DAVE – A Detect-and-Verify Paradigm for Low-Shot Counting

Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, Matej Kristan
Faculty of Computer and Information Science, University of Ljubljana, Slovenia
jer.pelhan@fri.uni-lj.si

Abstract

Low-shot counters estimate the number of objects corresponding to a selected category, based on only few or no exemplars annotated in the image. The current state-of-the-art estimates the total counts as the sum over the object location density map, but does not provide individual object locations and sizes, which are crucial for many applications. This is addressed by detection-based counters, which, however fall behind in the total count accuracy. Furthermore, both approaches tend to overestimate the counts in the presence of other object classes due to many false positives. We propose DAVE, a low-shot counter based on a detect-and-verify paradigm, that avoids the aforementioned issues by first generating a high-recall detection set and then verifying the detections to identify and remove the outliers. This jointly increases the recall and precision, leading to accurate counts. DAVE outperforms the top density-based counters by ∼20% in the total count MAE, it outperforms the most recent detection-based counter by ∼20% in detection quality and sets a new state-of-the-art in zero-shot as well as text-prompt-based counting. The code and models are available on GitHub.

1. Introduction

Low-shot counting considers estimating the number of target objects in an image, based only on a few annotated exemplars (few-shot) or even without providing the exemplars (zero-shot). Owing to the emergence of focused benchmarks [22, 26], there has been a surge in low-shot counting research recently. The current state-of-the-art low-shot counters are all density-based [6, 26, 28, 38]. This means that they estimate the total count by summing over an estimated object presence density map. Only recently, few-shot detection-based methods emerged [22] that estimate the counts as the number of detected objects.

Density-based methods substantially outperform the detection-based counters in total count estimation, but they do not provide detailed outputs such as object locations and sizes. The latter are however important in many downstream tasks such as bio-medical analysis [35, 41], where explainability is crucial for human expert verification as well as for subsequent analyses. There is thus a large applicability gap between the density-based and detection-based low-shot counters.

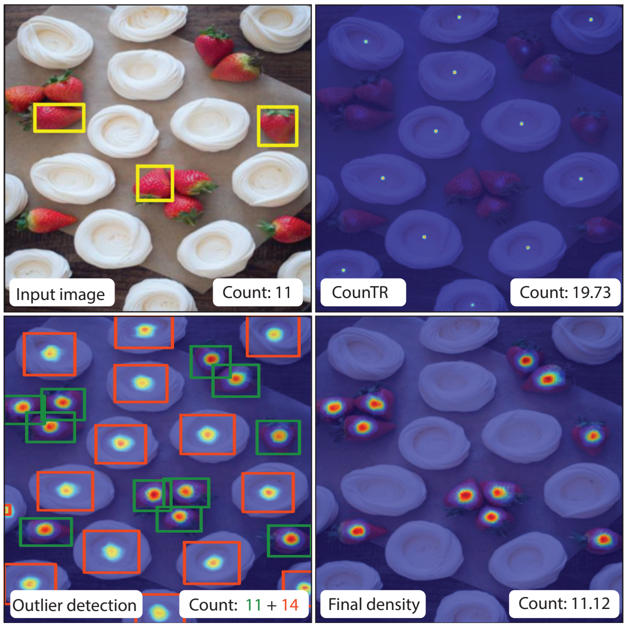

Figure 1. Despite considering exemplars (yellow boxes), the state-of-the-art (e.g., CounTR [16]) is prone to false activations on incorrect objects, leading to corrupted counts. DAVE avoids this issue by *detecting* all candidates (red and green boxes), *verifying* them, removing the outliers (red boxes), and correcting the final density map, thus jointly improving detection and count estimation.

Furthermore, both density-based and detection-based counters are prone to failure in scenes with several object types (Figure 1). The reason lies in the specificity-generalization tradeoff. Obtaining a high recall requires generalizing over the potentially diverse appearances of the selected object type instances in the image. However, this also leads to false activations on objects of other categories (false positives), leading to a reduced precision and count overestimation. A possible solution is to train on multiple-class images [22], however, this typically leads to a reduced recall and underestimated counts.

We address the aforementioned issues by proposing a low-shot counter DAVE, which combines the benefits of density-based and detection-based formulations, and introduces a novel detect-and-verify paradigm. DAVE tackles the specificity-generalization issues of the existing counters by applying a two-stage pipeline (Figure 1). In the first, *detection* stage, DAVE leverages density-based estimation to obtain a high-recall set of candidate detections, which however may contain false positives. This is addressed by the second, *verification* stage, where outliers are identified and rejected by analyzing the candidate appearances, thus increasing the detection precision. Regions corresponding to the outliers are then removed from the location density map estimated in the first stage, thus improving the density-based total count estimates as well. In addition, we extend DAVE to text-prompt-based and to a zero-shot scenario, which makes DAVE the first zero-shot as well as text-prompt detection-capable counter.

The primary contribution of the paper is the detect-and-verify paradigm for low-shot counting that simultaneously achieves high recall and precision. The proposed architecture is the first to extend to all low-shot counting scenarios. DAVE uniquely merges the benefits of both density and detection-based counting and is the first zero-shot-capable counter with detection output. DAVE outperforms all state-of-the-art density-based counters on the challenging benchmark [26], including the longstanding winner [6], achieving a relative 20% MAE and 43% RMSE total-count error reductions. It also outperforms all state-of-the-art detection-based counters on the recent benchmark FSCD147 [22] by ∼20% in detection metrics, as well as in the total count estimation by 38% MAE. Furthermore, it sets a new state-of-the-art in text-prompt-based counting. The zero-shot DAVE variant outperforms all zero-shot density-based counters and delivers detection accuracy on-par with the most recent *few-shot* counters. DAVE thus simultaneously outperforms both density-based and detection-based counters in a range of counting setups.

2. Related Work

Object counting emerged as detection-based counting of objects belonging to specific classes, such as vehicles [5], cells [8], people [17], and polyps [41]. To address poor performance in densely populated regions, density-based methods [3, 4, 29–31] emerged as an alternative.

All these methods rely on the availability of large datasets to train category-specific models, which, however are not available in many applications.

Class-agnostic approaches addressed this issue by test-time adaptation to various object categories with minimal supervision. Early representatives [19] and [37] proposed predicting the density map by applying a siamese matching network to compare image and exemplar features. Recently, the FSC147 dataset [26] was proposed to encourage the development of few-shot counting methods. Famnet [26] proposed a test-time adaptation of the backbone to improve density map estimation. BMNet+ [28] improved localization by jointly learning representation and a non-linear similarity metric. A self-attention mechanism was applied to reduce the intra-class appearance variability. SAFECount [38] introduced a feature enhancement module, improving generalization capabilities. CounTR [16] used a vision transformer [7] for image feature extraction and a convolutional encoder to extract exemplar features. An interaction module based on cross-attention was proposed to fuse both, image and exemplar features. LOCA [6] proposed an object prototype extraction module, which combined exemplar appearance and shape with an iterative adaptation.

All few-shot counting methods require few annotated exemplars to specify the object class. With the recent development of large language models (e.g. [23]) text-prompt-based counting methods emerged. Instead of specifying exemplars by bounding box annotations, these methods use text descriptions of the target object class. ZeroCLIP [36] proposed text-based construction of prototypes, which are used to select relevant image patches acting as exemplars for counting. CLIPCount [15] leveraged CLIP [23] for image-text alignment and introduced patch-text contrastive loss for learning the visual representations used for density prediction. Several works [13, 25] address the extreme case in which no exemplars are provided and the task is to count the majority class objects (i.e., zero-shot counting).

With minimal architectural changes, the recent few-shot methods [6, 16] also demonstrated a remarkable zero-shot counting performance. A common drawback of density-based counters is that they do not provide object locations.

To address the aforementioned limitation of density-based counters, the first few shot counting and detection method [22] has been recently proposed by extending a transformer-based object detector [2] with an ability to detect objects specified by exemplars. However, the detection-based counter falls far behind in total count estimation compared with the best density-based counters.

3. Counting by detection and verification

Formally, given an input image $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$ and a set of k exemplar bounding boxes $\boldsymbol{B}^{\mathrm{E}} = \{b_i\}_{i=1:k}$ denoting object exemplars, a low-shot detection counter is required to report bounding boxes $\boldsymbol{B}^{P} = \{b_i\}_{i=1:N_P}$ of all detected objects of the same category and their estimated count.

In the following we present the new detect-and-verify few-shot counting and detection method (DAVE), which consists of two stages. In the first, *detection*, stage (Section 3.1), candidate regions are estimated by pursuing a high

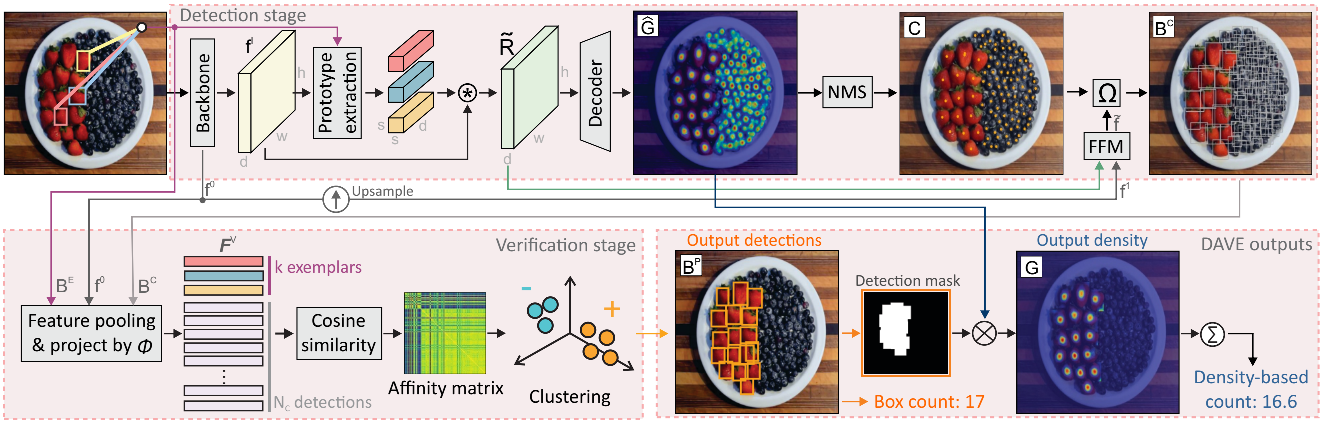

Figure 2. The proposed DAVE architecture consists of two stages, (i) *detection* and (ii) *verification*, and outputs detected objects as well as an improved location density map. NMS denotes non-maxima suppression, FFM is a feature fusion module, Ω is a bounding box regression head and ϕ is the verification feature extraction network.

recall, potentially including false positive detections, i.e., objects belonging to an incorrect category. In the second, *verification*, stage (Section 3.2) the candidate regions are analyzed to identify and reject the outliers, thus increasing the detection precision. The outliers are used to update the density map, thus also improving density-based count estimation, which may compensate for missed objects and can differ from the number of detections N_P. A detailed DAVE architecture is shown in Figure 2.

3.1. Detection stage

The aim of this stage is to predict candidate bounding boxes $\boldsymbol{B}^C = \{b_i\}_{i=1:N_c}$ with a high recall. Detection is thus split into first estimating the object centers $\boldsymbol{C} = \{(x_c^i, y_c^i)\}_{i=1:N_c}$, and then predicting the corresponding bounding box parameters. We re-purpose the architecture of the recent low-shot counter LOCA [6] for estimating the object location density map $\tilde{\mathbf{G}}$, from which we obtain the center locations $\boldsymbol{C}$ by non-maxima suppression.[1] Briefly, the location density estimation architecture is the following: the input image is first encoded by ResNet-50 [12], followed by a transformer, generating the representation $\mathbf{f}^I \in \mathbb{R}^{h \times w \times d}$. Exemplar prototypes are constructed using the OPE module [6] and correlated with $\mathbf{f}^I$ resulting in a similarity tensor $\tilde{\mathbf{R}} \in \mathbb{R}^{h \times w \times d}$. A decoder is then applied on $\tilde{\mathbf{R}}$ to obtain the final 2D location density map $\hat{\mathbf{G}} \in \mathbb{R}^{H_0 \times W_0}$. We refer the reader to [6] for more details.

Next, features are constructed for regressing the bounding box parameters for each detected center. The feature construction pipeline is designed to ensure that final features reflect the objectness information specific to the class selected by the exemplars. Features from the second, third, and fourth blocks of the backbone are resized to 64×64 pixels, concatenated along the channel dimension and reduced to d channels using a 3×3 convolution, i.e. $\mathbf{f}^0 \in \mathbb{R}^{h \times w \times d}$. The features are then upsampled to match the input image size ($\mathbf{f}^1 \in \mathbb{R}^{H_0 \times W_0 \times d}$). Next, the selected object category shape information is injected by fusing $\mathbf{f}^1$ and the upscaled similarity tensor $\tilde{\mathbf{R}}$ using the feature fusion module (FFM) [39], i.e., $\tilde{\mathbf{f}} = \mathrm{FFM}(\mathbf{f}^1, \tilde{\mathbf{R}})$.

The constructed features $\tilde{\mathbf{f}}$ are then fed into a bounding box regression head $\Omega(\cdot)$ akin to [14, 40], which predicts for each location a distance to the left, right, top and bottom bounding box edge of the underlying object. The network $\Omega(\cdot)$ consists of two 3×3 convolutional layers with d and 4 channels with GroupNorm [33] and ReLU operations in between, and predicts a dense bounding box map $\mathbf{v} \in \mathbb{R}^{H_0 \times W_0 \times 4}$. The object candidate bounding boxes $\boldsymbol{B}^C$ are thus obtained by reading out the corresponding values from $\mathbf{v}$ at locations $\boldsymbol{C}$.

3.2. Verification stage

In practice, the candidate detections $\boldsymbol{B}^C$ retain a high recall, but are also contaminated by false positives. The goal of the verification stage is thus to increase the precision by analysing the appearance of the detections and rejecting the outliers. First, a verification feature vector $\mathbf{f}_i^{\mathrm{v}}$ is extracted for each detected bounding box b_i as follows. The backbone features $\mathbf{f}^0$ are pooled into a feature tensor $\mathbf{f}_i \in \mathbb{R}^{s \times s \times d}$ and transformed by a shallow network $\phi(\cdot)$ consisting of two 1×1 convolutions with d channels and a BatchNorm and ReLu activation in between. The verification features are also extracted for the annotated exemplars, leading to $N_C + k$ features in total, i.e., $\boldsymbol{F}^{\mathrm{V}} = \{\mathbf{f}_i^{\mathrm{v}}\}_{i=1:(N_C+k)}$.

The verification features are then clustered by unsupervised clustering. Specifically, spectral clustering [21] is applied to an affinity matrix computed from cosine similarities between pairs of features in $\boldsymbol{F}^{\mathrm{V}}$, yielding several clus-

[1] The minimal distance between two peaks in NMS is set to 1 to maximize the detection rate.

ters. Object candidate detections belonging to clusters with at least one exemplar are kept, while the other are labelled as outliers and removed, yielding the final set of N_P object detections $\boldsymbol{B}^P = \{b_i\}_{i=1:N_P}$. Finally, the density map $\hat{\mathbf{G}}$ from the detection stage is updated by setting all values outside of the detected bounding boxes to zero, yielding $\mathbf{G}$, from which the improved density-based count is estimated (Figure 2).

3.3. Zero-shot and prompt-based adaptation

Zero shot counting. DAVE is easily adapted to a zero-shot setup in which exemplars are not provided and the task is to count and detect the majority-class objects. First, the location density prediction part is replaced by its zero-shot variant [6] to account for the absence of exemplars. The detection stage and most of the verification stage remain unchanged. The only change in the verification stage is the cluster selection method: all clusters whose size is at least 45%[2] of the largest cluster are kept as positive detections and the rest are identified as outliers. This is to account for the possibility that clusters may break up due to the absence of exemplars specifying the level of appearance similarity.

Prompt-based counting. Zero-shot DAVE is extended to the prompt-based counting setup, in which the target object class is specified by a text prompt. The only modification is the cluster selection protocol in the verification stage. The text prompt embedding is extracted by CLIP and compared to the CLIP embedding of each identified cluster. The latter is obtained by masking the image regions outside the bounding boxes corresponding to the cluster and computing the CLIP embedding. Cosine distances between the text embedding and individual cluster embeddings are computed, and clusters with less than 85%[2] of the highest prompt-to-cluster similarity are identified as outliers.

3.4. Training

Few-shot counting datasets typically contain centers of all objects annotated and the bounding boxes available for only $k = 3$ exemplars. We formulate the training to adhere to these restrictions. The object centers can be used to train location density prediction network. Since DAVE employs LOCA [6] for the initial density prediction, we use the publicly available pretrained version of LOCA, and train only the free parameters of the detection and verification stages in two phases.

In the first phase, the detection stage (i.e., the FFM and $\Omega(\cdot)$) is trained by a bounding box loss evaluated on the available ground truth exemplar bounding boxes, i.e., $\mathcal{L}_{box} = \sum_{i=1}^{k=3} 1 - \mathrm{GIoU}(\mathbf{v}(x^c, y^c), b_i^{\mathrm{GT}})$, where $(x_c^{(i)}, y_c^{(i)})$ are locations in the central regions of the ground truth bounding boxes b_i^{GT} and $\mathrm{GIoU}(\cdot)$ is the generalized intersection over union [27]

In the second phase, the verification feature extraction network $\phi(\cdot)$ is trained. Training examples are generated by stitching together a pair of images with annotated exemplar objects of different classes. The stitched image thus contains $2 \times 3 = 6$ bounding boxes, yielding two sets of features extracted by $\phi(\cdot)$, corresponding to the two sets of exemplars: $\{\mathbf{z}_j^1\}_{j=1:3}$ and $\{\mathbf{z}_j^2\}_{j=1:3}$. The verification network $\phi(\cdot)$ is then trained by a contrastive loss [32]:

$$\mathcal{L}_{cos} = \begin{cases} 1 - c(z^{i_1}_{j_1}, z^{i_2}_{j_2}), & i_1 = i_2 \\ \max(0, c(z^{i_1}_{j_1}, z^{i_2}_{j_2}) - \lambda), & \text{else,} \end{cases}$$

where $c(z^{i_1}_{j_1}, z^{i_2}_{j_2})$ is the cosine similarity between a pair of features, and λ is the margin.

4. Experiments

4.1. Implementation details

Preprocessing. Following [16], the input image is resized such that the mean of the exemplars width and height is between 50 and 10 pixels. In the zero-shot setup, the method is bootstrapped, with applying the first pass to estimate the object sizes and then applying the second pass with the resizing as in the few-shot case.

Training. In the first training stage, the feature fusion module FFM and the box regression head $\Omega(\cdot)$ are trained for 50 epochs by AdamW [18] with learning rate 10^{-4}, weight decay 10^{-4} and batch size 8. The size of input images is kept fixed ($H_0 = W_0 = 512$) by zero-padding. In the second stage, the verification feature extraction network $\phi(\cdot)$ is trained for 50 epochs by AdamW [18] with the learning rate 10^{-5}, the weight decay 10^{-4}, and the batch size 64.

4.2. Density-based counting performance

DAVE is compared with the density-based state-of-the-art counters. For consistent comparison, density-based count estimation is considered in DAVE as well, i.e., the count is estimated by summation of the output location density map $\mathbf{G}$ (Section 3.2). The methods are evaluated on the challenging FSC147 [26], which contains 6135 images of 147 object classes, split into 3659 training, 1286 validation and 1190 test images. The object classes are disjoint across the splits to reflect realistic applications where the target object category is unseen during training. In each image, three exemplars are annotated with bounding boxes and all target objects by point annotations. The standard evaluation protocol [26, 28, 38] with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) is followed.

Few-shot counting. In few-shot counting, all three exemplars are considered as the input. DAVE is compared with the most recent state-of-the-art density-based counters: LOCA [6], CounTR [16], SAFECount [38], BMNet+ [28], VCN [24], CFOCNet [37], MAML [10], FamNet [26], and CFOCNet [37]. Results are summarized in Table 1.

[2] Extensive analysis shows robustness to this hyperparameter.

DAVE outperforms all few-shot density-based counters by a large margin. It outperforms the long-standing winner LOCA [6] by 13% and 20% in MAE on validation and test sets, respectively. It achieves a relative improvement of 14% and a remarkable 43% RMSE on the validation and test sets, respectively, setting a solid new state-of-the-art.

Table 1. Few-shot density-based counting on the FSC147 [26].

	Validation set		Test set	
Method	MAE	RMSE	MAE	RMSE
GMN [20]	29.66	89.81	26.52	124.57
MAML [10]	25.54	79.44	24.90	112.68
FamNet [26]	23.75	69.07	22.08	99.54
CFOCNet [37]	21.19	61.41	22.10	112.71
BMNet+ [28]	15.74	58.53	14.62	91.83
VCN [24]	19.38	60.15	18.17	95.60
SAFECount [38]	15.28	47.20③	14.32	85.54③
CounTR [16]	13.13③	49.83	11.95③	91.23
LOCA [6]	10.24②	32.56②	10.79②	56.97②
DAVE	8.91①	28.08①	8.66①	32.36①

To verify the source of performance improvements, we visualize DAVE density predictions and compare them with the recent state-of-the-art methods (Figure 3). We observe that other methods often count objects of an incorrect category (columns 1, 2, 3, 4, 5, 6, 7) or structures in the background texture (columns 8, 9, 10). This indicates that related methods over-generalize localization features, which increases the recall at the cost of reduced precision. DAVE, however, retains the high recall, while successfully identifying the outliers and suppressing the corresponding activations in the density map, thus improving precision. This indicates the strong benefits of the proposed detect-and-verify paradigm for density-based counting.

One-shot counting. In the one-shot counting setup, a single exemplar is considered. Comparison with the recent state-of-the-art methods GMN [20], CFOCNet [37], FamNet [26], BMNet+ [28], CounTR [16], and LOCA [6] is reported in Table 2. DAVE excels in one-shot counting, surpassing the previous best-performing methods by 5% and 10% MAE, and 9% and 12% RMSE on the validation and test set, respectively. Results indicate that the detect-and-verify paradigm helps to fully utilize the meaningful information from the only available exemplar, leading to performance improvements.

Prompt-based counting. The prompt-based modification of DAVE from Section 3.3 (denoted here as DAVE$_{\text{prm}}$) is compared with the recent state-of-the-art prompt-based counters ZeroClip [36], CounTX [1] and CLIP-Count [15]. Results in Table 3 show that DAVE$_{\text{prm}}$ outperforms the best counter (CounTX [1]) by 12% and 5% MAE and 17% and 3% RMSE on validation and test sets, respectively. DAVE thus sets a solid new state-of-the-art in this setup.

Table 2. One-shot density-based counting on the FSC147 [26].

	Validation set		Test set	
Method	MAE	RMSE	MAE	RMSE
GMN [20]	29.66	89.81	26.52	124.57
CFOCNet [37]	27.82	71.99	28.60	123.96
FamNet [26]	26.55	77.01	26.76	110.95
BMNet+ [28]	17.89	61.12	16.89	96.65
CounTR [16]	13.15③	49.72 ③	12.06②	90.01③
LOCA [6]	11.36②	38.04②	12.53③	75.32②
DAVE	10.79①	34.55①	11.29①	66.36①

Table 3. Prompt-based counting on the FSC147 [26].

	Validation Set		Test Set	
Method	MAE	RMSE	MAE	RMSE
ZeroClip [36]	26.93	88.63	22.09	115.17
CLIP-Count [15]	18.79③	61.18②	17.78③	106.62②
CounTX [1]	17.70②	63.61③	15.73②	106.88③
DAVE$_{\text{prm}}$	15.48①	52.57①	14.90①	103.42①

Zero-shot counting. The zero-shot modification of DAVE from Section 3.3 (denoted here as DAVE$_{\text{0-shot}}$) is compared with the best zero-shot counters LOCA [6], CounTR [16], RepRPN-C [25] and RCC [13]. The results in Table 4 show that DAVE$_{\text{0-shot}}$ outperforms the state-of-the-art method LOCA [6], by a significant margin of 11% and 7% MAE on validation and test set, respectively, and outperforms all state-of-the-art in RMSE.

Table 4. Zero-shot density-based counting on the FSC147 [26].

	Validation Set		Test Set	
Method	MAE	RMSE	MAE	RMSE
RepRPN-C [25]	29.24	98.11	26.66	129.11
RCC [13]	17.49	58.81③	17.12	104.53③
CounTR [16]	17.40②	70.33	14.12①	108.01
LOCA [6]	17.43③	54.96②	16.22③	103.96②
DAVE$_{\text{0-shot}}$	15.54①	52.67①	15.14②	103.49①

4.3. Detection performance

Few-shot detection. Few-shot detectIon performance is evaluated on the FSCD147 [22] dataset, which has been recently extended from FSC147 [26] by annotating all objects with bounding boxes. We follow the standard evaluation protocol [22] with Average Precision (AP) and Average Precision at IoU=50 (AP50) as the main performance measures. DAVE is compared with the most recent few-shot detection-based counter C-DETR [22] as well as adapted few-shot detectors FSDetView [34], AttRPN [9] from [22].

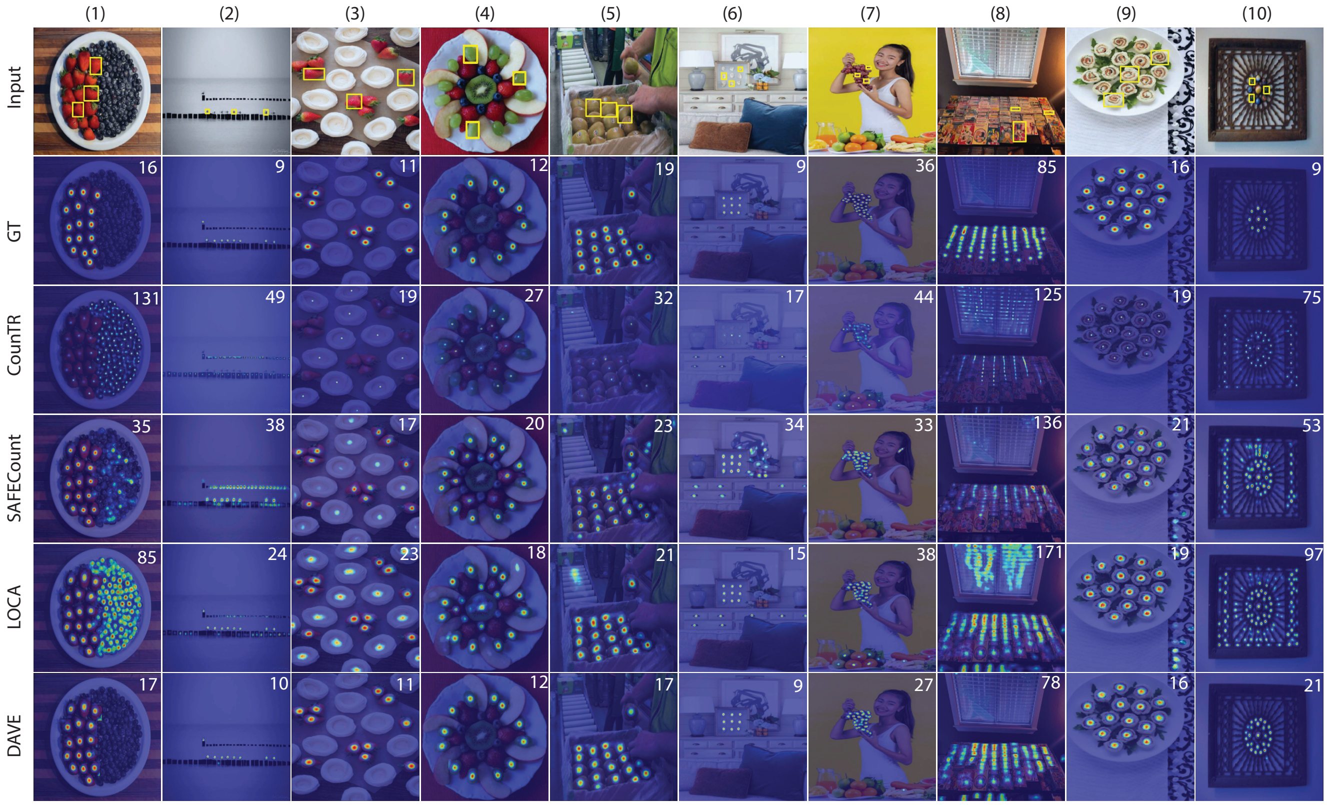

Figure 3. Qualitative comparison of DAVE with LOCA [6], SAFECount [38] and CounTR [16]. The first two columns show the input images and the ground truth (GT), while the predicted densities are shown in the rest.

Table 5. Detection performance on FSCD147 [22].

	Validation Set		Test Set	
Method	AP	AP50	AP	AP50
FSDetView-PB [34]	-	-	13.41	32.99
FSDetView-RR [34]	-	-	17.21	33.70
AttRPN-RR [9]	-	-	18.53	35.87
AttRPN-PB [9]	-	-	20.97③	37.19③
C-DETR [22]	17.27②	41.90②	22.66②	50.57②
DAVE	24.20①	61.08①	26.81①	62.82①

Results in Table 5 show that DAVE sets a new state-of-the-art in all measures on both validation and test splits. On the validation split, DAVE outperforms the most recent C-DETR [22] by 40% and 45% in AP and AP50, respectively, and outperforms C-DETR on the test split by 18% and 24% in AP and AP50, respectively.

The high AP50 and AP indicate that DAVE retrieves more objects with less false positives, and that localization of the detected objects is more accurate (see Figure 4, rows 1 and 2). DAVE also performs comparatively well in high-density regions with small objects, which are very challenging for the current state-of-the-art (Figure 4, rows 3 and 4). Compared to the best methods, DAVE better learns the appearance of targets composed of fine-grained objects, leading to improved detections (e.g., bowls of pills in Figure 4, row 5). These results speak of a substantial potential of the detect-and-verify approach for accurate localization.

We further evaluate DAVE on two recent datasets FSCD-LVIS [22] and FSCD-LVIS$_{\text{uns}}$ [22]. Both datasets are created from the LVIS [11] dataset containing 6196 images with 377 classes. In FSCD-LVIS [22] dataset, some classes in test set appear also in the training set. The second dataset, FSCD-LVIS$_{\text{uns}}$ [22] ensures that test set does not contain classes observed during training. Results in Table 6 show that DAVE outperforms the top method C-DETR by 37% and 55% w.r.t. AP and AP50, respectively on the FSCD-LVIS. On FSCD-LVIS$_{\text{uns}}$, DAVE also substantially outperforms the best method by 7% and 25% in AP and AP50, respectively.

Zero-shot detection. To the best of our knowledge, DAVE$_{\text{0-shot}}$ is the first zero-shot method capable of counting and detection. We thus compare it with the best counting and detection method C-DETR [22], which however is not zero-shot, since it requires three input exemplars. Results in Table 7 reveal excellent performance of DAVE$_{\text{0-shot}}$. On the validation split, it outperforms C-DETR [22] by 12% in AP50 and delivers comparable performance on the test split. While it achieves equally robust detection (AP50) as

Table 6. Detection on FSCD-LVIS/FSCD-LVIS$_{\mathrm{uns}}$ [22] test sets.

	FSCD-LVIS		FSCD-LVIS$_{\mathrm{uns}}$	
Method	AP	AP50	AP	AP50
FSDetView-RR [34]	1.96	6.70	0.89	2.38
FSDetView-PB [34]	2.72	7.57	1.03	2.89
AttRPN-RR [9]	3.28	9.44	2.52	7.86
AttRPN-PB [9]	4.08③	11.15③	3.15③	7.87③
C-DETR [22]	4.92②	14.49②	3.85②	11.28②
DAVE	6.75①	22.51①	4.12①	14.16①

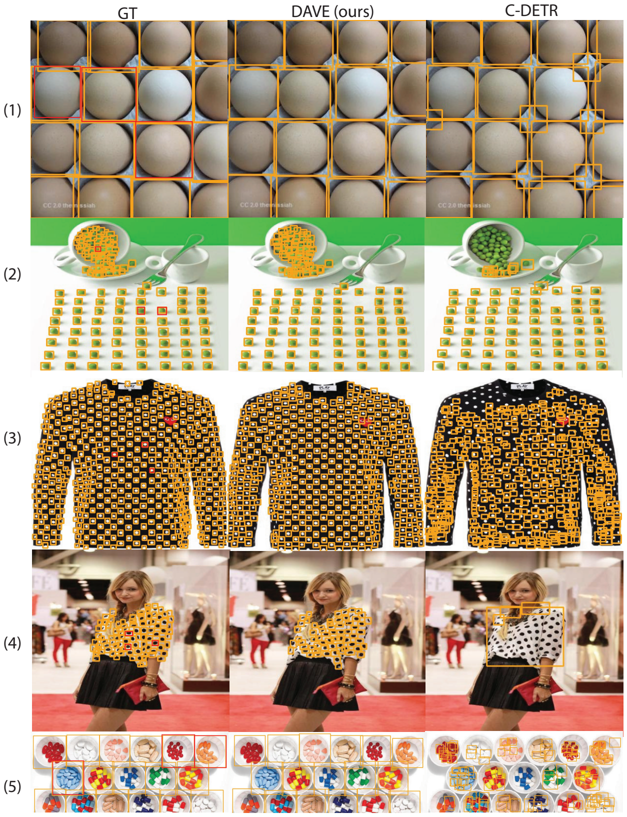

Figure 4. DAVE localization performance in challenging situations compared with the current best method C-DETR [22].

C-DETR, the localization is slightly less accurate (lower AP). Nevertheless, this is a remarkable result, considering C-DETR requires annotated exemplars as input, while DAVE$_{\text{0-shot}}$ does not. The experiment validates the generality of the proposed detect-and-verify paradigm for all low-shot counting tasks (few- and zero-shot).

Few-shot detection counting. The previous experiments analyzed the accuracy of detections. To further analyze the detection capability, we measure the accuracy of count estimation when approximated by the number of detected bounding boxes. In the following, we use the superscript DAVE$^{\text{box}}$ to distinguish the results from the density-based count estimation. Results are reported in Table 8. DAVE$^{\text{box}}$ outperforms all state-of-the-art by a significant margin, in particular, it outperforms C-DETR by 38% in MAE and 40% in RMSE. This further confirms the remarkable detection performance compared to the most recent detection-based methods. Note that the DAVE$^{\text{box}}$ not only outperforms all detection-based counters, but also all published density-based counters in terms of MAE, including LOCA [6] (Table 1), which have been up to now unchallenged by the detection-based counters.

Table 7. Without requiring exemplars, DAVE$_{\text{0-shot}}$ performs on par or better than C-DETR with three input exemplars.

	Validation Set		Test Set	
Method	AP	AP50	AP	AP50
C-DETR (*3-shot*) [22]	17.27①	41.90②	22.66①	50.57①
DAVE$_{\text{0-shot}}$	16.31②	46.87①	18.55②	50.08②

Table 8. Few-shot detection-based counting on FSC147 [26].

	Validation Set		Test Set	
Method	MAE	RMSE	MAE	RMSE
FSDetView-RR [34]	-	-	37.83	146.56
FSDetView-PB [34]	-	-	37.54	147.07
AttRPN-RR [9]	-	-	32.70	141.07③
AttRPN-PB [9]	-	-	32.42③	141.55
C-DETR [22]	20.38②	82.45②	16.79②	123.56②
DAVE$^{\text{box}}$	9.75①	40.30①	10.45①	74.51①

4.4. Ablation study

Impact of mixed-class training. We first verify whether false positives in state-of-the-art methods could be reduced by simply training on images with multiple object categories. The current top low-shot counter LOCA [6] is thus retrained on multi-class images, in which a FSCD147 [22] training image is concatenated with another image containing objects from a different class for hard negative training examples, as described in Section 3.4. We denote this version by LOCA$_{\text{mul}}$ and also include CounTR [16] in the comparison since it already applies such a training setup. We also construct a subset of FSCD147 composed of images containing objects from different classes[3] (denoted as FSCD147$_{\text{mul}}$), to expose the sensitivity of a counting method to other-class objects. Table 9 shows that the counting performance of LOCA on FSCD147$_{\text{mul}}$ is significantly lower compared to FSCD147, confirming that multi-class images are highly challenging. Training LOCA on multi-class images (LOCA$_{\text{mul}}$) substantially reduces the average error on FSCD147$_{\text{mul}}$ by 43%, but increases it by 28% on FSC147. This is likely due to LOCA$_{\text{mul}}$ compensating for

[3] Images were obtained from the test and evaluation splits of FSCD147.

the improved multi-class performance by a reduced overall performance. However, DAVE demonstrates excellent performance on both datasets and also consistently outperforms both LOCA versions and CounTR by large margins.

Impact of cluster selection. We compare the prompt-based DAVE$_{\text{prm}}$ with DAVE$_{\text{0-shot}}$ to demonstrate the impact of the cluster selection method. Notably, DAVE$_{\text{prm}}$ selects the clusters by comparing them with text prompts, while DAVE$_{\text{0-shot}}$ applies majority voting (Section 3.3). While the performance of the two methods is comparable on average, DAVE$_{\text{prm}}$ substantially outperforms on FSCD147$_{\text{mul}}$. This result is presented in Table 9 (bottom) and indicates that prompt-based cluster selection is particularly important on images with multiple classes to resolve the object category ambiguity.

Table 9. Performance in presence of objects from multiple classes.

	FSCD147			FSCD147$_{\text{mul}}$	
	MAE(↓)	RMSE(↓)	AP50(↑)	MAE(↓)	RMSE(↓)
LOCA [6]	10.79②	56.97②	-	21.28	43.67
LOCA$_{\text{mul}}$ [6]	12.63	78.95	-	13.25②	22.57②
CounTR [16]	11.95③	91.23③	-	14.56③	27.41③
DAVE	8.66①	32.36①	61.08①	3.05①	4.94①
DAVE$_{\text{0-shot}}$	15.54	103.49	50.08	12.86	23.21
DAVE$_{\text{prm}}$	14.90	103.42	50.24	6.46	10.72

Table 10. DAVE architecture analysis on FSCD147 [26].

	MAE(↓)	RMSE(↓)	AP(↑)	AP50(↑)
DAVE	**8.91**	**28.08**	**24.19**	**61.08**
DAVE$_{\overline{\phi}}$	9.41	29.91	24.11	60.85
DAVE$_{\overline{\text{R}}}$	8.97	28.12	19.50	53.57
DAVE$_{\text{cat}}$	8.99	28.18	18.49	51.72
DAVE$_{\text{sum}}$	8.95	28.13	20.74	55.54

Architecture design. Finally, we evaluate the DAVE architectural design decisions. First, we analyze the impact of using the prototype correlation response tensor $\tilde{R}$ in the box regression step. Table 10 shows that removing $\tilde{R}$ (DAVE$_{\overline{\text{R}}}$) results in a substantial drop of 12% and 9% in AP and AP50. This verifies the importance of fusion with $\tilde{R}$, which contains size and shape information of the selected objects, considerably improving the localization accuracy of DAVE detections. In particular, for target objects composed of smaller objects, this information is crucial for accurate bounding box prediction (Figure 4, last row). To evaluate the importance of the feature fusion module (FFM), we replace it with sum (DAVE$_{\text{sum}}$), and concatenation (DAVE$_{\text{cat}}$). Both replacements result in a detection performance drop of 9% and 15% AP, respectively. To validate the importance of robust appearance features in the verification stage, we remove the feature projection network $\phi(\cdot)$ and perform clustering directly on the backbone features (DAVE$_{\overline{\phi}}$). The errors of DAVE$_{\overline{\phi}}$ increase by 6% in MAE and 7% RMSE.

Limitations. DAVE outputs detections (i.e., bounding boxes), as well as total counts estimated from the density. To expose limitations, we inspect the discrepancy between the total count estimates and the number of detections with respect to the number of objects in the image (Figure 5). The discrepancy is most apparent for images with very large object counts, which typically contain many small objects packed together (i.e., extremely dense regions). Further error reductions are thus expected by improving DAVE detection stage in the presence of extreme density. The limitation is common to all low-shot counters, and we defer this to future research.

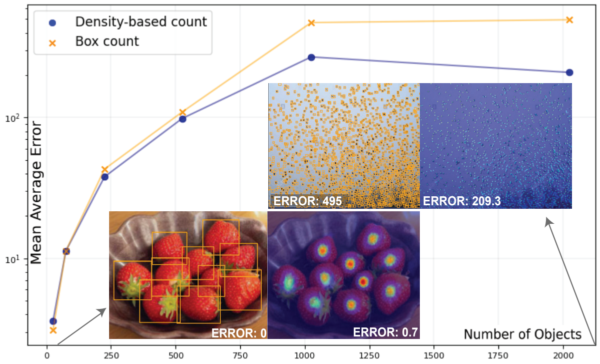

Figure 5. DAVE density-based and box-count accuracy with respect to the number of objects in the image.

5. Conclusion

We presented a novel low-shot object counting and detection method DAVE, that narrows the performance gap between density-based and detection-based counters. DAVE spans the entire low-shot spectrum, also covering text-prompt setups, and is the first method capable of zero-shot detection-based counting. This is achieved by the novel detect-and-verify paradigm, which increases the recall as well as precision of the detections.

Extensive analysis demonstrates that DAVE sets a new state-of-the-art in total count estimation, as well as in detection accuracy on several benchmarks with comparable complexity to related methods, running 110ms/image. In particular, DAVE outperforms the long-standing top low-shot counter [6], as well as the recent detection-based counter [22]. In a zero-shot setup, DAVE outperforms all density-based counters and delivers detections on par with the most recent few-shot counter that requires at least few annotations. DAVE also sets a new state-of-the-art in prompt-based counting. In our future work, we plan to explore interactive counting with the human in the loop and improve detection in extremely dense regions.

Acknowledgements. This work was supported by Slovenian research agency program P2-0214 and projects J2-2506, L2-3169 and Z2-4459.

References

[1] N. Amini-Naieni, K. Amini-Naieni, T. Han, and A. Zisserman. Open-world text-specified object counting. In *BMVC*, 2023. 5

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[3] Zhi-Qi Cheng, Qi Dai, Hong Li, Jingkuan Song, Xiao Wu, and Alexander G Hauptmann. Rethinking spatial invariance of convolutional networks for object counting. In *CVPR*, pages 19638–19648, 2022. 2

[4] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *CVPR*, 2019. 2

[5] Zhe Dai, Huansheng Song, Xuan Wang, Yong Fang, Xu Yun, Zhaoyang Zhang, and Huaiyu Li. Video-based vehicle counting framework. *IEEE Access*, 7:64460–64470, 2019. 2

[6] Nikola Djukic, Alan Lukezic, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. In *ICCV*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[8] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019. 2

[9] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, pages 4013–4022, 2020. 5, 6, 7

[10] 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. 4, 5

[11] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 6

[12] 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

[13] Michael Hobley and Victor Prisacariu. Learning to count anything: Reference-less class-agnostic counting with weak supervision. *arXiv preprint arXiv:2205.10203*, 2022. 2, 5

[14] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015. 3

[15] Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. *ACM Multimedia 2023*, 2023. 2, 5

[16] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. In *BMVC*. BMVA Press, 2022. 1, 2, 4, 5, 6, 7, 8

[17] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *CVPR*, 2019. 2

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 4

[19] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *ACCV*, pages 669–684. Springer, 2018. 2

[20] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 669–684. Springer, 2019. 5

[21] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *NeurIPS*, 14, 2001. 3

[22] Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. In *ECCV*, pages 348–365. Springer, 2022. 1, 2, 5, 6, 7, 8

[23] Alec Radford, Ilya Sutskever, Jong Wook Kim, Gretchen Krueger, and Sandhini Agarwal. Clip: connecting text and images. *OpenAI. https://openai. com/blog/clip/*, 2021. 2

[24] Viresh Ranjan and Minh Hoai. Vicinal counting networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4221–4230, 2022. 4, 5

[25] Viresh Ranjan and Minh Hoai Nguyen. Exemplar free class agnostic counting. In *Proceedings of the Asian Conference on Computer Vision*, pages 3121–3137, 2022. 2, 5

[26] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *CVPR*, pages 3394–3403, 2021. 1, 2, 4, 5, 7, 8

[27] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. In *CVPR*, 2019. 4

[28] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *CVPR*, pages 9529–9538, 2022. 1, 2, 4, 5

[29] Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *CVPR*, pages 19618–19627, 2022. 2

[30] Jia Wan and Antoni Chan. Modeling noisy annotations for crowd counting. *NeurIPS*, 33:3386–3396, 2020.

[31] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *NeurIPS*, 33:1595–1607, 2020. 2

[32] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, pages 2495–2504, 2021. 4

[33] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, pages 3–19, 2018. 3

[34] Yang Xiao, Vincent Lepetit, and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. *IEEE TPAMI*, 45(3):3090–3106, 2022. 5, 6, 7

[35] 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. 1

[36] Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *CVPR*, pages 15548–15557, 2023. 2, 5

[37] Shuo-Diao Yang, Hung-Ting Su, Winston H Hsu, and Wen-Chin Chen. Class-agnostic few-shot object counting. In *WACV*, pages 870–878, 2021. 2, 4, 5

[38] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *WACV*, pages 6315–6324, 2023. 1, 2, 4, 5, 6

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018. 3

[40] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *ACM MM*, pages 516–520, 2016. 3

[41] Vitjan Zavrtanik, Martin Vodopivec, and Matej Kristan. A segmentation-based approach for polyp counting in the wild. *Engineering Applications of Artificial Intelligence*, 88: 103399, 2020. 1, 2