Exemplar Free Class Agnostic Counting

Viresh Ranjan[1] and Minh Hoai Nguyen[1,2]

[1] Stony Brook University, USA
[2] VinAI Research, Vietnam

Abstract. We tackle the task of Class Agnostic Counting, which aims to count objects in a novel object category at test time without any access to labeled training data for that category. All previous class agnostic counting methods cannot work in a fully automated setting, and require computationally expensive test time adaptation. To address these challenges, we propose a visual counter which operates in a fully automated setting and does not require any test time adaptation. Our proposed approach first identifies exemplars from repeating objects in an image, and then counts the repeating objects. We propose a novel region proposal network for identifying the exemplars. After identifying the exemplars, we obtain the corresponding count by using a density estimation based Visual Counter. We evaluate our proposed approach on FSC-147 dataset, and show that it achieves superior performance compared to the existing approaches. Our code and models are available at: https://github.com/Viresh-R/ExemplarFreeCounting.git.

1 Introduction

In recent years, visual counters have become more and more accurate at counting objects from specialized categories such as human crowd [11, 49, 12, 25], cars [26], animals [3], and cells [2, 47, 13]. Most of these visual counters treat counting as a class-specific regression task, where a class-specific mapping is learned to map from an input image to the corresponding object density map, and the count is obtained by summing over the density map. However, this approach does not provide a scalable solution for counting objects from a large number of object categories because these visual counters can count only a single category at a time, and it also requires hundreds of thousands [49] to millions of annotated training objects [44, 38] to achieve reasonably accurate performance for each category. A more scalable approach for counting objects from many categories is to use class-agnostic visual counters [24, 30], which can count objects from many categories. But the downside of not having a predefined object category is that these counters require a human user to specify what they want to count by providing several exemplars for the object category of interest. As a result, these class-agnostic visual counters cannot be used in any fully automated systems. Furthermore, these visual counters need to be adapted to each new visual category [24] or each test image [30], leading to slower inference.

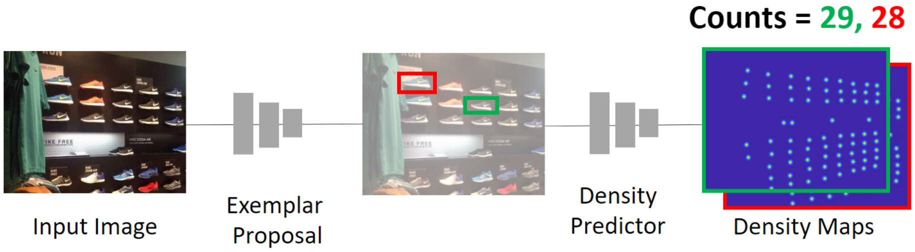


Fig. 1. Exemplar Free Class Agnostic Counter. Given an image containing instances of objects from unseen object categories, our proposed approach first generates exemplars from the repeating classes in the image using a novel region proposal network. Subsequently, a density predictor network predicts separate density maps for each of the exemplars. The total count for any exemplar, i.e. the number of times the object within the exemplar occurs in the image, is obtained by summing all the values in the density map corresponding to that exemplar.

In this paper, we present the first exemplar-free class-agnostic visual counter that is capable of counting objects from many categories, even for novel categories that have neither annotated objects at training time nor exemplar objects at testing time. Our visual counter does not require any human user in its counting process, and this will be very crucial for building fully automated systems in various applications in wildlife monitoring, healthcare and visual anomaly detection. For example, this visual counter can be used to alert environmentalists when a herd of animals with significant size pass by an area monitored by a wildlife camera. Another example is to use this visual counter to monitor for critical health conditions when any certain type of cells outgrows the other types. Unlike existing class-agnostic counters [24, 30], our approach does not use any test time adaptation or finetuning.

At this point, a reader might wonder if it is possible to identify all possible exemplars in an image automatically by using a class-agnostic object detector such as a Region Proposal Network (RPN) [33], and run an existing class-agnostic visual counter using the detected exemplars to count all objects in all categories. Although this approach does not require a human's input during the counting process, it can be computationally expensive. This is because the RPNs usually produce a thousand or more of object proposals. And this in turn requires executing the class-agnostic visual counter at least a thousand times, a time-consuming and computationally demanding process.

To avoid this expensive procedure, we develop in this paper a novel convolutional network architecture called **Rep**etitive **R**egion **P**roposal **N**etwork (RepRPN), which can be used to automatically identify few exemplars from the most frequent classes in the image. RepRPN is used at the first stage of our proposed two-stage visual counting algorithm named RepRPN-Counter. We use a density estimation based Visual Counter as the second stage of the RepRPN-Counter, which predicts a separate high resolution density map for each exemplar. Given an input image, RepRPN considers multiple region proposals, and compute the objectness and repetition scores for each proposal. The repetition score of a proposal is defined as the number of times the object contained within the proposal occurs in the image. The proposals with the highest repetition scores are chosen as the exemplars, and the second stage density predictor estimates the density maps only for the chosen exemplars with high repetition scores. This exemplar generation procedure relies on the underlying assumption that in an image containing different classes with varying counts, the classes of interest are the ones having larger counts. Compared to the traditional RPN [33], RepRPN is better suited for visual counting task, since it can significantly reduce the training and inference time for any two-stage counter. Furthermore, RepRPN can serve as a fast visual counter for applications which can tolerate some margin of error and do not require the localization information conveyed by density maps. Note that the second stage predictor of our visual counter estimates a separate density map for each of the chosen exemplars.

While training RepRPN-Counter, another technical challenge that we need to overcome is the lack of proper annotated data. The only dataset suitable for training class-agnostic visual counters is FSC-147 [30], which contains annotation for a single object category in each image, and may contain unannotated objects from other categories. To obtain annotation for unannotated objects in the FSC-147 dataset, we propose a novel knowledge transfer strategy where we use a RepRPN trained on a large scale object detection dataset [19] and a density prediction network [30] trained on FSC-147 as teacher networks.

In short, the contributions of this paper are threefold: (1) we develop the first exemplar free class agnostic visual counter for novel categories that have neither annotated objects at training time nor exemplar objects at testing time; (2) we develop a novel architecture to simultaneously estimate the objectness and repetition scores of each proposal; (3) we propose a knowledge transfer strategy to handle unannotated objects in the FSC-147 dataset.

2 Related Work

Visual Counting. Most previous methods for visual counting focus on specific categories [48, 41, 31, 1, 25, 49, 28, 4, 35, 17, 22, 6, 29, 37, 20, 45, 39, 42, 43]. These visual counters can count a single category at a time, and require training data with hundreds of thousands [49] to millions of annotated instances [44] for every visual category, which are expensive to collect. These visual counters cannot generalize to new categories at test time, and hence, cannot handle our class agnostic counting task. To reduce the expensive annotation cost, some of these methods focus on designing unsupervised [22] and semi-supervised tasks [23] for visual counting. However, these methods still require a significant amount of annotations and training time for each new category.

Class Agnostic Counting. Most related to ours is the previous works on class agnostic counting [24, 30], which build counters that can be trained to count novel classes using relatively small number of examples from the novel classes.

Lu and Zisserman [24] proposed a Generic Matching Network (GMN) for class-agnostic counting, which follows a two-stage training framework where the first stage is trained on a large-scale video object tracking data, and the second stage consists of adapting GMN to a novel object class. GMN uses labeled data from the novel object class during the second stage, and only works well if several dozens to hundreds of examples are available for the adaptation. Few-shot Adaptation and Matching Network (FamNet) [30] is a recently proposed class agnostic few-shot visual counter which generalizes to a novel category at test time given only a few exemplars from the category. However, FamNet is an interactive visual counter which requires an user to provide the exemplars from the test image. Both GMN and FamNet require test time adaptation for each new class or test image, leading to slower counting procedures.

Zero-Shot Object Detection. Also related to ours is the previous work on zero-shot object detection [5, 50, 27]. Most of these approaches [5, 27] use a region proposal network to generate class-agnostic proposals, and map the features from the proposals to a semantic space where they can be directly compared with semantic word embeddings of novel object classes. However, all of these zero-shot detection approaches require access to the semantic word embeddings for the test classes, and cannot work for our class agnostic counting task where the test classes are not known a priori.

Few-Shot Learning. Also related to ours is the previous works on few-shot learning [16, 15, 36, 8, 32], which aim to adapt classifiers to novel categories based on only a few labeled examples. One of the meta learning based few-shot approaches, Model Agnostic Meta Learning (MAML) [8], has been adapted for class-agnostic counting [30]. MAML focuses on learning parameters which can adapt to novel classes at test time by doing only a few gradient descent steps. Although these few-shot methods reduce the labeled data needed to generalize to new domains, most of these approaches cannot be used for our class agnostic counting task due to the unavailability of labeled data from the novel test class.

3 Proposed Approach

We propose an exemplar-free class-agnostic visual counter called RepRPN-Counter. Given an image containing one or more repetitive object categories, RepRPN-Counter predicts a separate density map for each of the repetitive categories. The object count for the repetitive categories can be obtained by simply summing up the corresponding density map. For a category that is counted, RepRPN-Counter also provides the bounding box for an example from the category.

RepRPN-Counter consists of two key components: 1) a Repetitive Region Proposal Network (RepRPN) for identifying exemplars from repetitive objects in an image, along with their approximate count; and 2) a Density Prediction Network (DPN) that predicts a density map corresponding to any exemplar produced by the RepRPN.

For the rest of this section, we will describe the architecture of RepRPN in Sec. 3.1, the architecture of RepRPN-Counter in Sec. 3.2, the knowledge transfer

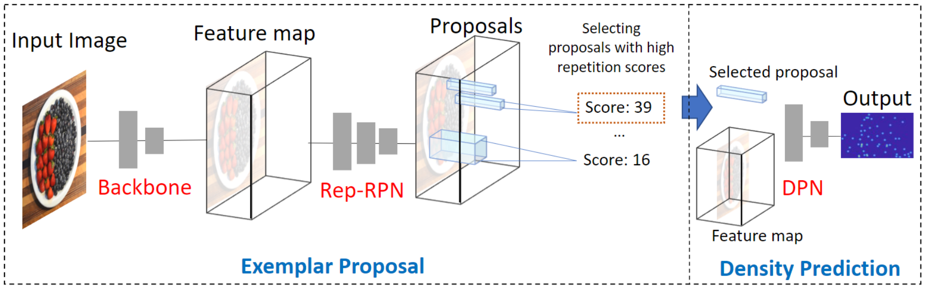


Fig. 2. RepRPN-Counter is a two-stage Exemplar Free Class Agnostic Counter. RepRPN-Counter has two key components: 1) Repetitive Region Proposal Network (RepRPN) and 2) Density Prediction Network (DPN). RepRPN predicts repetition score and objectness score for every proposal. Repetition score is used to select few proposals, called exemplars, from the repeating classes in the image. The DPN predicts a separate density map for any proposal selected by the RepRPN. The total count for any proposal, i.e. the number of times the object within the proposal occurs in the image, is obtained by summing all the values in the density map corresponding to that proposal. The DPN ignores proposals which are less likely to contain repetitive objects, so as to reduce the time required for training and evaluation. To keep things simple, we have shown the density prediction step for a single proposal. In reality, several density maps are predicted by the DPN, one for every selected proposal.

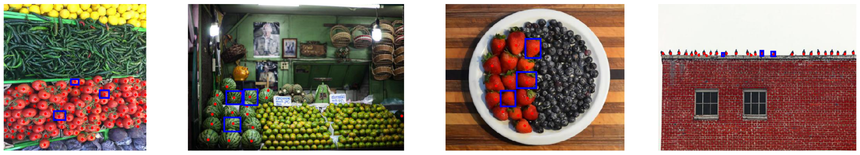


Fig. 3. Missing labels in the FSC-147 dataset. Each image in the dataset comes with bounding box annotations for the exemplar objects(shown in blue), and dot annotations for all objects belonging to the same category as the exemplar. For each image, objects of only a single class are annotated. We present a knowledge transfer strategy to deal with incomplete annotation.

approach for handling incomplete annotation in Sec. 3.3, and the overall training strategy in Sec. 3.4

3.1 Repetitive Region Proposal Networks

Repetitive Region Proposal Network (RepRPN) proposes exemplars from the repetitive object classes in an image. RepRPN takes as input convolutional feature representation of an image computed by the Resnet-50 backbone [10], and predicts proposal bounding boxes along with objectness and repetition scores for every proposal at every anchor location. The objectness score is the probability of the proposal belonging to any object class and not the background class. The repetition score refers to the number of times the object within the proposal occurs in the image. For example, consider an image with m cats and n oranges.

The RepRPN should predict m as the repetition score for any cat proposal, and n as the repetition score for any orange proposal. The repetition score is used to select exemplars from the repetitive classes in the image, i.e. the proposals with the highest repetition score are chosen as the exemplars. The original RPN formulation [33] uses a fixed window around an anchor location to predict the proposal boxes and objectness scores. However, this fixed sized window does not cover the entire image, and it does not contain sufficient information to predict the repetition score. This has been verified in our experiment where an RPN using only fixed-size window over convolutional features was unable to predict the repetition score. Predicting repetition score would require access to information from the entire image. To obtain this global information efficiently, we make use of the Encoder Self-Attention layers [40]. Given a feature vector at any location in the convolutional feature map, self-attention layers can pool information from *similar* vectors from the entire image, and can be used to estimate repetition score at any anchor location. To apply self-attention, we first transform the convolutional features into a sequence of length n: $S \in R^{n \times d}$. To preserve positional information, we concatenate appropriate $\frac{d}{2}$-dimensional row and column embeddings, resulting in d dimensional positional embeddings which are added with the sequence S. We refer to the resulting embeddings as $X \in R^{n \times d}$.

Given the sequence X, the self-attention layer first transforms X into query (X_Q), key (X_K), and value (X_V) matrices by multiplying X with matrices W_Q, W_K, and W_V:

$$X_Q = XW_Q, \quad X_K = XW_K, \quad X_V = XW_V. \tag{1}$$

The self-attention layer outputs a new sequence U where the i^{th} element in the output sequence is obtained as a weighted average of the value sequence, and the weights are decided based on the similarity between the i^{th} query element and the key sequence. The output sequence U is computed as follows:

$$U = softmax(X_Q X_K^T) X_V. \tag{2}$$

Tensor U will be reshaped into a tensor U' that has the same spatial dimensions as the input convolutional feature map. Tensor U' will be forwarded to the bounding box regression, objectness prediction, and repetition score prediction heads. Each prediction head consists of a single 1×1 convolutional layer. At each anchor location in the image, we consider k anchors boxes. For each anchor box, we predict an objectness score, a repetition score, and bounding box coordinates. The repetition score is used to identify proposals containing the repetitive objects in the image, i.e. proposals with a large repetition score contain repetitive objects.

3.2 RepRPN-Counter

As shown in Fig. 2, RepRPN-Counter consists of a Resnet-50 feature backbone, a RepRPN proposal network, and a Density Prediction Network (DPN). The

RepRPN and the DPN share the same feature backbone. The RepRPN provides the DPN with the bounding box locations of the proposals with large repetition scores, also called exemplars, and the DPN predicts a separate density map for each exemplar. DPN is trained and evaluated on only the chosen exemplars, and not all the proposals, so as to reduce the training and inference time. Similar to the previous works on class-agnostic counting [24, 30], DPN combines the convolutional features of an exemplar, with the convolutional features of the entire image to predict the density map for the exemplar. The exemplar features are obtained by performing ROI pooling on the convolutional features computed by the backbone, at the locations defined by the exemplar bounding boxes. The exemplar features are correlated with the image features, and the resulting correlation map is propagated through the DPN. The DPN is a fully convolutional network consisting of five convolutional layers and three upsampling layers (more architecture details are provided in the Supplementary submission), and the predicted density maps have the same spatial dimensions as the the input image. Note that the DPN predicts several density maps, one for each exemplar. The overall count for an object class pertaining to an exemplar is obtained by simply summing all the values in the density map corresponding to the exemplar. The DPN is not evaluated on the proposals with a low repetition score. For such proposals, the repetition score can be used as the final count.

3.3 Knowledge transfer for handling missing labels

The only existing dataset consisting of images of densely populated objects from many visual categories that can be used for training class agnostic visual counters is FSC-147 [30]. However, it is not trivial to train RepRPN-Counter on FSC-147 because of the missing labels in the dataset. FSC-147 comes with two types of annotations for each image: a few exemplar bounding boxes to specify the object category to be counted, and dot annotations for all of the objects belonging to the same category as the specified exemplars. However, an image may contain objects from another category that has not been annotated, as shown in Fig. 3. Given the missing labels, forcing RepRPN-Counter to predict zero count for the unannotated objects may degrade the performance of the counter.

We use knowledge transfer from teacher networks to address the incomplete annotation issue. We first train a RepRPN on the MSCOCO object detection dataset [19]. The MSCOCO training set consists of over 82K natural images from 80 visual categories, and the RPNs trained on this large dataset have been shown to generalize to previously unseen classes, thereby proving useful for tasks like zero-shot object detection [27]. We use the RepRPN trained on MSCOCO as a teacher network for generating the target labels for the objectness scores and the repetition scores for those proposals not intersecting with the annotated objects in the FSC-147 dataset. To get the target density maps corresponding to the unannotated proposals, we use the pretrained class-agnostic visual counter FamNet [30], which can predict the density map for a novel object class given only a single exemplar. When needed, an unannotated proposal is fed into FamNet, and the output of FamNet is used as the target density map for training the proposed network RepRPN-Counter.

3.4 Training Objective

RepRPN-Counter is trained in two stages. The first stage consists of training the RepRPN. Once trained, the RepRPN is kept frozen and used to generate exemplars for the density estimation network DPN. The second stage of training consists of training the DPN to predict the density map for every exemplar.

Training objective for RepRPN. For the i^{th} anchor box, the outputs of the RepRPN are the objectness score y_i, the bounding box coordinates b_i, and the repetition score c_i. Let the corresponding ground truth labels be y_i^*, b_i^*, c_i^*. We follow the same protocol as used in Faster RCNN [33] for obtaining the binary objectness label y_i^*, and the same parameterization for the bounding box coordinates b_i. c_i^* is the number of times the object within the anchor box, if any, occurs in the image. Since predicting c_i requires access to global information about the image, RepRPN makes use of self-attentional features as described in Sec. 3.1. The training loss for the i^{th} anchor box is:

$$\mathcal{L}_{RepRPN} = \lambda \mathcal{L}_{cls}(y_i, y_i^*) + \lambda \mathcal{L}_{reg}(b_i, b_i^*) + \mathcal{L}_{reg}(c_i, c_i^*), \qquad (3)$$

where $\mathcal{L}_{cls}$ is the binary cross entropy loss, and $\mathcal{L}_{reg}$ is the smooth L_1 loss. When training RepRPN on the FSC-147 dataset, the labels y_i^* and c_i^* for the positive anchors are obtained using the ground-truth annotation of FSC-147. Note that in the FSC-147 dataset, only three exemplars per image are annotated with bounding boxes, while the rest of the objects are annotated with a dot around their center. We obtain the bounding boxes for all the dot annotated objects by placing a bounding box of the average exemplar size around each of the dots. For anchors not intersecting with any of the annotated bounding boxes in FSC-147, y_i^* and c_i^* labels are obtained using a teacher RepRPN, which has been pre-trained on the MSCOCO dataset [19].

Training objective for DPN. Given an exemplar bounding box b_i and the feature map U for an input image I of size $H \times W$, the density prediction network DPN predicts a density map $Z_{b_i} = f(U, b_i)$ of size $H \times W$. The training objective for the DPN is based on the mean square error:

$$\mathcal{L}_{mse}(Z_{b_i}, Z^*) = \frac{1}{HW} \sum_{r=1}^{H} \sum_{c=1}^{W} (Z_{b_i}(r,c) - Z^*(r,c))^2,$$

where Z^* is the target density map corresponding to Z_{b_i}. If the exemplar b_i intersects with any annotated object, Z^* is obtained by convolving a Gaussian kernel with the corresponding dot annotation map. Note that Gaussian blurred dot annotation maps are commonly used for training density estimation based visual counters [49, 12, 28, 24, 21]. For cases where b_i does not intersect with any annotated object, we use the pretrained FamNet [30] as a teacher network for obtaining Z^*. The FamNet teacher can predict a density map, given an exemplar b_i and an input image I.

3.5 Implementation details

For training, we use Adam optimizer [14] with a learning rate of 10^{-5} and batch size of one. We use the first four convolutional blocks from the ImageNet pretrained ResNet-50 [10] as the backbone. We keep the backbone frozen during training, since finetuning the backbone would yield poor results. This is because the backbone has feature maps suitable for detecting a large number of classes, and finetuning the backbone leads to specialization towards FSC-147 training classes, resulting in poor performance on the novel test classes.

The weights of the RepRPN and DPN are initialized from a zero mean univariate Gaussian with standard deviation of 10^{-3}. RepRPN uses five self-attention transformer layers, each with eight heads. For training the RepRPN, we use four anchors sizes of $32, 64, 128, 256$ and three aspect ratios of $0.5, 1, 2$. We sample a batch of 96 anchors from each image during training. Training is done for 1000 epochs.

4 Experiments

4.1 Dataset

We perform experiments on the recently proposed FSC147 dataset [30], which was originally proposed for the exemplar based class-agnostic counting task. The FSC147 dataset consists of 6135 images from 147 visual categories, which are split into train, val, and test splits comprising of 89, 29, and 29 classes respectively. There are no common categories between the train, val, and test sets. The mean and maximum counts for images in the dataset are 56 and 3701, respectively. We train our model on the train set, and evaluate it on the test and val sets. Each image comes with annotations for a single object category of interest only, which consists of several exemplar bounding boxes and complete dot annotation for the objects of interest in the image. Since our goal is to build an exemplar free counter, unlike previous methods [24, 30], we do not use human annotated exemplars as an input to our counter.

4.2 Evaluation Metrics

We use the Top-k version of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to compare the performance of the different visual counters. MAE and RMSE are defined as follows. $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$; $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, where n is the number of test images, and y_i and $\hat{y}_i$ are the ground truth and predicted counts. MAE and RMSE are the most commonly used metrics for counting task [49, 25, 24, 30]. However, RepRPN-Counter predicts several density maps and corresponding counts, one for each selected proposal. Given k predicted counts from k proposals, we compute Top-k MAE and RMSE by first selecting those proposals from the top k proposals which have an IoU ratio of at least 0.3 with any ground truth boxes, and average the counts

Table 1. Comparing RepRPN-Counter to class-agnostic counters. FamNet, GMN and MAML are exemplar based class-agnostic counters which have been adapted and trained for the exemplar-free setting, where a RPN is used for generating exemplars. We report the Top-1, Top-3 and Top-5 MAE and RMSE metrics on the val and test sets of FSC-147 dataset. RepRPN-Counter consistently outperforms the competing approaches.

Method	MAE (Val set)			RMSE (Val set)			MAE (Test set)			RMSE (Test Set)		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
GMN	43.25	40.96	39.02	114.52	108.47	106.06	43.35	39.72	37.86	145.34	142.81	141.39
MAML	34.96	33.16	32.44	**98.83**	101.80	101.08	37.38	33.27	31.47	133.89	131.00	129.31
FamNet(pretrained)	47.66	42.85	39.52	125.54	121.59	116.08	50.89	42.70	39.38	150.52	146.08	143.51
FamNet	34.51	33.17	32.15	99.87	99.31	98.75	35.81	33.32	32.27	133.57	132.52	131.46
RepRPN-Counter	**31.69**	**30.40**	**29.24**	100.31	**98.73**	**98.11**	**28.32**	**27.45**	**26.66**	**128.76**	**129.69**	**129.11**

corresponding to the selected proposals to get the predicted count $\hat{y}_i$. In case none of the k proposals intersect with any ground truth boxes, we simply average all of the k counts to get $\hat{y}_i$.

4.3 Comparison with class-agnostic visual counters

We compare our proposed RepRPN-Counter with the previous class-agnostic counting methods [24, 30, 8] on the task of counting objects from novel classes. We do not compare with class-specific counters [49, 25] because such counters cannot handle novel classes at test time. Furthermore, these counters require hundreds [49] or thousands [44, 38] of images per category during training, while FSC-147 dataset contains an average of only 41 images per category.

GMN [24], FamNet [30], and MAML [8, 30] are exemplar based counters which can predict density map for any unseen object category based on few exemplars of the object category from the same image. These counters were originally proposed to work with human provided exemplars as an input to the counter. In order to make these exemplar based counters work in our exemplar free setup, we modify GMN, FamNet, and MAML based visual counters by replacing human provided exemplars with RPN [33] generated exemplars. We use the RPN of Faster RCNN [33] to generate the proposals for the competing approaches, and use the top k proposals with the highest objectness score as the exemplars. For fair comparison, both the RPNs used with the competing approaches as well as the RepRPN are pre-trained on the MSCOCO dataset [19]. We do not use MSCOCO to train the DPN. We train the competing approaches and our proposed approach on the train set of FSC-147, and report the results on the val and test sets of FSC-147. We also compare our method with a pre-trained version of FamNet originally trained on the few-shot counting task. Following [30], all of the methods are trained with three proposals, and evaluated with 1, 3, and 5 proposals. We report the Top-1, Top-3 and Top-5 MAE and RMSE values in Table 1.

As can be seen from Table 1, our method RepRPN-Counter outperforms all of the competing methods. The pre-trained FamNet performs the worst, even

Table 2. Comparing RepRPN-Counter with pre-trained object detectors, on Val-COCO and Test-COCO subsets of FSC-147, which only contain COCO classes. Pre-trained object detectors are available for these COCO classes. For RepRPN-Counter, we use the density map corresponding to the proposal with the highest repetition score. Without access to any labeled data from these COCO classes, our proposed approach outperforms all of the object detectors which are trained using the entire COCO train set containing a large number of images from these COCO classes

	Val-COCO Set		Test-COCO Set	
Method	MAE	RMSE	MAE	RMSE
Faster R-CNN	52.79	172.46	36.20	79.59
RetinaNet	63.57	174.36	52.67	85.86
Mask R-CNN	52.51	172.21	35.56	80.00
Detr	58.35	175.97	45.51	96.57
RepRPN-Counter (Ours)	**50.72**	**160.95**	**25.29**	**56.98**

though it was trained on the same FSC-147 training set. This shows that simply combining pre-trained exemplar based class agnostic counters with RPN-based exemplars does not provide a reasonable solution for the exemplar-free setting. When retrained specifically for the exemplar-free setting, the performance of FamNet significantly improves when compared to its pre-trained version. GMN performs worse than the other baselines, possibly due to the need for more examples for the adaptation process. This observation was earlier reported for exemplar based class-agnostic counting task as well [30].

4.4 Comparison with object detectors

One approach to counting is to use a detector and count the number of detections in an image. However, it requires thousands of examples to train an object detector, and the detector-based counters cannot be used for novel object classes. That being said, we compare RepRPN-Counter with object detectors on a subset of COCO categories from the validation and test sets of FSC-147. These subsets are called Val-COCO and Test-COCO, containing 277 and 282 images respectively. We compare our approach with the official implementations [46] of MaskRCNN [9], FasterRCNN [34], RetinaNet [18], and Detr [7]. The results are shown in Table 2. Without any access to labeled data from these COCO classes, our proposed method still outperforms the object detectors that have been trained using the entire COCO train set containing thousands of images from these COCO classes. Detr [7] performs worse than some of the earlier object detectors because Detr uses a fixed number of query slots (usually 100), which limits the maximum number of objects it can detect, while FSC-147 has images containing thousands of objects.

Table 3. Comparing RepRPN with RPN, on the test set of FSC-147. Using RepRPN instead of RPN leads to significant boost in performance

	RPN		RepRPN	
Method	MAE	RMSE	MAE	RMSE
GMN	43.35	145.34	32.17	137.29
MAML	37.38	133.89	32.09	141.03
FamNet (pre-trained)	50.89	150.52	38.64	144.27
FamNet	35.81	133.57	32.94	132.82

4.5 Comparing RepRPN with RPN

We are also interested in checking if RepRPN can boost the performance of class-agnostic visual counters other than RepRPN-Counter. For this experiment, we replace RPN [33] with RepRPN for GMN, FamNet, and MAML, and report the Top-1 MAE and RMSE scores on the FSC-147 test set in Table 3. Using RepRPN instead of RPN leads to significant boost in the performance for all class-agnostic visual counters. This suggests that RepRPN is much better suited for the exemplar proposal for exemplar free counting task in comparison to RPN. Also, RepRPN works well with different types of class agnostic counters, including the proposed RepRPN-Counter, GMN, FamNet, and MAML.

4.6 Ablation Studies

Our proposed RepRPN-Counter consists of two primary components: the RepRPN for exemplar proposal and the DPN for density prediction. Furthermore, our proposed knowledge transfer approach allows us to deal with unannotated objects in the FSC-147 dataset. In Table 4, we analyze the contribution of these components on the overall performance. The RepRPN baseline uses the repetition score as the final count. We propose to use RepRPN with DPN, but one can replace RepRPN by RPN [33] to get the method RPN+DPN. One can assume there are no unannotated objects in the FSC-147 dataset, and train our proposed RepRPN-Counter on FSC-147 without any knowledge transfer. As can be seen from Table 4, all the components of RepRPN-Counter are useful, and the best results are obtained when all the components are present. RPN+DPN performs much worse than RepRPN+DPN, which shows that RepRPN is better suited for our counting task than RPN.

4.7 Qualitative Results

In Fig. 4, we present a few input images, the proposal with the highest repetition score generated by RepRPN for each image, and the corresponding density map generated by the density prediction network. RepRPN-Counter performs well on the first three test cases. But it fails on the last one, because the aspect ratio of the chosen proposal is very different from the majority of the objects of interest.

Table 4. Analyzing individual components of RepRPN-Counter on the overall performance on the test set of FSC-147. RPN + DPN refers to the case where we replace RepRPN from our proposed approach with the RPN from Faster RCNN. As can be seen, RepRPN is a critical component of our proposed approach, and replacing it with RPN decreases the performance significantly. RepRPN+DPN-NoKT refers to the method when we do not use any knowledge transfer, which leads to a drop in performance. This shows the usefulness of the proposed knowledge transfer strategy.

Method	MAE			RMSE		
	Top1	Top3	Top5	Top1	Top3	Top5
RepRPN+DPN (proposed)	28.32	27.45	26.66	128.76	129.69	129.11
RepRPN+DPN-NoKT (no knowledge transfer)	29.52	28.80	28.42	132.76	131.03	130.82
RPN+DPN	35.81	33.32	32.27	133.57	132.52	131.46
RepRPN (without DPN)	29.60	29.18	28.95	136.25	136.21	136.26

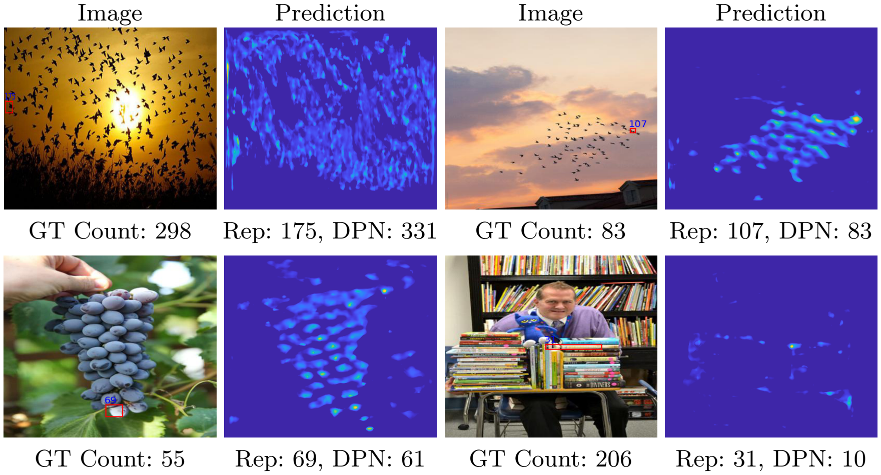


Image | Prediction | Image | Prediction

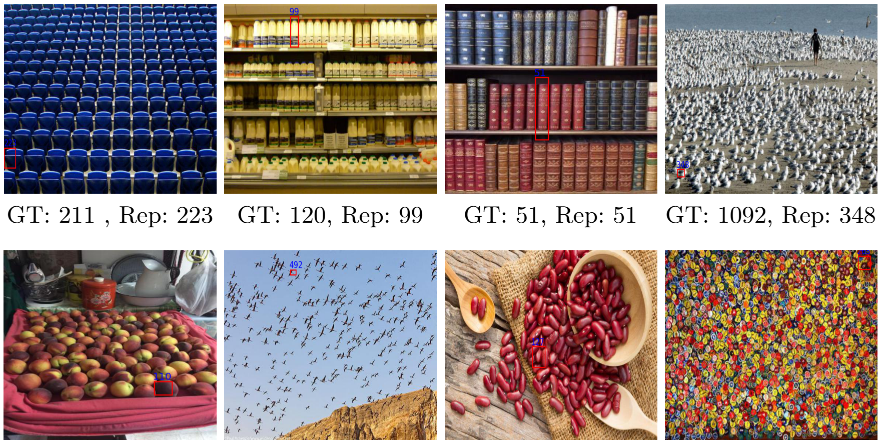
GT Count: 298 | Rep: 175, DPN: 331 | GT Count: 83 | Rep: 107, DPN: 83

GT Count: 55 | Rep: 69, DPN: 61 | GT Count: 206 | Rep: 31, DPN: 10

Fig. 4. Input images and the density maps predicted by RepRPN-Counter. Also shown in red are the selected proposals. Rep is the repetition score predicted by RepRPN, while DPN is the count obtained by summing the final density map.

In Fig. 5, we show the RepRPN proposal with the highest repetition score for several images from the Val and Test set of FSC-147. First three examples in each row contain test cases where the repetition score is close to the groundtruth count. The last example in each row shows test case which proved to be harder for RepRPN. RepRPN does not perform well in some cases when the objects are extremely small in size. Since RepRPN, and RPN in general, uses a fixed set of anchor sizes and aspect ratios, they may fail at detecting extremely small objects. It is also difficult for RepRPN to handle extreme variation in scale within the image, as evident from the failure cases.

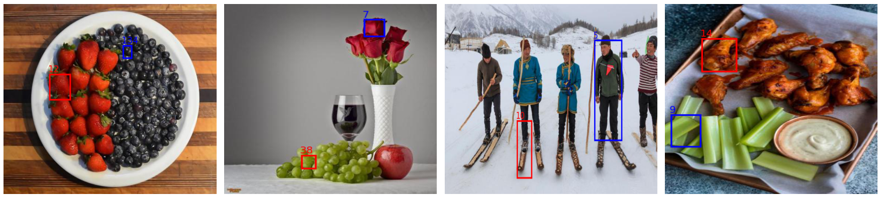
RepRPN-Counter can be used for multi-class Class Agnostic counting task, i.e. counting multiple object classes in an image. In Fig. 6, we show few images from the Val and Test set of FSC-147 having at least two object classes, and the



GT: 211 , Rep: 223 GT: 120, Rep: 99 GT: 51, Rep: 51 GT: 1092, Rep: 348

GT: 108, Rep: 110 GT: 402, Rep: 492 GT: 161, Rep: 127 GT: 1228, Rep: 263

Fig. 5. Selected proposal (shown in red) and corresponding repetition score (Rep) predicted by RepRPN. The first three examples in each row are success cases for RepRPN, and the predicted repetition score is close to the ground truth count. The last example in each row shows a failure case.



C_{red}: 12, C_{Blue}: 134 C_{red}: 38, C_{Blue}: 7 C_{red}: 10, C_{Blue}: 5 C_{red}: 14, C_{Blue}: 9

Fig. 6. Counting multiple classes in an image using RepRPN-Counter. Shown are RepRPN proposals from two of the most frequent classes in an image, and the corresponding counts predicted by RepRPN-Counter. C_{red}/C_{blue} is the predicted count for the proposal shown in red/blue.

counts predicted by RepRPN-Counter for the two most frequent classes in the image. RepRPN-Counter provides a reasonable count estimate for both classes.

5 Conclusions

In this paper, we tackled the task of Exemplar Free Class Agnostic Counting. We proposed RepRPN-Counter, the first exemplar free class agnostic counter capable of handling previously unseen categories at test time. Our two-stage counter consists of a novel region proposal network for finding the exemplars from repetitive object classes, and a density estimation network to estimate the density map corresponding to each exemplar. We also showed that our region proposal network can significantly improve the performance of the previous state-of-the-art class-agnostic visual counters.

References

1. Abousamra, S., Hoai, M., Samaras, D., Chen, C.: Localization in the crowd with topological constraints. In: AAAI (2021) 3
2. Arteta, C., Lempitsky, V., Noble, J.A., Zisserman, A.: Detecting overlapping instances in microscopy images using extremal region trees. Medical image analysis **27**, 3–16 (2016) 1
3. Arteta, C., Lempitsky, V., Zisserman, A.: Counting in the wild. In: ECCV (2016) 1
4. Babu Sam, D., Sajjan, N.N., Venkatesh Babu, R., Srinivasan, M.: Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In: CVPR (2018) 3
5. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 384–400 (2018) 4
6. Cao, X., Wang, Z., Zhao, Y., Su, F.: Scale aggregation network for accurate and efficient crowd counting. In: ECCV (2018) 3
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 11
8. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks (2017) 4, 10
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017) 11
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 5, 9
11. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: CVPR (2013) 1
12. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: ECCV (2018) 1, 8
13. Khan, A., Gould, S., Salzmann, M.: Deep convolutional neural networks for human embryonic cell counting. In: ECCV. Springer (2016) 1
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
15. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop (2015) 4
16. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015) 4
17. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: CVPR (2018) 3
18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017) 11
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 3, 7, 8, 10
20. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: CVPR (2019) 3
21. Liu, X., van de Weijer, J., Bagdanov, A.D.: Leveraging unlabeled data for crowd counting by learning to rank. In: CVPR (2018) 8

22. Liu, X., Van De Weijer, J., Bagdanov, A.D.: Leveraging unlabeled data for crowd counting by learning to rank. In: CVPR (2018) 3
23. Liu, Y., Liu, L., Wang, P., Zhang, P., Lei, Y.: Semi-supervised crowd counting via self-training on surrogate tasks. In: European Conference on Computer Vision. pp. 242–259. Springer (2020) 3
24. Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: ACCV (2018) 1, 2, 3, 4, 7, 8, 9, 10
25. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: ICCV (2019) 1, 3, 9, 10
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: ECCV (2016) 1
27. Rahman, S., Khan, S., Porikli, F.: Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In: Asian Conference on Computer Vision. pp. 547–563. Springer (2018) 4, 7
28. Ranjan, V., Le, H., Hoai, M.: Iterative crowd counting. In: ECCV (2018) 3, 8
29. Ranjan, V., Shah, M., Nguyen, M.H.: Crowd transformer network. arXiv preprint arXiv:1904.02774 (2019) 3
30. 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) 1, 2, 3, 4, 7, 8, 9, 10, 11
31. Ranjan, V., Wang, B., Shah, M., Hoai, M.: Uncertainty estimation and sample selection for crowd counting. In: ACCV (2020) 3
32. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016) 4
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015) 2, 3, 6, 8, 10, 12
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS (2015) 11
35. Sam, D.B., Surya, S., Babu, R.V.: Switching convolutional neural network for crowd counting. In: CVPR (2017) 3
36. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: One-shot learning with memory-augmented neural networks (2016) 4
37. Shi, M., Yang, Z., Xu, C., Chen, Q.: Revisiting perspective information for efficient crowd counting. In: CVPR (2019) 3
38. Sindagi, V.A., Yasarla, R., Patel, V.M.: Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. arXiv preprint arXiv:2004.03597 (2020) 1, 10
39. Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Wu, Y.: Rethinking counting and localization in crowds: A purely point-based framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3365–3374 (2021) 3
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, **L**., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017) 6
41. Wan, J., Chan, A.: Adaptive density map generation for crowd counting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1130–1139 (2019) 3
42. Wan, J., Liu, Z., Chan, A.B.: A generalized loss function for crowd counting and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1974–1983 (2021) 3

43. Wang, C., Song, Q., Zhang, B., Wang, Y., Tai, Y., Hu, X., Wang, C., Li, J., Ma, J., Wu, Y.: Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3234–3242 (2021) 3
44. Wang, Q., Gao, J., Lin, W., Li, X.: Nwpu-crowd: A large-scale benchmark for crowd counting. arXiv preprint arXiv:2001.03360 (2020) 1, 3, 10
45. Wang, Q., Gao, J., Lin, W., Yuan, Y.: Learning from synthetic data for crowd counting in the wild. In: CVPR (2019) 3
46. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2 (2019) 11
47. Xie, W., Noble, J.A., Zisserman, A.: Microscopy cell counting and detection with fully convolutional regression networks. Computer methods in biomechanics and biomedical engineering: Imaging & Visualization **6**(3), 283–292 (2018) 1
48. Zhang, A., Yue, L., Shen, J., Zhu, F., Zhen, X., Cao, X., Shao, L.: Attentional neural fields for crowd counting. In: ICCV (2019) 3
49. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: CVPR (2016) 1, 3, 8, 9, 10
50. Zhu, P., Wang, H., Saligrama, V.: Zero shot detection. IEEE Transactions on Circuits and Systems for Video Technology **30**(4), 998–1010 (2019) 4