Class-Agnostic Object Counting Robust to Intraclass Diversity

Shenjian Gong[1], Shanshan Zhang*[,1], Jian Yang[1], Dengxin Dai[2], and Bernt Schiele[2]

[1] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology {shenjiangong,shanshan.zhang,csjyang}@njust.edu.cn
[2] MPI Informatics {ddai,schiele}@mpi-inf.mpg.de

Abstract. Most previous works on object counting are limited to pre-defined categories. In this paper, we focus on class-agnostic counting, i.e., counting object instances in an image by simply specifying a few exemplar boxes of interest. We start with an analysis on intraclass diversity and point out three factors: color, shape and scale diversity seriously hurts counting performance. Motivated by this analysis, we propose a new counter robust to high intraclass diversity, for which we propose two effective modules: Exemplar Feature Augmentation (EFA) and Edge Matching (EM). Aiming to handle diversity from all aspects, EFA generates a large variety of exemplars in the feature space based on the provided exemplars. Additionally, the edge matching branch focuses on the more reliable cue of shape, making our counter more robust to color variations. Experimental results on standard benchmarks show that our Robust Class-Agnostic Counter (RCAC) achieves state-of-the-art performance. The code is publicly available at https://github.com/Yankeegsj/RCAC.

Keywords: object counting, few-shot learning

1 Introduction

Object counting, i.e., estimating the number of object instances of a certain category in a given image, has a wide range of applications such as video surveillance and agriculture. However, most methods in previous works can only count pre-defined categories, such as people [16,13], animals [3], plants [18,22] and cars [17]. For most existing works, each model is typically trained for one category with a large amount of labeled data. They have two notable limitations. On one hand, we need to train multiple models if we are required to count objects of various categories, which is computationally expensive and inconvenient. On the other hand, such models cannot be adapted to unseen categories at test time. But in practice, it is desirable to develop counting methods that are more general and flexible, which are extendable to any arbitrary new category at test time.

*Corresponding author

To this end, class-agnostic object counting is more suited for real applications and has been investigated recently. Interactive Object Counting (IOC) [2] addresses the counting task with human interaction. The user is asked to annotate a small number of objects with dots and the algorithm learns a codebook and partitions all pixels into object and background groups. This process is repeated until the results are satisfactory. In contrast, some more recent works [15,19] formulate counting as a matching problem, turning out to be more effective and efficient. Generic Matching Network (GMN) [15] learns the matching function from concatenation of query image and exemplar box features to a similarity heatmap. When adapting the model to a novel category, only a fraction of parameters need to be optimized. Few-shot adaptation & matching Network (FamNet) [19] computes the correlation maps between exemplar box and image features and then predicts the density map.

However, the current best performance is still far from satisfactory. For example, the average ground truth count on the FSC-147 validation set is 63.54, while the mean average error (MAE) of the current top method FamNet [19] is as high as 24. In order to understand the limitations of current methods, we analyze failure cases and find that objects of interest in the same image may differ in color, shape and scale, which largely hinders counting performance. A detailed analysis can be found in Sec. 3. It has been shown by FamNet [19] that it is helpful to provide more diverse exemplar boxes. Yet the exemplar boxes are provided by annotators subjectively and thus the diversity cannot be guaranteed; also, the number of provided exemplar boxes is limited, potentially not covering all instances. To address this problem, in this paper, we aim to develop a new counting method, which is more robust to intraclass diversity. Specifically, we propose two effective modules. One the one hand, we apply exemplar augmentation in the feature space to handle high diversity in various aspects. One the other hand, we introduce an additional matching branch that uses edge features to deal with diversity in color.

To summarize, the main contributions of our work are as follows: (1) We analyze the top-performing class-agnostic counting method FamNet [19], showing that intra-class diversity is a key factor decreasing counting performance, and point out the diversity comes from three aspects: color, shape and scale. (2) Two modules are proposed to overcome the high diversity challenge. The exemplar feature augmentation module increases the exemplar diversity so as to achieve more effective matching with a wide range of instances. Moreover, the additional matching branch using edge features focuses on the more reliable cue of shape, down-weighing some less reliable cues, including background and object colors. (3) Experimental results on two related datasets show that our method achieves state-of-the-art results for class-agnostic counting, outperforming previous methods by a large margin; also, since no test time adaptation is employed, our method is more convenient to apply.

2 Related Work

In this section, we first briefly review recent works on class-aware object counting and then focus on class-agnostic object counting methods.

Class-Aware Object Counting. Most object counting methods are limited to pre-defined categories, e.g., people, animals and cars. Generally, they can be divided into two groups. One of them is detection based counting [4,9,12]. Each of them applies an object detector on the given image, and then counts the number of bounding boxes. However, it is hard to choose a proper threshold for the detection confidence to select out reasonable boxes; and object detectors usually perform poorly at crowds. The other group is regression based counting [19,22,16,13,5,6]. These methods estimate a density map for each image, and counting is achieved by summing up the pixel values. For both kinds of methods, box or point annotations for all persons are required at training time, which are rather expensive. Class-aware object counters perform well on trained categories but they cannot be adapted to a new category at test time. Also, it is expensive to obtain rich training annotations.

Class-Agnostic Object Counting. Similar to class-aware object counting, a straightforward way for class-agnostic object counting is to apply a few-shot object detector [10,7,11] on the given image. But the major disadvantage is that it is tricky to choose a proper detection score threshold for counting; also, the detectors usually fail at crowded scenes. In contrast, regression based methods are cleaner and expected to achieve higher performance.

Some early regression based works perform pixel-wise classification. For example, IOC [2] learns a codebook from a few dot annotations marked by the user, so as to distinguish object and background pixels. Few-Shot Sequential Approach (FSSA) [24] uses the extracted prototype features to classify each pixel as one of the object classes present in the support set or as background. More recently, counting is formulated as a matching problem, which becomes more effective and efficient. GMN [15] proposed a class-agnostic counting approach consisting of three modules, namely embedding module, matching module and adaptation module. The exemplar box and query image features extracted from the embedding module are concatenated and fed to the matching module to predict a similarity heatmap. The adaption module is used to adapt to a new domain and is the only module needs to be updated for adaptation. FamNet [19] and Class-agnostic Few-shot Object Counting Network (CFOCNet) [25] are most related to our work. They both take correlation matching maps between the exemplar box and query images and then predict the density maps based on them. FamNet performs additional fine-tuning at test time. Model Agnostic Meta Learning (MAML) [8] based few-shot approaches also fine-tune some parameters to make the model better adapt to novel classes. In this paper, we also employ correlation maps for matching. The major difference is that, we propose new modules against high diversity aiming for more effective matching.

3 Analysing Intraclass Diversity for Counting

In this section, we aim to analyse the impact of intraclass diversity on counting performance.

We choose the method of FamNet* as our baseline, which is a simpler version of FamNet [19] without test-time adaptation. It has been shown [19] that test-time adaptation only brings minimal improvements and thus we do not consider it here. The

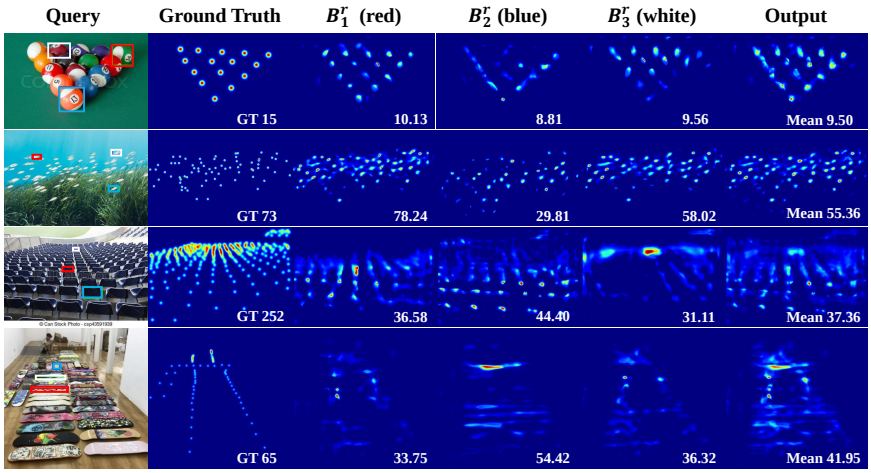

Fig. 1. Some Failure cases of FamNet* [19] from the FSC-147 dataset. At each row, from left to right, We show each query image, its ground truth density map, estimated density map given each exemplar box and the final average density map. $B^{r,1}$, $B^{r,2}$, $B^{r,3}$ are shown in red, blue and white bounding boxes respectively. The numbers indicate the ground truth counting number or estimated counting results. Colorful for better visualization.

pipeline of FamNet* is as follows (shown as the black arrows in Fig. 2): the query image is fed to the backbone network (ResNet-50) for feature extraction, which is trained on ImageNet and not updated during training; multi-scale features for each exemplar box are obtained by performing ROI pooling on the feature maps from the third and fourth ResNet-50 blocks; the query image features also come from the third and fourth blocks; correlation maps are calculated by taking each exemplar box feature as a convolution kernel, which is applied to the query image feature maps; the density map is then predicted by a shallow subnet consisting of 5 conv layers using the correlation matching maps as input.

We start with analyzing failure cases for the FamNet* we trained on the FSC-147. We pick those samples with relative errors higher than 20% and do visual inspection. The relative error is calculated as absolute prediction error divided by the ground truth count. By observing the above samples, we find three typical factors that affect the performance: high diversity w.r.t. color, scale and shape. In Fig. 1 we show some failure cases from the FSC-147 dataset. Each image is provided with three exemplar boxes (B_1^r, B_2^r, B_3^r), each generates a density map and the final output density map is obtained by averaging the above three. The counting number (shown at the right bottom) of each density map is calculated by summing up all pixel values on it.

The color diversity comes from two aspects: the foreground objects and the background. For the query image in row 1, there is a high color difference among the object instances. Although the provided three exemplar boxes are of different colors, they still fail to cover all colors of different objects. Similarly, in row 2, we can also see color

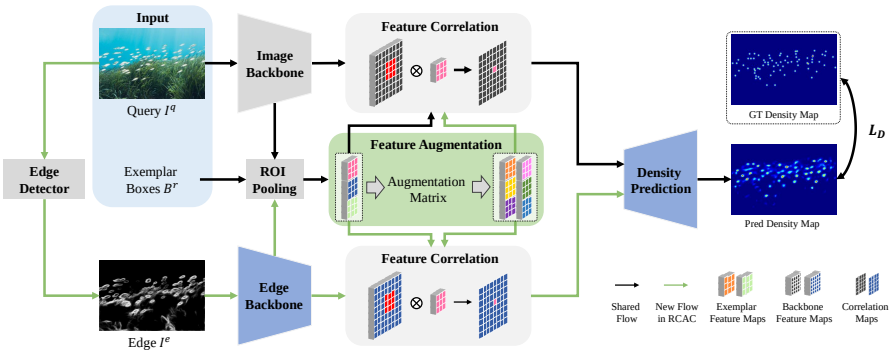

Fig. 2. Pipeline of our proposed method – Robust Class-Agnostic Counter (RCAC). Given an input query image I_q along with several exemplar boxes B^r, we first apply an edge detector to obtain a gray-scale edge image, and then we have a two-stream scheme using the RGB and edge images for matching in parallel. Specifically, the RGB and edge images are fed into separate backbone networks for feature extraction; exemplar box features are cropped from the full feature maps via ROI pooling, which are augmented via our proposed feature augmentation module; after that, feature correlation layer takes feature maps of each exemplar feature as convolution kernel to calculate the correlation map on the full feature maps; correlation maps come from the same exemplar goes through the density prediction module to generate one density map, and the final predicted density map is obtained by averaging all the density maps. Need to note that only the edge backbone and density prediction module represented with blue trapezoid are optimized during training. The black arrows indicate the shared flows between ours and the baseline method.

difference inside object boxes. The fishes are all white but their background regions differ in color. The high diversity w.r.t. color results in large counting errors.

For the query image in row 3, the chairs distribute from the near to the distant, showing large variance in scale. Although the provided three exemplar boxes are of different scales, they are not able to cover the full scale range of all instances. This challenging scenario makes the predicted counting number (37.36) become much lower than the ground truth (252).

For the query image in row 4, the skateboards show different shapes caused by different orientations. We can see many of them are put horizontally, but some are put vertically, e.g. those ones close to the blue box. Also, they are of different scales from the near to the distant. Moreover, the color varies a lot across different instances. This example is representative that different factors may happen at the same time, leading to very challenging scenarios.

We further provide some quantitative analysis regarding the impact of diversity. First, from the full validation set we select three subsets with high diversity w.r.t. color, scale and shape respectively in the following way. For each image, we compute the variance values w.r.t. color (represented by hue), scale (represented by area) and shape (represented by aspect ratio) based on the provided exemplar boxes. Then we select top hundreds of images with highest scale variance as the scale diversity subset, similar for

color and shape. A comparison of results of FamNet* on the full validation set and three diverse subsets is shown in Tab. 1. We can find that compared to the full validation set, diversity w.r.t. color, scale and shape all bring a significant performance drop. Especially for the subsets with high color and scale diversity, the performance drops by ~ 20 pp w.r.t. MAE.

The above analysis indicates that counting performance is highly affected by the diversity of object instances. Therefore, we are aiming to develop a new counting method robust to high intraclass diversity. Qualitative results of our method on high intraclass diversity images is shown in Fig. 4.

4 Our Robust Class-Agnostic Counter

In this section, we first introduce the setting of few-shot counting. After that, we provide the pipeline of our method, followed by detailed description of two new modules: exemplar feature augmentation and edge matching.

4.1 Problem Formulation

We follow the few-shot setting from our baseline method FamNet* [19]: given a query image ($I^q \in \mathbb{R}^{3\times H\times W}$) and K exemplar bounding boxes ($B^r \in \mathbb{R}^{K\times 4}$) that locate the reference instances belonging to the same category, the task is to predict the density map $\widetilde{Y}$ of the query image and the counting number is calculated by summing up all pixel values of $\widetilde{Y}$.

4.2 Pipeline

The overall pipeline is depicted in Fig. 2. We have a two-stream feature extraction, obtaining feature maps for each input RGB query image and its gray-scale edge image. For each stream, the exemplar features are cropped from the full feature maps via ROI pooling and then augmented via our proposed exemplar feature augmentation module. After that, correlation maps are calculated by taking each exemplar box features as a convolution kernel, which is applied to the entire query image feature maps. Then the correlation maps generated by the same exemplar box from two streams are concatenated and sent to the density prediction module, which outputs one density map for each exemplar. The final density map is obtained by averaging all density maps and the counting number is calculated by summing up all pixel values.

Feature extraction. For the RGB image (I^q) stream, we use the ImageNet pre-trained ResNet-50 backbone, obtaining the query feature maps F^q. Please note we take the output of two layers as feature maps (3rd and 4th blocks of ResNet-50), denoted as $F^q_i,\ i \in \{1, 2\}$, and the network is frozen during training. For the edge image (I^e) stream, we use a light version of VGG net, which is initialized randomly and optimized during training.

ROI pooling. This operation (ROI) crops the exemplar feature maps based on exemplar boxes B^r. The feature maps of the k-th exemplar box B^r_k are obtained as follows:

$$F^{q,r}_{k,i} = ROI(F^q_i, B^r_k),\ i \in \{1, 2\}\,. \tag{1}$$

Each exemplar feature map is first resized to the same size based on the maximal exemplar box. And then we resize each exemplar feature map by 0.9 and 1.1 to obtain multi-scale features, following FamNet*. In this way, we obtain multi-level and multi-scale feature maps for each exemplar:

$$\begin{aligned} F_{k,i,s}^{q,r} &= Resize(F_{k,i}^{q,r}, s), \\ i \in \{1,2\},\ s &\in \{0.9, 1.0, 1.1\}. \end{aligned} \tag{2}$$

Feature correlation. The correlation maps $M = \{M_1^q, M_2^q, ...M_K^q\}$ are obtained by convolving the query image and each exemplar box and are used for density prediction. The process is denoted as:

$$\begin{aligned} M_{k,i,s}^q = Conv(F_i^q, kernel), kernel = F_{k,i,s}^r, \\ i \in \{1,2\},\ s \in \{0.9, 1.0, 1.1\}, \end{aligned} \tag{3}$$

where $Conv$ denotes the convolution operation that correlate the exemplar features with the query features to obtain multiple correlation maps. After convolution, for each exemplar, we append the obtained 6 correlation maps (2×3: two-level (output of 3rd and 4th blocks of ResNet-50) and three-scale (0.9, 1.0, 1.1) features) to M_k^q for density prediction.

Density prediction. For the k-th exemplar, given M_k^q from the previous step, the density prediction module ($\mathcal{D}$) predicts a relevant density map. The final density map is obtained by averaging K density maps.

$$\widetilde{Y} = Mean(\mathcal{D}(M_1^q), \mathcal{D}(M_2^q), ..., \mathcal{D}(M_K^q)) \tag{4}$$

Optimization. Our objective is to minimize the difference between Y and $\widetilde{Y}$:

$$L_D = \sum \left\| Y - \widetilde{Y} \right\|^2. \tag{5}$$

4.3 Exemplar Feature Augmentation (EFA)

To obtain more exemplars for robust prediction, we propose to apply exemplar feature augmentation to generate other latent exemplar features. To be specific, given $B^r \in \mathbb{R}^{K\times 4}$, we compute a weighted sum of these K features with a weight vector $\vec{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)$, $\sum_{i=1}^{K} \alpha_i = 1$. Besides the original K exemplar features, we generate additional N augmented features using an augmentation matrix $\nu \in \mathbb{R}^{N\times K}$, consisting of N different weight vectors. The nth augmentation feature is denoted as:

$$\begin{aligned} F_{K+n,i}^{q,r} &= \sum_k \nu(n,k) F_{k,i}^{q,r}. \\ i \in \{1,2\}, k \in \{1,2,...,K\}, n &\in \{1,2,...,N\}. \end{aligned} \tag{6}$$

For instance, when the weight vector is equal to $(1, 0, ..., 0)$, the augmented feature is the same as $F_{1,i}^{q,r}$. Please note that we sample $\vec{\alpha}$ with a multinomial dirichlet distribution.

In this way, we obtain a larger set of density maps, and the final density map can be formulated as:

$$\widetilde{Y} = Mean(\mathcal{D}(M_1^q), \mathcal{D}(M_2^q), ..., \mathcal{D}(M_{K+N}^q)). \tag{7}$$

Imagine that we want to count objects of various colors, but only three samples are given. EFA is like creating new samples of different colors in the feature space via combining the provided exemplars. In this way, objects with various colors can be better matched; and similarly, the intraclass diversity w.r.t. shape and scale can be also handled.

Dirichlet distribution. In machine learning, one common distribution called Beta distribution is denoted as:

$$Beta(\alpha \mid \theta_1, \theta_2) = \frac{\Gamma(\theta_1)\Gamma(\theta_2)}{\Gamma(\theta_1 + \theta_2)} \alpha^{\theta_1 - 1}(1-\alpha)^{\theta_2 - 1}, \tag{8}$$

where Γ represents Gamma function. Dirichlet distribution generalizes the Beta distribution to a multinomial distribution. It is expressed as:

$$\begin{aligned} &Dirichlet(\{\alpha_1,\alpha_2, ..., \alpha_K\} \mid \{\theta_1, \theta_2, ..., \theta_K\}) \\ &= \frac{\prod_{i=1}^{K} \Gamma(\theta_i)}{\Gamma(\sum_{i=1}^{K} \theta_i)} \prod_{i=1}^{K} \alpha_i^{\theta_i - 1}, \end{aligned} \tag{9}$$

where $\sum_{i=1}^{K} \alpha_i = 1$ and $\alpha_i \geq 0$. We choose the multinomial Dirichelet distribution as it meets the following requirements: (1) The sum of all weights equals to 1, so that the value level of the augmented features and the original features remain unchanged. (2) The number of exemplars can vary. (3) The diversity of sampled weights is high. Fig. 3 shows the sampling probability of dirichlet distribution with different θ. First of all, we treat K exemplars equally, therefore, the parameters of the distribution satisfy the condition that θ_i are equal. In addition, the original sample occupies three vertices of the triangle shown in Fig. 3. Expect not to generate features similar to the original, we adopt the sampling distribution with the maximum sampling probability for the average fusion of the K exemplars. Meanwhile, in order to make the sampling area large, the center sampling probability should not be too large. Based on the above considerations, we adopt the dirichlet distribution with $\theta_i = 2$.

4.4 Edge Matching (EM)

Different object instances may differ in color, while shape is a more reliable cue across instances, leading to more robust counting. On the other hand, edge is a kind of class-agnostic knowledge, which will not bring category bias. To allow our model more focus on the shape cue, we introduce an additional stream for matching, where edge features are used instead of RGB features.

The gray-scale edge image we use in this paper is generated by the RCF model [14] trained on the BSDS500 dataset [1]. We obtain one edge image for each RGB image. For instance, in Fig. 2, I^e is predicted from I^q with the trained RCF model.

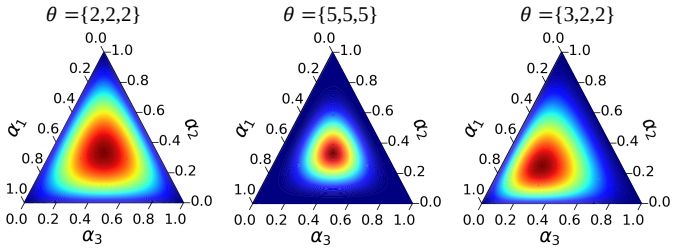

Fig. 3. Dirichlet sampling distributions with different configuration parameters.

The structure of the edge stream is the same as the RGB stream. The only difference is that we use a shallower network for edge feature extraction. Since the gray-scale edge image is much lighter than the RGB image, we employ a VGG-like net [23] with a smaller number of channels as the edge backbone, and update it during training.

In the same way as depicted in Sec 4.2, for the k-th exemplar, we get 6 edge correlation maps and append them to M_k^e for the edge branch. Finally, the density prediction module takes M_k^q and M_k^e as input and then predicts the corresponding density map. The final density map can be computed as:

$$\widetilde{Y} = Mean(\mathcal{D}(M_1^q, M_1^e), ..., \mathcal{D}(M_K^q, M_K^e)) \tag{10}$$

5 Experiments

In this section, we first describe the datasets and evaluation metrics we use, followed by implementation details; then we show our experimental results with comparisons to the state-of-the-art; finally, we perform ablation studies.

5.1 Datasets

FSC-147 [19] is a recently proposed dataset for class-agnostic counting. It consists of 6135 images with 147 object categories, from animals and plants to vehicles and toys. The number of counted objects in a single picture varies greatly, ranging from 7 to 3731, and the average number is 56. Approximate center of each object instance is annotated with a dot to generate the ground truth density map. On each image, three object instances with bounding boxes are selected as exemplars. The training, validation and test sets consist of 3659 images (89 categories), 1286 images (29 categories), and 1190 images (29 categories), respectively.

CARPK [9] is a car counting dataset which contains 459 images collected from different parking lots taken by drone cameras. There are nearly 90,000 cars in total and each instance is annotated with one bounding box. We use the center points of bounding boxes to get density maps. Moreover, same with [19], a set of 12 bounding boxes from the training set are sampled randomly as exemplars used for all the training and test images.

5.2 Evaluation Metrics

Following previous works [19,15], we adopt Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as evaluation metrics. They are formulated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\sum Y_i - \sum \widetilde{Y}_i\right|, \tag{11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\sum Y_i - \sum \widetilde{Y}_i\right|^2}, \tag{12}$$

where N is the number of test images; $\sum Y_i$, $\sum \widetilde{Y}_i$ represent ground truth and predicted counts.

5.3 Implementation Details

The architectures of RGB image backbone (ResNet-50) and the density prediction module are the same as [19]. For the edge backbone, we use the block of a Conv2d layer (with a 3×3 kernel) in VGG [23] as the basic unit. The number of channels of Conv2d layers are: [16, 16 (s=2), 32, 32 (s=2), 64, 64, 64 (s=2), 128, 128, 128, 128 (s=2), 128, 128], where 's' denotes the stride of each unit with 1 for default. In fact, our VGG-like network is quite light, even lighter than Resnet18. The numbers of parameters of our VGG-like net and Resnet18 are 0.92M and 2.78M respectively. Following [19], we generate the ground truth density maps using an Adaptive Gaussian kernel. No data augmentation is applied in all experiments. For FSC-147, we set $K = 3$, and we generate $N = 7$ augmentation features. For CARPK, there are totally 12 exemplars, but at each iteration we randomly take $K = 5$ and generate $N = 25$ augmentation features for matching. We train the network with Adam optimizer, and the learning rate is set to 10^{-5} and our model converges at 500 th epoch. All experiments are conducted on a single NVIDIA RTX 2080TI GPU with 11GB of VRAM and our code is implemented with Pytorch.

5.4 Comparisons with State-of-the-Art Methods

FSC-147 dataset. As shown in Tab. 2, we compare our method with previous published class-agnostic counting methods on the FSC-147 dataset.

Table 1. Comparison on high-diversity subsets (w.r.t. MAE).

Subset	FamNet*	EFA	Δ	EM	Δ
Full Val Set	24.32	23.08	1.24	23.29	1.03
Color Diversity	42.32	38.43	3.89	35.84	6.84
Scale Diversity	42.65	38.95	3.70	39.09	3.56
Shape Diversity	29.68	27.82	1.86	28.07	1.61

From the results in Tab. 2, we have the following observations. (1) Generally, regression based counting methods (GMN [15], MAML [8], FamNet [19] and Ours) perform better than detection based approaches (FR [11], FSOD [7]). (2) Our method outperforms the baseline method FamNet [19] by a large margin. In particular, on the validation set, the gain is 3.21 pp w.r.t. MAE; on the test set, the improvement is as large as 17.68 pp. These improvements demonstrate the effects of our proposed two new modules. (3) Our method surpasses all existing methods, defining a new state-of-the-art on the FSC-147 dataset.

Table 2. Comparison of our method and previous methods on the FSC-147 dataset.

Method	Val Set		Test Set	
	MAE	RMSE	MAE	RMSE
Mean	53.38	124.53	47.55	147.67
Median	48.68	129.70	47.73	152.46
FR [11]	45.45	112.53	41.64	141.04
FSOD [7]	36.36	115.00	32.53	140.65
Pre-trained GMN [15]	60.56	137.78	62.69	159.67
GMN [15]	29.66	89.81	26.52	124.57
MAML [8]	25.54	79.44	24.90	112.68
CFOCNet [25]	27.82	71.99	28.60	123.96
FamNet [19]	23.75	69.07	22.08	99.54
RCAC (Ours)	**20.54**	**60.78**	**20.21**	**81.86**

We further observe the improvements of our method to the baseline on high diversity images. We show the effects of two modules on high-diversity subsets from the comparison in Tab. 1. As stated in our abstract and introduction, EFA handles all kinds of diversity, while EM focuses on handling color diversity, indicating that our method is more robust to high intraclass diversity.

Additionally, Fig. 4 shows some qualitative results on the FSC-147 dataset. In row 1, our method obtains stronger responses and a more accurate count number at the scenario of high shape diversity led by severe occlusion. In row 2, our method produces cleaner density maps with less noises at the background regions than the baseline by handling color diversity. Inside each exemplar box, the background colors are dominant, resulting in noisy responses at background regions on the baseline density map. In row 3, our method produces more balanced density maps across different scales than the baseline by handling scale diversity. In row 4, our method produces more uniform density maps across different foreground color diversity.

CARPK dataset. Similar to [19], we further verify our method on the CARPK dataset, due to the lack of class-agnostic counting datasets. The experiments are implemented under the same few-shot setting. Since there is only one category for CARPK, it is considered rather a simple version of class-agnostic object counting. The results are shown in Tab. 3. Our model outperforms all previous approaches except GMN, which

Table 3. Comparison of car counting performance on the CARPK dataset. *GMN uses extra images of cars from the ILSVRC video dataset for training. "Fine-tuned" denotes whether the models are further fine-tuned on CARPK.

Method	Fine-tuned	MAE	RMSE
YOLO [9,20]	✓	48.89	57.55
Faster RCNN [9,21]	✓	47.45	57.39
One-look Regression [9,17]	✓	59.46	66.84
Faster RCNN (RPN-small) [9,21]	✓	24.32	37.62
Spatially Regularized RPN [9]	✓	23.80	36.79
GMN* [15]	✓	7.48	9.90
FamNet [19]	✓	18.19	33.66
RCAC (Ours)	✓	**13.62**	**19.08**
FamNet [19]	×	28.84	44.47
RCAC (Ours)	×	**17.98**	**24.21**

uses external training images of cars from the ILSVRC video dataset. It is notable that our approach improves over FamNet by 4.57 pp w.r.t MAE and 14.58 pp w.r.t RMSE. These results indicate that our method generalizes well to different datasets.

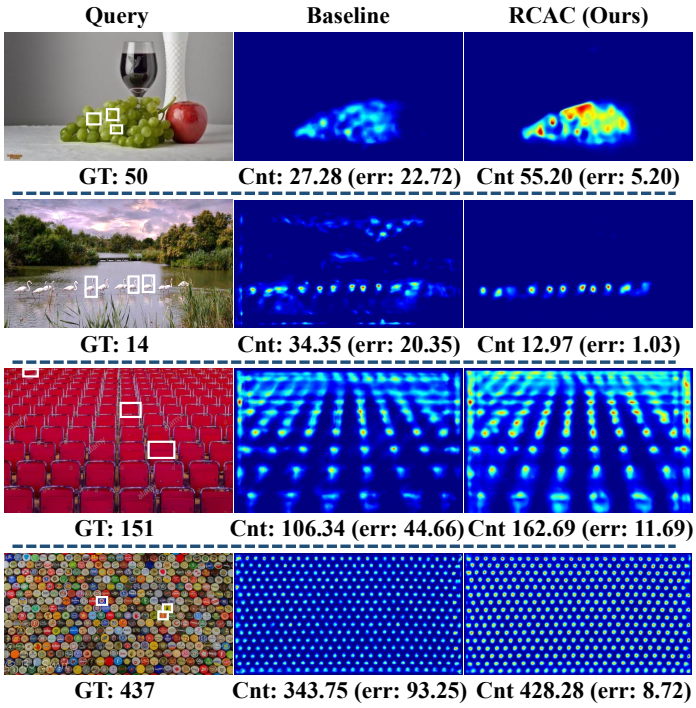

Fig. 4. Qualitative results of different methods on high intraclass diversity images from FSC-147. Zoom in and colorful for better visualization.

5.5 Ablation Studies

In the following, we conduct some ablation studies to analyze our proposed exemplar feature augmentation and edge matching modules. All experiments are conducted on the validation set of FSC-147.

Effects of two new modules. As shown in Tab. 4, the performance is improved by 1.24 pp w.r.t MAE (from 24.32 to 23.08) when exemplar feature augmentation is employed. On the other hand, we also observe a remarkable improvement of 1.03 pp w.r.t MAE (from 24.32 to 23.29) from edge matching. Moreover, we obtain a total gain of 3.78 pp w.r.t MAE by adding both modules. These results indicate the effects of two proposed modules.

Table 4. Effects of two proposed components.

Exemplar Feature Augmentation	Edge Matching	MAE	RMSE
×	×	24.32	70.94
✓	×	23.08	67.23
×	✓	23.29	63.35
✓	✓	20.54	60.78

Impact of Dirichlet distribution parameter θ. From Tab. 5, we can see our method obtains consistent improvements to the baseline by using exemplar feature augmentation no matter which sampling parameter we choose. By analyzing Fig. 3 and Tab. 5 simultaneously, we find it works better to set θ_i evenly ($\{2,2,2\}$ vs. $\{3,2,2\}$) such that we have a high probability to include the average fusion of the K exemplars. Also, it helps to have a larger sampling area ($\{2,2,2\}$ vs. $\{5,5,5\}$ such that more diverse combinations can be generated. Finally we set θ to $\{2,2,2\}$ for all our experiments as it performs the best.

Table 5. Effect of different dirichlet distribution parameters.

θ	MAE	RMSE
-	23.29	63.35
{ 3,2,2 }	21.44	63.40
{ 5,5,5 }	20.46	61.70
{ 2,2,2 }	20.54	60.78

Impact of feature augmentation quantity. We analyze how the value of N affects the performance of our exemplar feature augmentation strategy. As shown in Tab. 6, we find that in general larger N leads to better performance and it does not seem to saturate at $N = 7$. Due to the limited memory of our NVIDIA RTX 2080TI GPU, we select $N = 7$ for our experiments. But we expect to achieve better performance with a large value.

Table 6. Impact of different values for augmentation quantity N.

Augmentation Quantity N	MAE	RMSE
0	23.29	63.35
1	22.19	62.99
3	21.21	62.07
5	21.14	62.92
7	20.54	60.78

Inference time analysis. To verify the efficiency of our RCAC, we compare the inference time of our RCAC with FamNet* and FamNet in Tab. 7. We can see our RCAC runs slightly slower than our baseline FamNet* (75ms vs. 47ms) due to additional computations for edge detection. In order to accelerate our RCAC, we replace RCF with Sobel operators, which reduces the inference time at a cost of small performance drop. Please note our RCAC (w/ Sobel) still outperforms FamNet* by $\sim$ 2pp at a similar speed; and compared to previous top method FamNet, our RCAC (w/ RCF) not only obtains better performance (by $\sim$ 3pp), but also runs much faster (75ms vs. 3,900ms).

Table 7. Inference time analysis. "T" represents the inference time.

Method	$K = 3, N = 0$	
	MAE	T (ms)
FamNet*	24.32	47
FamNet	23.75	3,900
RCAC (w/ Sobel)	22.41	59
RCAC (w/ RCF)	20.94	75

In the supplementary material, we provide more ablation studies on the impact of the effect of using edge images at the 2nd branch, effect of number of exemplars, qualitative results of augmented exemplars and application of EFA in another task.

6 Conclusion

In this paper, we analyze failure cases of previous top-performing class-agnostic object counter and find high intraclass diversity in the query image has an adverse effect on counting performance. To solve this problem, we propose two novel modules: exemplar feature augmentation and edge matching. They make our counter robust to high intraclass diversity. Extensive experiments have demonstrated the effectiveness and robustness of our method.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62172225), Fundamental Research Funds for the Central Universities (No. 30920032201) and the "111" Program B13022.

References

1. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. PAMI **33**(5), 898–916 (2010) 8
2. Arteta, C., Lempitsky, V., Noble, J.A., Zisserman, A.: Interactive object counting. In: ECCV. pp. 504–518 (2014) 2, 3
3. Arteta, C., Lempitsky, V., Zisserman, A.: Counting in the wild. In: ECCV. pp. 483–498 (2016) 1
4. Chattopadhyay, P., Vedantam, R., Selvaraju, R.R., Batra, D., Parikh, D.: Counting everyday objects in everyday scenes. In: CVPR. pp. 1135–1144 (2017) 3
5. Cholakkal, H., Sun, G., Khan, F.S., Shao, L.: Object counting and instance segmentation with image-level supervision. In: CVPR. pp. 12397–12405 (2019) 3
6. Cholakkal, H., Sun, G., Khan, S., Khan, F.S., Shao, L., Van Gool, L.: Towards partial supervision for generic object counting in natural scenes. PAMI (2020) 3
7. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: CVPR. pp. 4013–4022 (2020) 3, 11
8. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML. pp. 1126–1135 (2017) 3, 11
9. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: ICCV. pp. 4145–4153 (2017) 3, 9, 12
10. Hsieh, T.I., Lo, Y.C., Chen, H.T., Liu, T.L.: One-shot object detection with co-attention and co-excitation. In: NIPS. pp. 2725–2734 (2019) 3
11. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: ICCV. pp. 8420–8429 (2019) 3, 11
12. Laradji, I.H., Rostamzadeh, N., Pinheiro, P.O., Vazquez, D., Schmidt, M.: Where are the blobs: Counting by localization with point supervision. In: ECCV. pp. 547–562 (2018) 3
13. Liu, Y., Wen, Q., Chen, H., Liu, W., Qin, J., Han, G., He, S.: Crowd counting via cross-stage refinement networks. IEEE TIP **29**, 6800–6812 (2020) 1, 3
14. Liu, Y., Cheng, M.M., Hu, X., Bian, J.W., Zhang, L., Bai, X., Tang, J.: Richer convolutional features for edge detection. PAMI **41**(8), 1939–1946 (2019) 8
15. Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: ACCV. pp. 669–684 (2018) 2, 3, 10, 11, 12
16. Mo, H., Ren, W., Xiong, Y., Pan, X., Zhou, Z., Cao, X., Wu, W.: Background noise filtering and distribution dividing for crowd counting. IEEE TIP **29**, 8199–8212 (2020) 1, 3
17. 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. pp. 785–800 (2016) 1, 12
18. Rahnemoonfar, M., Sheppard, C.: Deep count: fruit counting based on deep simulated learning. Sensors **17**(4), 905 (2017) 1
19. Ranjan, V., Sharma, U., Nguyen, T., Hoai, M.: Learning to count everything. In: CVPR. pp. 3394–3403 (2021) 2, 3, 4, 6, 9, 10, 11, 12
20. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016) 12
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS **28**, 91–99 (2015) 12
22. Ribera, J., Guera, D., Chen, Y., Delp, E.J.: Locating objects without bounding boxes. In: CVPR. pp. 6479–6489 (2019) 1, 3
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 9, 10

24. Sokhandan, N., Kamousi, P., Posada, A., Alese, E., Rostamzadeh, N.: A few-shot sequential approach for object counting. arXiv preprint arXiv:2007.01899 (2020) 3
25. Yang, S.D., Su, H.T., Hsu, W.H., Chen, W.C.: Class-agnostic few-shot object counting. In: WACV. pp. 870–878 (2021) 3, 11