Few-shot Object Counting and Detection

Thanh Nguyen[1⋆], Chau Pham[1⋆], Khoi Nguyen[1], and Minh Hoai[1,2]

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

Abstract. We tackle a new task of few-shot object counting and detection. Given a few exemplar bounding boxes of a target object class, we seek to count and detect all objects of the target class. This task shares the same supervision as the few-shot object counting but additionally outputs the object bounding boxes along with the total object count. To address this challenging problem, we introduce a novel two-stage training strategy and a novel uncertainty-aware few-shot object detector: Counting-DETR. The former is aimed at generating pseudo ground-truth bounding boxes to train the latter. The latter leverages the pseudo ground-truth provided by the former but takes the necessary steps to account for the imperfection of pseudo ground-truth. To validate the performance of our method on the new task, we introduce two new datasets named FSCD-147 and FSCD-LVIS. Both datasets contain images with complex scenes, multiple object classes per image, and a huge variation in object shapes, sizes, and appearance. Our proposed approach outperforms very strong baselines adapted from few-shot object counting and few-shot object detection with a large margin in both counting and detection metrics. The code and models are available at https://github.com/VinAIResearch/Counting-DETR.

Keywords: Few-shot Object Counting, Few-shot Object Detection

1 Introduction

This paper addresses a new task of Few-Shot object Counting and Detection (FSCD) in crowded scenes. Given an image containing many objects of multiple classes, we seek to count and detect all objects of a target class of interest specified by a few exemplar bounding boxes in the image. To facilitate few-shot learning, in training, we are only given the supervision of few-shot object counting, i.e., dot annotations for the approximate centers of all objects and a few exemplar bounding boxes for object instances from the target class. It is worth noting that the test classes may or may not be present in training classes. The problem setting is depicted in Fig. 1.

FSCD is different from Few-Shot Object Counting (FSC) and Few-Shot Object Detection (FSOD). Compared to FSC, FSCD has several advantages: (1) obtaining object bounding boxes "for free", which is suitable for quickly annotating bounding boxes for a new object class with a few exemplar bounding boxes;

⋆ Equal contribution

(a) Training   (b) Testing

Fig. 1. We address the task of few-shot counting and detection in a novel setting: (a) in training, each training image contains dot annotations for all objects and a few exemplar boxes. (b) In testing, given an image with a few exemplar boxes defining a target class, our goal is to count and detect all objects of that target class in the image.

(2) making the result of FSC more interpretable since bounding boxes are easier to verify than the density map. Compared to FSOD which requires bounding box annotation for all objects in the training phase of the base classes, FSCD uses significantly less supervision, i.e., only a few exemplar bounding boxes and dot annotations for all objects. This is helpful in crowded scenes where annotating accurate bounding boxes for all objects is ambiguously harder and significantly more expensive than the approximate dot annotation.

Consequently, FSCD is more challenging than both FSC and FSOD. FSCD needs to detect and count all the objects as for FSOD, but it is only trained with the supervision of FSC. This invalidates of most of available approaches used in these problems without significant changes in network architecture or loss function. Specifically, it is not trivial to extend the density map produced by FSC approaches to predict the object bounding boxes; and it is hard to train a few-shot object detector with few exemplar bounding boxes of the base classes.

A naive approach for FSCD is to extend FamNet [32], a density-map-based approach for FSC, whose counting number is obtained by summing over the predicted density map. To extend FamNet to detect objects, one can use a regression function on top of the features extracted from the peak locations (whose density values are highest in their respective local neighborhoods), the features extracted from the exemplars, and the exemplar boxes themselves. The process of this naive approach is illustrated in Fig. 2a. However, this approach has two limitations due to: 1) the imperfection of the predicted density map, and 2) the non-discriminative peak features. In the former, the density value is high in the environment locations whose color is similar to those of the exemplars, or the density map is peak-indistinguishable when the objects are packed in a dense region as depicted in Fig. 2b. In the latter, the extracted features are trained with counting objective (not object detection) so that they cannot represent for different shapes, sizes, and orientations, as illustrated in Fig. 2c.

To address the aforementioned limitations, we propose a new point-based approach, named Counting-DETR, treating objects as points. In particular, counting and detecting objects is equivalent to counting and detecting points, and the object bounding box is predicted directly from point features. Counting-DETR is based on an object detector, Anchor DETR [43], with improvements to better address FSCD. **First**, inspired by [5] we adopt a two-stage training strategy: (1) Counting-DETR is trained to generate pseudo ground-truth (GT) bounding
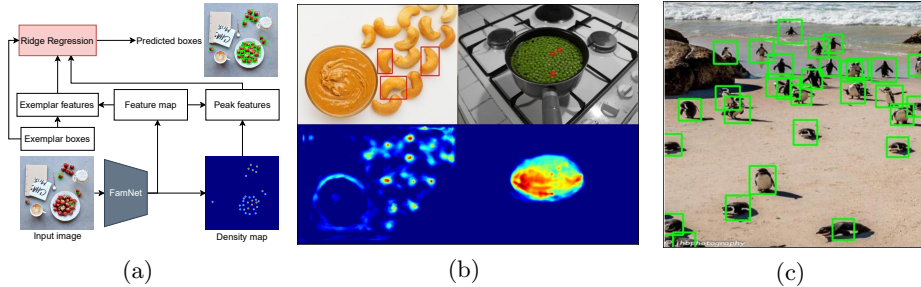
Fig. 2. Limitations of a naive approach for FSCD by extending FamNet [32] with a regression function for object detection. (a) Processing pipeline of this approach: a regressor takes as input exemplar boxes with their features, and features at peak density locations to predict bounding boxes for the peak locations. (b) Limitation 1: poor quality of the density map predicted by FamNet when the exemplars share similar appearance with background or densely packed region. The first row presents the input images with a few exemplars each, the second row presents the corresponding density map predicted by FamNet. (c) Limitation 2: Non-discriminative peak features cannot represent objects with significant differences in shape and size. The green boxes are predicted from the features extracted at the annotated dots.

boxes given the annotated points of training images; (2) Counting-DETR is further fine-tuned on the generated pseudo GT bounding boxes to detect objects on test images. **Second,** since the generated pseudo GT bounding boxes are imperfect, we propose to estimate the uncertainty for bounding box prediction in the second stage. The estimated uncertainty regularizes learning such that lower box regression loss is incurred on the predictions with high uncertainty. The overview of Counting-DETR is illustrated in Fig. 3.

In short, the contributions of our paper are: (1) we introduce a new problem of few-shot object counting and detection (FSCD); (2) we introduce two new datasets, FSCD-147 and FSCD-LVIS; (3) we propose a two-stage training strategy to first generate pseudo GT bounding boxes from the dot annotations, then use these boxes as supervision for training our proposed few-shot object detector; and (4) we propose a new uncertainty-aware point-based few-shot object detector, taking into account the imperfection of pseudo GT bounding boxes.

2 Related Work

In this section, we review some related work on object counting and detection.

Visual counting focuses on some predefined classes such as car [27,15], cell [2,46], and human [29,16,21,8,41,37,17,51,49,31,33,1]. The methods can be grouped into two types: density-map-based and detection-based. The former, e.g., [40,24], predicts and sums over density map from input image to get the final results. The latter (e.g., [15,11]) counts the number of objects based on the detected boxes. The latter is better at justifying the counting number, however, it requires the
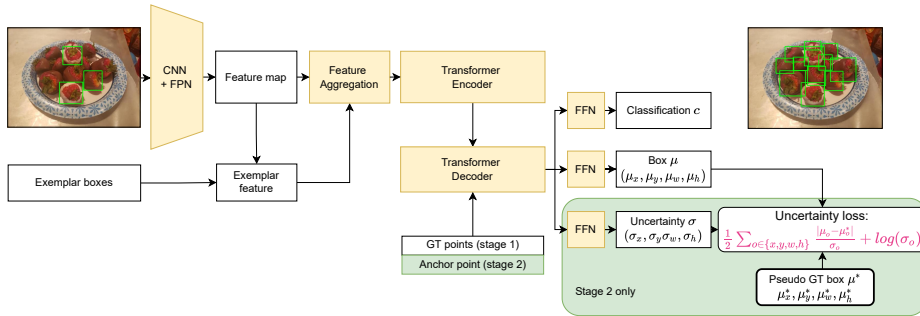
Fig. 3. The overview of our two-step training approach: (1) Counting-DETR is first trained on a few pairs of dot and bounding boxes and then used to predict pseudo GT boxes for the annotated dots; (2) Counting-DETR is trained to predict the object bounding boxes, with the prediction target being the pseudo GT boxes from the first stage. Specifically, the input image is first forwarded through a CNN+FPN backbone to extract its feature map. The exemplar features are extracted from their boxes to integrate with the feature map producing the exemplars-integrated feature map. This feature map is then taken as input to the encoder-decoder transformer along with either the annotated dots in the first stage or the anchor points in the second stage for foreground/background classification and bounding box regression. In the second stage, the estimated uncertainty is used to regularize the training with a new uncertainty loss to account for the imperfection of the pseudo GT bounding boxes.

GT bounding boxes for training and its performance is exceeded by that of the former, especially for images of crowded scenes.

Few-shot counting (FSC) counts the number of objects in an image with some exemplar bounding boxes of a new object class. Since the number of exemplar boxes is so small that an object detector cannot be reliably learned, prior methods on FSC are all based on density-map regression. GMN [26] formulates object counting as object matching in video tracking such that a class agnostic counting module can be pretrained on a large-scale video object tracking dataset (ImageNet VID [36]). FamNet [32] correlates the features extracted from a few exemplars with the feature map to obtain the density map for object counting. VCN [30] improves upon [32] by augmenting the input image with different styles to make the counting more robust. LaoNet [22] combines self-attention and cross attention in the transformer to aggregate features from the exemplar to the image to facilitate density map prediction. ICFR [48] proposes an iterative framework to progressively refine the exemplar-related features, thus producing a better density map than a single correlation in [32]. However, these approaches do not output object bounding boxes. An extension for object detection from these approaches is depicted in Fig. 2a, but it has several limitations as illustrated Fig. 2b and Fig. 2c. Whereas, our approach Counting-DETR effectively predicts object bounding boxes along with the object count with only the supervision of FSC.

Object detection methods include anchor-based approaches such as Faster-RCNN [34] and Retina Net [23], point-based approaches such as like FCOS [38] and Center-Net [52], and transformer-based approaches such as DETR [3], Point DETR [5] and Anchor DETR [43]. DETR is the first approach to apply transformer architecture [39] to object detection. Anchor DETR improves the convergence rate and performance of DETR by learnable anchor points representing the initial prediction of the objects in the image. However, these methods require thousands of bounding box annotations on some predefined classes for training and cannot generalize well on a new class in testing with a few box exemplars as in our few-shot setting. Point DETR [5] alleviates this requirement using two separate detectors: teacher (i.e., Point-DETR) and student (i.e., FCOS). The former learns from a small set of fully annotated boxes to generate pseudo-GT bounding boxes of a large amount of point-annotated images. Then the latter is trained with these pseudo-GT boxes to predict the bounding boxes of the test images. This approach is complicated and does not take into account the imperfect pseudo-GT bounding boxes. In contrast, our Counting-DETR is a unified single network with the uncertainty-aware bounding prediction.

Few-shot object detection (FSOD) approaches are mostly based on Faster-RCNN [34] and can be divided into two subgroups based on episodic training [19,47,50,45,6] and fine-tuning [4,42,44,7]. The former leverages episodic training technique to mimic the evaluation setting in the training whereas the latter fine-tunes some layers while keeping the rest unchanged to preserve the knowledge learned from the training classes. However, all FSOD approaches require box annotations for all objects of the base classes in training. It is not the case in our setting where only a few exemplar bounding boxes are given in training. To address this problem, we propose a two-stage training strategy wherein the first stage, the pseudo GT bounding boxes for all objects are generated from the given exemplar boxes and the dot annotations.

Object detection with uncertainty accounts for the uncertainty in the input image due to blurring or indistinguishable boundaries between objects and the background. Prior work [14,20] assumes the object bounding boxes are characterized by a Gaussian distribution whose mean and standard deviation are predicted by a network trained with an uncertainty loss function derived from maximum the likelihood between the predicted distribution and the GT boxes. [10,9] apply the uncertainty loss for 3D object detection. Our uncertainty loss shares some similarities with prior work, however, we use the uncertainty loss to account for the imperfection of the pseudo GT bounding boxes (not the input image) such that the Laplace distribution works significantly better than the prior Gaussian distribution as shown in experiments.

3   Proposed Approach

Problem definition: In training, we are given a set of images containing multiple object categories. For each image I, a few exemplar bounding boxes B_k,

$k = 1, \ldots, K$ where K is the number of exemplars, and the dot annotations for all object instances of a target class are annotated. This kind of supervision is the same as in few-shot object counting. In testing, given a query image with bounding boxes for a few exemplar objects in the target class, our goal is to detect and count all instances of the target class in the query image.

To address this problem, we propose a novel uncertainty-aware point-based few-shot object detector, named Counting-DETR trained with a novel two-stage training strategy to first generate pseudo GT bounding boxes from dot annotations and then train Counting-DETR on the generated pseudo GT boxes to predict bounding boxes of a new object class defined by a few bounding box exemplars in testing. The overview of Counting-DETR is illustrated in Fig. 3.

3.1   Feature Extraction and Feature Aggregation

Feature extraction: A CNN backbone is used to extract feature map $F^I \in \mathbb{R}^{H \times W \times D}$ from the input image I where H, W, D are height, width, and number of channels of the feature map, respectively. We then extract the exemplar feature vectors $f_k^B \in \mathbb{R}^{1 \times D}$, at the center of the exemplar bounding boxes B_k. Finally, the exemplar feature vector $f^B \in \mathbb{R}^{1 \times D}$ is obtained by averaging these feature vectors, or $f^B = \frac{1}{K} \sum_k f_k^B$.

Feature aggregation: We integrate the exemplar feature f^B to the feature map of the image F^I to produce the exemplar-integrated feature map F^A:

$$F^A = W_{proj} * [F^I; F^I \otimes f^B], \tag{1}$$

where $*, \otimes, [\cdot; \cdot]$ are the convolution, channel-wise multiplication, and concatenation operations, respectively. $W_{proj} \in \mathbb{R}^{2D \times D}$ is a linear projection weight. The first term in the concatenation preserves the original information of the feature map, while the second term aims at enhancing features at locations whose appearance are similar to those of the exemplars and suppressing the others.

3.2   The Encoder-Decoder Transformer

Inspired by DETR [3], we design our transformer of Counting-DETR to take as input the exemplar-integrated feature map F^A and M query points $\{p_m\}_{m=1}^M$ and predict the bounding box b_m for each query point p_m. The queries are the 2D points representing the initial guesses for the object locations rather than the learnable embeddings to achieve a faster training rate as shown in [43]. Thus, Counting-DETR is point-based approach that leverages the given dot annotations as the queries to predict the pseudo GT bounding boxes. Also, the transformer consists of two sub-networks: encoder and decoder. The former aims at enhancing features among the input set of features with the self-attention operation. The latter allows all the query points to interact with the enhanced features from the encoder with the cross-attention operation, thus capturing global information.

Next, the decoder is used to: (1) predict the classification score s representing the presence or absence of the object at a particular location, (2) regress the object's bounding box μ represented by the offset x, y from the GT object center to the query point along with its size w, h. Following [43], first, the Hungarian algorithm is used to match each of the GT bounding boxes with its corresponding predicted bounding boxes. Then for each pair of matched GT and predicted bounding boxes, the focal loss [23] and the combination of L_1 loss and GIoU loss [35] are used as training loss functions. In particular, at each query point, the following loss is computed:

$$L_{\text{DETR}} = \lambda_1 \text{Focal}(s, s^*) + \lambda_2 L_1(\mu, \mu^*) + \lambda_3 \text{GIoU}(\mu, \mu^*), \tag{2}$$

where s^*, μ^* are the GT class label and bounding box, respectively. $\lambda_1, \lambda_2, \lambda_3$ are the coefficients of focal, L_1, and GIoU loss functions, respectively.

Notably, Counting-DETR also estimates the uncertainty σ when training under the supervision of the imperfect pseudo GT bounding boxes $\tilde{\mu}$. This uncertainty is used to regularize the learning of bounding box μ such that a lower loss is incurred at the prediction with high uncertainty. We propose to use the following uncertainty loss:

$$L_{\text{uncertainty}} = \frac{1}{2} \sum_{o \in \{x,y,w,h\}} \frac{|\mu_o - \tilde{\mu}_o|}{\sigma_o} + \log \sigma_o, \tag{3}$$

where σ is the estimated uncertainty. This loss is derived from the maximum likelihood estimation (MLE) between the predicted bounding box distribution characterized by a Laplace distribution and the pseudo GT bounding box as evidence. Another option is the Gaussian distribution, however, in the experiments, we show that the Gaussian has the inferior performance to that of Laplace, and is even worse than the variant that does not employ uncertainty estimation.

3.3 The Two-stage Training Strategy

The proposed few-shot object detector, Counting-DETR, can only be trained with the bounding box supervision for all objects. However, we only have bounding box annotation for a few exemplars and the point annotation for all objects as the setting of FSCD. Hence, we propose a two-stage training strategy as follows.

In Stage 1, we first pretrain Counting-DETR on a few exemplar bounding boxes with their centers as the query points (as described in Sec. 3.2). Subsequently, the pretrained network is used to predict the pseudo GT bounding boxes on the training images with the dot annotations as the query points. It is worth noting that, in this stage, we have the GT exemplar center as queries and their corresponding bounding boxes as supervision, so we do not use the Hungarian matching, uncertainty estimation, and uncertainty loss, i.e., we only use L_{DETR} in Eq. (2) to train our Counting-DETR. The visualization of some generated pseudo-GT boxes is illustrated in Fig. 4.

In Stage 2, the generated pseudo GT bounding boxes on the training images are used to fine-tune the pretrained Counting-DETR. The fine-tuned model is
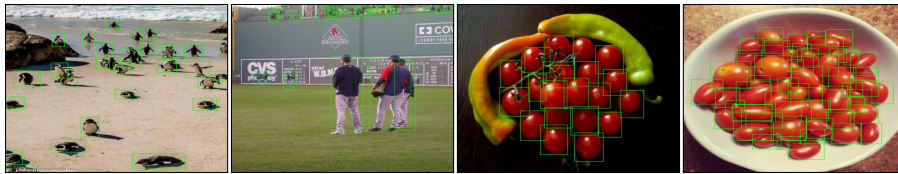
Fig. 4. Examples of pseudo GT bounding boxes generated by the 1-st stage our method.

(a) FSCD-147   (b) FSCD-LVIS
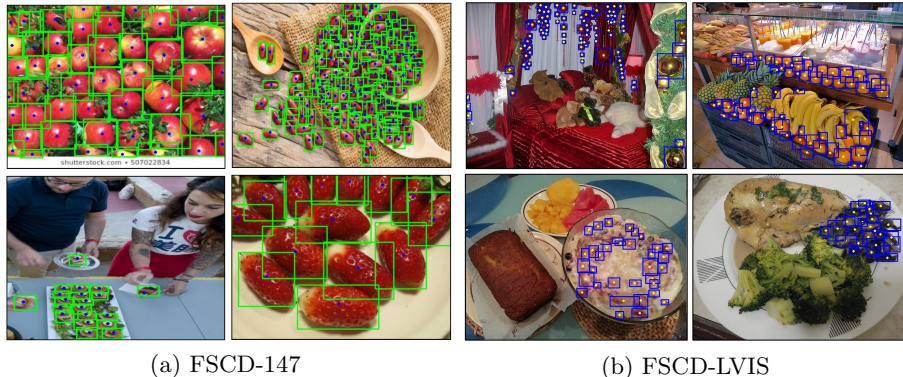
Fig. 5. Sample images from our datasets and annotated bounding boxes.

then used to make predictions on the test images with the uniformly sampled anchor points as queries. Different from Stage 1, the supervision is the imperfect pseudo GT bounding boxes, hence, we additionally leverage the uncertainty estimation branch with uncertainty loss to train. Particularly, we use the following loss to train Counting-DETR in this stage:

$$L_{\text{combine}} = L_{\text{DETR}} + \lambda_4 L_{\text{uncertainty}}, \tag{4}$$

where λ_4 is the coefficient of $L_{\text{uncertainty}}$.

4   New Datasets for Few-shot Counting and Detection

A contribution of our paper is the introduction of two new datasets for few-shot counting and detection. In this section, we will describe these datasets.

4.1   The FSCD-147 Dataset

The FSC-147 dataset [32] was recently introduced for the few-shot object counting task with 6135 images across a diverse set of 147 object categories. In each image, the dot annotation for all objects and three exemplar boxes are provided. However, this dataset does not contain bounding box annotations for all objects. For evaluation purposes, we extend the FSC-147 dataset by providing bounding box annotations for all objects of the val and test sets. We name the new dataset FSC**D**-147. To be consistent with the counting, an object will be annotated with

Table 1. Comparison between the FSCD-147 and FSCD-LVIS datasets

		Number of images			
Dataset	#classes	total	train	val	test
FSCD-147	147	6135	3659	1286	1190
FSCD-LVIS	372	6195	4000	1181	1014

Table 2. Number of images for each bin of the FSCD-LVIS dataset

Counting range	20-30	30-40	40-50	50-60	60-70	>70
# images in range	2628	1212	684	402	286	959

its bounding box only if it has a dot annotation. Fig. 5a shows some samples of the FSCD-147 dataset. It is worth noting that annotating bounding boxes for many objects in crowded scenes of FSC-147 is a laborious process, and this is a significant contribution of our paper.

4.2 The FSCD-LVIS Dataset

Although the FSC-147 dataset contains images with a large number of objects in each image, the scene of each image is rather simple. Each image of FSC-147 shows the target object class so clearly that one can easily know which object class to count without having to specify any exemplars as shown in Fig. 1 and Fig. 5a. For real-world deployment of methods for few-shot counting and detection, we introduce a new dataset called FSCD-LVIS. Specifically, the scene is more complex with multiple object classes with multiple object instances each as illustrated in Fig. 5b. Without providing the exemplars for the target class, one cannot definitely guess which the target class is.

The FSCD-LVIS dataset contains 6196 images and 377 classes, extracted from the LVIS dataset [12]. For each image, we filter out all instances with an area smaller than 20 pixels, or a width or a height smaller than 4 pixels. The comparison between the FSCD-LVIS and FSC-147 datasets is shown in Tab. 1. The histogram of the number of labeled objects per image is illustrated in Tab. 2. The LVIS dataset has the box annotations for all objects, however, to be consistent with the setting of FSCD, we randomly choose three annotated bounding boxes of a selected object class as the exemplars for each image in the training set of FSCD-LVIS.

5 Experimental Results

Metrics. For object counting, we use Mean Average Error (MAE) and Root Mean Squared Error (RMSE), which are standard measures used in the counting literature. Besides, the Normalized Relative Error (NAE) and Squared Relative Error (SRE) are also adopted. In particular, $\text{MAE} = \frac{1}{J}\sum_{j=1}^{J}|c_j^* - c_j|$; $\text{RMSE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(c_j^* - c_j)^2}$; $\text{NAE} = \frac{1}{J}\sum_{j=1}^{J}\frac{|c_j^* - c_j|}{c_j^*}$; $\text{SRE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\frac{(c_j^* - c_j)^2}{c_j^*}}$ where J

Table 3. Ablation study on each component's contribution to the final results

Combination		Counting				Detection	
Pseudo box	Uncertainty	MAE ($\downarrow$)	RMSE($\downarrow$)	NAE($\downarrow$)	SRE ($\downarrow$)	AP($\uparrow$)	AP50($\uparrow$)
✓	✓	**20.38**	**82.45**	**0.19**	**3.38**	**17.27**	**41.90**
✗	✓	29.74	104.04	0.26	4.44	11.37	29.98
✓	✗	23.57	93.54	0.21	3.77	14.19	36.34
✗	✗	31.36	105.76	0.27	4.60	10.81	28.76

Table 4. Performance of Counting-DETR with different types of anchor points

	Counting				Detection	
Anchor type	MAE ($\downarrow$)	RMSE($\downarrow$)	NAE($\downarrow$)	SRE ($\downarrow$)	AP($\uparrow$)	AP50($\uparrow$)
Learnable	25.20	**81.94**	0.25	3.92	16.46	38.34
Fixed grid (proposed)	**20.38**	82.45	**0.19**	**3.38**	**17.27**	**41.90**

is the number of test images, c_j^* and c_j are GT and the predicted number of objects for image j, respectively. Unlike the absolute errors MAE and RMSE, the relative errors NAE and SRE reflect the practical usage of visual counting, i.e., with the same number of wrong objects counted (e.g., 10), it is more serious for images having a smaller number of objects (e.g., 20) than the ones having larger numbers of objects (e.g., 200).

For object detection, we use mAP and AP50. They are the average precision metrics with the IoU threshold between predicted and GT boxes for determining a correct prediction ranging from 0.5 to 0.95 for mAP and 0.5 for AP50.

Implementation details. We implement our approach, baselines, and ablations in PyTorch [28]. Our backbone network is ResNet-50 [13] with the frozen Batch Norm layer [18]. We extract exemplar features f_k^B from exemplar boxes B_k from Layer 4 of the backbone. Our transformer network shares the same architecture as that of Anchor DETR [43] with the new uncertainty estimation and is trained with additional uncertainty loss as described in Sec. 3.2, while keeping the rest intact with six layers for both encoder and decoder. We use AdamW optimizer [25] with the learning rate of 10^{-5} for the backbone and 10^{-4} for the transformer to train Counting-DETR in 30 epochs with a batch size of one. We use the following training loss coefficients $\lambda_1 = 2, \lambda_2 = 5, \lambda_3 = 2, \lambda_4 = 2$, which were tuned based on the validation set. Also, the number of exemplar boxes is set to $K = 3$, as in FamNet [32] for a fair comparison.

5.1 Ablation Study

We conduct several experiments on the validation data of FSCD-147 to study the contribution of various components of our method.

Pseudo Box and Uncertainty Loss. From Tab. 3, we see that pseudo GT boxes (pseudo box) generated from our first stage are much better than the

Table 5. Performance of Counting-DETR with different numbers of anchor points

# anchor points	Counting				Detection	
	MAE ($\downarrow$)	RMSE($\downarrow$)	NAE($\downarrow$)	SRE ($\downarrow$)	AP($\uparrow$)	AP50($\uparrow$)
100	30.22	113.24	0.24	4.55	11.26	28.62
200	26.76	103.11	0.22	4.15	14.06	34.33
300	23.57	93.54	0.21	3.77	14.19	36.34
400	22.62	88.45	0.21	3.69	14.91	37.30
500	21.72	85.20	0.20	3.52	16.03	39.66
600	**20.38**	**82.45**	**0.19**	**3.38**	**17.27**	**41.90**
700	21.19	83.70	0.22	3.47	15.10	37.85

boxes generated by Ridge Regression (-6 in AP, +9 in MAE). Without using uncertainty loss (similar to Point DETR [5]), the performance drops substantially (-3 in AP, +3 in MAE). That justifies the effectiveness of our uncertainty loss. Without using both of them, the performance gets worst (-7 in AP, +11 in MAE). These results demonstrate the important contribution of our proposed pseudo GT box generation and uncertainty loss.

Types of anchor points. As described in Sec. 3.2, we follow the design of Anchor DETR whose anchor points can either be learnable or fixed-grid. The results of these two types are shown in Table 4, we can see that the fixed-grid anchor points are comparable to the learnable anchor points on counting metrics, but better on the detection metrics. Thus, the fixed-grid anchor points are chosen for the Counting-DETR.

Numbers of anchor points. Tab. 5 presents the results with different numbers of anchor points M. Both the detection and counting results increase as the number of anchor points increases, and they reach the highest points when the number of anchor points is $M = 600$. Hence, we choose 600 anchor points for Counting-DETR.

Types for the uncertainty loss. Instead of using the Laplace distribution as described in Sec. 3.2, we use Gaussian distribution to derive the uncertainty loss: $L_{\text{uncertainty}}^{\text{Gaussian}} = \frac{1}{2} \sum_{o \in \{x,y,w,h\}} \frac{(\mu_o - \mu_o^*)^2}{\sigma_o^2} + \log \sigma_o^2$. This loss is similar to [14]. The results are shown in Tab. 6. The uncertainty loss derived from the Gaussian distribution yields the worst results among the variants, even worse than the variant without using any uncertainty loss. On the contrary, our proposed uncertainty loss derived from the Laplace distribution gives the best results.

5.2   Comparison to Prior Work

Since there is no existing method for the new FSCD task, we compare Counting-DETR with several strong baselines adapted from few-shot object counting and few-shot object detection: FamNet [32]+RR, FamNet [32]+MLP, Attention-RPN [6]+RR, and FSDetView [45]+RR. Other few-shot object detectors are not chosen due to the unavailability of the source code or the requirement for fine-tuning

Table 6. Ablation study for the uncertainty loss

	Counting				Detection	
Distribution type	MAE ($\downarrow$)	RMSE($\downarrow$)	NAE($\downarrow$)	SRE ($\downarrow$)	AP($\uparrow$)	AP50($\uparrow$)
W/o uncertainty loss	23.20	92.87	0.21	3.77	13.86	35.67
Gaussian loss	24.46	94.20	0.22	3.84	14.03	34.91
Laplacian loss (proposed)	**20.38**	**82.45**	**0.19**	**3.38**	**17.27**	**41.90**

Table 7. Comparison with strong baselines on the FSCD-147 test set

	Counting				Detection	
Method	MAE ($\downarrow$)	RMSE ($\downarrow$)	NAE($\downarrow$)	SRE ($\downarrow$)	AP($\uparrow$)	AP50($\uparrow$)
FamNet [32]+RR	22.09	**99.55**	0.44	6.45	9.44	29.73
FamNet [32]+MLP	22.09	**99.55**	0.44	6.45	1.21	6.12
Attention-RPN [6]+RR box	32.70	141.07	0.38	5.27	18.53	35.87
FSDetView [45]+RR box	37.83	146.56	0.48	5.47	13.41	32.99
Attention-RPN [6]+pseudo box	32.42	141.55	0.38	5.25	20.97	37.19
FSDetView [45]+pseudo box	37.54	147.07	0.44	5.40	17.21	33.70
Counting-DETR (proposed)	**16.79**	123.56	**0.19**	**5.23**	**22.66**	**50.57**

Exemplar boxes   GT boxes   Predicted boxes   Predicted density map
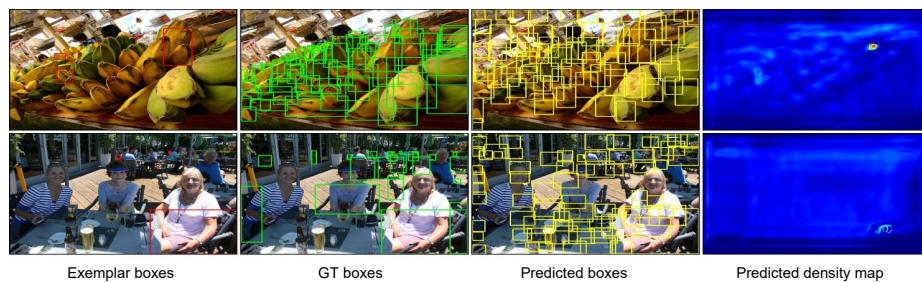
Fig. 6. Results of FamNet on the FSCD-LVIS dataset. The objects of interest for Row 1 and 2 are bananas and chairs, respectively. FamNet fails to output good density maps for images containing objects with huge variations in shape and size.

on the whole novel classes together (see Sec. 2). It is different from our setting, where each novel class is processed independently in a separate image. Fam-Net+RR is a method that uses Ridge Regression on top of the density map predicted by FamNet as depicted in Fig. 2a. FamNet+MLP is similar to Fam-Net+RR but replaces the ridge regression with a two-layer MLP with the Layer norm. Attention-RPN and FSDetView are detection-based methods, which require GT bounding boxes for all objects to train, thus, we generate the pseudo GT bounding boxes using either (1) the FamNet+RR with the features extracted from the dot annotations of training images instead of peak locations (called RR box) or (2) our first stage of training as described in Sec. 3.3 (called pseudo box). Tab. 7 and Tab. 8 show the comparison on the test sets of FSCD-147 and FSCD-LVIS, respectively.

Table 8. Comparison with strong baselines on the FSCD-LVIS test set

Method	Counting				Detection	
	MAE (↓)	RMSE (↓)	NAE(↓)	SRE (↓)	AP(↑)	AP50(↑)
FamNet [32]+RR	60.53	84.00	1.82	14.58	0.84	2.04
Attention-RPN [6]+RR box	61.31	64.10	1.02	6.94	3.28	9.44
FSDetView [45]+RR box	26.81	33.18	0.56	4.51	1.96	6.70
Attention-RPN [6]+pseudo box	62.13	65.16	1.07	7.21	4.08	11.15
FSDetView [45]+pseudo box	24.89	31.34	0.54	4.46	2.72	7.57
Counting-DETR	**18.51**	**24.48**	**0.45**	**3.99**	**4.92**	**14.49**

Table 9. Comparison on FSCD-LVIS with unseen test classes

Method	Counting				Detection	
	MAE (↓)	RMSE (↓)	NAE(↓)	SRE (↓)	AP(↑)	AP50(↑)
FamNet [32]+RR	68.45	93.31	2.34	17.41	0.07	0.30
Attention-RPN [6]+RR box	35.55	42.82	1.21	7.47	2.52	7.86
FSDetView [45]+RR box	28.56	39.72	0.73	4.88	0.89	2.38
Attention-RPN [6]+pseudo box	39.16	46.09	1.34	8.18	3.15	7.87
FSDetView [45]+pseudo box	28.99	40.08	0.75	4.93	1.03	2.89
Counting-DETR	**23.50**	**35.89**	**0.57**	**4.17**	**3.85**	**11.28**

On FSCD-147, our method significantly outperforms others with a large margin for object detection. For counting, compared to a density-based approach like FamNet, Counting-DETR achieves worse results in RMSE metric but with much better results in other counting metrics MAE, NAE, and SRE. FamNet+MLP seems to overfit to the exemplar boxes so it performs the worst in detection.

On FSCD-LVIS, our method outperforms all others for both detection and counting tasks. This is because the image in FSCD-LVIS is much more complicated than those in FSCD-147, i.e., multiple object classes per image and significant differences in object size and shape. Also, the class of interest is usually packed and occluded by other classes, so the density map cannot be reliably predicted as shown in Fig. 6. More interestingly, we also evaluate the performance of Counting-DETR and other baselines on a special test set of unseen classes of the FSCD-LVIS dataset to show their generalizability to unseen classes during training in Tab. 9. It can be seen that our approach performs the best while the FamNet+RR performs the worst.

Fig. 7 shows the qualitative comparison between our approach and the other methods, including FSDetView [45], Attention-RPN [6], and FamNet [32]+RR. Our method can successfully detect the objects of interest while other methods cannot, as shown in the first four rows of Fig. 7. The last row is a failure case for all methods, due to object truncation, perspective distortion, and scale variation.

Exemplar boxes   FSDetView   Attention RPN   FamNet+RR   Counting-DETR
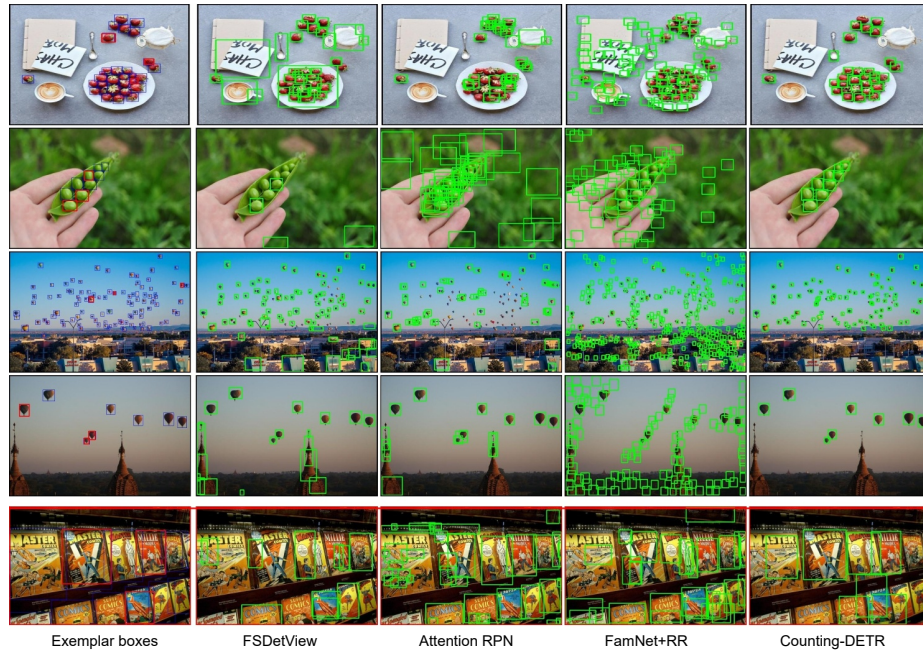
Fig. 7. Qualitative comparison. Each row shows an example with the exemplars (red) and GT bounding boxes (blue) on the first column. The first four rows show the superior performance of ours over the others, while the last row is a failure case. In the first row, our method can distinguish between the target class and other foreground classes, while other methods confuse between different foreground classes. In the next three rows, exemplar objects either share similar color with environment or contain many background pixels. These conditions lead to either under detect or over detect where other methods are unsure about if detected objects are foreground or background. The last row shows a failure case for all methods due to the huge variation in the scale, distortion, and truncation of the objects.

6   Conclusions

We have introduced a new task of few-shot object counting and detection that shares the same supervision with few-shot object counting but additionally predict object bounding boxes. To address this task, we have collected two new datasets, adopted a two-stage training strategy to generate pseudo bounding boxes for training, and developed a new uncertainty-aware few-shot object detector to adapt to the imperfection of pseudo label. Extensive experiments on the two datasets demonstrate that the proposed approach outperforms strong baselines adapted from few-shot object counting and few-shot object detection.

References

1. Abousamra, S., Hoai, M., Samaras, D., Chen, C.: Localization in the crowd with topological constraints. In: Proceedings of AAAI Conference on Artificial Intelligence (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), discrete Graphical Models in Biomedical Image Analysis 3

3. 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 (ECCV). Springer (2020) 5, 6

4. Chen, H., Wang, Y., Wang, G., Qiao, Y.: Lstd: A low-shot transfer detector for object detection. In: Proceedings of the AAAI conference on artificial intelligence (2018) 5

5. Chen, L., Yang, T., Zhang, X., Zhang, W., Sun, J.: Points as queries: Weakly semi-supervised object detection by points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8823–8832 (2021) 2, 5, 11

6. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 5, 11, 12, 13

7. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized few-shot object detection without forgetting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 5

8. Fang, Y., Zhan, B., Cai, W., Gao, S., Hu, B.: Locality-constrained spatial transformer network for video crowd counting. arXiv preprint arXiv:1907.07911 (2019) 3

9. Feng, D., Rosenbaum, L., Timm, F., Dietmayer, K.: Labels are not perfect: Improving probabilistic object detection via label uncertainty. arXiv preprint arXiv:2008.04168 (2020) 5

10. Feng, D., Wang, Z., Zhou, Y., Rosenbaum, L., Timm, F., Dietmayer, K., Tomizuka, M., Zhan, W.: Labels are not perfect: Inferring spatial uncertainty in object detection. IEEE Transactions on Intelligent Transportation Systems (2021) 5

11. Goldman, E., Herzig, R., Eisenschtat, A., Goldberger, J., Hassner, T.: Precise detection in densely packed scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3

12. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) (2019) 9

13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2016) 10

14. He, Y., Zhu, C., Wang, J., Savvides, M., Zhang, X.: Bounding box regression with uncertainty for accurate object detection. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition (CVPR) (2019) 5, 11

15. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal networks. In: The IEEE International Conference on Computer Vision (ICCV). IEEE (2017) 3

16. Hu, D., Mou, L., Wang, Q., Gao, J., Hua, Y., Dou, D., Zhu, X.X.: Ambient sound helps: Audiovisual crowd counting in extreme conditions. arXiv preprint (2020) 3
17. 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: Proceedings of the european conference on computer vision (ECCV) (2018) 3
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning(ICML). PMLR (2015) 10
19. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 5
20. Lee, Y., Hwang, J.w., Kim, H.I., Yun, K., Kwon, Y.: Localization uncertainty estimation for anchor-free object detection. arXiv preprint arXiv:2006.15607 (2020) 5
21. Lian, D., Li, J., Zheng, J., Luo, W., Gao, S.: Density map regression guided detection network for rgb-d crowd counting and localization. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 3
22. Lin, H., Hong, X., Wang, Y.: Object counting: You only need to look at one. arXiv preprint arXiv:2112.05993 (2021) 4
23. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2017) 5, 7
24. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 3
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 10
26. Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: Asian Conference on Computer Vision (ACCV) (2018) 4
27. 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: European conference on computer vision (ECCV). Springer (2016) 3
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NeurIPS) 32, pp. 8024–8035. Curran Associates, Inc. (2019) 10
29. Peng, D., Sun, Z., Chen, Z., Cai, Z., Xie, L., Jin, L.: Detecting heads using feature refine net and cascaded multi-scale architecture. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE (2018) 3
30. Ranjan, V., Hoai, M.: Vicinal counting networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4221–4230 (2022) 4
31. Ranjan, V., Le, H., Hoai, M.: Iterative crowd counting. In: Proceedings of European Conference on Computer Vision (ECCV) (2018) 3
32. 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 (CVPR) (2021) 2, 3, 4, 8, 10, 11, 12, 13

33. Ranjan, V., Wang, B., Shah, M., Hoai, M.: Uncertainty estimation and sample selection for crowd counting. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2020) 3

34. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems (NeurIPS) **28** (2015) 5

35. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019) 7

36. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015) 4

37. Sindagi, V.A., Yasarla, R., Patel, V.M.: Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 3

38. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV) (2019) 5

39. 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 (NeurIPS (2017) 5

40. Wang, B., Liu, H., Samaras, D., Hoai, M.: Distribution matching for crowd counting. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 3

41. Wang, Q., Gao, J., Lin, W., Li, X.: Nwpu-crowd: A large-scale benchmark for crowd counting and localization. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **43**(6), 2141–2149 (Jun 2021) 3

42. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957 (2020) 5

43. Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor detr: Query design for transformer-based detector. arXiv preprint arXiv:2109.07107 (2021) 2, 5, 6, 7, 10

44. Wu, J., Liu, S., Huang, D., Wang, Y.: Multi-scale positive sample refinement for few-shot object detection. In: European Conference on Computer Vision (ECCV). Springer (2020) 5

45. Xiao, Y., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. In: European Conference on Computer Vision (ECCV). Springer (2020) 5, 11, 12, 13

46. 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) 3

47. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn : Towards general solver for instance-level low-shot learning. In: Proc. of IEEE International Conference on Computer Vision (ICCV) (2019) 5

48. You, Z., Yang, K., Luo, W., Lu, X., Cui, L., Le, X.: Iterative correlation-based feature refinement for few-shot counting. arXiv preprint arXiv:2201.08959 (2022) 4

49. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 3

50. Zhang, G., Luo, Z., Cui, K., Lu, S.: Meta-detr: Few-shot object detection via unified image-level meta-learning. arXiv preprint arXiv:2103.11731 (2021) 5
51. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3
52. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. In: arXiv preprint arXiv:1904.07850 (2019) 5