



A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation

Vishwanath A. Sindagi^{a,**}, Vishal M. Patel^b

^aDept. of Electrical and Computer Engineering, 94 Brett Road, Piscataway, NJ 08854, USA

^bDept. of Electrical and Computer Engineering, 94 Brett Road, Piscataway, NJ 08854, USA

ABSTRACT

Estimating count and density maps from crowd images has a wide range of applications such as video surveillance, traffic monitoring, public safety and urban planning. In addition, techniques developed for crowd counting can be applied to related tasks in other fields of study such as cell microscopy, vehicle counting and environmental survey. The task of crowd counting and density map estimation is riddled with many challenges such as occlusions, non-uniform density, intra-scene and inter-scene variations in scale and perspective. Nevertheless, over the last few years, crowd count analysis has evolved from earlier methods that are often limited to small variations in crowd density and scales to the current state-of-the-art methods that have developed the ability to perform successfully on a wide range of scenarios. The success of crowd counting methods in the recent years can be largely attributed to deep learning and publications of challenging datasets. In this paper, we provide a comprehensive survey of recent Convolutional Neural Network (CNN) based approaches that have demonstrated significant improvements over earlier methods that rely largely on hand-crafted representations. First, we briefly review the pioneering methods that use hand-crafted representations and then we delve in detail into the deep learning-based approaches and recently published datasets. Furthermore, we discuss the merits and drawbacks of existing CNN-based approaches and identify promising avenues of research in this rapidly evolving field.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Crowd counting aims to count the number of people in a crowded scene where as density estimation aims to map an input crowd image to its corresponding density map which indicates the number of people per pixel present in the image (as illustrated in Fig. 1) and the two problems have been jointly addressed by researchers. The problem of crowd counting and density estimation is of paramount importance and it is essential for building higher level cognitive abilities in crowded scenarios such as crowd monitoring [15] and scene understanding [87, 115]. Crowd analysis has attracted significant attention from researchers in the recent past due to a variety of reasons. Exponential growth in the world population and the resulting urbanization has led to an increased number of activities such

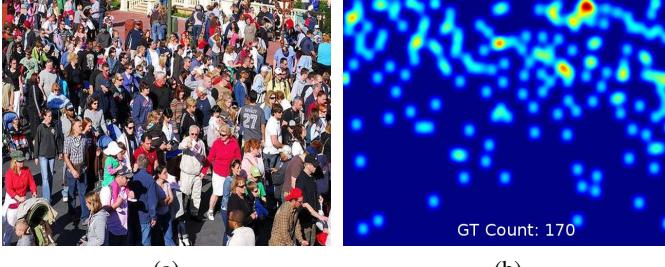
as sporting events, political rallies, public demonstrations etc. (shown in Fig. 2), thereby resulting in more frequent crowd gatherings in the recent years. In such scenarios, it is essential to analyze crowd behavior for better management, safety and security.

Like any other computer vision problem, crowd analysis comes with many challenges such as occlusions, high clutter, non-uniform distribution of people, non-uniform illumination, intra-scene and inter-scene variations in appearance, scale and perspective making the problem extremely difficult. Some of these challenges are illustrated in Fig. 2. The complexity of the problem together with the wide range of applications for crowd analysis has led to an increased focus by researchers in the recent past.

Crowd analysis is an inherently inter-disciplinary research topic with researchers from different communities (such as sociology [68, 10], psychology [5], physics [13, 38], biology [72, 110], computer vision and public safety) have addressed

****Corresponding author:**

e-mail: vishwanath.sindagi@rutgers.edu (Vishwanath A. Sindagi)



(a)

(b)

Fig. 1: Illustration of density map estimation. (a) Input image (b) Corresponding density map with count.



(a)

(b)



Fig. 2: Illustration of various crowded scenes and the associated challenges. (a) Parade (b) Musical concert (c) Public demonstration (d) Sports stadium. High clutter, overlapping of subjects, variation in scale and perspective can be observed across images.

the issue from different viewpoints. Crowd analysis has a variety of critical applications of inter-disciplinarian nature:

Safety monitoring: The widespread usage of video surveillance cameras for security and safety purposes in places such as sports stadiums, tourist spots, shopping malls and airports has enabled easier monitoring of crowd in such scenarios. However, traditional surveillance algorithms may break down as they are unable to process high density crowds due to limitations in their design. In such scenarios, we can leverage the results of algorithms specially designed for crowd analysis related tasks such as behavior analysis [83, 48], congestion analysis [114, 40], anomaly detection [56, 14] and event detection [8].

Disaster management: Many scenarios involving crowd gatherings such as sports events, music concerts, public demonstrations and political rallies face the risk of crowd related disasters such as stampedes which can be life threatening. In such cases, crowd analysis can be used as an effective tool for early overcrowding detection and appropriate management of crowd, hence, eventual aversion of any disaster [1, 3].

Design of public spaces: Crowd analysis on existing public spots such as airport terminals, train stations, shopping malls and other public buildings [23, 90] can reveal important design shortcomings from crowd safety and convenience point of view. These studies can be used for design of public spaces that are optimized for better safety and crowd movement [62, 2].

Intelligence gathering and analysis: Crowd counting techniques can be used to gather intelligence for further analysis and inference. For instance, in retail sector, crowd counting can be used to gauge people’s interest in a product in a store and this information can be used for appropriate product placement [58, 67]. Similarly, crowd counting can be used to measure queue lengths to optimize staff numbers at different times of the day. Furthermore, crowd counting can be used to analyze pedestrian flow at signals at different times of the day and this information can be used for optimizing signal-wait times [9].

Virtual environments: Crowd analysis methods can be used to understand the underlying phenomenon thereby enabling us to establish mathematical models that can provide accurate simulations. These mathematical models can be further used for simulation of crowd phenomena for various applications such as computer games, inserting visual effects in film scenes and designing evacuation plans [36, 74].

Forensic search: Crowd analysis can be used to search for sus-

pects and victims in events such as bombing, shooting or accidents in large gatherings. Traditional face detection and recognition algorithms can be speeded up using crowd analysis techniques which are more adept at handling such scenarios [47, 7].

These variety of applications has motivated researchers across various fields to develop sophisticated methods for crowd analysis and related tasks such as counting [15, 16, 20, 41, 17, 85, 35, 41], density estimation [52, 19, 111, 107, 75, 99, 11], segmentation [46, 27], behaviour analysis [6, 86, 22, 115, 114, 103], tracking [77, 116], scene understanding [87, 115] and anomaly detection [63, 56]. Among these, crowd counting and density estimation are a set of fundamental tasks and they form basic building blocks for various other applications discussed earlier. Additionally, methods developed for crowd counting can be easily extended to counting tasks in other fields such as cell microscopy [99, 97, 52, 20], vehicle counting [70], environmental survey [31, 105], etc.

Over the last few years, researchers have attempted to address the issue of crowd counting and density estimation using a variety of approaches such as **detection-based counting**, **clustering-based counting** and **regression-based counting** [61]. The initial work on regression-based methods mainly use hand-crafted features and the more recent works use Convolutional Neural Network (CNN) based approaches. The CNN-based approaches have demonstrated significant improvements over previous hand-crafted feature-based methods, thus, motivating more researchers to explore CNN-based approaches further for related crowd analysis problems. In this paper, we review various single image crowd counting and density estimation methods with a specific focus on recent CNN-based approaches.

Researchers have attempted to provide a comprehensive survey and evaluation of existing techniques for various aspects of

crowd analysis [105, 30, 44, 55, 117]. Zhan *et al.* [105] and Junior *et al.* [44] were among the first ones to study and review existing methods for general crowd analysis. Li *et al.* [55] surveyed different methods for crowded scene analysis tasks such as crowd motion pattern learning, crowd behavior, activity analysis and anomaly detection in crowds. More recently, Zitouni *et al.* [117] evaluated existing methods across different research disciplines by inferring key statistical evidence from existing literature and provided suggestions towards the general aspects of techniques rather than any specific algorithm. While these works focussed on the general aspects of crowd analysis, researchers have studied in detail crowd counting and density estimation methods specifically [61, 81, 79]. Loy *et al.* [61] provided a detailed description and comparison of video imagery-based crowd counting and evaluation of different methods using the same protocol. They also analyzed each processing module to identify potential bottlenecks to provide new directions for further research. In another work, Ryan *et al.* [79] presented an evaluation of regression-based methods for crowd counting across multiple datasets and provided a detailed analysis of performance of various hand-crafted features. Recently, Saleh *et al.* [81] surveyed two main approaches which are direct approach (i.e., object based target detection) and indirect approach (e.g. pixel-based, texture-based, and corner points based analysis).

Though existing surveys analyze various methods for crowd analysis and counting, they however cover only traditional methods that use hand-crafted features and do not take into account the recent advancements driven primarily by CNN-based approaches [87, 39, 113, 11, 85, 97, 4, 98, 111, 107, 70, 88] and creation of new challenging crowd datasets [106, 107, 111]. While CNN-based approaches have achieved drastically lower error rates, the creation of new datasets has enabled learning of more generalized models. To keep up with the rapidly advancing research in crowd counting, we believe it is necessary to analyze these methods in detail in order to understand the trends. Hence, in this paper, we provide **a survey of recent state-of-the-art CNN-based approaches for crowd counting and density estimation for single images.**

Rest of the paper is organized as follows: Section 2 briefly reviews the traditional crowd counting and density estimation approaches with an emphasis on the most recent methods. This is followed by a detailed survey on CNN-based methods along with a discussion on their merits and drawbacks in Section 3. In Section 5, recently published challenging datasets for crowd counting are discussed in detail along with results of the state-of-the-art methods. We discuss several promising avenues for achieving further progress in Section 6. Finally, concluding remarks are made in Section 7.

2. Review of traditional approaches

Various approaches have been proposed to tackle the problem of crowd counting in images [41, 19, 52, 107, 111] and videos [12, 35, 77, 21]. Loy *et al.* [61] broadly classified traditional crowd counting methods based on the approach into the following categories: (1) Detection-based approaches,

(2) Regression-based approaches, and (3) Density estimation-based approaches.

Since the focus of this work is on CNN-based approaches, in this section, we briefly review the detection and regression-based approaches using hand-crafted features for the sake of completeness. In addition, we present a review of the recent traditional methods [41, 52, 75, 99, 102] that have not been analyzed in earlier surveys.

2.1. Detection-based approaches

Most of the initial research was focussed on detection style framework, where a sliding window detector is used to detect people in the scene [26] and this information is used to count the number of people [54]. Detection is usually performed either in the monolithic style or parts-based detection. Monolithic detection approaches [25, 51, 94, 28] typically are traditional pedestrian detection methods which train a classifier using features (such as Haar wavelets [95], histogram oriented gradients [25], edgelet [100] and shapelet [80]) extracted from a full body. Various learning approaches such as Support Vector Machines, boosting [96] and random forest [34] have been used with varying degree of success. Though successful in low density crowd scenes, these methods are adversely affected by the presence of high density crowds. Researchers have attempted to address this issue by adopting part-based detection methods [29, 57, 101], where one constructs boosted classifiers for specific body parts such as the head and shoulder to estimate the people counts in a designated area [54]. In another approach using shape learning, Zhao *et al.* [112] modelled humans using 3D shapes composed of ellipsoids, and employed a stochastic process to estimate the number and shape configuration that best explains a given foreground mask in a scene. Ge and Collins [35] further extended the idea by using flexible and practical shape models.

2.2. Regression-based approaches

Though parts-based and shape-based detectors were used to mitigate the issues of occlusion, these methods were not successful in the presence of extremely dense crowds and high background clutter. To overcome these issues, researchers attempted to count by regression where they learn a mapping between features extracted from local image patches to their counts [16, 78, 20]. By counting using regression, these methods avoid dependency on learning detectors which is a relatively complex task. These methods have two major components: low-level feature extraction and regression modelling. A variety of features such as foreground features, edge features, texture and gradient features have been used for encoding low-level information. Foreground features are extracted from foreground segments in a video using standard background subtraction techniques. Blob-based holistic features such as area, perimeter, perimeter-area ration, etc. have demonstrated encouraging results [15, 20, 78]. While these methods capture global properties of the scene, local features such as edges and texture/gradient features such as local binary pattern (LBP), histogram oriented gradients (HOG), gray level co-occurrence matrices (GLCM) have been used to further improve the results.

Once these global and local features are extracted, different regression techniques such as linear regression [71], piecewise linear regression [15], ridge regression [20], Gaussian process regression and neural network [64] are used to learn a mapping from low-level feature to the crowd count.

In a recent approach, Idrees *et al.* [41] identified that no single feature or detection method is reliable enough to provide sufficient information for accurate counting in the presence of high density crowds due to various reasons such as low resolution, severe occlusion, foreshortening and perspective. Additionally, they observed that there exists a spatial relationship that can be used to constrain the count estimates in neighboring local regions. With these observations in mind, they proposed to extract features using different methods that capture different information. By treating densely packed crowds of individuals as irregular and non-homogeneous texture, they employed Fourier analysis along with head detections and SIFT interest-point based counting in local neighborhoods. The count estimates from this localized multi-scale analysis are then aggregated subject to global consistency constraints. The three sources, i.e., Fourier, interest points and head detection are then combined with their respective confidences and counts at localized patches are computed independently. These local counts are then globally constrained in a multi-scale Markov Random Field (MRF) framework to get an estimate of count for the entire image. The authors also introduced an annotated dataset (UCF_CC_50) of 50 images containing 64000 humans.

Chen *et al.* [19] introduced a novel cumulative attribute concept for learning a regression model when only sparse and imbalanced data are available. Considering that the challenges of inconsistent features along with sparse and imbalanced (encountered during learning a regression function) are related, cumulative attribute-based representation for learning a regression model is proposed. Specifically, features extracted from sparse and imbalanced image samples are mapped onto a cumulative attribute space. The method is based on the notion of discriminative attributes used for addressing sparse training data. This method is inherently capable of handling imbalanced data.

2.3. Density estimation-based approaches

While the earlier methods were successful in addressing the issues of occlusion and clutter, most of them ignored important spatial information as they were regressing on the global count. In contrast, Lempitsky *et al.* [52] proposed to learn a linear mapping between local patch features and corresponding object density maps, thereby incorporating spatial information in the learning process. In doing so, they avoided the hard task of learning to detect and localize individual object instances by introducing a new approach of estimating image density whose integral over any region in the density map gives the count of objects within that region. The problem of learning density maps is formulated as a minimization of a regularized risk quadratic cost function. A new loss function appropriate for learning density maps is introduced. The entire problem is posed as a convex optimization task which they solve using cutting-plane optimization.

Observing that it is difficult to learn a linear mapping, Pham *et al.* [75] proposed to learn a non-linear mapping between lo-

cal patch features and density maps. They used random forest regression from multiple image patches to vote for densities of multiple target objects to learn a non-linear mapping. In addition, they tackled the problem of large variation in appearance and shape between crowded image patches and non-crowded ones by proposing a crowdedness prior and they trained two different forests corresponding to this prior. Furthermore, they were able to successfully speed up the estimation process for real-time performance by proposing an effective forest reduction that uses permutation of decision trees. Apart from achieving real-time performance, another advantage of their method is that it requires relatively less memory to build and store the forest.

Similar to the above approach, Wang and Zou [99] identified that though existing methods are effective, they were inefficient from computational complexity point of view. To this effect, they proposed a fast method for density estimation based on subspace learning. Instead of learning a mapping between dense features and their corresponding density maps, they learned to compute the embedding of each subspace formed by image patches. Essentially, they exploited the relationship between images and their corresponding density maps in the respective feature spaces. The feature space of image patches are clustered and examples of each subspace are collected to learn its embedding. Their assumption that local image patches and their corresponding density maps share similar local geometry enables them to learn locally linear embedding using which the density map of an image patch can be estimated by preserving the geometry. Since, implementing locally linear embedding (LLE) is time-consuming, they divided the feature spaces of image patches and their counterpart density maps into subspaces, and computed the embedding of each subspace formed by image patches. The density map of input patch is then estimated by simple classification and mapping with the corresponding embedding matrix.

In a more recent approach, Xu and Qiu [102] observed that the existing crowd density estimation methods used a smaller set of features thereby limiting their ability to perform better. Inspired by the ability of high-dimensional features in other domains such as face recognition, they proposed to boost the performances of crowd density estimation by using a much extensive and richer set of features. However, since the regression techniques used by earlier methods (based on Gaussian process regression or Ridge regression) are computationally complex and are unable to process very high-dimensional features, they used random forest as the regression model whose tree structure is intrinsically fast and scalable. Unlike traditional approaches to random forest construction, they embedded random projection in the tree nodes to combat the curse of dimensionality and to introduce randomness in the tree construction.

3. CNN-based methods

The success of CNNs in numerous computer vision tasks has inspired researchers to exploit their abilities for learning non-linear functions from crowd images to their corresponding density maps or corresponding counts. A variety of CNN-based

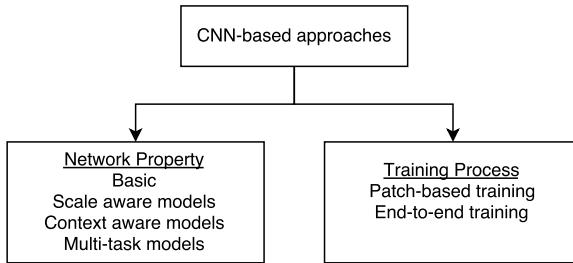


Fig. 3: Categorization of existing CNN-based approaches.

methods have been proposed in the literature. We broadly categorize these methods based on property of the networks and training approach as shown in Fig. 3. Based on the property of the networks, we classify the approaches into the following categories:

- **Basic CNNs:** Approaches that involve basic CNN layers in their networks fall into this category. These methods are amongst initial deep learning approaches for crowd counting and density estimation.
- **Scale-aware models:** The basic CNN-based approaches evolved into more sophisticated models that were robust to variations in scale. This robustness is achieved through different techniques such as multi-column or multi-resolution architectures.
- **Context-aware models:** Another set of approaches attempted to incorporate local and global contextual information present in the image into the CNN framework for achieving lower estimation errors.
- **Multi-task frameworks:** Motivated by the success of multi-task learning for various computer vision tasks, various approaches have been developed to combine crowd counting and estimation along with other tasks such as foreground-background subtraction and crowd velocity estimation.

In an yet another categorization, we classify the CNN-based approaches based on the inference methodology into the following two categories:

- **Patch-based inference:** In this approach, the CNNs are trained using patches cropped from the input images. Different methods use different crop sizes. During the prediction phase, a sliding window is run over the test image and predictions are obtained for each window and finally aggregated to obtain total count in the image.
- **Whole image-based inference:** Methods in this category perform a whole-image based inference. These methods avoid computationally expensive sliding windows.

Table 1 presents a categorization of various CNN-based crowd counting methods based on their network property and inference process.

3.1. Survey of CNN-based methods

In this section, we review various CNN-based crowd counting and density estimation methods along with their merits and

Table 1: Categorization of existing CNN-based approaches.

Method	Category	
	Network property	Inference process
Fu <i>et al.</i> [33]	Basic	Patch-based
Wang <i>et al.</i> [98]	Basic	Patch-based
Zhang <i>et al.</i> [107]	Multi-task	Patch-based
Boominathan <i>et al.</i> [11]	Scale-aware	Patch-based
Zhang <i>et al.</i> [111]	Scale-aware	Whole image-based
Walach and Wolf [97]	Basic	Patch-based
Onoro <i>et al.</i> [70]	Scale-aware	Patch-based
Shang <i>et al.</i> [85]	Context-aware	Whole image-based
Sheng <i>et al.</i> [89]	Context-aware	Whole image-based
Kumagai <i>et al.</i> [50]	Scale-aware	Patch-based
Marsden <i>et al.</i> [65]	Scale-aware	Whole image-based
Mundhenk <i>et al.</i> [69]	Basic	Patch-based
Artetta <i>et al.</i> [4]	Multi-task	Patch-based
Zhao <i>et al.</i> [113]	Multi-task	Patch-based
Sindagi <i>et al.</i> [92]	Multi-task	Whole image-based
Sam <i>et al.</i> [82]	Scale-aware	Patch-based
Kang <i>et al.</i> [113]	Basic	Patch-based

drawbacks.

Wang *et al.* [98] and Fu *et al.* [33] were among the first ones to apply CNNs for the task of crowd density estimation. Wang *et al.* proposed an end-to-end deep CNN regression model for counting people from images in extremely dense crowds. They adopted AlexNet network [49] in their architecture where the final fully connected layer of 4096 neurons is replaced with a single neuron layer for predicting the count. Besides, in order to reduce false responses background like buildings and trees in the images, training data is augmented with additional negative samples whose ground truth count is set as zero. In a different approach, Fu *et al.* proposed to classify the image into one of the five classes: very high density, high density, medium density, low density and very low density instead of estimating density maps. Multi-stage ConvNet from the works of Sermanet *et al.* [84] was adopted for better shift, scale and distortion invariance. In addition, they used a cascade of two classifiers to achieve boosting in which the first one specifically samples misclassified images whereas the second one reclassifies rejected samples.

Zhang *et al.* [107] analyzed existing methods to identify that their performance reduces drastically when applied to a new scene that is different from the training dataset. To overcome this issue, they proposed to learn a mapping from images to crowd counts and to adapt this mapping to new target scenes for cross-scene counting. To achieve this, they first learned their network by alternatively training on two objective functions: crowd count and density estimation which are related objectives. By alternatively optimizing over these objective functions one is able to obtain better local optima. In order to adapt this network to a new scene, the network is fine-tuned using training samples that are similar to the target scene. It is important to note that the network is adapted to new target scenes without any extra label information. The overview of their approach is shown in Fig. 4. Also, in contrast to earlier methods

that use the sum of Gaussian kernels centered on the locations of objects, a new method for generating ground truth density map is proposed that incorporates perspective information. In doing so, the network is able to perform perspective normalization thereby achieving robustness to scale and perspective variations. Additionally, they introduced a new dataset for the purpose of evaluating cross-scene crowd counting. The network is evaluated for cross-scene crowd counting as well as single scene crowd counting and superior results are demonstrated for both scenarios.

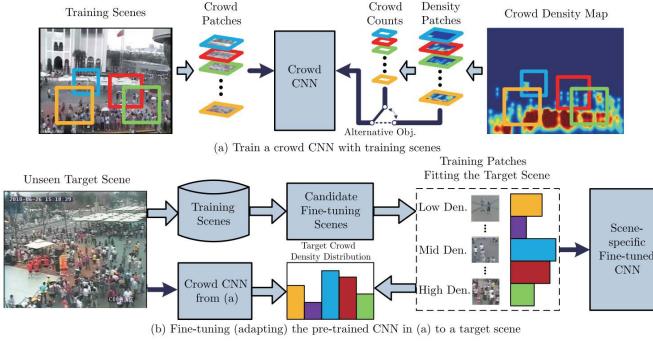


Fig. 4: Overview of cross scene crowd counting proposed by Zhang *et al.* [107].

Inspired by the success of cross-scene crowd counting [107], Walach and Wolf [97] performed layered boosting and selective sampling. Layered boosting involves iteratively adding CNN layers to the model such that every new layer is trained to estimate the residual error of the earlier prediction. For instance, after the first CNN layer is trained, the second CNN layer is trained on the difference between the estimation and ground truth. This layered boosting approach is based on the notion of Gradient Boosting Machines (GBM) [32] which are a subset of powerful ensemble techniques. An overview of their boosting approach is presented in Fig. 5. The other contribution made by the authors is the use of sample selection algorithm to improve the training process by reducing the effect of low quality samples such as trivial samples or outliers. According to the authors, the samples that are correctly classified early on are trivial samples. Presenting such samples for training even after the networks have learned to classify them tends to introduce bias in the network for such samples, thereby affecting its generalization performance. Another source of training inefficiency is the presence of outliers such as mislabeled samples. Apart from affecting the network's performance, these samples increase the training time. To overcome this issue, such samples are eliminated out of the training process for a number of epochs. The authors demonstrated that their method reduces the count estimation error by 20% to 30% over existing state-of-the-art methods at that time on different datasets.

In contrast to the above methods that use patch-based training, Shang *et al.* [85] proposed an end-to-end count estimation method using CNNs (Fig. 6). Instead of dividing the image into patches, their method takes the entire image as input and directly outputs the final crowd count. As a result, computations on overlapping regions are shared by combining multiple

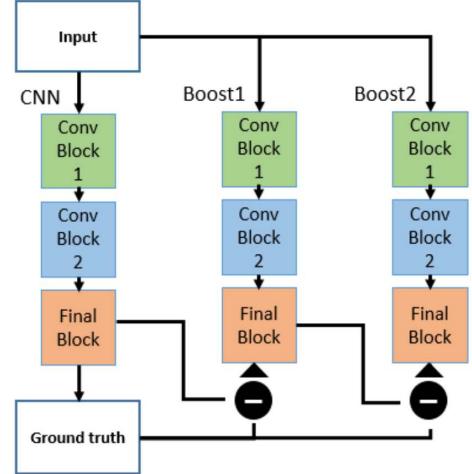


Fig. 5: Overview of learning to count using boosting by Walach and Wolf [97].

stages of processing leading to a reduction of complexity. The network simultaneously learns to estimate local counts and can be viewed as learning a patch level counting model which enables faster training. By doing so, contextual information is incorporated into the network, enabling it to ignore background noises and achieve better performance. The network is composed of three parts: (1) Pre-trained GoogLeNet model [93], (2) Long-short time memory (LSTM) decoders for local count, and (3) Fully connected layers for the final count. The network takes an image as input and computes high-dimensional CNN feature maps using the GoogleNet network. Local blocks in these high-dimensional features are decoded into local count using a LSTM unit. A set of fully connected layers after the LSTM unit map the local counts into global count. The two counting objectives are jointly optimized during training.

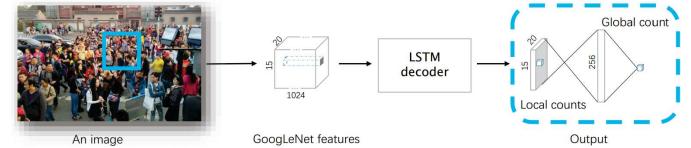


Fig. 6: Overview of the end-to-end counting method proposed by Shang *et al.* [85]. GoogLeNet is used to compute high-dimensional features which are further decoded into local counts using LSTM units.

In an effort to capture semantic information in the image, Boominathan *et al.* [11] combined deep and shallow fully convolutional networks to predict the density map for a given crowd image. The combination of two networks enables one to build a model robust to non-uniform scaling of crowd and variations in perspective. Furthermore, an extensive augmentation of the training dataset is performed in two ways. Patches from the multi-scale image representation are sampled to make the system robust to scale variations. Fig. 7 shows overview of this method.

In another approach, Zhang *et al.* [111] proposed a multi-column based architecture (MCNN) for images with arbitrary

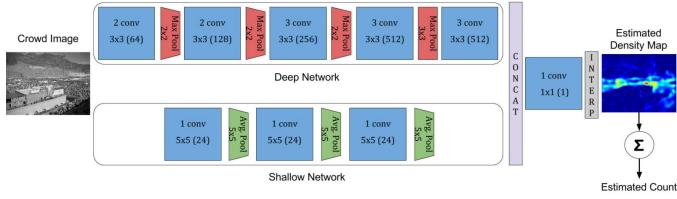


Fig. 7: Overview of counting method proposed by Boominathan *et al.* [11]. A deep network is used in combination with a shallow network to address scale variations across images.

crowd density and arbitrary perspective. Inspired by the success of multi-column networks for image recognition [24], the proposed method ensures robustness to large variation in object scales by constructing a network that **comprises of three columns corresponding to filters with receptive fields of different sizes (large, medium, small)** as shown in Fig. 8. These different columns are designed to cater to different object scales present in the images. Additionally, a new method for generating ground truth crowd density maps is proposed. In contrast to existing methods that either use sum of Gaussian kernels with a fixed variance or perspective maps, Zhang *et al.* proposed to take into account perspective distortion by estimating spread parameter of the Gaussian kernel based on the size of the head of each person within the image. However, it is impractical to estimate head sizes and their underlying relationship with density maps. Instead they used an important property observed in high density crowd images that the head size is related to distance between the centers of two neighboring persons. The spread parameter for each person is data-adaptively determined based on its average distance to its neighbors. Note that the ground truth density maps created using this technique incorporate distortion information without the use of perspective maps. Finally, considering that existing crowd counting datasets do not cater to all the challenging situations encountered in real world scenarios, a new **ShanghaiTech crowd datasets is constructed**. This new dataset includes 1198 images with about 330,000 annotated heads.

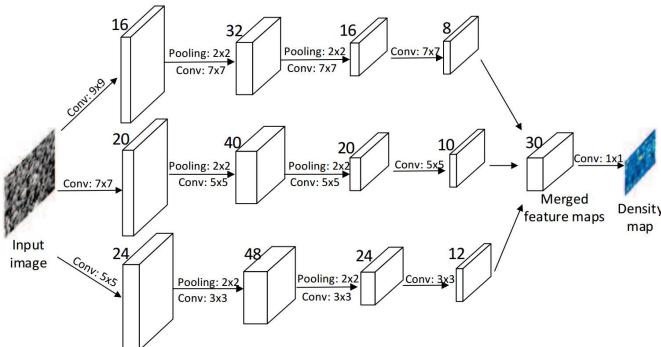


Fig. 8: Overview of single image crowd counting via multi-column network by Zhang *et al.* [111].

Similar to the above approach, Onoro and Sastre [70] developed a scale aware counting model called Hydra CNN that is able to estimate object densities in a variety of crowded sce-

narios without any explicit geometric information of the scene. First, a deep fully-convolutional neural network (which they call as Counting CNN) with six convolutional layers is employed. Motivated by the observation of earlier work [107, 61] that incorporating perspective information for geometric correction of the input features results in better accuracy, geometric information is incorporated into the Counting CNN (CCNN). To this effect, they developed Hydra CNN that learns a multi-scale non-linear regression model. As shown in Fig. 9 **the network consists of 3 heads and a body with each head learning features for a particular scale**. Each head of the Hydra-CNN is constructed using the CCNN model whose outputs are concatenated and fed to the body. The body consists of a set of two fully-connected layers followed by a rectified linear unit (ReLU), a dropout layer and a final fully connected layer to estimate the object density map. While the different heads extract image descriptors at different scales, the body learns a high-dimensional representation that fuses the multi-scale information provided by the heads. This network design of Hydra CNN is inspired by the work of Li *et al.* [53]. Finally, the network is trained with pyramid of image patches extracted at multiple scales. The authors demonstrated through their experiments that the Hydra CNN is able to perform successfully in scenarios and datasets with significant variations in the scene.

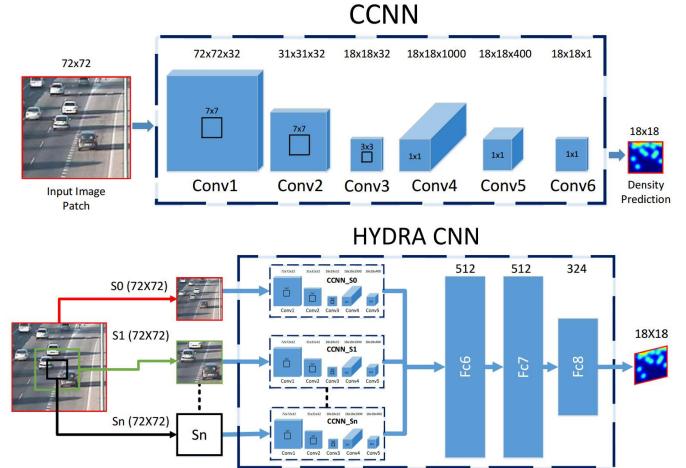


Fig. 9: Overview of Hydra-CNN by Onoro *et al.* [70].

Instead of training all regressors of a multi-column network [111] on all the input patches, Sam *et al.* [82] argue that **better performance is obtained by training regressors with a particular set of training patches by leveraging variation of crowd density within an image**. To this end, they proposed a switching CNN that **cleverly** selects an optimal regressor suited for a particular input patch. As shown in Fig. 10, the proposed network consists of multiple independent regressors similar to multi-column network [111] with different receptive fields and a switch classifier. The switch classifier is trained to select the optimal regressor for a particular input patch. Independent CNN crowd density regressors are trained on patches sampled from a grid in a given crowd scene. The switch classifier and the independent regressors are alternatively trained. The authors describe multiple stages of training their network. First, the independent

regressors are pretrained on image patches to minimize the Euclidean distance between the estimated density map and ground truth. This is followed by a differential training stage where, the count error is factored in to improve the counting performance by back-propagating a regressor with the minimum count error for a given training patch. After training the multiple regressors, a switch classifier based on VGG-16 architecture [91] is trained to select an optimal regressor for accurate counting. Finally, the switch classifier and CNN regressors are co-adapted in the coupled training stage.

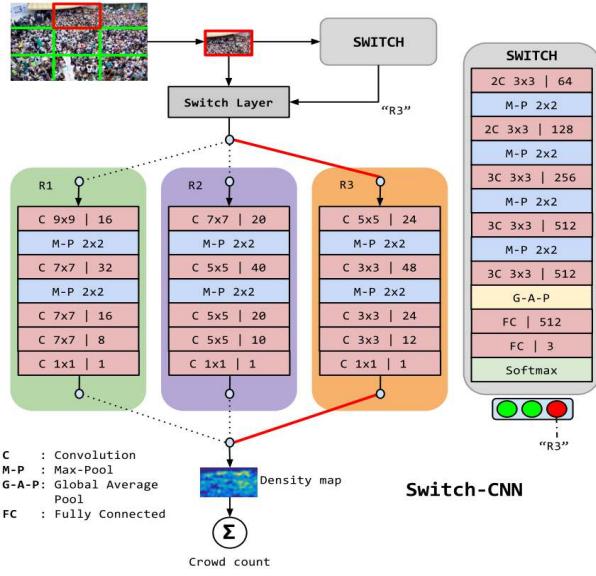


Fig. 10: Overview of Switching CNN by Sam *et al.* [82].

While the above methods concentrated on incorporating scale information in the network, Sheng *et al.* in [89] proposed to integrate semantic information by learning locality-aware feature sets. Noting that earlier methods that use hand-crafted features ignored key semantic and spatial information, the authors proposed a new image representation which incorporates semantic attributes as well as spatial cues to improve the discriminative power of feature representations. They defined semantic attributes at the pixel level and learned semantic feature maps via deep CNN. The spatial information in the image is encoded using locality-aware features in the semantic attribute feature map space. The locality-aware features (LAF) are built on the idea of spatial pyramids on neighboring patches thereby encoding spatial context and local information. The local descriptors from adjacent cells are then encoded into image representations using weighted VLAD encoding method.

Similar to [111, 70], Kumagai *et al.* [50], based on the observation that a single predictor is insufficient to appropriately predict the count in the presence of large appearance changes, proposed a Mixture of CNNs (MoCNN) that are specialized to a different scene appearances. As shown in Fig. 11, the architecture consists of a mixture of expert CNNs and a gating CNN that adaptively selects the appropriate CNN among the experts according to the appearance of the input image. For prediction, the expert CNNs predict crowd count in the image while the

gating CNN predicts appropriate probabilities for each of the expert CNNs. These probabilities are further used as weighting factors to compute the weighted average of the counts predicted by all the expert CNNs.

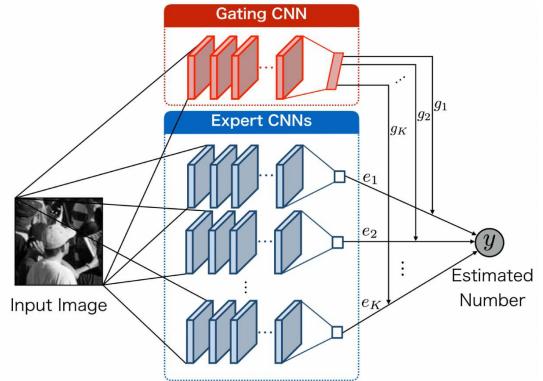


Fig. 11: Overview of MoC (Mixture of CNN) for crowd counting by Kumagai *et al.* [50].

Motivated by the success of scale aware models [111, 70], Marsden *et al.* [65] proposed to incorporate scale into the models with much less number of model parameters. Observing that the earlier scale aware models [111, 70] are difficult to optimize and are computationally complex, Marsden *et al.* [65] proposed a single column fully convolutional network where the scale information is incorporated into the model using a simple yet effective multi-scale averaging step during prediction without any increase in the model parameters. The method addresses the issues of scale and perspective changes by feeding multiple scales of test image into the network during prediction phase. The crowd count is estimated for each scale and the final count is obtained by taking an average of all the estimates. Additionally, a new training set augmentation scheme is developed to reduce redundancy among the training samples. In contrast to the earlier methods that use randomly cropped patches with high degree of overlap, the training set in this work is constructed using the four image quadrants as well as their horizontal flips ensuring no overlap. This technique avoids potential overfit when the network is continuously exposed to the same set of pixels during training, thereby improving the generalization performance of the network. In addition, the generalization performance of the proposed method is studied by measuring cross dataset performance.

Inspired by the superior results achieved by simultaneous learning of related tasks [76, 104], Sindagi *et al.* [92] and Marsden *et al.* [66] explored multi-task learning to boost individual task performance. Marsden *et al.* [66] proposed a Resnet-18 [37] based architecture for simultaneous crowd counting, violent behaviour detection and crowd density level classification. The network consists of initial 5 convolutional layers of Resnet18 including batch normalisation layers and skip connections form the primary module. The convolutional layers are followed by a set of task specific layers. Finally, sum of all the losses corresponding to different tasks is minimized. Additionally, the authors constructed a new 100 image dataset specifi-



Fig. 12: Overview of Fully Convolutional Network for crowd counting by Marsden *et al.* [65].

cally designed for multi-task learning of crowd count and behaviour. In a different approach, Sindagi *et al.* [92] proposed a cascaded CNN architecture to incorporate learning of a high-level prior to boost the density estimation performance. Inspired by [18], the proposed network simultaneously learns to classify the crowd count into various density levels and estimate density map (as shown in Fig. 13). Classifying crowd count into various levels is equivalent to coarsely estimating the total count in the image thereby incorporating a high-level prior into the density estimation network. This enables the layers in the network to learn globally relevant discriminative features. Additionally, in contrast to most recent work, they make use of transposed convolutional layers to generate high resolution density maps.

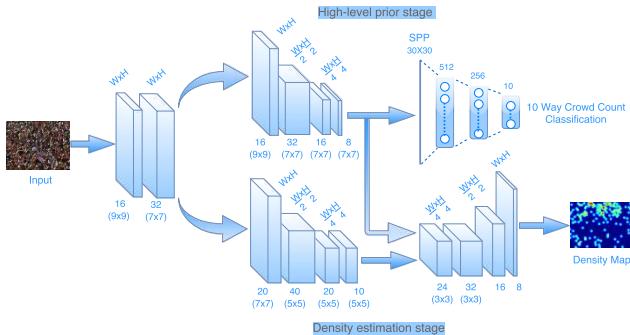


Fig. 13: Overview of Cascaded Multi-task CNN by Sindagi *et al.* [92].

In a recent work, Kang *et al.* [45] explored maps generated by density estimation methods for the purpose of various crowd analysis tasks such as counting, detection and tracking. They performed a detailed analysis of the effect of using full-resolution density maps on the performance of these tasks. They demonstrated through their experiments that full resolution density maps improved the performance of localization tasks such as detection and tracking. Two different approaches

are considered for generating full-resolution maps. In the first approach, a sliding window based CNN regressor is used for pixel-wise density prediction. In the second approach, Fully Convolutional Networks [60] along with skip connections are used to learn a non-linear mapping between input image and the corresponding density map.

In a slightly different application context of counting, Mundhenk *et al.* [69] and Arteta *et al.* [4] proposed to count different types of objects such as cars and penguins respectively. Mundhenk *et al.* [69] addressed the problem of automated counting of automobiles from satellite/aerial platforms. Their primary contribution is the creation of a large diverse set of cars from overhead images. Along with the large dataset, they present a deep CNN-based network to recognize the number of cars in patches. The network is trained in a classification setting where the output of the network is a class that is indicative of the number of objects in the input image. Also, they incorporated contextual information by including additional regions around the cars in the training patches. Three different networks based on AlexNet [49], GoogLeNet [93] and ResNet [37] with Inception are evaluated. For a different application of counting penguins in images, Arteta *et al.* [4] proposed a deep multi-task architecture for accurate counting even in the presence of labeling errors. The network is trained in a multi-task setting where, the tasks of foreground-background subtraction and uncertainty estimation along with counting are jointly learned. The authors demonstrated that the joint learning especially helps in learning a counting model that is robust to labeling errors. Additionally, they exploited scale variations and count variability across the annotations to incorporate scale information of the object and prediction of annotation difficulty respectively into the model. The network was evaluated on a newly created Penguin dataset.

Zhao *et al.* addressed a higher level cognitive task of counting people that cross a line in [113]. Though the task is a video-based application, it comprises of a CNN-based model that is trained with pixel-level supervision maps similar to single image crowd density estimation methods, making it a relevant approach to be included in this article. Their method consists of a two-phase training scheme (as shown in Fig. 14) that decomposes original counting problem into two sub-problems: estimating crowd density map and crowd velocity map where the two tasks share the initial set of layers enabling them to learn more effectively. The estimated crowd density and crowd velocity maps are then multiplied element-wise to generate the crowd counting maps. Additionally, they contributed a large-scale dataset for evaluating crossing-line crowd counting algorithms, which includes 5 different scenes, 3,100 annotated frames and 5,900 annotated pedestrians.

4. Discussion

With a variety of methods discussed in Section 3, we analyze various advantages and disadvantages of the broad approaches followed by these methods in this section.

Zhang *et al.* [107] were among the first ones to address the problem of adapting models to new unlabelled datasets using a simple and effective method based on finding similar patches

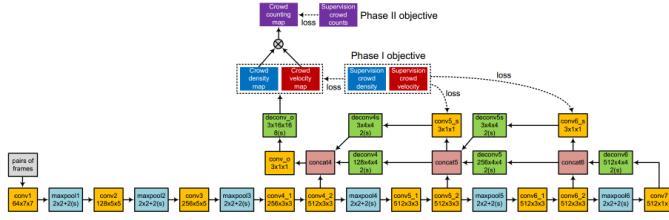


Fig. 14: Overview of the method proposed by Zhao *et al.* [113] for counting people crossing a line.

across datasets. However, their method is heavily dependent on accurate perspective maps which may not be necessarily available for all the datasets. Additionally, the use of 72×72 sized patches for training and evaluation ignores global context which is necessary for accurate estimation of count. Walach *et al.* [97] successfully addressed training inefficiencies in earlier methods using a layered boosting approach and a simple sample selection method. However, similar to Zhang *et al.* [107], their method involves patch-based training and evaluation resulting in loss of global context information along with inefficiency during evaluation due to the use of a sliding window approach. Additionally, these methods tend to ignore scale variance among the dataset assuming that their models will implicitly learn the invariance.

In an effort to explicitly model scale invariance, several methods involving combination of networks were proposed ([111, 70, 82, 50, 11]). While these methods demonstrated significant improvements in the performance using multiple column networks and a combination of deep and shallow networks, the invariance achieved is limited by the number of columns present in the network and receptive field sizes which are chosen based on the scales present in the dataset. Additionally, these methods do not explicitly model global context information which is crucial for a task such as crowd counting. In a different approach, Marsden *et al.* [65] attempt to address the scale issue by performing a multi-scale averaging during the prediction phase. While being simple and effective, it results in an inefficient inference stage. Additionally, these methods do not explicitly encode global context present in an image which can be crucial for improving the count performance. To this end, few approaches model local and global context [89, 85] by considering key spatial and semantic information present in the image.

In an entirely different approach, few methods [66, 92] take advantage of multi-task learning and incorporate high-level priors into the network. For instance, Sindagi *et al.* [92] simultaneously learn density estimation and a high-level prior in the form of crowd count classification. While they demonstrated high performance gain by learning an additional task of crowd density level classification, the number of density levels is dataset dependent and it needs to be carefully chosen based on the density levels present in the dataset.

5. Datasets and results

A variety of datasets have been created over the last few years driving researchers to create models with better generalization abilities. While the earlier datasets usually contain low density crowd images, the most recent ones focus on high density crowd thus posing numerous challenges such as scale variations, clutter and severe occlusion. The creation of these large scale datasets has motivated recent approaches to develop methods that cater to such challenges. In this section, we review five key datasets [15, 20, 41, 107, 111] followed by a discussion on the results of CNN-based approaches and recent traditional methods that were not included in the earlier surveys.

5.1. Datasets

UCSD dataset: The UCSD dataset [15] was among the first datasets to be created for counting people. The dataset was collected from a video camera at a pedestrian walkway. The dataset consists of 2000 frames of size 238×158 from a video sequence along with ground truth annotations of each pedestrian in every fifth frame. For the rest of the frames, linear interpolation is used to create the annotations. A region-of-interest is also provided to ignore unnecessary moving objects such as trees. The dataset contains a total of 49,885 pedestrian instances and it is split into training and test set. While the training set contains frames with indices 600 to 1399, the test set contains the remaining 1200 images. This dataset has relatively low density crowd with an average of around 15 people in a frame and since the dataset was collected from a single location, there is no variation in the scene perspective across images.

Mall dataset: Considering little variation in the scene type in the UCSD dataset, Chen *et al.* in [20] collected a new Mall dataset with diverse illumination conditions and crowd densities. The dataset was collected using a surveillance camera installed in a shopping mall. Along with having various density levels, it also has different activity patterns (static and moving crowds). Additionally, the scene contained in the dataset has severe perspective distortion resulting in large variations in size and appearance of objects. The dataset also presents the challenge of severe occlusions caused by the scene objects, e.g. stall, indoor plants along the walking path. The video sequence in the dataset consists of 2000 frames of size 320×240 with 6000 instances of labelled pedestrians. The first 800 frames are used for training and the remaining 1200 frames are used for evaluation. In comparison to the UCSD dataset, the Mall dataset has relatively higher crowd density images. However, both the datasets do not have any variation in the scene perspective across images since they are a part of a single continuous video sequence.

UCF_CC_50 dataset: The UCF_CC_50 [41] is the first truly challenging dataset constructed to include a wide range of densities and diverse scenes with varying perspective distortion. The dataset was created from publicly available web images. In order to capture diversity in the scene types, the authors collected images with different tags such as concerts, protests, stadiums and marathons. It contains a total of 50 images

Table 2: Summary of various datasets.

Dataset	No. of images	Resolution	Min	Ave	Max	Total count
UCSD [15]	2000	158x238	11	25	46	49,885
Mall [20]	2000	320x240	13	-	53	62,325
UCF_CC_50 [41]	50	Varied	94	1279	4543	63,974
WorldExpo '10 [106, 107]	3980	576x720	1	50	253	199,923
ShanghaiTech Part A [111]	482	Varied	33	501	3139	241,677
ShanghaiTech Part B [111]	716	768x1024	9	123	578	88,488

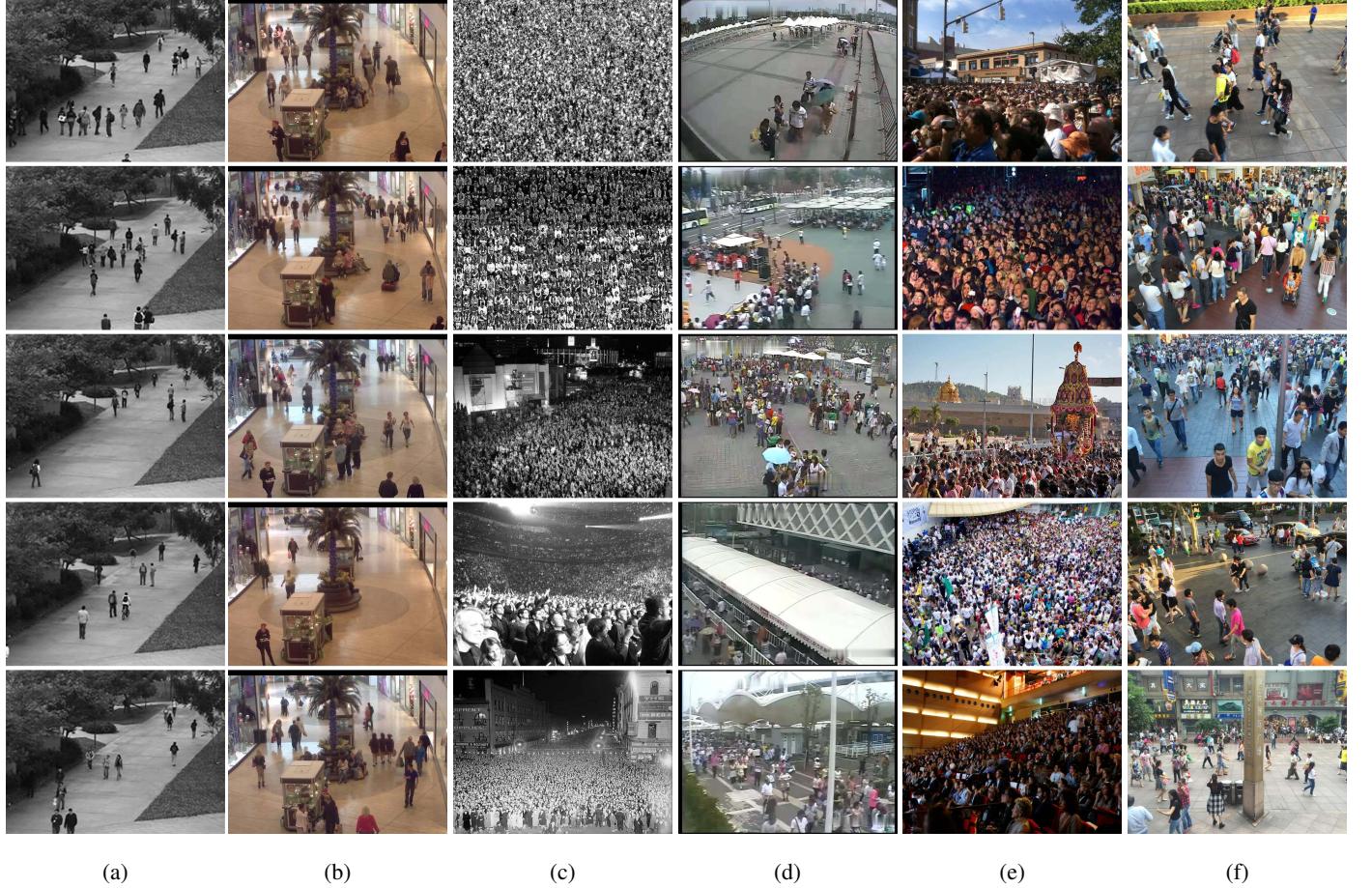


Fig. 15: Sample images from various datasets. (a) UCSD [15] (b) Mall [20] (c) UCF_CC_50 [41] (d) WorldExpo '10 [107] (e) Shanghai Tech Part A [111] (f) SHanghai Tech Part B [111]. It can be observed that in the case of UCSD and Mall dataset , the images come from the same video sequence providing no variation in perspective across images.

of varying resolutions with an average of 1280 individuals per image. A total of 63075 individuals were labelled in the entire dataset. The number of individuals varies from 94 to 4543 indicating a large variation across the images. The only drawback of this dataset is that only a limited number of images are available for training and evaluation. Considering the low number of images, the authors defined a cross-validation protocol for training and testing their approach where the dataset was divided into sets of 10 and a five fold cross-validation is performed. The challenges posed by this dataset are so enormous that even the results of recent CNN-based state-of-the-art approaches on this dataset are far from optimal.

WorldExpo '10 dataset: Since some of the earlier approaches and datasets focussed primarily on single scene counting, Zhang *et al.* [107] introduced a dataset for the purpose of cross-scene crowd counting. The authors attempted to perform a data-driven cross-scene crowd counting for which they collected a new large-scale dataset that includes 1132 annotated video sequences captured by 108 surveillance cameras, all from Shanghai 2010 WorldExpo event. Large diversity in the scene types is ensured by collecting videos from cameras having disjoint bird views. The dataset consists of a total of 3980 frames of size 576×720 with 199923 labelled pedestrians. The dataset is split into two parts: training set consisting of 1,127 one-minute long video sequences from 103 scenes and

test set consisting of 5 one-hour long video sequences from 5 different scenes. Each test scene consists of 120 labelled frames with the crowd count varying from 1 to 220. Though an attempt is made to capture diverse scenes with varying density levels, the diversity is limited to only 5 scenes in the test set and the maximum crowd count is limited to 220. Hence, the dataset is not sufficient enough for evaluating approaches designed for extremely dense crowds in a variety of scenes.

Shanghai Tech dataset: Zhang *et al.* [111] introduced a new large-scale crowd counting dataset consisting of 1198 images with 330,165 annotated heads. The dataset is among the largest ones in terms of the number of annotated people and it contains two parts: Part A and Part B. Part A consists of 482 images that are randomly chosen from the Internet whereas Part B consists of images taken from the streets of metropolitan areas in Shanghai. Part A has considerably larger density images as compared to Part B. Both the parts are further divided into training and evaluation sets. The training and test of Part A has 300 and 182 images, respectively, whereas that of Part B has 400 and 316 images, respectively. The dataset successfully attempts to create a challenging dataset with diverse scene types and varying density levels. However, the number of images for various density levels are not uniform making the training and evaluation biased towards low density levels. Nevertheless, the complexities present in this dataset such as varying scales and perspective distortion has created new opportunities for more complex CNN network designs.

Sample images from the five datasets are shown in Fig. 15. The datasets are also summarized in Table 2. It can be observed that the UCSD and the Mall dataset have relatively low density images and typically focus on single scene type. In contrast, the other datasets have significant variations in the density levels along with different perspectives across images.

5.2. Discussion on results

Results of the recent traditional approaches along with CNN-based methods are tabulated in Table 3. The count estimation errors are reported directly from the respective original works. The following standard metrics are used to compare different methods:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - y'_i|, \quad (1)$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |y_i - y'_i|^2}, \quad (2)$$

where MAE is mean absolute error, MSE is mean squared error, N is the number of test samples, y_i is the ground truth count and y'_i is the estimated count corresponding to the i^{th} sample. We make the following observations regarding the results:

- In general, CNN-based methods outperform the traditional approaches across all datasets.
- While the CNN-based methods are especially effective in large density crowds with a diverse scene conditions, the

traditional approaches suffer from high error rates in such scenarios.

- Among the CNN-based methods, most performance improvement is achieved by scale-aware and context-aware models. It can be observed from Table 3 that a reduction in count error is largely driven by the increase in the complexity of CNN models (due to addition of context and scale information).
- While the multi-column CNN architecture [111] achieves the state-of-the-art results on 3 datasets: UCSD, World-Expo '10 and ShanghaiTech, the CNN-boosting approach by [97] achieves the best results on the Mall dataset. The best results on the UCF_CC_50 dataset are achieved by joint local and global count approach [85] and Hydra-CNN [70].
- The work in [97] suggests that layered boosting can achieve performances that are comparable to scale aware models.
- The improvements obtained by selective sampling in [98] and [97] suggests that it helps to obtain unbiased performance.
- Whole image-based methods such as Zhang *et al.* [111] and Shang *et al.* [85] are less computationally complex from the prediction point of view and they have proved to achieve better results over patch-based techniques.
- Finally, techniques such as layered boosting and selective sampling [70, 99] not only improve the estimation error but also reduce the training time significantly.

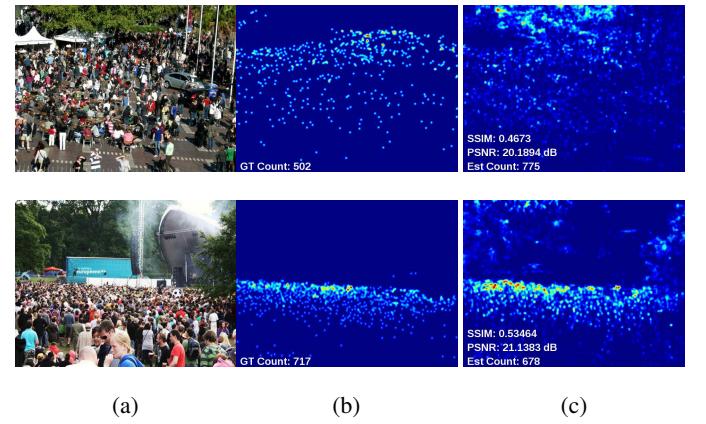


Fig. 16: Results of Zhang *et al.* [111] on ShanghaiTech dataset. (a) Input image(b) Ground-truth density map (c) Estimated density maps. It can be observed that though the method is able to accurate estimation of crowd count, the estimated density maps are of poor quality.

6. Future research directions

Based on the analysis of various methods and results from Section 3 and 5 and the trend of other developments in computer vision, we believe that CNN-based deeper architectures will dominate further research in the field of crowd counting and density estimation. We make the following observations regarding future trends in research on crowd counting:

Table 3: Comparison of results on various datasets. The CNN-based approaches provide significant improvements over traditional approaches that rely on hand-crafted representations. Further, among the CNN-based methods, scale aware and context aware approaches tend to achieve lower count error.

Approach type	Dataset	UCSD		Mall		UCF CC 50		WorldExpo '10		Shanghai Tech-A		Shanghai Tech-B	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Traditional approaches	Multi-source multi-scale Idrees <i>et al.</i> [41]					468.0	590.3						
	Cumulative Attributes Chen <i>et al.</i> [19]	2.07	6.86	3.43	17.07								
	Density learning Lempitsky <i>et al.</i> [52]	1.7				493.4	487.1						
	Count forest Pham <i>et al.</i> [75]	1.61	4.40	2.5	10.0								
	Exemplar density Wang <i>et al.</i> [99]	1.98	1.82	2.74	2.10								
	Random projection forest Xu <i>et al.</i> [102]	1.90	6.01	3.22	15.5								
CNN-based approaches	Cross-scene Zhang <i>et al.</i> [107]	1.60	3.31			467.0	498.5	12.9		181.8	277.7	32.0	49.8
	Deep + shallow Boominathan <i>et al.</i> [11]					452.5							
	M-CNN Zhang <i>et al.</i> [111]	1.07	1.35			377.6	509.1	11.6		110.2	173.2	26.4	41.3
	CNN-boosting Walach and Wolf [97]	1.10		2.01		364.4							
	Hydra-CNN Onoro <i>et al.</i> [70]					333.7	425.2						
	Joint local & global count Shang <i>et al.</i> [85]					270.3		11.7					
	MoCNN Kumagai <i>et al.</i> [50]			2.75	13.4	361.7	493.3						
	FCN Marsden <i>et al.</i> [65]					338.6	424.5			126.5	173.5	23.76	33.12
	CNN-pixel Kang <i>et al.</i> [45]	1.12	2.06			406.2	404.0	13.4					
	Weighted V-LAD Sheng <i>et al.</i> [89]	2.86	13.0	2.41	9.12								
	Cascaded-MTL Sindagi <i>et al.</i> [92]					322.8	341.4			101.3	152.4	20.0	31.1
	Switching-CNN Sam <i>et al.</i> [82]	1.62	2.10			318.1	439.2	9.4		90.4	135.0	21.6	33.4

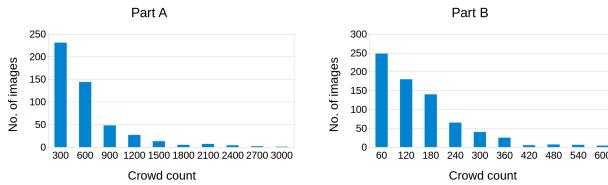


Fig. 17: Distribution of crowd counts in ShanghaiTech dataset. It can be observed that the dataset is highly imbalanced.

- Given the requirement of large datasets for training deep networks, collection of large scale datasets (especially for extremely dense crowds) is essential. Though many datasets exist currently, only one of them (The UCF_CC_50 [41]) caters to large density crowds. However, the size of the dataset is too small for training deeper networks. Though Shanghai Tech [111] attempts to capture large density crowds, the number of images per density level is non-uniform with a large number of images available for low density levels and very few samples for high density levels (as shown in Fig. 17).
- Considering the difficulty of training deep networks for new scenes, it would be important to explore how to leverage from models trained on existing sources. Most of the existing methods retrain their models on a new scene
- and it is impractical to do so in real world scenarios as it would be expensive to obtain annotations for every new scene. Zhang *et al.* [107] attempted to address this issue by performing a data driven training without the need of labelled data for new scenes. In an another approach, Liu *et al.* [59] considered the problem of transfer learning for crowd counting. A model adaptation technique for Gaussian process counting model was introduced. Considering the source model as a prior and the target dataset as a set of observations, the components are combined into a predictive distribution that captures information in both the source and target datasets. However, the idea of transfer learning or domain adaptation [73] for crowd scenes is relatively unexplored and is a nascent area of research.
- Most crowd counting and density estimation methods have been designed for and evaluated either only on single images or videos. Combining the techniques developed separately for these methods is a non-trivial task. Development of low-latency methods that can operate in real-time for counting people in crowds from videos is another interesting problem to be addressed in future.
- Another key issue ignored by earlier research is that the quality of estimated crowd density maps. Many existing CNN-based approaches have a number of max-pooling layers in their networks compelling them to regress on down-sampled density maps. Also, most methods optimize over traditional Euclidean loss which is known to

- have certain disadvantages [43]. Regressing on down-sampled density maps using Euclidean loss results in low quality density maps. Fig. 16 demonstrates the results obtained using the state-of-the-art method [111]. It can be observed that though accurate count estimates are obtained, the quality of the density maps is poor. As a result, these poor quality maps adversely affect other higher level cognition tasks which depend on them. Recent work on style-transfer [108], image de-raining [109] and image-to-image translation [42] have demonstrated promising results from the use of additional loss functions such as adversarial loss and perceptual loss. In principle, density estimation can be considered as an image-to-image translation problem and it would be interesting to see the effect of these recent loss functions. Generating high quality density maps along with low count estimation error would be another important issue to be addressed in the future.
5. Finally, considering advancements by scale-aware [111, 70] and context-aware models [85], we believe designing networks to incorporate additional contextual and scale information will enable further progress.

7. Conclusion

This article presented an overview of recent advances in CNN-based methods for crowd counting and density estimation. In particular, we summarized various methods for crowd counting into traditional approaches (that use hand-crafted features) and CNN-based approaches. The CNN-based approaches are further categorized based on the training process and the network property. Obviously all the literature on crowd counting cannot be covered, hence, we have chosen a representative subset of the latest approaches for a detailed analysis and review. We also reviewed the results demonstrated by various traditional and CNN-based approaches to conclude that CNN-based methods are more adept at handling large density crowds with variations in object scales and scene perspective. Additionally, we observed that incorporating scale and contextual information in the CNN-based methods drastically improves the estimation error. Finally, we identified some of the most compelling challenges and issues that confront research in crowd counting and density estimation using computer vision and machine learning approaches.

Acknowledgement This work was supported by US Office of Naval Research (ONR) Grant YIP N00014-16-1-3134.

References

- [1] Abdelghany, A., Abdelghany, K., Mahmassani, H., Alhalabi, W., 2014. Modeling framework for optimal evacuation of large-scale crowded pedestrian facilities. *European Journal of Operational Research* 237, 1105–1118.
- [2] Al-Kodmany, K., 2013. Crowd management and urban design: New scientific approaches. *Urban Design International* 18, 282–295.
- [3] Almeida, J.E., Rossetti, R.J., Coelho, A.L., 2013. Crowd simulation modeling applied to emergency and evacuation simulations using multi-agent systems. *arXiv preprint arXiv:1303.4692* .
- [4] Arteta, C., Lempitsky, V., Zisserman, A., 2016. Counting in the wild, in: *European Conference on Computer Vision*, Springer. pp. 483–498.
- [5] Aveni, A.F., 1977. The not-so-lonely crowd: Friendship groups in collective behavior. *Sociometry* , 96–99.
- [6] Bandini, S., Gorrini, A., Vizzari, G., 2014. Towards an integrated approach to crowd analysis and crowd synthesis: A case study and first results. *Pattern Recognition Letters* 44, 16–29.
- [7] Barr, J.R., Bowyer, K.W., Flynn, P.J., 2014. The effectiveness of face detection algorithms in unconstrained crowd scenes, in: *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, IEEE. pp. 1020–1027.
- [8] Benabbas, Y., Ihaddadene, N., Djeraba, C., 2010. Motion pattern extraction and event detection for automatic visual surveillance. *EURASIP Journal on Image and Video Processing* 2011, 163682.
- [9] Bernal, E.A., Li, Q., Loce, R.P., 2014. System and method for video-based detection of drive-offs and walk-offs in vehicular and pedestrian queues. *US Patent App.* 14/279,652.
- [10] Blumer, H., 1951. Collective behavior. *New outline of the principles of sociology* , 166–222.
- [11] Boominathan, L., Kruthiventi, S.S., Babu, R.V., 2016. Crowdnet: A deep convolutional network for dense crowd counting, in: *Proceedings of the 2016 ACM on Multimedia Conference*, ACM. pp. 640–644.
- [12] Brostow, G.J., Cipolla, R., 2006. Unsupervised bayesian detection of independent motion in crowds, in: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, IEEE. pp. 594–601.
- [13] Castellano, C., Fortunato, S., Loreto, V., 2009. Statistical physics of social dynamics. *Reviews of modern physics* 81, 591.
- [14] Chaker, R., Al Aghbari, Z., Junejo, I.N., 2017. Social network model for crowd anomaly detection and localization. *Pattern Recognition* 61, 266–281.
- [15] Chan, A.B., Liang, Z.S.J., Vasconcelos, N., 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking, in: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, IEEE. pp. 1–7.
- [16] Chan, A.B., Vasconcelos, N., 2009. Bayesian poisson regression for crowd counting, in: *2009 IEEE 12th International Conference on Computer Vision*, IEEE. pp. 545–551.
- [17] Chan, A.B., Vasconcelos, N., 2012. Counting people with low-level features and bayesian regression. *IEEE Transactions on Image Processing* 21, 2160–2177.
- [18] Chen, J.C., Kumar, A., Ranjan, R., Patel, V.M., Alavi, A., Chellappa, R., 2016. A cascaded convolutional neural network for age estimation of unconstrained faces, in: *International Conference on BTAS*, IEEE. pp. 1–8.
- [19] Chen, K., Gong, S., Xiang, T., Change Loy, C., 2013. Cumulative attribute space for age and crowd density estimation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2467–2474.
- [20] Chen, K., Loy, C.C., Gong, S., Xiang, T., 2012. Feature mining for localised crowd counting, in: *European Conference on Computer Vision*.
- [21] Chen, S., Fern, A., Todorovic, S., 2015. Person count localization in videos from noisy foreground and detections, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1364–1372.
- [22] Cheng, Z., Qin, L., Huang, Q., Yan, S., Tian, Q., 2014. Recognizing human group action by layered model with multiple cues. *Neurocomputing* 136, 124–135.
- [23] Chow, W.K., Ng, C.M., 2008. Waiting time in emergency evacuation of crowded public transport terminals. *Safety Science* 46, 844–857.
- [24] Ciregan, D., Meier, U., Schmidhuber, J., 2012. Multi-column deep neural networks for image classification, in: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE. pp. 3642–3649.
- [25] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, IEEE. pp. 886–893.
- [26] Dollar, P., Wojek, C., Schiele, B., Perona, P., 2012. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence* 34, 743–761.
- [27] Dong, L., Parameswaran, V., Ramesh, V., Zoghlaei, I., 2007. Fast crowd segmentation using shape indexing, in: *2007 IEEE 11th International Conference on Computer Vision*, IEEE. pp. 1–8.
- [28] Enzweiler, M., Gavrila, D.M., 2009. Monocular pedestrian detection: Survey and experiments. *IEEE transactions on pattern analysis and ma-*

- chine intelligence 31, 2179–2195.
- [29] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* 32, 1627–1645.
- [30] Ferryman, J., Ellis, A.L., 2014. Performance evaluation of crowd image analysis using the pets2009 dataset. *Pattern Recognition Letters* 44, 3–15.
- [31] French, G., Fisher, M., Mackiewicz, M., Needle, C., 2015. Convolutional neural networks for counting fish in fisheries surveillance video, in: British Machine Vision Conference Workshop, BMVA Press.
- [32] Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* , 1189–1232.
- [33] Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., Zhu, C., 2015. Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence* 43, 81–88.
- [34] Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V., 2011. Hough forests for object detection, tracking, and action recognition. *IEEE transactions on pattern analysis and machine intelligence* 33, 2188–2202.
- [35] Ge, W., Collins, R.T., 2009. Marked point processes for crowd counting, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 2913–2920.
- [36] Gustafson, S., Arumugam, H., Kanyuk, P., Lorenzen, M., 2016. Mure: fast agent based crowd simulation for vfx and animation, in: ACM SIGGRAPH 2016 Talks, ACM. p. 56.
- [37] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
- [38] Henderson, L.F., 1971. The Statistics of Crowd Fluids 229, 381–383. doi:10.1038/229381a0.
- [39] Hu, Y., Chang, H., Nian, F., Wang, Y., Li, T., 2016. Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation* 38, 530–539.
- [40] Huang, L., Chen, T., Wang, Y., Yuan, H., 2015. Congestion detection of pedestrians using the velocity entropy: A case study of love parade 2010 disaster. *Physica A: Statistical Mechanics and its Applications* 440, 200–209.
- [41] Idrees, H., Saleemi, I., Seibert, C., Shah, M., 2013. Multi-source multi-scale counting in extremely dense crowd images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2547–2554.
- [42] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition.
- [43] Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer. pp. 694–711.
- [44] Junior, J.C.S.J., Musse, S.R., Jung, C.R., 2010. Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine* 27, 66–77.
- [45] Kang, D., Ma, Z., Chan, A.B., 2017. Beyond counting: Comparisons of density maps for crowd analysis tasks-counting, detection, and tracking. arXiv preprint arXiv:1705.10118 .
- [46] Kang, K., Wang, X., 2014. Fully convolutional neural networks for crowd segmentation. arXiv preprint arXiv:1411.4464 .
- [47] Klontz, J.C., Jain, A.K., 2013. A case study on unconstrained facial recognition using the boston marathon bombings suspects. Michigan State University, Tech. Rep 119, 1.
- [48] Ko, T., 2008. A survey on behavior analysis in video surveillance for homeland security applications, in: Applied Imagery Pattern Recognition Workshop, 2008. AIPR'08. 37th IEEE, IEEE. pp. 1–8.
- [49] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.
- [50] Kumagai, S., Hotta, K., Kurita, T., 2017. Mixture of counting cnns: Adaptive integration of cnns specialized to specific appearance for crowd counting. arXiv preprint arXiv:1703.09393 .
- [51] Leibe, B., Seemann, E., Schiele, B., 2005. Pedestrian detection in crowded scenes, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE. pp. 878–885.
- [52] Lempitsky, V., Zisserman, A., 2010. Learning to count objects in images, in: Advances in Neural Information Processing Systems, pp. 1324–1332.
- [53] Li, G., Yu, Y., 2015. Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5455–5463.
- [54] Li, M., Zhang, Z., Huang, K., Tan, T., 2008. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-and-shoulder detection, in: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, IEEE. pp. 1–4.
- [55] Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S., 2015. Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 367–386.
- [56] Li, W., Mahadevan, V., Vasconcelos, N., 2014. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence* 36, 18–32.
- [57] Lin, S.F., Chen, J.Y., Chao, H.X., 2001. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 31, 645–654.
- [58] Lipton, A.J., Venetianer, P.L., Haering, N., Brewer, P.C., Yin, W., Zhang, Z., Yu, L., Hu, Y., Myers, G.W., Chosak, A.J., et al., 2015. Video analytics for retail business process monitoring. US Patent 9,158,975.
- [59] Liu, B., Vasconcelos, N., 2015. Bayesian model adaptation for crowd counts, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4175–4183.
- [60] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.
- [61] Loy, C.C., Chen, K., Gong, S., Xiang, T., 2013. Crowd counting and profiling: Methodology and evaluation, in: Modeling, Simulation and Visual Analysis of Crowds. Springer, pp. 347–382.
- [62] Lu, L., Chan, C.Y., Wang, J., Wang, W., 2016. A study of pedestrian group behaviors in crowd evacuation based on an extended floor field cellular automaton model. *Transportation Research Part C: Emerging Technologies* .
- [63] Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N., 2010. Anomaly detection in crowded scenes., in: CVPR, p. 250.
- [64] Marana, A., Costa, L.d.F., Lotufo, R., Velastin, S., 1998. On the efficacy of texture analysis for crowd monitoring, in: Computer Graphics, Image Processing, and Vision, 1998. Proceedings. SIBGRAPI'98. International Symposium on, IEEE. pp. 354–361.
- [65] Marsden, M., McGuinness, K., Little, S., O'Connor, N.E., 2016. Fully convolutional crowd counting on highly congested scenes. arXiv preprint arXiv:1612.00220 .
- [66] Marsden, M., McGuinness, K., Little, S., O'Connor, N.E., 2017. Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. arXiv preprint arXiv:1705.10698 .
- [67] Mongeon, M.C., Loce, R.P., Shreve, M.A., 2015. Busyness detection and notification method and system. US Patent App. 14/625,960.
- [68] Moussaïd, M., Perozo, N., Garnier, S., Helbing, D., Theraulaz, G., 2010. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS one* 5, e10047.
- [69] Mundhenk, T.N., Konjevod, G., Sakla, W.A., Boakye, K., 2016. A large contextual dataset for classification, detection and counting of cars with deep learning, in: European Conference on Computer Vision, Springer. pp. 785–800.
- [70] Onoro-Rubio, D., López-Sastre, R.J., 2016. Towards perspective-free object counting with deep learning, in: European Conference on Computer Vision, Springer. pp. 615–629.
- [71] Paragios, N., Ramesh, V., 2001. A mrf-based approach for real-time subway monitoring, in: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, IEEE. pp. I–1034.
- [72] Parrish, J.K., Edelstein-Keshet, L., 1999. Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science* 284, 99–101.
- [73] Patel, V.M., Gopalan, R., Li, R., Chellappa, R., 2015. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine* 32, 53–69.
- [74] Perez, H., Hernandez, B., Rudomin, I., Ayguade, E., 2016. Task-based crowd simulation for heterogeneous architectures, in: Innovative Research and Applications in Next-Generation High Performance Com-

- puting. IGI Global, pp. 194–219.
- [75] Pham, V.Q., Kozakaya, T., Yamaguchi, O., Okada, R., 2015. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3253–3261.
- [76] Ranjan, R., Patel, V., Chellappa, R., 2016. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on PAMI* .
- [77] Rodriguez, M., Laptev, I., Sivic, J., Audibert, J.Y., 2011. Density-aware person detection and tracking in crowds, in: 2011 International Conference on Computer Vision, IEEE, pp. 2423–2430.
- [78] Ryan, D., Denman, S., Fookes, C., Sridharan, S., 2009. Crowd counting using multiple local features, in: Digital Image Computing: Techniques and Applications, 2009. DICTA'09., IEEE, pp. 81–88.
- [79] Ryan, D., Denman, S., Sridharan, S., Fookes, C., 2015. An evaluation of crowd counting methods, features and regression models. *Computer Vision and Image Understanding* 130, 1–17.
- [80] Sabzeydani, P., Mori, G., 2007. Detecting pedestrians by learning shapelet features, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, pp. 1–8.
- [81] Saleh, S.A.M., Suandi, S.A., Ibrahim, H., 2015. Recent survey on crowd density estimation and counting for visual surveillance. *Engineering Applications of Artificial Intelligence* 41, 103–114.
- [82] Sam, D.B., Surya, S., Babu, R.V., 2017. Switching convolutional neural network for crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- [83] Saxena, S., Brémond, F., Thonnat, M., Ma, R., 2008. Crowd behavior recognition for video surveillance, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, pp. 970–981.
- [84] Serbanet, P., Chintala, S., LeCun, Y., 2012. Convolutional neural networks applied to house numbers digit classification, in: Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE, pp. 3288–3291.
- [85] Shang, C., Ai, H., Bai, B., 2016. End-to-end crowd counting via joint learning local and global count, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, pp. 1215–1219.
- [86] Shao, J., Change Loy, C., Wang, X., 2014. Scene-independent group profiling in crowd, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2219–2226.
- [87] Shao, J., Kang, K., Loy, C.C., Wang, X., 2015. Deeply learned attributes for crowded scene understanding, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 4657–4666.
- [88] Shao, J., Loy, C.C., Kang, K., Wang, X., 2016. Slicing convolutional neural network for crowd video understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5620–5628.
- [89] Sheng, B., Shen, C., Lin, G., Li, J., Yang, W., Sun, C., 2016. Crowd counting via weighted vlad on dense attribute feature maps. *IEEE Transactions on Circuits and Systems for Video Technology* .
- [90] Sime, J.D., 1995. Crowd psychology and engineering. *Safety science* 21, 1–14.
- [91] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .
- [92] Sindagi, V., Patel, V., 2017. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, in: Advanced Video and Signal Based Surveillance (AVSS), 2017 IEEE International Conference on, IEEE.
- [93] Szegedy, C., Liu, W., Jia, Y., Serbanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.
- [94] Tuzel, O., Porikli, F., Meer, P., 2008. Pedestrian detection via classification on riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence* 30, 1713–1727.
- [95] Viola, P., Jones, M.J., 2004. Robust real-time face detection. *International journal of computer vision* 57, 137–154.
- [96] Viola, P., Jones, M.J., Snow, D., 2005. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision* 63, 153–161.
- [97] Walach, E., Wolf, L., 2016. Learning to count with cnn boosting, in: European Conference on Computer Vision, Springer, pp. 660–676.
- [98] Wang, C., Zhang, H., Yang, L., Liu, S., Cao, X., 2015. Deep people counting in extremely dense crowds, in: Proceedings of the 23rd ACM international conference on Multimedia, ACM, pp. 1299–1302.
- [99] Wang, Y., Zou, Y., 2016. Fast visual object counting via example-based density estimation, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, pp. 3653–3657.
- [100] Wu, B., Nevatia, R., 2005. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, IEEE, pp. 90–97.
- [101] Wu, B., Nevatia, R., 2007. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision* 75, 247–266.
- [102] Xu, B., Qiu, G., 2016. Crowd density estimation based on rich features and random projection forest, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp. 1–8.
- [103] Yi, S., Wang, X., Lu, C., Jia, J., Li, H., 2016. L0 regularized stationary-time estimation for crowd analysis. *IEEE transactions on pattern analysis and machine intelligence* .
- [104] Yu, J., Zhang, B., Kuang, Z., Lin, D., Fan, J., 2017. iprivity: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security* 12, 1005–1016.
- [105] Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L.Q., 2008. Crowd analysis: a survey. *Machine Vision and Applications* 19, 345–357.
- [106] Zhang, C., Kang, K., Li, H., Wang, X., Xie, R., Yang, X., 2016a. Data-driven crowd understanding: A baseline for a large-scale crowd dataset. *IEEE Transactions on Multimedia* 18, 1048–1061.
- [107] Zhang, C., Li, H., Wang, X., Yang, X., 2015. Cross-scene crowd counting via deep convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833–841.
- [108] Zhang, H., Dana, K., 2017. Multi-style generative network for real-time transfer. arXiv preprint .
- [109] Zhang, H., Sindagi, V., Patel, V.M., 2017. Image de-raining using a conditional generative adversarial network. arXiv preprint arXiv:1701.05957 .
- [110] Zhang, H.P., Beer, A., Florin, E.L., Swinney, H.L., 2010. Collective motion and density fluctuations in bacterial colonies. *Proceedings of the National Academy of Sciences* 107, 13626–13630.
- [111] Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y., 2016b. Single-image crowd counting via multi-column convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 589–597.
- [112] Zhao, T., Nevatia, R., Wu, B., 2008. Segmentation and tracking of multiple humans in crowded environments. *IEEE transactions on pattern analysis and machine intelligence* 30, 1198–1211.
- [113] Zhao, Z., Li, H., Zhao, R., Wang, X., 2016. Crossing-line crowd counting with two-phase deep neural networks, in: European Conference on Computer Vision, Springer, pp. 712–726.
- [114] Zhou, B., Tang, X., Wang, X., 2015. Learning collective crowd behaviors with dynamic pedestrian-agents. *International Journal of Computer Vision* 111, 50–68.
- [115] Zhou, B., Wang, X., Tang, X., 2012. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, pp. 2871–2878.
- [116] Zhu, F., Wang, X., Yu, N., 2014. Crowd tracking with dynamic evolution of group structures, in: European Conference on Computer Vision, Springer, pp. 139–154.
- [117] Zitouni, M.S., Bhaskar, H., Dias, J., Al-Mualla, M.E., 2016. Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques. *Neurocomputing* 186, 139–159.