Learning to Count Anything: Reference-less Class-agnostic Counting with Weak Supervision

Michael A. Hobley, Victor A. Prisacariu
Active Vision Lab, University of Oxford
{mahobley,victor}@robots.ox.ac.uk

Abstract

Class-agnostic counting methods enumerate objects of an unseen class. However, most require reference images to define the type of object to be counted. Reference-less class-agnostic counting attempts to do without such examples, but current methods still require instance-level supervision and reference images during training. In this paper, we present the first class-agnostic counting method which needs neither instance annotations nor reference images at training or inference time. We demonstrate that a simple projection from a globally contextual feature space is superior to other reference-less methods and is competitive with methods that use reference images.

Project page: countinganything.active.vision

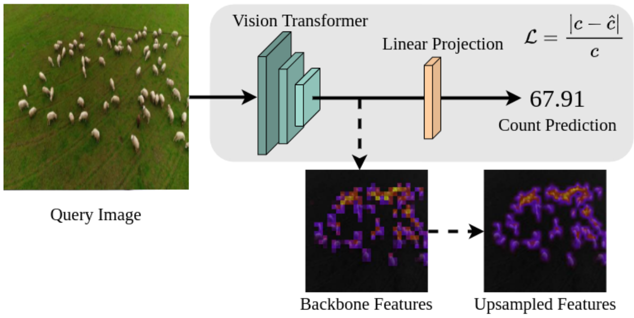

Figure 1. **The RCC pipeline.** Our method learns to count objects of novel classes without reference images, using only the ground truth count, c, as supervision. We also visualise these features to show they suffice to localise instances of the counted object.

1. Introduction

Counting is the most fundamental quantitative analysis skill. It is also one of the first abstract tasks people learn. Once learnt, the concept is simple: to find the number of instances of an object class. Significantly, whereas people can generally count objects without a prior understanding or example of the type of the object to be counted, most automated methods cannot. Most counting methods are class-specific, requiring the appearance of objects to be the same during inference-time and training. Class-agnostic methods can count previously unseen object types but generally require a reference image to specify the class to be counted and instance-wise annotations during training. Concurrent methods [19,22] have aimed to count objects of an arbitrary semantic class without the need for an inference-time reference image to define the type. However, they rely on complex pipelines and still require point-level supervision and either bounding boxes or reference images during training.

In this paper, unlike the methods described above, we use the same simple image-to-scalar pipeline during training as is available at inference time. Our method enumerates objects of an arbitrary type (**class-agnostic**), does not require example images to define the countable type during inference or training (**reference-less**), and uses only scalar labels rather than instance-level supervision (**weakly-supervised**). Class-agnostic counting methods can be broadly used on new domains and are adaptable to dynamic, real-world applications where visual appearance may not be consistent. Reference-less and weakly-supervised methods are cheaper and less laborious as they do not require ongoing human labelling.

We identify that the ability to count is fundamentally object-repetition recognition. We demonstrate that vision transformers are perfectly suited to this task because of their use of global receptive self-attention. We prove experimentally that given good globally contextual features, enumerating instances is simple, requiring only a single linear projection, a trivial loss function, and minimal supervision.

Our key contributions are as follows: i) We propose RCC, a reference-less class-agnostic counter, and show it can be trained with weak-supervision (i.e. without instance annotations), ii) We demonstrate that RCC outperforms other reference-less class-agnostic counting methods and is competitive with current methods that train with images as a prior and full point-level supervision.

2. Related Work

Counting approaches are most commonly regression-based. These methods aim to regress a single global count [7,31,32] or a pixel-level density map prediction, which can be enumerated by integration [5, 23, 33] or instance detection [2, 3, 8]. A regressed density map can also be used as a rudimentary object detector, followed by cross-correlation with a reference to find instances of the desired class [28].

Class-specific counting methods focus on enumerating the instances of a single or small set of known classes [5, 12, 14, 33]. These methods have limited to no capacity to adapt to novel classes and so require new data and retraining for each unique type of object, which is expensive, time-consuming, and difficult to gather.

To avoid these costs, Lu et al. [21] proposed class-agnostic counting, a framework in which test-time classes are not present during training. However, this and most subsequent class-agnostic methods [23, 28, 34] require a prior of object class at test time, in the form of reference images. Class-agnostic counting has so far generally been achieved by creating a sufficiently general feature space and applying some form of matching to the whole feature map [27, 34] or to proposed regions of interest [23, 28].

Concurrent works, RepRPN [22] and CounTR [19], do not require a reference image at inference time. This removes the need for intervention during deployment. RepRPN is a two-step method which proposes regions likely to contain an object of interest and then uses them for a reference-based density map regression method. CounTR uses a large vision-transformer encoder-decoder to regress a density map of instance locations. As it is trained in a mixed few/zero-shot way, applying understanding gained from reference-based examples to reference-less cases, we group it with other methods that use references during training. As RepRPN and CounTR use density map regression, they require instance-wise supervision during training. Since reference-less class-agnostic method encounter an ambiguity of what to count in any given image with multiple types, they are only suited to single-class environments.

Weak Supervision. Current class-agnostic counting methods aim to localise instances before enumerating them, requiring expensive to gather point-level annotations for supervision. Weakly-supervised counting methods aim to generate an accurate scalar count with minimal [17, 26] or no point-level annotations [4, 31, 35]. Prior to this work, there are no weakly-supervised class-agnostic methods.

3. Method

While RepRPN [22] and CounTR [19] address the problem of uninformed, generalised object counting, they require point-level annotations at training time and example images or bounding boxes as supervision, which are costly to gather. In this paper, we show that reference-less counting is simple enough that it can be trained in the same pipeline as is available at inference time, without point-level annotations or example images. We recognise the problem of uninformed, generalised object counting is at its core a repetition recognition task: 'Is this object repeated elsewhere?'. This requires understanding the global context of an image, to which vision transformers, especially their use of attention, are perfectly suited. Given a feature space that is general and globally aware, we find that regressing a count is simple and requires only a single linear projection without point-level supervision and with minimal training.

Vision transformer backbone. The crucial global context afforded by transformers stems from their use of an attention mechanism. Attention generates a new feature from linear projections of all other features based on their similarity. Our vision transformer, g, inspired by Vaswani et al. [30], uses multiple attention heads to generate informative, diverse features with a global receptive field. We use a small vision transformer backbone (ViT-S) inspired by DeiT-S [29]. This was chosen due to its efficiency and to provide a good comparison to other counting methods. ViT-S has a similar number of parameters (21M vs 23M), throughput (1237im/sec vs 1007im/sec), and supervised ImageNet performance (79.3% vs 79.8%) as ResNet-50 [29], which is used by many contemporary counting methods. It should be noted that this vision transformer configuration limits the resolution of our input image to (224×224) as opposed to the ∼(384×384) of most ResNet-based methods.

As even data efficient transformers require large amounts of data to train and the datasets on which we evaluate are small, we initialised our transformer backbone, using weights from Caron et al. [6]. This self-supervised pre-training gives the network an understanding of meaningful image features without supervision or exposure to our limited datasets, minimising the chance of overfitting.

Count Regression. We directly regress a single, scalar count prediction, $\hat{c}$, for the whole input image, x, from the latent features of $g(x)$, using a learnt linear projection, F, such that $\hat{c} = F{\cdot}g(x)$. Although it may seem that a location-based loss should aid in training, we found that point-level annotations and Gaussian density maps were unnecessary and often detrimental in a class-agnostic setting. Since positional annotations are often arbitrarily placed and contain, at most, limited information about the size and shape of an object, identifying different parts of all correct objects would be punished by a positional loss function. This arbitrary punishment hinders the network's understanding of the counting task. We allow the network to develop its own conceptual representation of the task by regressing an estimate for the count directly. To this end, we use one of the simplest possible loss functions, the absolute percentage error, defined as: $\mathcal{L} = \text{Absolute Percentage Error} = |c - \hat{c}|/c$, where c and $\hat{c}$ are the ground truth and the predicted count respectively. This was a slight improvement over Absolute Error as it limited the disproportionate effect of very high-density images on the gradients. We found that the network, without the imposition of human ideas of positional salience, learnt to implicitly localise instances of the counted class in a meaningful way, as seen in Figure 2 and further discussed in Section 4.

Our linear count projection, projects from a $p \times d$ feature space to a scalar count prediction, where d is the dimensionality of the transformer features and p is the number of patches of the vision transformer. We found $d = 384$ and $p = 28^2$ sufficient to achieve competitive results while also being lightweight enough to be trainable on a single 1080Ti.

Tiling Augmentation. In-order to improve our method's ability to enumerate high-density problems without requring a complex spatial loss [24] or non-maximal suppression [25], we use a tiling augmentation. This tiles resized versions of the input image into a (2×2) grid for 50% of instances. This augmentation improved our methods MAE and RMSE by about 10% and 20% respectively.

4. Results

In this section, we show that RCC dramatically outperforms other reference-less methods and is competitive with reference-based methods. We demonstrate RCC's generalisability by applying it to another visually distinct domain. As it may be of real-world utility, we then show that our features are sufficient to localise the instances in an image. Finally, we validate specific components of our method.

Benchmarking Methods. We evaluate our method against two trivial baseline methods (predicting the training-set mean or median count), two few-shot detection methods (FR [16], FSOD [10]), six reference-based class-agnostic counting methods (GMN [21], MAML [11], FamNet [23], CFOCNet [34], BMNet [27], LaoNet [18]), and two methods that do not need reference images at inference time (RepRPN [22], CounTR [19]), see Table 1. We also include reference-less modifications to reference-based methods, denoted by a *. To ensure fair comparison, we also train competitive methods using the same backbone as our method, ViT-S, denoted by † in the tables. As in previous works [22,23,27], we use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to evaluate performance.

FSC-147: Few and Zero-Shot Comparisons. We significantly outperform RepRPN and all reference-less versions of reference-based methods on FSC-147 [23] without using point-level annotations. We achieve comparable results to zero-shot CounTR with its original ViT-B backbone, which is significantly larger ($\sim 8\times$ the parameters) than ours and is trained using reference images. We significantly outperform it when it is trained with the same backbone as RCC or when it is trained without reference images. We are also competitive with previous few-shot or reference-based methods on FSC-147 without using reference images, point-level annotations, or adaptations during training or inference, see Table 1. This result demonstrates that the combination of an architecture well-suited for counting and a simple, count-centred objective function can learn a meaningful conceptual representation of counting without any location-based information. We found that our method, like others, has poor performance on images with over 1000 objects. This is likely due to images of this density making up only 0.16% of the training set, as well as the low patch resolution of our method not being able to separate them. When the images over 1000 were removed (0.39% of the validation set and 0.17% of the test set), the MAE improved by 15.9% and 19.4% respectively.

We found that methods that were modified to use ViT-S in place of their ResNet-50 backbones performed worse; this is likely due to the significant decrease in input image resolution. The decrease in performance of CounTR when replacing ViT-Base backbone with ViT-S is unsurprising as it has decreased latent feature dimensionality, fewer transformer blocks, and lower input image resolution.

CARPK: Cross-Dataset Generalisability. To confirm our cross-dataset generalisability, we test our model on CARPK [15], a dataset comprised of birds-eye views of parking lots which significantly differ from the appearance of any objects in FSC-147. We exclude the "car" class from our pre-training. We significantly outperform FamNet and are competitive with BMNet and CounTR (0-shot).

Feature Visualisation. We show that our method is able to localise instances of the counted objects in Figure 2. We achieve this by either taking most significant value from a singular value decomposition of the counting features weighted by the linear projection weights or by using a trained localisation head. This head, comprised of 3 Conv-ReLU-Upsample blocks, generates a pixel-wise density map prediction from a frozen backbones patch-wise features. This backbone and localisation head trained on FSC-147 is generalisable to other domains like those in the CARPK and ShanghaiTech [36] datasets, see Figure 2.

Ablation Studies. We demonstrate the importance of the global context provided by vision transformer by evaluating RCC with ResNet-50 [13] and ConvNeXt [20] backbones

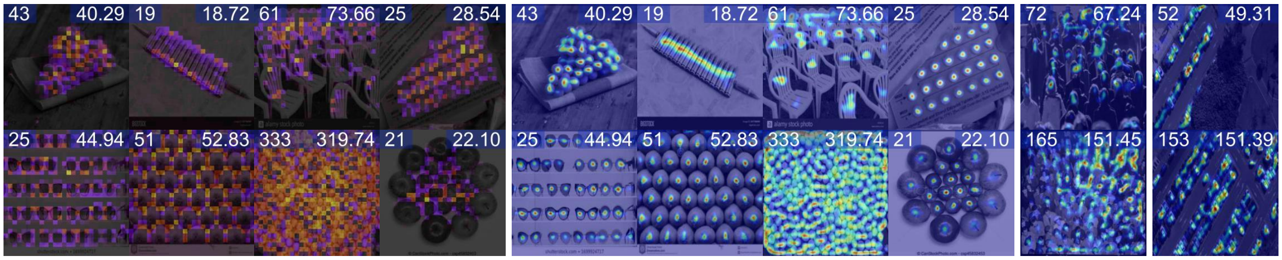

Figure 2. **Localisation of latent counting features.** Columns 1-4: Patch-wise principal component of the counting backbone features, Columns 5-8, 9, 10: Pixel-wise localisation predictions of unseen dataset classes from FSC-147, CARPK, and ShanghaiTech respectively. The count and prediction of each image are in the top left and the top right respectively.

Method			Val Set		Test Set	
			MAE	RMSE	MAE	RMSE
Mean			53.38	124.53	47.55	147.67
Median			48.68	129.7	47.73	152.46
Reference-based training						
GMN	(1-shot)	2018	29.66	89.81	26.52	124.57
FamNet	(3-shot)	2021	24.32	70.94	22.56	101.54
FamNet+	(3-shot)	2021	23.75	69.07	22.08	99.54
FamNet†	(3-shot)	2021	37.90	109.76	33.51	137.79
FamNet+†	(3-shot)	2021	37.77	109.04	33.18	137.20
CFOCNet	(3-shot)	2021	21.19	61.41	22.10	112.71
LaoNet	(1-shot)	2021	17.11	56.81	15.78	97.15
BMNet	(3-shot)	2022	19.06	67.95	16.71	103.31
BMNet+	(3-shot)	2022	15.74	58.53	14.62	91.83
BMNet†	(3-shot)	2022	19.29	68.58	18.34	115.31
BMNet+†	(3-shot)	2022	17.21	60.18	16.90	107.66
CounTR	(3-shot)	2022	**13.13**	**49.83**	**11.95**	**91.23**
CounTR†	(3-shot)	2022	22.10	73.08	19.52	111.50
CounTR	(0-shot)	2022	**17.40**	**70.33**	**14.12**	**108.01**
CounTR†	(0-shot)	2022	28.96	90.66	25.58	110.39
Reference-less training						
MAML*	(0-shot)	2017	32.44	101.08	31.47	129.31
GMN*	(0-shot)	2018	39.02	106.06	37.86	141.39
FamNet*	(0-shot)	2021	32.15	98.75	32.27	131.46
RepRPN	(0-shot)	2022	29.24	98.11	26.66	129.11
CounTR†*	(0-shot)	2022	33.15	100.80	28.95	125.31
RCC(ours)	(0-shot)		**17.49**	**58.81**	**17.12**	**104.53**

Table 1. **Comparison to SOTA methods on FSC-147.** We outperform other reference-less methods and are competitive with methods which use reference images and test-time adaptation. † denotes methods trained using the same backbone as ours. * indicates a reference-less modification of a reference-based method.

Method		MAE	RMSE
Mean		65.63	72.26
Median		67.88	74.58
Reference-based training			
FamNet	(3-shot)	28.84	44.47
BMNet	(3-shot)	14.61	24.60
BMNet+	(3-shot)	10.44	13.77
CounTR	(3-shot)	**10.20**	**12.67**
CounTR	(0-shot)	11.33	14.60
Reference-less training			
RCC (ours)	(0-shot)	**12.31**	**15.40**

Table 2. **Generalisation performance on CARPK.** Models were trained on FSC-147 images.

Backbone	Projection		Simple head		Complex head	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
ResNet-50	24.99	75.87	26.51	110.36	24.02	89.31
ConvNeXt	21.58	87.21	23.07	111.70	24.60	134.66
ViT-Small	**14.23**	**43.83**	17.46	78.60	15.22	52.24

Table 3. **Performance of different feature backbones and count regression heads on the test-set of FSC-133 [1].**

in place of our transformer. We initialised both architectures with standard weights pre-trained on ImageNet [9]. This gives a significant unfair advantage to these backbones but acts as a useful best-case for these architectures. When trained comparably, Resnet-50 has a similar accuracy and ConvNeXt has a higher classification accuracy than our architecture [20]. As seen in Table 3, even with the advantageous supervised pre-training, the attention-less features perform significantly worse than the self-supervised vision transformer. Table 3 also compares our scalar projection with simple (1 conv & 2 linear layers) and more complex (4 conv & 3 linear layers) counting head architectures. The convolutional heads perform worse than our projection on all three backbones. It appears that their extra computational capacity is not only unnecessary but is in fact detrimental because it overfits to the training classes.

5. Conclusion

In this work, we present RCC, the first reference-less class-agnostic counting method that does not require training time reference images and that can be trained without point-level annotations. We show that, even without the point-level supervision, RCC is superior to other reference-less methods and is competitive with reference-based class-agnostic counting approaches. RCC is cheaper and less laborious to scale than other methods as it does not require the same ongoing human labelling.

References

[1] Anonymous. FSC-133: FSC-147 without errors, ambiguities, and repeated images. *arXiv preprint*, 2022. 4

[2] Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Interactive object counting. In *European conference on computer vision*, pages 504–518. Springer, 2014. 2

[3] Olga Barinova, Victor Lempitsky, and Pushmeet Kholi. On detection of multiple object instances using hough transforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1773–1784, 2012. 2

[4] Matthias von Borstel, Melih Kandemir, Philip Schmidt, Madhavi K Rao, Kumar Rajamani, and Fred A Hamprecht. Gaussian process density counting from weak supervision. In *European Conference on Computer Vision*, pages 365–380. Springer, 2016. 2

[5] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2

[7] Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *2009 IEEE 12th international conference on computer vision*, pages 545–551. IEEE, 2009. 2

[8] Hisham Cholakkal, Guolei Sun, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Luc Van Gool. Towards partial supervision for generic object counting in natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[10] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020. 3

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 3

[12] Hyojun Go, Junyoung Byun, Byeongjun Park, Myung-Ae Choi, Seunghwa Yoo, and Changick Kim. Fine-grained multi-class object counting. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 509–513. IEEE, 2021. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[14] Jeroen Hoekendijk, Benjamin Kellenberger, Geert Aarts, Sophie Brasseur, Suzanne SH Poiesz, and Devis Tuia. Counting using deep learning regression gives value to ecological surveys. *Scientific reports*, 11(1):1–12, 2021. 2

[15] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017. 3

[16] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019. 3

[17] Yinjie Lei, Yan Liu, Pingping Zhang, and Lingqiao Liu. Towards using count-level weak supervision for crowd counting. *Pattern Recognition*, 109:107616, 2021. 2

[18] Hui Lin, Xiaopeng Hong, and Yabin Wang. Object counting: You only need to look at one. *arXiv preprint arXiv:2112.05993*, 2021. 3

[19] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721*, 2022. 1, 2, 3

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 3, 4

[21] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *Asian Conference on Computer Vision*, 2018. 2, 3

[22] Viresh Ranjan and Minh Hoai. Exemplar free class agnostic counting. *arXiv preprint arXiv:2205.14212*, 2022. 1, 2, 3

[23] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021. 2, 3

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[26] Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, and R Venkatesh Babu. Almost unsupervised learning for dense crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8868–8875, 2019. 2

[27] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9529–9538, 2022. 2, 3

[28] Negin Sokhandan, Pegah Kamousi, Alejandro Posada, Eniola Alese, and Negar Rostamzadeh. A few-shot sequential approach for object counting. *arXiv preprint arXiv:2007.01899*, 2020. 2

[29] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[31] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1299–1302, 2015. 2

[32] Meng Wang and Xiaogang Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *CVPR 2011*, pages 3401–3408. IEEE, 2011. 2

[33] Weidi Xie, J Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, 6(3):283–292, 2018. 2

[34] Shuo-Diao Yang, Hung-Ting Su, Winston H Hsu, and Wen-Chin Chen. Class-agnostic few-shot object counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 870–878, 2021. 2, 3

[35] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 2

[36] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 3