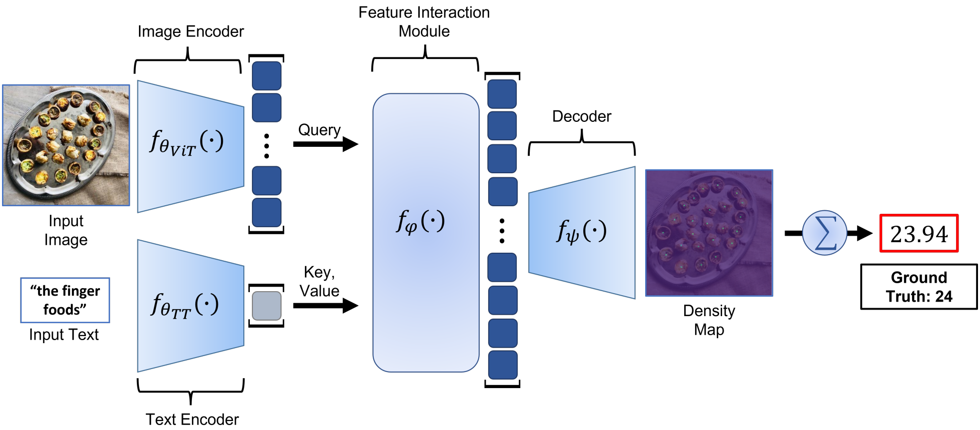

Figure 2: The CounTX architecture. The input image and class description are encoded by a vision transformer and a text transformer respectively. The image features are then passed to the feature interaction module to compute the query vectors, and the text feature is passed in to compute the key and value vectors. The output of the feature interaction module is reshaped to a spatial feature map that is upsampled in the decoder module to produce a density map with the same height and width as the input image and entries that sum to the object count.

this, we introduce CounTX, a novel transformer-based architecture that directly determines a density map from each input image and class description. This density map can be summed to estimate the object count. In section 3.2, the inspiration for the CounTX architecture is explained, and each of its modules are described in detail. In section 3.3, the motivation and methods used for pre-training and fine-tuning CounTX are outlined.

3.2 Architecture

In this section, we describe the CounTX architecture, illustrated in Figure 2. CounTX is inspired by the class-agnostic object counting framework, CounTR, of Chang *et al.* [13]. Like CounTR, CounTX includes a transformer-based image encoder, $f_{\theta_{ViT}}$, and a transformer based feature interaction module, f_{φ}. While the feature interaction module in CounTR is used to mine the similarities between image patches and visual exemplars, the feature interaction module in CounTX is used to mine the similarities between image patches and class descriptions. In this way, CounTX is a natural extension from open-world object counting using image exemplars to open-world object counting using text descriptions.

For an image, X, with class description, t, an estimate, $\hat{y}$, of the number of objects described by t in X can be obtained from CounTX as:

$$\hat{Y} = f_{\psi}(f_{\varphi}(f_{\theta_{ViT}}(X), f_{\theta_{TT}}(t))) \tag{2}$$

where $\hat{Y}$ is the generated density map, with entries that sum to the object count (i.e., $\hat{y} = \sum_{p,q}(\hat{Y})_{p,q}$). Next, we describe each module in equation 2 in detail.

Image Encoder ($f_{\theta_{ViT}}$). For the image encoder, $f_{\theta_{ViT}}$, the CLIP vision transformer ViT-B-16 [11, 22] is used. This image backbone has been contrastively pretrained with the text encoder, $f_{\theta_{TT}}$, on image-text pairs in LAION-2B [25]. The ViT-B-16 model has a patch size of 16×16, 12 layers, and a final embedding dimension of 512. Only the patch tokens output by this image encoder are used, and the CLS token is discarded.

Text Encoder ($f_{\theta_{TT}}$). For the text encoder, $f_{\theta_{TT}}$, the CLIP text transformer [11, 22] contrastively pretrained with the ViT-B-16 image encoder, $f_{\theta_{ViT}}$, on LAION-2B [25] is used. $f_{\theta_{TT}}$ has a context length of 77, 12 layers, and a final embedding dimension of 512. While $f_{\theta_{ViT}}$ transforms each input image into a spatial map of 512-dimensional feature vectors corresponding to the image patches, $f_{\theta_{TT}}$ transforms a class description into a single 512-dimensional feature vector.

Feature Interaction Module (f_φ). To fuse the information captured by the image features $f_{\theta_{ViT}}(X)$ and the text feature $f_{\theta_{TT}}(t)$, two transformer decoder layers with embedding dimensions of 512 are used in f_φ. f_φ uses the image features $f_{\theta_{ViT}}(X)$ to compute the query vectors and the text feature $f_{\theta_{TT}}(t)$ to compute the key and value vectors. The cross-attention mechanisms in f_φ allow the model to leverage any similarities between the image patches and the class description preserved by $f_{\theta_{ViT}}(X)$ and $f_{\theta_{TT}}(t)$.

Decoder (f_ψ). Before being passed to the decoder, f_ψ, the output of the feature interaction module, f_φ, is reshaped into a spatial feature map with 512 channels. Each channel is upsampled using bilinear interpolation to 24×24 pixels. The resized maps are then passed through a convolutional layer with 256 filters and upsampled to increase their height and width by a factor of two four times. This progressive four-block convolution and upsampling operation results in a feature map with the same height and width as the input image and 256 channels. These channels are combined into a single-channel density map using a 1×1 convolution. This density map is summed to estimate the object count.

3.3 Training Procedure

The image and text encoders were pre-trained on abundant image-text pairs using CLIP [22]. Therefore, prior to fine-tuning CounTX on the counting task, the image encoder and the text encoder map the input images and class descriptions to a joint text-image embedding space, aiding the feature interaction module in comparing data from the two modalities. Thus, the image and text encoders are first initialized with their pre-trained weights from CLIP. The text encoder is then frozen, while the image encoder is fine-tuned with the rest of the model on the counting task. As shown in section 4.4, this combination of fine-tuning and freezing the image and text encoders produces the best performance. The augmentation and scalable mosaicing schemes used in [15] are employed during fine-tuning. The mean squared per-pixel error between the predicted density map, $\hat{Y}$, and the ground truth density map, Y, as shown in equation 3, is averaged over all the images in each batch to compute the total loss for optimization.

$$\mathcal{L}(\hat{Y}, Y) = \frac{1}{H \times W} \sum_{p,q} ((\hat{Y})_{p,q} - (Y)_{p,q})^2 \tag{3}$$

4 Experiments

4.1 Datasets & Metrics

Datasets. CounTX is evaluated on FSC-147 [24], a class-agnostic object counting dataset containing 6135 images. The FSC-147 training set contains images from 89 classes, while the validation and test sets each contain images from 29 classes. The classes in the training, validation, and test sets are disjoint, making FSC-147 an open-world dataset. FSC-147 offers three visual exemplars for each image, but CounTX does not use them.

While FSC-147 includes class names, these class names are not natural language responses to the question "what object should be counted?" Furthermore, some of the class names do not accurately describe the object being counted. We construct a set of descriptions from FSC-147 suitable for our setting that transform the FSC-147 class names to responses to the question "what object should be counted?" and correct any mistakes that we find. We name this set of descriptions 'FSC-147-D'. For instance, the original class name for image 3696.jpg in FSC-147 does not indicate that the pastries, not the candy pieces on top of the pastries, should be counted. The original class name for image 4231.jpg in FSC-147 is incorrect as the cupcakes are supposed to be counted, not the tray holding them. These class names are changed in FSC-147-D as shown in Figure 3.

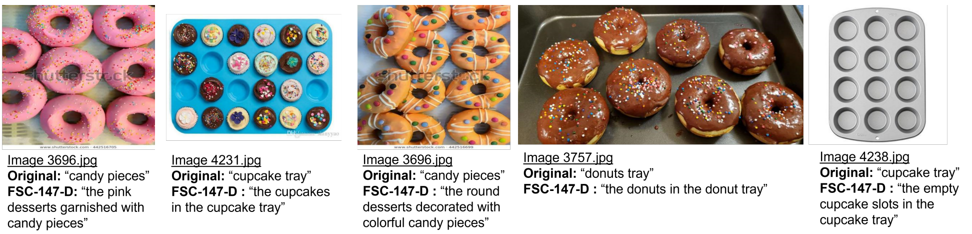

Figure 3: Examples of changes made to FSC-147 to construct FSC-147-D, a dataset for the open-world text-specified object counting setting.

In addition to FSC-147, CounTX is evaluated qualitatively on a subset of CountBench [21], a new text-image benchmark with a total of 540 images containing between two to ten instances of a particular object, where each image's caption reflects this number. CountBench is not a benchmark for text-specified object counting as its captions contain the number of objects to be counted. Thus, for qualitative evaluation, the CountBench captions were replaced with responses to the question "what object should be counted?" For example, the original CountBench caption for the leftmost image in Figure 5 is "a set of six enamelled, gilt silver espresso spoons, tillander, helsinki 1955-56," which was replaced with "the gilt silver espresso spoons" for the text-specified object counting setting.

Metrics. The Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) are used to measure the performance of CounTX. The MAE and RMSE are given by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \quad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{4}$$

where N is the number of test images, $\hat{y}_i$ is the predicted count for image X_i, and y_i is the ground truth count for image X_i.

Method	Year	Published	How to Specify the Class	Validation		Test	
				MAE	RMSE	MAE	RMSE
RepRPN-Counter [23]	2022	✓	None	31.69	100.31	28.32	128.76
RCC [9]	2022	✗	None	20.39	64.62	21.64	103.47
CounTR (0-shot) [15]	2022	✓	None	17.40	70.33	14.12	108.01
LOCA (0-shot) [7]	2022	✗	None	17.43	54.96	16.22	103.96
Patch-selection [28]	2023	✓	Text (class name)	26.93	88.63	22.09	115.17
CounTX (FSC-147-D)	**2023**	-	**Text (class name)**	17.70	**63.61**	**15.73**	106.88
CounTX (FSC-147-D)	**2023**	-	**Text (FSC-147-D)**	**17.10**	65.61	15.88	**106.29**
CounTR (3-shot) [15]	2022	✓	3 Visual Exemplars	13.13	49.83	11.95	91.23
LOCA (3-shot) [7]	2022	✗	3 Visual Exemplars	10.24	32.56	10.79	56.97

Table 1: State-of-the-art performance on FSC-147 for exemplar-free, exemplar-based, and text-based methods. Note that CounTX is trained on FSC-147-D, but evaluated under two different settings: specifying classes with class names (from the original FSC-147) or with descriptions (from FSC-147-D). CounTR and LOCA are grayed out because they use visual exemplars, which provide more information than class descriptions.

4.2 Implementation

Training. Each training image is cropped with a random square window with the same height as the original image and resized to 224×224 pixels. The images are then normalized before being passed through the model. To construct the ground truth density map for each image X_i, a Gaussian filter is applied to its corresponding binary map, Y_i in Equation 1, such that the filtered map still sums to the object count. For optimization, the loss defined in Equation 3 averaged over each batch is minimized. Following [15], the density map values are scaled by 60, and errors contributed by pixels at certain positions are dropped with a 20 % probability. Scaling by a factor of 60 prevents the model from generating a density map of zeros by increasing the penalty for this solution. CounTX is trained on images from the FSC-147 training set and text descriptions from FSC-147-D. We use the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, a batch size of 8, and a learning rate of 6.25×10^{-6} that is warmed up for 10 epochs and then decayed with a half-cycle cosine schedule. The model at the epoch with the smallest mean absolute error on the validation set over 1000 epochs is selected.

Inference. Following [15], a square sliding window is scanned over the image with a stride of 128 pixels. As in training, each square sliding window of the image is resized to 224×224 pixels and normalized before being passed through the model. The density map for overlapping regions is computed using the averaging technique in [15].

4.3 Comparison to State-of-the-art

CounTX is evaluated on the FSC-147 dataset and compared against 0-shot exemplar-free methods that count instances of the dominant class, 3-shot exemplar-based methods, and text-based methods. As shown in Table 1, CounTX trained on FSC-147-D achieves a new state-of-the-art performance across all measures on FSC-147 for text-specified object counting, significantly outperforming Patch-selection [28], both when evaluating with descriptions from FSC-147-D or with class names from the original FSC-147.

4.4 Ablation Study

Freezing vs. Fine-tuning. Different freezing and fine-tuning strategies are compared for both the image encoder and the text encoder. As shown in Table 2, freezing the text encoder and fine-tuning the image encoder on the counting task results in the best performance on FSC-147 across all measures.

Image Encoder Frozen	Text Encoder Frozen	Validation		Test	
		MAE	RMSE	MAE	RMSE
Yes	Yes	37.92	103.57	37.37	130.36
Yes	No	33.34	100.58	37.61	131.2
No	No	17.73	68.19	16.39	107.65
No	**Yes**	**17.10**	**65.61**	**15.88**	**106.29**

Table 2: Performance of different freezing and fine-tuning strategies on FSC-147.

4.5 Qualitative Results

FSC-147 Test Set Image Mosaics. To verify that CounTX uses the class description to count objects, pairs of images in the FSC-147 test set are stitched together, and CounTX is tasked to predict the counts of different classes in the same mosaicked image. In Figure 4, a few examples of when the model clearly distinguished between classes using the class description are presented.

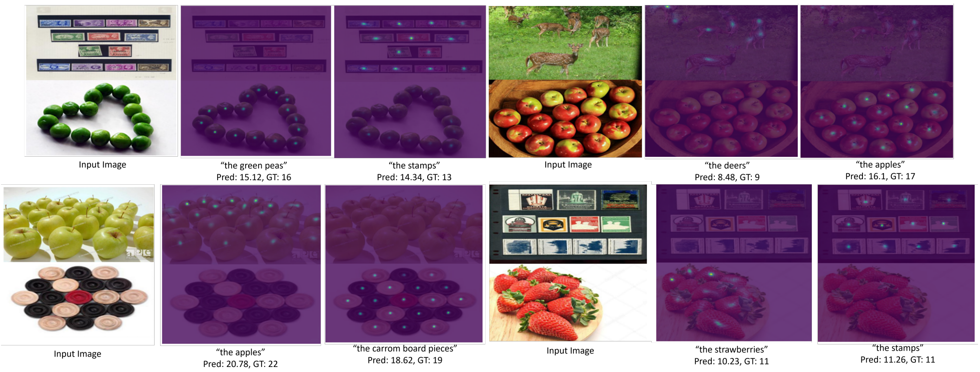

Figure 4: Evaluation on composite images. CounTX uses the class description to identify the object to count. This is clear by how the density map highlights only the regions specified by the class description in each example.

CountBench Density Maps. To further investigate CounTX's generalization abilities to images with small numbers of class instances, responses to the question "what object should be counted?" were constructed for a subset of CountBench [21], and CounTX was applied to this subset. A few of the density maps generated by CounTX are included in Figure 5 and more examples appear in the supplementary material.

FSC-147 Test Set Density Maps. In Figure 6, examples of density maps produced by CounTX when applied to the FSC-147 test set are presented. Each density map is overlaid

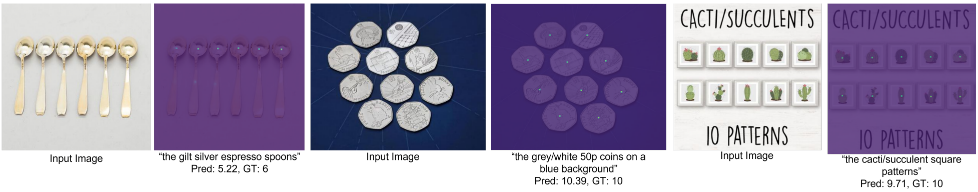

Figure 5: Density maps produced by CounTX when applied to the CountBench subset.

on top of its corresponding image. The class description, predicted count, and ground truth count are also provided.

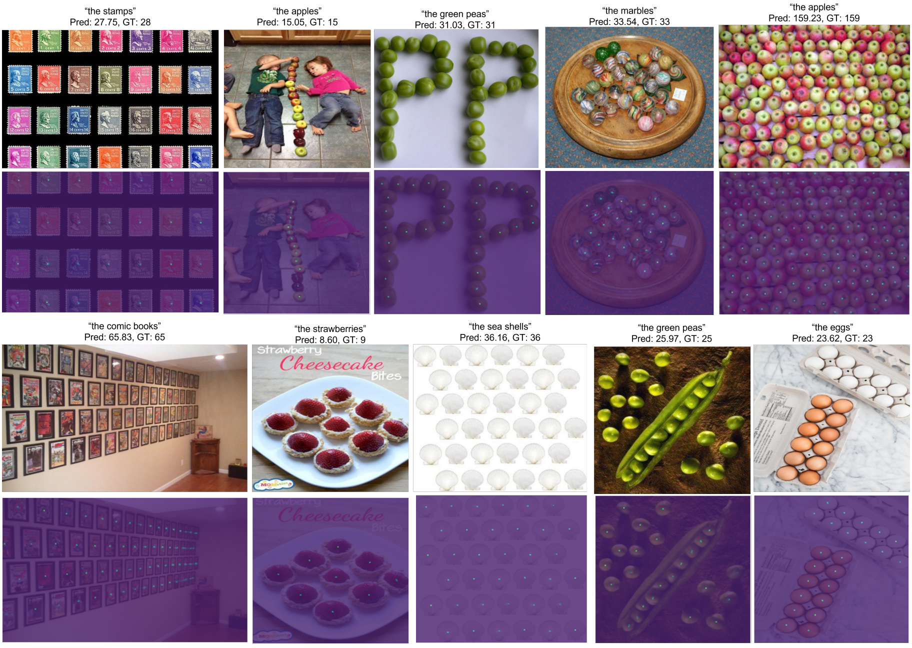

Figure 6: Density maps produced by CounTX when applied to the FSC-147 test set.

5 Conclusion & Future Work

This paper proposes CounTX, an open-world object counting model that accepts an image and an *arbitrary* object class description, and directly uses these inputs to predict the object count. The paper also presents FSC-147-D, an augmented version of FSC-147 (a standard benchmark for class-agnostic counting) with class descriptions. Trained on FSC-147-D, CounTX demonstrates state-of-the-art performance across all measures on FSC-147 for methods that use text to specify the class.

It has been mentioned multiple times in the literature [20, 21, 31] that CLIP models, such as the image encoder used for CounTX, lack the spatial awareness needed for tasks such as counting. This is because, during their pre-training, features from CLIP models may not need to capture the rich structural information in images. Therefore, future work would include replacing the image encoder in CounTX with a more generally trained image backbone that also maps images to a joint text-image embedding space. Some possible visual backbones include a LiT DinoV2 image encoder [20, 31], or a LiT vision transformer trained with self-supervised patch reconstruction to improve its spatial awareness. The CounTX framework together with these spatially aware image encoders could be extended to models that answer other visual questions that require an understanding of spatial layout. These might include queries about the area, shape, and structure of objects in a scene.

Acknowledgement

The authors would like to thank Chang Liu for his extensive support of the CounTR implementation, and Weidi Xie for insightful discussions. T. Han would like to thank Yuki M Asano and Lukas Knobel for helpful discussions. This research is funded by an AWS Studentship, the Reuben Foundation, the AIMS CDT program at the University of Oxford, EPSRC Programme Grant VisualAI EP/T028572/1, and a Royal Society Research Professorship RP \R1 \191132.

References

[1] Carlos Arteta, Victor S. Lempitsky, Julia Alison Noble, and Andrew Zisserman. Interactive object counting. In *Proc. ECCV*, 2014.

[2] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *Proc. ECCV*, 2016.

[3] Deepak Babu Sam, Abhinav Agarwalla, Jimmy Joseph, Vishwanath A. Sindagi, R. Venkatesh Babu, and Vishal M. Patel. Completely self-supervised crowd counting via distribution matching. In *Proc. ECCV*, 2022.

[4] Olga Barinova, Victor S. Lempitsky, and Pushmeet Kohli. On detection of multiple object instances using hough transforms. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010.

[5] Siu-Yeung Cho, Tommy W. S. Chow, and Chi-Tat Leung. A neural-based crowd estimation by hybrid global learning algorithm. In *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 1999.

[6] Chaitanya Desai, Deva Ramanan, and Charless C. Fowlkes. Discriminative models for multi-class object layout. In *IJCV*. Kluwer Academic Publishers, 2011.

[7] Nikola Djukic, Alan Lukezic, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. *arXiv preprint arXiv:2211.08217*, 2022.

[8] Shenjian Gong, Shanshan Zhang, Jiansheng Yang, Dengxin Dai, and Bernt Schiele. Class-agnostic object counting robust to intraclass diversity. In *Proc. ECCV*, 2022.

[9] Michael A. Hobley and Victor Adrian Prisacariu. Learning to count anything: Reference-less class-agnostic counting with weak supervision. *arXiv preprint arXiv:2205.10203*, 2022.

[10] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proc. ICCV*, 2017.

[11] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021.

[12] Dan Kong, D. Gray, and Hai Tao. A viewpoint invariant approach for crowd counting. In *18th International Conference on Pattern Recognition (ICPR)*, 2006.

[13] Victor S. Lempitsky and Andrew Zisserman. Learning to count objects in images. In *NeurIPS*, 2010.

[14] Wei Lin, Kunlin Yang, Xinzhu Ma, Junyu Gao, Lingbo Liu, Shinan Liu, Jun Hou, Shuai Yi, and Antoni Chan. Scale-prior deformable convolution for exemplar-guided class-agnostic counting. In *Proc. BMVC*, 2022.

[15] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. CounTR: Transformer-based generalised visual counting. In *Proc. BMVC*, 2022.

[16] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *Proc. ACCV*, 2018.

[17] Aparecido Nilceu Marana, Sergio A. Velastín, Luciano da Fontoura Costa, and R.A. Lotufo. Estimation of crowd density using image processing. In *IEE Colloquium on Image Processing for Security Applications*, 1997.

[18] T. Nathan Mundhenk, Goran Konjevod, Wesam A. Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *Proc. ECCV*, 2016.

[19] Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. In *Proc. ECCV*, 2022.

[20] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[21] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

[23] Viresh Ranjan and Minh Hoai Nguyen. Exemplar free class agnostic counting. In *Proc. ACCV*, 2023.

[24] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proc. CVPR*, 2021.

[25] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks Track*, 2022.

[26] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and ZHIGUO CAO. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proc. CVPR*, 2022.

[27] Weidi Xie, J. Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. In *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. Taylor & Francis, 2018.

[28] Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *Proc. CVPR*, 2023.

[29] Shuo-Diao Yang, Hung-Ting Su, Winston H. Hsu, and Wen-Chin Chen. Class-agnostic few-shot object counting. In *Proc. WACV*, 2021.

[30] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *Proc. WACV*, 2023.

[31] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proc. CVPR*, 2022.