

DeepSetNet: Predicting Sets with Deep Neural Networks

S. Hamid Rezaatofghi Vijay Kumar B G Anton Milan
Ehsan Abbasnejad Anthony Dick Ian Reid
School of Computer Science, The University of Adelaide, Australia

Abstract

This paper addresses the task of set prediction using deep learning. This is important because the output of many computer vision tasks, including image tagging and object detection, are naturally expressed as sets of entities rather than vectors. As opposed to a vector, the size of a set is not fixed in advance, and it is invariant to the ordering of entities within it. We define a likelihood for a set distribution and learn its parameters using a deep neural network. We also derive a loss for predicting a discrete distribution corresponding to set cardinality. Set prediction is demonstrated on the problem of multi-class image classification. Moreover, we show that the proposed cardinality loss can also trivially be applied to the tasks of object counting and pedestrian detection. Our approach outperforms existing methods in all three cases on standard datasets.

1. Introduction

Deep neural networks have shown state-of-the-art performance on many computer vision problems, including semantic segmentation [33], visual tracking [31], image captioning [21], scene classification [22], and object detection [27]. However, traditional convolutional architectures require a problem to be formulated in a certain way: in particular, they are designed to predict a *vector* (or a matrix, or a tensor in a more general sense), that is either of a fixed length or whose size depends on the input (*cf.* fully convolutional architectures).

For example, consider the task of scene classification where the goal is to predict the label (or category) of a given image. Modern approaches typically address this by a series of convolutional layers, followed by a number of fully connected layers, which are finally mapped to predict a fixed-sized vector [22, 39, 41]. The length of the predicted vector corresponds to the number of candidate categories, *e.g.* 1,000 for the ImageNet challenge [37]. Each element is a score or probability of a particular category, and the final prediction is a probability distribution over all categories. This strategy is perfectly admissible if one expects to find

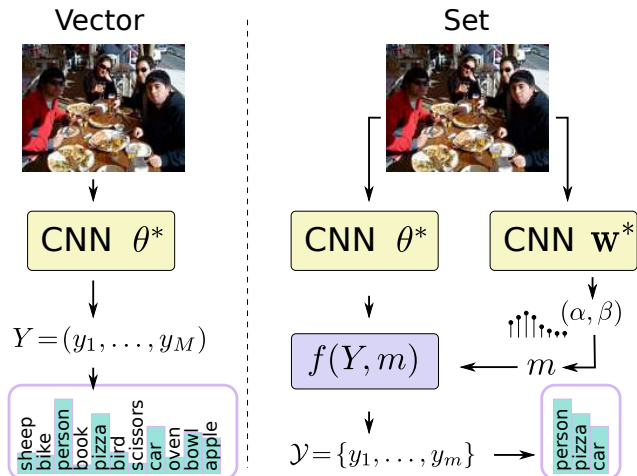


Figure 1: **Left:** Traditional CNNs learn parameters θ^* to predict a fixed vector Y . **Right:** In contrast, we propose to train a separate CNN to learn a parameter vector w^* , which is used to predict the set cardinality of a particular output.

exactly one or at least the same number of categories across all images. However, natural images typically show multiple entities (*e.g.* table, pizza, person, *etc.*), and what is perhaps more important, this number differs from image to image. During evaluation, this property is not taken into account. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) only counts an error if the “true” label is not among the top-5 candidates. Another strategy to account for multiple classes is to fix the number to a certain value for all test instances, and report precision and recall by counting false positive and false negative predictions, as was done in [15, 47]. Arguably, both methods are suboptimal because in a real-world scenario, where the correct labelling is unknown, the prediction should in fact not only rank all labels according to their likelihood of being present, but also to report *how many* objects (or labels) are actually present in one particular image. Deciding how many objects are present in an image is a crucial part of human perception and scene understanding but is missing from our current evaluation of automated image understanding methods.

As a second example, let us consider pedestrian detec-

tion. The parallel to scene classification that we motivated above is that, once again, in real scenarios, the number of people in a particular scene is not known beforehand. The most common approach is to assign a confidence score to a number of region candidates [7, 11, 14, 36], which are typically selected heuristically by thresholding and non-maxima suppression. We argue that it is important not to simply discard the information about the actual number of objects at test time, but to exploit it while selecting the subset of region proposals.

The examples above motivate and underline the importance of *set prediction* in certain applications. It is important to note that, in contrast to vectors, a set is a collection of elements which is *invariant under permutation* and the size of a set is *not fixed* in advance. To this end, we use a principled definition of a set as the union of cardinality distribution and family of joint distributions over each cardinality value. In summary, our main contributions are as follows: (i) Starting from the mathematical definition of a set distribution, we derive a loss that enables us to employ existing machine learning methodology to learn this distribution from data. (ii) We integrate our loss into a deep learning framework to exploit the power of a multi-layer architecture. (iii) We present state-of-the-art results for multi-label image classification and pedestrian detection on standard datasets and competitive results on the task of object counting.

2. Related Work

A sudden success in multiple applications including voice recognition [17], machine translation [40] and image classification [22], has sparked the deployment of deep learning methods throughout numerous research areas. Deep convolutional (CNN) and recurrent (RNN) neural networks now outperform traditional approaches in tasks like semantic segmentation [6, 25], image captioning [21] or object detection [27]. Here, we will briefly review some of the recent approaches to image classification and object detection and point out their limitations.

Image or scene classification is a fundamental task of understanding photographs. The goal here is to predict a scene label for a given image. Early datasets, such as Caltech-101 [10], mostly contained one single object and could easily be described by one category. Consequently, a large body of literature focused on single-class prediction [22, 38, 49, 30]. However, real-world photographs typically contain a collection of multiple objects and should therefore be captioned with multiple tags. Surprisingly, there exists rather little work on multi-class image classification that makes use of deep architectures. Gong *et al.* [16] combine deep CNNs with a top- k approximate ranking loss to predict multiple labels. Wei *et al.* [48] propose a Hypotheses-Pooling architecture that is specifically designed to handle multi-label output. While both methods

open a promising direction, their underlying architectures largely ignore the correlation between multiple labels. To address this limitation, recently, Wang *et al.* [47] combined CNNs and RNNs to predict a number of classes in a sequential manner. RNNs, however, are not suitable for set prediction mainly for two reasons. First, the output represents a sequence rather than a set and is thus highly dependent on the prediction order, as was shown recently by Vinyals *et al.* [42]. Second, the final prediction may not result in a feasible solution (*e.g.* it may contain the same element multiple times), such that post-processing or heuristics such as beam search must be employed [43, 47]. Here we show that our approach not only guarantees to always predict a valid set, but also outperforms previous methods.

Pedestrian detection can also be viewed as a classification problem. Traditional approaches follow the sliding-window paradigm [44, 7, 46, 11, 2], where each possible (or rather plausible) image region is scored independently to contain a person or not. More recent methods like Fast R-CNN [14] or the single-shot multi-box detector (SSD) [27] learn the relevant image features rather than manually engineering them, but retain the sliding window approach.

All the above approaches require some form of post-processing to suppress spurious detection responses that originate from the same person. This is typically addressed by non-maximum suppression (NMS), a greedy optimisation strategy with a fixed overlap threshold. Recently, several alternatives have been proposed to replace the greedy NMS procedure, including sequential head detection with LSTMs [18], a global optimisation approach to NMS [34, 23], or even learning NMS end-to-end using CNNs [19]. None of the above methods, however, explicitly consider the number of objects while selecting the final set of boxes. Contrary to existing pedestrian detection approaches, we incorporate the cardinality into the NMS algorithm itself and show an improvement over the state of the art on two benchmarks. Note that the idea of considering the number of pedestrians can be applied in combination with any of the aforementioned detection techniques to further improve their performances.

It is important to bear in mind that unlike many existing approaches that learn to count [1, 5, 12, 20, 24, 35, 50, 51], our main goal is not object counting. Rather, we derive a formulation for set prediction using deep learning. Estimating the cardinality of objects and thereby their count is a byproduct of our approach. To demonstrate the effectiveness of our formulation, we also conduct experiments on the task of object counting, outperforming many recent methods on the widely used USCD dataset.

3. Random Vectors vs. Random Finite Sets

In statistics, a continuous random variable y is a variable that can take an infinite number of possible values. A

continuous random vector can be defined by stacking several continuous random variables into a fixed length vector, $Y = (y_1, \dots, y_m)$. The mathematical function describing the possible values of a continuous random vector and their associated joint probabilities is known as a probability density function (PDF) $p(Y)$ such that $\int p(Y)dY = 1$.

In contrast, a random finite set (RFS) \mathcal{Y} is a finite-set valued random variable $\mathcal{Y} = \{y_1, \dots, y_m\}$. The main difference between an RFS and a random vector is that for the former, the number of constituent variables, m , is random and the variables themselves are random and unordered. Throughout the paper, we use $\mathcal{Y} = \{y_1, \dots, y_m\}$ for a set with unknown cardinality, $\mathcal{Y}_m = \{y_1, \dots, y_m\}_{||}$ for a set with known cardinality and $Y = (y_1, \dots, y_m)$ for a vector with known dimension.

A statistical function describing a finite-set variable \mathcal{Y} is a combinatorial probability density function $p(\mathcal{Y})$ which consists of a discrete probability distribution, the so-called cardinality distribution, and a family of joint probability densities on both the number and the values of the constituent variables. Similar to the definition of a PDF for a random variable, the PDF of an RFS must sum to unity over all possible cardinality values and all possible element values and their permutations [28]. The PDF of an m -dimensional random vector can be defined in terms of an RFS as

$$p(y_1, \dots, y_m) \triangleq \frac{1}{m!} p(\{y_1, \dots, y_m\}_{||}). \quad (1)$$

The normalisation factor $m! = \prod_{k=1}^m k$ appears because the probability density for a set $\{y_1, \dots, y_m\}_{||}$ must be equally distributed among all the $m!$ possible permutations of the vector [28].

Conventional machine learning approaches, such as Bayesian learning and convolutional neural networks, have been proposed to learn the optimal parameters θ^* of the distribution $p(Y|\mathbf{x}, \theta^*)$ which maps the input vector \mathbf{x} to the output vector Y . In this paper, we instead propose an approach that can learn a pair (θ^*, \mathbf{w}^*) of parameter vectors for a set distribution that allow one to map the input vector \mathbf{x} into the output set \mathcal{Y} , i.e. $p(\mathcal{Y}|\mathbf{x}, \theta^*, \mathbf{w}^*)$. The additional parameters \mathbf{w}^* define a PDF over the set cardinality, as we will explain in the next section.

4. Deep Set Network

Let us begin by defining a training set $\mathcal{D} = \{\mathcal{Y}_i, \mathbf{x}_i\}$, where each training sample $i = 1, \dots, n$ is a pair consisting of an input feature $\mathbf{x}_i \in \mathbb{R}^l$ and an output (or label) set $\mathcal{Y}_i = \{y_1, y_2, \dots, y_{m_i}\}$, $y_k \in \mathbb{R}^d$, $m_i \in \mathbb{N}^0$. In the following we will drop the instance index i for better readability. Note that $m := |\mathcal{Y}|$ denotes the cardinality of set \mathcal{Y} . The

probability of a set \mathcal{Y} is defined as

$$p(\mathcal{Y}|\mathbf{x}, \theta, \mathbf{w}) = p(m|\mathbf{x}, \mathbf{w}) \times m! \times U^m \times p(y_1, y_2, \dots, y_m|\mathbf{x}, \theta), \quad (2)$$

where $p(m|\cdot, \cdot)$ and $p(y_1, y_2, \dots, y_m|\cdot, \cdot)$ are respectively a cardinality distribution and a symmetric joint probability density distribution of the elements. U is the unit of hypervolume in \mathbb{R}^d , which makes the joint distribution unitless [28, 45]. θ denotes the parameters that estimate the joint distribution of set element values for a fixed cardinality,¹ while \mathbf{w} represents the collection of parameters which estimate the cardinality distribution of the set elements.

The above formulation represents the probability density of a set which is very general and completely independent from the choices of both cardinality and spatial distributions. It is thus straightforward to transfer it to many applications that require the output to be a set. However, to make the problem amenable to mathematical derivation and implementation, we adopt two assumptions: *i*) the outputs (or labels) in the set are independent and identically distributed (*i.i.d.*) and *ii*) their cardinality follows a Poisson distribution with parameter λ . Thus, we can write the distribution as

$$p(\mathcal{Y}|\mathbf{x}, \theta, \mathbf{w}) = \int p(m|\lambda) p(\lambda|\mathbf{x}, \mathbf{w}) d\lambda \times m! \times U^m \times \left(\prod_{k=1}^m p(y_k|\mathbf{x}, \theta) \right). \quad (3)$$

4.1. Posterior distribution

To learn the parameters θ and \mathbf{w} , we first define the posterior distribution over them as

$$\begin{aligned} p(\theta, \mathbf{w}|\mathcal{D}) &\propto p(\mathcal{D}|\theta, \mathbf{w}) p(\theta) p(\mathbf{w}) \\ &\propto \prod_{i=1}^n \left[\int p(m_i|\lambda) p(\lambda|\mathbf{x}_i, \mathbf{w}) d\lambda \times m_i! \right. \\ &\quad \left. U^{m_i} \left(\prod_{k=1}^{m_i} p(y_k|\mathbf{x}_i, \theta) \right) \right] p(\mathbf{x}_i) p(\theta) p(\mathbf{w}). \end{aligned} \quad (4)$$

A closed form solution for the integral in Eq. (4) can be obtained by using conjugate priors:

$$\begin{aligned} m &\sim \mathcal{P}(m; \lambda) \\ \lambda &\sim \mathcal{G}(\lambda; \alpha(\mathbf{x}, \mathbf{w}), \beta(\mathbf{x}, \mathbf{w})) \\ \alpha(\mathbf{x}, \mathbf{w}), \beta(\mathbf{x}, \mathbf{w}) &> 0 \quad \forall \mathbf{x}, \mathbf{w} \\ \theta &\sim \mathcal{N}(\theta; 0, \sigma_1^2 \mathbf{I}) \\ \mathbf{w} &\sim \mathcal{N}(\mathbf{w}; 0, \sigma_2^2 \mathbf{I}), \end{aligned}$$

¹This is also known as *spatial distribution of points* in point process statistics.

where $\mathcal{P}(\cdot, \lambda)$, $\mathcal{G}(\cdot; \alpha, \beta)$, and $\mathcal{N}(\cdot; 0, \sigma^2 \mathbf{I})$ represent respectively a Poisson distribution with parameters λ , a Gamma distribution with parameters (α, β) and a zero mean normal distribution with covariance equal to $\sigma^2 \mathbf{I}$.

We assume that the cardinality follows a Poisson distribution whose mean, λ , follows a Gamma distribution, with parameters which can be estimated from the input data \mathbf{x} . Note that the cardinality distribution in Eq. (2) can be replaced by any other discrete distribution. For example, it is a valid assumption to model the number of objects in natural images by a Poisson distribution [5]. Thus, we could directly predict λ to model this distribution by formulating the cardinality as $p(m|\mathbf{x}, \mathbf{w}) = \mathcal{P}(m; \lambda(\mathbf{x}, \mathbf{w}))$. However, this would limit the model's expressive power because two visually entirely different images with the same number of objects would be mapped to the same λ . Instead, to allow for uncertainty of the mean, we model it with another distribution, which we choose to be Gamma for mathematical convenience. Consequently, the integrals in $p(\boldsymbol{\theta}, \mathbf{w}|\mathcal{D})$ are simplified and form a negative binomial distribution,

$$\text{NB}(m; a, b) = \frac{\Gamma(m+a)}{\Gamma(m+1)\Gamma(a)} \cdot (1-b)^a b^m, \quad (5)$$

where Γ is the Gamma function. Finally, the full posterior distribution can be written as

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{w}|\mathcal{D}) &\propto \prod_{i=1}^n \left[\text{NB}\left(m_i; \alpha(\mathbf{x}_i, \mathbf{w}), \frac{1}{1+\beta(\mathbf{x}_i, \mathbf{w})}\right) \right. \\ &\quad \left. \times m_i! \times U^{m_i} \times \left(\prod_{k=1}^{m_i} p(y_k|\mathbf{x}_i, \boldsymbol{\theta}) \right) \right] p(\boldsymbol{\theta}) p(\mathbf{w}). \end{aligned} \quad (6)$$

4.2. Learning

For simplicity, we use a point estimate for the posterior $p(\boldsymbol{\theta}, \mathbf{w}|\mathcal{D})$, i.e. $p(\boldsymbol{\theta}, \mathbf{w}|\mathcal{D}) = \delta(\boldsymbol{\theta} = \boldsymbol{\theta}^*, \mathbf{w} = \mathbf{w}^*|\mathcal{D})$, where $(\boldsymbol{\theta}^*, \mathbf{w}^*)$ are computed using the following MAP estimator:

$$(\boldsymbol{\theta}^*, \mathbf{w}^*) = \arg \max_{\boldsymbol{\theta}, \mathbf{w}} \log(p(\boldsymbol{\theta}, \mathbf{w}|\mathcal{D})). \quad (7)$$

The optimisation problem in Eq. (7) can be decomposed w.r.t. the parameters $\boldsymbol{\theta}$ and \mathbf{w} . Therefore, we can learn them independently as

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} -\gamma_1 \|\boldsymbol{\theta}\| + \sum_{i=1}^n \sum_{k=1}^{m_i} \log(p(y_k|\mathbf{x}_i, \boldsymbol{\theta})) \quad (8)$$

and

$$\begin{aligned} \mathbf{w}^* = \arg \max_{\mathbf{w}} &\sum_{i=1}^n \left[\log \left(\frac{\Gamma(m_i + \alpha(\mathbf{x}_i, \mathbf{w}))}{\Gamma(m_i + 1)\Gamma(\alpha(\mathbf{x}_i, \mathbf{w}))} \right) \right. \\ &\quad \left. + \log \left(\frac{\beta(\mathbf{x}_i, \mathbf{w})^{\alpha(\mathbf{x}_i, \mathbf{w})}}{(1 + \beta(\mathbf{x}_i, \mathbf{w})^{\alpha(\mathbf{x}_i, \mathbf{w}) + m_i})} \right) \right] - \gamma_2 \|\mathbf{w}\|, \end{aligned} \quad (9)$$

where γ_1 and γ_2 are the regularisation parameters, proportional to the predefined covariance parameters σ_1 and σ_2 . These parameters are also known as weight decay parameters and commonly used in training neural networks.

The learned parameters $\boldsymbol{\theta}^*$ in Eq. (8) are used to map an input feature vector \mathbf{x} into an output vector Y . For example, in image classification, $\boldsymbol{\theta}^*$ is used to predict the distribution Y over all categories, given the input image \mathbf{x} . Note that $\boldsymbol{\theta}^*$ can generally be learned using a number of existing machine learning techniques. In this paper we rely on deep CNNs to perform this task.

To learn the highly complex function between the input feature \mathbf{x} and the parameters (α, β) , which are used for estimating the output cardinality distribution, we train a second deep neural network. Using neural networks to predict a discrete value may seem counterintuitive, because these methods at their core rely on the backpropagation algorithm, which assumes a differentiable loss. Note that we achieve this by describing the discrete distribution by continuous parameters α, β (Negative binomial $\text{NB}(\cdot, \alpha, \frac{1}{1+\beta})$), and can then easily draw discrete samples from that distribution. More formally, to estimate \mathbf{w}^* , we compute the partial derivatives of the objective function in Eq. (9) w.r.t. $\alpha(\cdot, \cdot)$ and $\beta(\cdot, \cdot)$ and use standard backpropagation to learn the parameters of the deep neural network.

We refer the reader to the supplementary material for the complete derivation of the partial derivatives, a more detailed derivation of the posterior in Eqs. (4)-(6) and the proof for decomposition of the MAP estimation in Eq. (7).

4.3. Inference

Having the learned parameters of the network $(\mathbf{w}^*, \boldsymbol{\theta}^*)$, for a test feature \mathbf{x}^+ , we use a MAP estimate to generate a set output as $\mathcal{Y}^* = \arg \max_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{D}, \mathbf{x}^+)$, where

$$p(\mathcal{Y}|\mathcal{D}, \mathbf{x}^+) = \int p(\mathcal{Y}|\mathbf{x}^+, \boldsymbol{\theta}, \mathbf{w}) p(\boldsymbol{\theta}, \mathbf{w}|\mathcal{D}) d\boldsymbol{\theta} d\mathbf{w}$$

and $p(\boldsymbol{\theta}, \mathbf{w}|\mathcal{D}) = \delta(\boldsymbol{\theta} = \boldsymbol{\theta}^*, \mathbf{w} = \mathbf{w}^*|\mathcal{D})$. Since the unit of hypervolume U in most practical application is unknown, to calculate the mode of the set distribution $p(\mathcal{Y}|\mathcal{D}, \mathbf{x}^+)$, we use sequential inference as explained in [28]. To this end, we first calculate the mode m^* of the cardinality distribution $m^* = \arg \max_m p(m|\mathbf{x}^+, \mathbf{w}^*)$, where $p(m|\mathbf{x}^+, \mathbf{w}^*) = \text{NB}\left(m; \alpha(\mathbf{x}^+, \mathbf{w}^*), \frac{1}{1+\beta(\mathbf{x}^+, \mathbf{w}^*)}\right)$. Then, we calculate the mode of the joint distribution for the given cardinality m^* as

$$\mathcal{Y}^* = \arg \max_{\mathcal{Y}_{m^*}} p(\{y_1, \dots, y_{m^*}\}|\mathbf{x}^+, \boldsymbol{\theta}^*). \quad (10)$$

To estimate the most likely set \mathcal{Y}^* with cardinality m^* , we use the first CNN with the parameters $\boldsymbol{\theta}^*$ which predicts $p(y_1, \dots, y_M|\mathbf{x}^+, \boldsymbol{\theta}^*)$, where M is the maximum cardinality of the set, i.e. $\{y_1, \dots, y_{m^*}\} \subseteq \{y_1, \dots, y_M\}$, $\forall m^*$.

Since the samples are *i.i.d.*, the joint probability maximised when the probability of each element in the set is maximised. Therefore, the most likely set \mathcal{Y}^* with cardinality m^* is obtained by ordering the probabilities of the set elements y_1, \dots, y_M as the output of the first CNN and choosing m^* elements with highest probability values.

Note that the assumptions in Sec. 4 are necessary to make both learning and inference computationally tractable and amenable to an elegant mathematical formulation. A major advantage of this approach is that we can use any state-of-the-art classifier/detector as the first CNN (θ^*) to further improve its performance. Modifying any of the assumptions, *e.g.* non-*i.i.d.* set elements, leads to serious mathematical complexities [28], and is left for future work.

5. Experimental Results

Our proposed method is best suited for applications that expect the solution to be in the form of a set, *i.e.* permutation invariant and of an unknown cardinality. To that end, we perform an experiment on multi-label image classification in Sec. 5.1. In addition, we explore our cardinality estimation loss on the object counting problem in Sec. 5.2 and then show in Sec. 5.3 how incorporating cardinality into a state-of-the-art pedestrian detector and formulating it as a set problem can boost up its performance.

5.1. Multi-Label Image Classification

As opposed to the more common and more studied problem of (single-label) image classification, the task here is rather to label a photograph with an arbitrary, a-priori unknown number of tags. We perform experiments on two standard benchmarks, the PASCAL VOC 2007 dataset [9] and the Microsoft Common Objects in Context (MS COCO) dataset [26].

Implementation details. In this experiment, similar to [47], we build on the 16-layers VGG network [39], pre-trained on the 2012 ImageNet dataset. We adapt VGG for our purpose by modifying the last fully connected prediction layer to predict 20 classes for PASCAL VOC, and 80 classes for MS COCO. We then fine-tune the entire network for each of these datasets using two commonly used losses for multi-label classification, *softmax* and *binary cross-entropy (BCE)*² [15, 47]. To learn both classifiers, we set the weight decay to $5 \cdot 10^{-4}$, with a momentum of 0.9 and a dropout rate of 0.5. The learning rate is adjusted to gradually decrease after each epoch, starting from 0.01 for *softmax* and from 0.001 for *binary cross-entropy*. The learned parameters of these classifiers correspond to θ^* for our proposed deep set network (*cf.* Eq. (8) and Fig. 1).

²Weighted Approximate Ranking (WARP) objective is another commonly used loss for multi-label classification. However, it does not perform as well as *softmax* and *binary cross-entropy* for the used datasets [47].

To learn the cardinality distribution, we use the same VGG-16 network as above and modify the final fully connected layer to predict 2 values followed by two weighted sigmoid activation functions for α and β . It is important to note, that the outputs must be positive to describe a valid Gamma distribution. We therefore also append two weighted sigmoid transfer functions with weights α_M, β_M to ensure that the values predicted for α and β are in a valid range. Our model is not sensitive to these parameters and we set their values to be large enough ($\alpha_M = 160$ and $\beta_M = 20$) to guarantee that the mode of the distribution can accommodate the largest cardinality existing in the dataset. We then fine-tune the network on cardinality distribution using the objective loss defined in Eq. (9). To train the cardinality CNN, we set a constant learning rate 0.001, weight decay $5 \cdot 10^{-12}$, momentum rate 0.9 and dropout 0.5.

Evaluation protocol. To evaluate the performance of the classifiers and our deep set network, we employ the commonly used evaluation metrics for multi-label image classification [15, 47]: *precision* and *recall* of the generated labels per-class (C-P and C-R) and overall (O-P and O-R). Precision is defined as the ratio of correctly predicted labels and total predicted labels, while recall is the ratio of correctly predicted labels and ground-truth labels. In case no predictions (or ground truth) labels exist, *i.e.* the denominator becomes zero, precision (or recall) is defined as 100%. To generate the predicted labels for a particular image, we perform a forward pass of the CNN and choose top- k labels according to their scores similar to [15, 47]. Since the classifier always predicts a fixed-sized prediction for all categories, we sweep k from 0 to the maximum number of classes to generate a precision/recall curve, which is common practice in multi-label image classification. However, for our proposed DeepSet Network, the number of labels per instance is predicted from the cardinality network. Therefore, prediction/recall is not dependent on value k and one single precision/recall value can be computed.

To calculate the per-class and overall precision/recall, their average values over all classes and all examples are computed, respectively. In addition, we also report the F1 score (the harmonic mean of precision and recall) averaged over all classes (C-F1) and all instances and classes (O-F1).

PASCAL VOC 2007. The Pascal Visual Object Classes (VOC) [9] benchmark is one of the most widely used datasets for detection and classification. It consists of 9963 images with a 50/50 split for training and test, where objects from 20 pre-defined categories have been annotated by bounding boxes. Each image may contain between 1 and 7 unique objects. We compare our results with a state-of-the-art classifier as described above. The resulting precision/recall plots are shown in Fig. 4(a) together with our proposed approach using the estimated cardinality. Note

			
GT: ---	motorcycle	chair, dining-table, book, tv, couch, potted-plant, vase	person, chair, car, dining-table, cup, knife, fork, pizza, wine-glass
Prediction: ---	motorcycle	chair, dining-table, book, tv, couch, potted-plant, vase	person, chair, car, dining-table, cup, knife, fork, pizza, wine-glass

Figure 2: Qualitative results of our multi-class image labelling approach. For each image, the ground truth tags and our predictions are denoted below. Note that we show the exact output of our network, without any heuristics or post-processing.



				
GT: chair, cup, book, keyboard, mouse	banana	person, toothbrush	teddy-bear	oven
Prediction: chair, cup, book, keyboard, mouse, tv	banana, bottle	person, toothbrush, cell-phone	teddy-bear, bird, car, person	oven, toaster, fridge, bowl, sink, microwave

Figure 3: Interesting failure cases of our method. The “spurious” TV class predicted on the left is an artifact in annotation because in many examples, computer monitors are actually labelled as TV. In other cases, our network can correctly reason about the number of objects or concepts in the scene, but is constrained by a fixed list of categories defined in the dataset.

that by enforcing the correct cardinality for each image, we are able to clearly outperform the baseline w.r.t. both measures. Note also that our prediction (+) can nearly replicate the oracle (*), where the ground truth cardinality is known. The mean absolute cardinality error of our prediction on PASCAL VOC is 0.32 ± 0.52 .

Microsoft COCO. Another popular benchmark for image captioning, recognition, and segmentation is the recent Microsoft Common Objects in Context (MS-COCO) [26]. The dataset consists of 123 thousand images, each labelled with per instance segmentation masks of 80 classes. The number of unique objects for each image can vary between 0 and 18. Around 700 images in the training dataset do not contain any of the 80 classes and there are only a handful of images that have more than 10 tags. The majority of the images contain between one and three labels. We use 82783 images as training and validation split (90% - 10%), and the remaining 40504 images as test data. We predict the cardinality of objects in the scene with a mean absolute error of 0.74 and a standard deviation of 0.86.

Fig. 4(b) shows a significant improvement of precision and recall and consequently the F1 score using our deep set

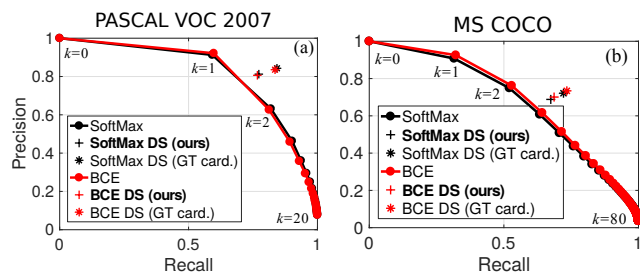


Figure 4: Experimental results on multi-label image classification. The baselines (solid curves) represent state-of-the-art classifiers, fine-tuned for each dataset, using two different loss functions. The methods are evaluated by choosing the top- k predictions across the entire dataset, for different k . Our approach predicts k and is thus evaluated only on one single point (+). It outperforms both classifiers significantly in terms of precision and recall and comes very close to the performance when the true cardinality is known (*).

network compared to the softmax and binary cross-entropy classifiers for all ranking values k . We also outperform the state-of-the-art multi-label classifier CNN-RNN [47], for the reported value of $k = 3$. Our results, listed in Tab. 1,

Table 1: Quantitative results for multi-label image classification on the MS COCO dataset.

Classifier	Eval.	C-P	C-R	C-F1	O-P	O-R	O-F1
Softmax	k=3	58.6	57.6	58.1	60.7	63.3	62.0
BCE	k=3	56.2	60.1	58.1	61.6	64.2	62.9
CNN-RNN [47]	k=3	66.0	55.6	60.4	69.2	66.4	67.8
Ours (Softmax)	k=m*	68.2	59.9	63.8	68.8	67.4	68.1
Ours (BCE)	k=m*	66.5	62.9	64.6	70.1	68.7	69.4

show around 7 percentage points improvement for the F1 score on top of the baseline classifiers and about 3 percentage points improvement compared to the state of the art on this dataset. Examples of perfect label prediction using our proposed approach are shown in Fig. 2. The deep set network can properly recognise images with no labels at all, as well as images with many tags. We also investigated failure cases where either the cardinality CNN or the classifier fails to make a correct prediction. We showcase some of these cases in Fig. 3. We argue here that some of the failure cases are simply due to a missed ground truth annotation, such as the left-most example, but some are actually semantically correct w.r.t. the cardinality prediction, but are penalized during evaluation because a particular object category is not available in the dataset. This is best illustrated in the second example in Fig. 3. Here, our network correctly predicts the number of objects in the scene, which is two, however, the can does not belong to any of the 80 categories in the dataset and is thus not annotated. Similar situations also appear in other images further to the right.

5.2. Object Counting

To show the robustness of our cardinality loss, we first evaluate our cardinality estimation on the common crowd counting application. To this end, we apply our approach on the widely used UCSD dataset [4] and compare our results to four state-of-the-art approaches [1, 32, 35, 50]. The UCSD dataset includes a 2000-frames long video sequence, captured by a fixed outdoor surveillance camera. In addition to the video, the region of interest (ROI), the perspective map of the scene and the location annotations of all pedestrians in each frame are also provided.

Implementation details. We build our cardinality network structure on top of the well-known AlexNet [22] architecture. However, we replace the first convolutional layer with a single channel filter to accept grayscale images as input, and the last fully connected layer with 2 layers outputs, similar to the case above (cf. Sec. 5.1). To estimate the counts, we calculate the mode of the negative binomial distribution.

As input, we use a grayscale image constructed by superimposing all region proposals and their scores generated by an off-the-shelf pedestrian detector (before non-maximum suppression). We use the multi-scale deep CNN approach (MS-CNN) [3] trained on the KITTI dataset [13] for our

Table 2: Count mean absolute error on UCSD dataset.

Method	max	downscale	upscale	min	overall
C-Forest [35]	1.43	1.30	1.59	1.62	1.49
IOC [1]	1.24	1.31	1.69	1.49	1.43
Cs-CCNN [50]	1.70	1.26	1.59	1.52	1.52
CCNN [32]	1.65	1.79	1.11	1.50	1.51
Hydra 2s [32]	2.22	1.93	1.37	2.38	1.98
Hydra 3s [32]	2.17	2.99	1.44	1.92	2.13
Ours	1.23	1.60	0.79	2.62	1.56

purpose. We found, that this input provides a stronger signal than the raw RGB images, yielding better results. Note that we process the input images with a pedestrian detector, however, we do not use any location or perspective information that is available for this dataset. During learning, we only rely on the object count for each image region.

We follow exactly the same data split used in [32] by conducting four different and separate experiments on maximal, downscale, upscale and minimal subsets in UCSD dataset. In order to train our network, similar to [32] we use data augmentation in each experiment by extracting 800 random patches from each training image and their corresponding ground truth counts. We also randomly flip each patch during training. To ensure that we can count all pedestrians from the entire image at test time, we choose the patch sizes to be exactly half of the image size (79×119 pixels) and then perform inference on the resulting 4 non-overlapping regions. The weights are initialised randomly and the network is trained for 100 epochs. All hyperparameters are set as in Sec. 5.1.

Results. Tab. 2 shows the mean absolute error between the predicted and the ground truth counts. We show competitive or superior performance in each experiment except for the ‘minimal’ subset. The main reason is that the training set size is too small (only 10 images) in this particular split and even data augmentation cannot generalize the cardinality model for the test images. Moreover, unlike other methods, we do not utilize any location information but only provide the object count as ground truth. Considering the overall performance, our approach outperforms state-of-the-art counting approaches that do not use the perspective map (Hydra 2s and 3s) and performs favourably compared to many existing methods that exploit localisation and perspective information.

Discussion. One obvious alternative for our proposed cardinality loss may seem to directly regress for m . This alternative, however, has two main drawbacks. First, it cannot be formulated within a Bayesian set framework to model uncertainty, and second, the regression loss does not yield a discrete distribution and hence does not fit the underlying mathematical foundation of this paper. Nevertheless, we have run the same experiments as above using a standard regression loss but did not reach the same performance. Using the regression loss we achieve a mean cardinality error

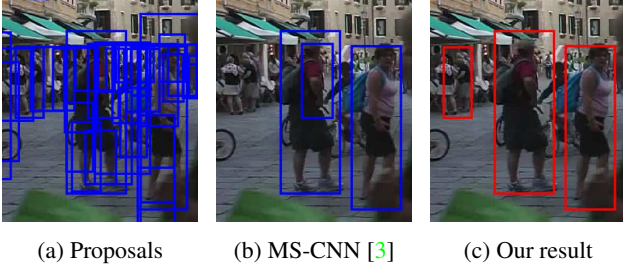


Figure 5: Example pedestrian detection result of our approach. To select relevant detection candidates from an overcomplete set of proposals (a), state-of-the-art methods rely on non-maximum suppression (NMS) with a fixed setting (b). We show that a better result can be achieved by adjusting the NMS threshold adaptively, depending on the number of instances in each image (3 in this case) (c).

(MCE) of 0.83 on MS COCO, while our loss yields an MCE of 0.74. This is also reflected in the O-F1 score which drops from 69.4 to 68.4 when directly regressing for m .

5.3. Pedestrian Detection

In this section, we cast the task of pedestrian detection as a set prediction problem and demonstrate that incorporating cardinality prediction (person count) can be beneficial to improve performance. To this end, we perform experiments on two widely used datasets, Caltech Pedestrians [8] and MOT16 from the MOTChallenge benchmark [29]. Recalling Eqs. (8) and (9), we need two networks with parameters \mathbf{w}^* and θ^* for cardinality estimation and detection proposals, respectively. For the cardinality network, we use the exact same architecture and setup as in Sec. 5.2 and train it on the training sets of these datasets. Note that it is not our intention to engineer a completely novel pedestrian detector here. Rather, for θ^* , we take an off-the-shelf state-of-the-art system (MS-CNN) [3] and show how it can be further improved by taking the cardinality prediction into account.

To generate the final detection outputs, most detectors often rely on non-maximum suppression (NMS), which greedily picks the boxes with highest scores and suppresses any boxes that overlap more than a pre-defined threshold T_O . This heuristic makes the solution more ad-hoc than what is expressed in our set formulation in Eq. (2). However, we are still able to improve the detector performance by adjusting this threshold for each frame separately. To obtain the final detection output, we use the prediction on the number of people (m^*) in the scene to choose an adaptive NMS threshold for each image. In particular, we start from the default value of T_O , and increase it gradually until the number of boxes reaches m^* . In the case if the number of final boxes is larger than m^* , we pick m^* boxes with the highest scores, which corresponds to the MAP set prediction as discussed in Sec. 4.3. To ensure a fair comparison, we also determine the best (global) value for $T_O = 0.4$ for

Table 3: Pedestrian detection results measured by F1 score (higher is better) and log-average miss rate (lower is better).

Method	F1-score \uparrow		MR \downarrow	
	Caltech	MOT16	Calt.	MOT16
MS-CNN [3]	51.61	59.04	60.9	82.8
MS-CNN-DS (ours)	52.15	61.86	60.4	81.7
MS-CNN-DS (GT card.)	52.28	62.42	60.3	81.5

the MS-CNN baseline. Fig. 5 demonstrates an example of the adjusted NMS threshold when considering the number of pedestrians in the image.

To quantify the detection performance, we adapt the same evaluation metrics and follow the protocols used on the Caltech detection benchmark [8]. The evaluation metrics used here are log-average miss rate (MR) over false positive per image. Additionally, we compute the F1 score (the harmonic mean of precision and recall). The F1 score is computed from *all* detections predicted from our DeepSet network and is compared with the *highest* F1 score along the MS-CNN precision-recall curve. To calculate MR, we concatenate all boxes resulted from our adaptive NMS approach and change the threshold over all scores from our predicted sets. Quantitative detection results are shown in Tab. 3. Note that we do not retrain the detector, but are still able to improve its performance by predicting the number of pedestrians in each frame in these two dataset.

6. Conclusion

We proposed a deep learning approach for predicting sets. To achieve this goal, we derived a loss for learning a discrete distribution over the set cardinality. This allowed us to use standard backpropagation for training a deep network for set prediction. We have demonstrated the effectiveness of this approach on crowd counting, pedestrian detection and multi-class image classification, achieving competitive or superior results in all three applications. As our network is trained independently, it can be trivially applied to any existing classifier or detector, to further improve performance.

Note that this decoupling is a direct consequence of our underlying mathematical derivation due to the *i.i.d.* assumptions, which renders our approach very general and applicable to a wide range of models. In future, we plan to extend our method to multi-class cardinality estimation and investigate models that do not make *i.i.d.* assumptions. Another potential avenue could be to exploit the Bayesian nature of the model to include uncertainty as opposed to relying on the MAP estimation alone.

Acknowledgments. This research was supported by the Australian Research Council through the Centre of Excellence in Robotic Vision, CE140100016, and through Laureate Fellowship FL130100102 to IDR.

References

- [1] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman. Interactive object counting. In *ECCV*, pages 504–518, 2014. [2](#), [7](#)
- [2] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool. Pedestrian detection at 100 frames per second. In *CVPR 2012*. [2](#)
- [3] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV 2016*. [7](#), [8](#)
- [4] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, pages 1–7, 2008. [7](#)
- [5] A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *CVPR*, pages 545–551, 2009. [2](#), [4](#)
- [6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015. [2](#)
- [7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR 2005*, pages 886–893. [2](#)
- [8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE T. Pattern Anal. Mach. Intell.*, 34, 2012. [8](#)
- [9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*. [5](#)
- [10] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE T. Pattern Anal. Mach. Intell.*, 28(4):594–611, Apr. 2006. [2](#)
- [11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE T. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. [2](#)
- [12] L. Fiaschi, U. Koethe, R. Nair, and F. A. Hamprecht. Learning to count with regression forest and structured labels. In *ICPR*, pages 2685–2688, 2012. [2](#)
- [13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI Vision Benchmark Suite. In *CVPR 2012*. [7](#)
- [14] R. Girshick. Fast R-CNN. In *ICCV 2015*. [2](#)
- [15] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013. [1](#), [5](#)
- [16] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *CoRR*, abs/1312.4894, 2013. [2](#)
- [17] A. Graves, A.-r. Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 6645–6649, 2013. [2](#)
- [18] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):17351780, Nov. 1997. [2](#)
- [19] J. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. In *CVPR 2017*, July 2017. [2](#)
- [20] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, 2013. [2](#)
- [21] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR 2016*. [1](#), [2](#)
- [22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*2012*, pages 1097–1105. [1](#), [2](#), [7](#)
- [23] D. Lee, G. Cha, M.-H. Yang, and S. Oh. Individualness and determinantal point processes for pedestrian detection. In *ECCV 2016*, pages 330–346. DOI: 10.1007/978-3-319-46466-4_20. [2](#)
- [24] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *NIPS*, pages 1324–1332, 2010. [2](#)
- [25] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR 2017*. [2](#)
- [26] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common objects in context. *arXiv:1405.0312 [cs]*, May 2014. arXiv: 1405.0312. [5](#), [6](#)
- [27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV 2016*, Lecture Notes in Computer Science, pages 21–37, 2016. DOI: 10.1007/978-3-319-46448-0_2. [1](#), [2](#)
- [28] R. P. Mahler. *Statistical multisource-multitarget information fusion*, volume 685. Artech House Boston, 2007. [3](#), [4](#), [5](#)
- [29] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. [8](#)
- [30] V. N. Murthy, V. Singh, T. Chen, R. Manmatha, and D. Comaniciu. Deep decision network for multi-class image classification. In *CVPR 2016*, June 2016. [2](#)
- [31] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR 2016*. [1](#)
- [32] D. Onoro-Rubio and R. J. López-Sastre. Towards perspective-free object counting with deep learning. In *ECCV*, pages 615–629, 2016. [7](#)
- [33] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV 2015*, Dec. 2015. [1](#)
- [34] T. T. Pham, S. Hamid RezaTofighi, I. Reid, and T.-J. Chin. Efficient point process inference for large-scale object detection. In *CVPR 2016*. [2](#)
- [35] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada. COUNT forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *ICCV*, pages 3253–3261, 2015. [2](#), [7](#)
- [36] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*2015*. [2](#)
- [37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252, 2015. [1](#)

- [38] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. 2
- [39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 5
- [40] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*2014*, pages 3104–3112. 2
- [41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 1
- [42] O. Vinyals, S. Bengio, and M. Kudlur. Order matters: Sequence to sequence for sets. *arXiv:1511.06391 [cs, stat]*, Nov. 2015. arXiv: 1511.06391. 2
- [43] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *NIPS*2015*, pages 2692–2700. 2
- [44] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004. 2
- [45] B.-N. Vo et al. Model-based classification and novelty detection for point pattern data. In *ICPR*, 2016. 3
- [46] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *CVPR 2010*. 2
- [47] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. CNN-RNN: A unified framework for multi-label image classification. In *CVPR*, June 2016. 1, 2, 5, 6, 7
- [48] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. CNN: Single-label to multi-label. *CoRR*, abs/1406.5726, 2014. 2
- [49] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV 2014*, pages 818–833. DOI: 10.1007/978-3-319-10590-1_53. 2
- [50] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, pages 833–841, 2015. 2, 7
- [51] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. 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. 2