Deep Learning for Crowd Counting: A Survey

Tjeng Wawan Cenggoro

Computer Science Department, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
wcenggoro@binus.edu

Abstract *- The growth of deep learning for crowd counting is immense in the recent years. This results in numerous deep learning model developed with huge multifariousness. This paper aims to capture a big picture of existing deep learning models for crowd counting. Hence, the development of novel models for future works can be accelerated.*

Keywords: *Deep learning, computer vision, crowd counting*

I. INTRODUCTION

Crowd counting is one of the computer vision problems that rapidly flourishes since the advent of deep learning. Since 2015, state-of-the-art techniques across all crowd counting dataset are dominated by the use of Convolutional Neural Networks (CNN) (LeCun et al., 1989), one of the deep learning models that excels in image processing.

Unfortunately, such a rapid growth resulting in a great number of variations arose in deep learning model for crowd counting. Therefore, an extensive survey for crowd counting with deep learning is essential to develop new ideas for future works in this field. This paper aims to provide a comprehensive survey for deep learning techniques that has been developed for crowd counting. To simplify the understanding of recent progress in deep learning for crowd counting, we group all techniques into several major categories. This will ease the researchers in this field to develop novel deep learning models for crowd counting.

II. PROBLEM FORMULATION

The problem of crowd counting in deep learning is typically tackled by density map regression. This approach was first proposed by in 2010 by Lempitsky and Zisserman (Lempitsky & Zisserman, 2010). The density map is generated from point annotations centered on the head of people in the image. Each point is projected to the density map by using Gaussian distribution centered on the point. By using this approach, the total count can be calculated by summing the value of all pixels in the density map. Figure 1 shows a sample of a crowd image and its density map.

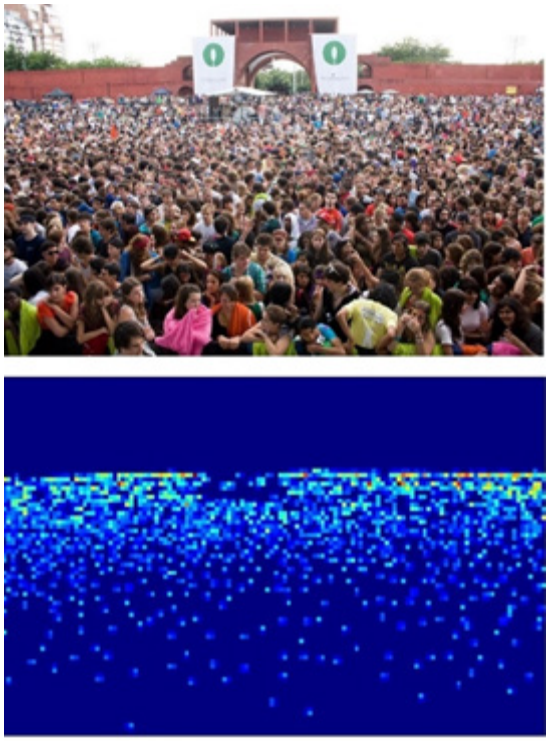

Figure 1 A sample of crowd image and its density map

To evaluate the quality of the prediction, Mean Absolute Error (MAE) and Mean Squared Error (MSE) are usually used. These metrics are formulated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|z_i - \hat{z}_i| \quad (1)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{z}_i)^2} \quad (2)$$

where, z_i is the value of a pixel in density map of i^{th} pixel in an image, $\hat{z}_i$ is the predicted value for the i^{th} pixel, and N is the total number of pixels in an image.

III. DATASETS

For benchmarking in crowd counting, there are four datasets that are frequently used: ShanghaiTech Part A and B, UCF_CC_50, and WorldExpo'10. There are also three other notable datasets but used less frequently in crowd counting research: UCSD, Mall, and UCF-QNRF. The statistics of these dataset are given in table 1.

Table 1 Crowd Counting Dataset Statistics

Dataset	#Images	Avg. #People
ShanghaiTech Part A	482	501.4
ShanghaiTech Part B	716	123.6
UCF_CC_50	50	1279.5
WorldExpo'10	3,980	50.2
UCF_QNRF	1,535	815.4
UCSD	2,000	±20.0
Mall	2,000	±30.0

In recent crowd counting research, ShanghaiTech Part A and B (Y. Zhang, Zhou, Chen, Gao, & Ma, 2016) are the most popular to be used for benchmarking. The images in Part A were collected by randomly crawling from internet. The Part B images were taken from a crowded area in Shanghai.

The second most popular dataset for crowd counting is UCF_CC_50 (Idrees, Saleemi, Seibert, & Shah, 2013). Despite comprising only 50 images, it has the most crowd density among other dataset with 1,279.5 average number of people per image. Thus, it poses a different challenge than the other crowd counting dataset. The images in this dataset were collected from internet.

In terms of number of images, WorldExpo'10 (C. Zhang, Li, Wang, & Yang, 2015) is currently the largest dataset for crowd counting with 3,980 images. Despite of the large number of images, the crowd density of this dataset is relatively low among other popular datasets. The images in this dataset were captured from 103 different scenes in Shanghai 2010 WorldExpo.

To compete with WorldExpo'10, UCF_QNRF dataset (Idrees et al., 2018) was developed as another massive dataset for crowd counting. Although it consists only 1,535 images, it has significantly more crowd density than WorldExpo'10. The resolution of each images is also about 14 times larger than WorldExpo'10. The images in this dataset were collected from internet.

Among the popular dataset, UCSD (Chan, Zhang-Sheng John Liang, & Vasconcelos, 2008) is the earliest dataset developed. It consists of 2,000 images with about 20 people captured in average. The images were captured from a single scene in University of California San Diego (UCSD).

Mall dataset (K. Chen, Loy, Gong, & Xiang, 2012) is developed after UCSD to capture more diversities in crowd density and environmental condition. It consists of 2,000 images with about 30 people captured in the images averagely. The images were captured from surveillance camera in a shopping mall.

Other than the datasets explained before, GCC dataset (Q. Wang, Gao, Lin, & Yuan, 2019) recently emerges as one of important dataset for crowd counting. GCC has a massive size of 15,212 images with average people count of 501.3. With such a huge size, GCC is applicable to develop a robust deep learning model. However, GCC is not captured from real camera, but synthetically generated from GTA V scenes. Therefore, this dataset is more suitable for deep learning model pretraining rather than for benchmarking.

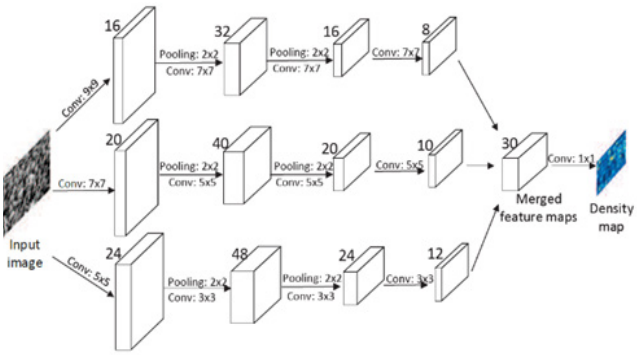

Figure 2 Multi-column CNN architecture

IV. CATEGORIZATION OF DEEP LEARNING MODELS FOR CROWD COUNTING

In this paper, we categorize various approaches in deep learning for crowd counting into six major categories:

1. Scale-aware CNN
2. Multi-tasking CNN
3. CNN with local context
4. CNN with ensemble learning
5. Generative Adversarial Networks (GAN) for crowd counting
6. Unsupervised/Semi supervised CNN

Section 1 to 6 respectively elaborates each category in detail. In section 7, we list several deep learning models that are not fit to the six categories.

1. Scale-Aware CNN

Encoding scale prior in a CNN is the most prominent approach to create a novel deep learning model nowadays. This trend is pioneered by the development of Multi Column CNN (MCNN) in 2016 (Y. Zhang et al., 2016). MCNN use three-columns CNN with different size of convolution kernels for different scaling. The illustration of MCNN architecture is given in figure 2.

The design of multiple CNN columns enables the model to be perspective-free, meaning that it does not rely on additional perspective map. Zhang et al. proved that MCNN is able to outperform the previous state-of-the-art approach that utilize perspective map. This fact is a huge leap in crowd counting research that the scale-aware model become a common theme even now.

In the same year, Oñoro-Rubio and López-Sastre emphasize more on the benefit of multi-column CNN as perspective-free model by introducing HydraCNN (Oñoro-Rubio & López-Sastre, 2016). This model is developed with three-columns CNN, which each column is fed with a patch with different scale in a same image. In 2016 as well,

Boominathan et al. developed CrowdNet (Boominathan, Kruthiventi, & Babu, 2016), which comprises deep and shallow column. The author argues that the depth difference allows each column to capture different scale variation.

The most powerful multi-column scale-aware model to date is Adaptive Scenario Discovery (ASD) (Wu et al., 2018) that was developed in 2018. It is currently the second-best model for UCF_CC_50 dataset. ASD uses two CNN columns that separately generates density map for sparse and congested case.The sparse column uses 4 convolution layers with 3x3 kernel size. On the other hand, the congested pathway uses 5 convolutional layers with 1x1, 9x9, 7x7, 7x7, and 3x3 kernel size in sequence, ended with a 2x2 max pooling layer. To fuse the result of the sparse and congested column, Wu et al. use an adaption module that is modeled after Squeeze and Excitation module from SENet (Hu, Shen, Albanie, Sun, & Wu, 2019).

In 2018 as well, there are three other crowd counting research that used multi-column scale-aware approach: Adaptively Fusing Predictions (AFP) and iterative counting CNN (ic-CNN). Each CNN column in AFP takes a same image with different scaling as input. Afterward, each column predicts its own attention map and intermediary density map. The final density map is generated by weighting each intermediary density map with the corresponding attention map and fusing them with a 1x1 convolutional layer. Meanwhile, ic-CNN use two CNN columns for high resolution (HR-CNN) and low resolution (LR-CNN). The predicted density map and convolutional features of LR-CNN is passed to HR-CNN. Given the image, LR-CNN density map, and LR-CNN convolutional feature, HR-CNN predict a density map with the same size as the ground truth density map.

The most recent model that belongs to multi-column scale-aware CNN category is Scale-Aware Attention Networks (SAAN) by Hossain et al. (Hossain, Hosseinzadeh, Chanda, & Wang, 2019). SAAN employs a three-columns architecture with different kernel size similar to MCNN. The generated density map of each column is weighted with the output of two other networks, Global Scale Attention (GSA) and Local Scale Attention (LSA). GSA provides three scalar weights for each column while LSA provides pixel-wise weighting for each column.

Although the scale-aware approach was pioneered by multi-column architecture, the recent trend is shifted to encode scale prior in single-column CNN. This approach allows faster processing time, as the total parameters are usually much smaller than multi-column approach. The most common idea of single-column scale-aware CNN is to design a module that is able to encode a variation of scale in an image, or in the other word, a scale-aware module. Therefore, a single-column CNN built with the module can be a scale-aware model. Figure 3 illustrates the scale pyramid module in SPN, one of the single-column scale-aware CNN that utilize the scale-aware module idea.

Surprisingly, single-column CNN approach happens to be not only faster, but also generally better in performance. In fact, there are two single-column models that achieve a state-of-the-art performance in at least one of the popular crowd counting datasets. These models are Scale Spatial Fully Connected Networks (SFCN) and Scale Pyramid Networks (SPN).

SFCN is currently one of the state-of-the-art models for ShanghaiTech Dataset Part B. It stacks a dilated convolutional layer (Yu & Koltun, 2015) and a Down Up Left Right (DULR) layer on top of the first three blocks of ResNet101 (He, Zhang, Ren, & Sun, 2016). The DULR layer is a stack of (1 x w) and (h x 1) convolutional layer. It was first introduced in PCC Net (Gao, Wang, & Li, 2019). By using DULR layer, SFCN is able to encode the difference in scale changes from the upper part of image to the lower part of image as well as from the left part of image to right part of image. SFCN also achieve its impressive performance by the help of pretraining from GCC dataset.

Meanwhile, SPN has a state-of-the-art performance on ShanghaiTech Dataset Part A. The author of SPN proposed a scale pyramid module that uses 4 parallel dilated convolution kernels with different dilation rate. The dilation rate of each kernel is 2, 4, 8, and 12 respectively. This scale pyramid module is the main building block of SPN. With the four size of dilation rate in scale pyramid module, the SPN is able to encode four variation of scale.

The first work that used single-column scale-aware architecture was done by Zeng et al. in 2017 (Zeng, Xu, Cai, Qiu, & Zhang, 2017). They proposed Multi-Scale CNN (MSCNN), which layers are built with a module called as multi-scale blob. This module is an inception-like module (Szegedy et al., 2015) but with 3x3, 5x5, 7x7, and 9x9 kernels for each path in the module. In a similar way to SPN, the four different kernel size is able to encode different variation of scale.

In the following year after MSCNN introduction, there are three works that used single-column scale-aware CNN for crowd counting: Single Column Networks (SCNet), Scale Aggregation Networks (SANet), and Scale-adaptive CNN (SaCNN). In SCNet (Z. Wang et al., 2018), a residual fusion module is introduced. This module comprises stacked of

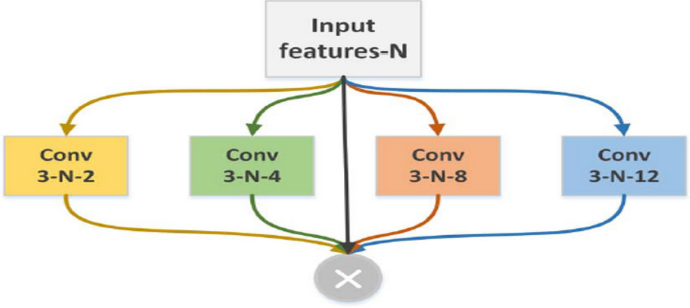

Figure 3 Illustration of scale pyramid module in SPN

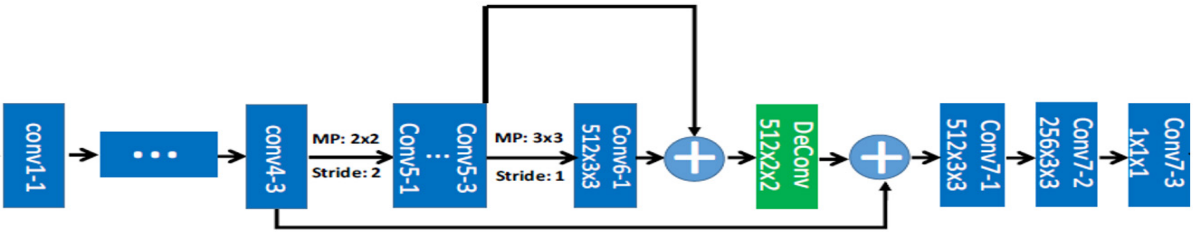

Figure 4 Illustration of SaCNN architecture

Nested Dilated Convolutional Layer (NDL), which are connected with a residual connection. The idea of NDL is similar to scale pyramid module of SPN. The difference is that NDL uses 3 different kernel size instead of 4.

Meanwhile, the author of SANet (Cao, Wang, Zhao, & Su, 2018) proposed a similar module to multi scale blob called as scale aggregation module. The module instead uses kernels with the size of 1x1, 3x3, 5x5, and 7x7 respectively.

Different from other single-column scale-aware CNN, SaCNN (L. Zhang, Shi, & Chen, 2018) tackles multi-scale encoding not by using scale-aware module. Instead, it leverages residual connection (He et al., 2016) from earlier layer to later layer. Because the deeper the layer encodes the larger area, combining output from different depths is tantamount to fusing information from different scale in an image. Specifically, SaCNN uses residual connection from layer 5_3 to layer 6_1 of VGG16 (Simonyan & Zisserman, 2015). To further enhance the scale adaptive effect, the Zhang et al. added a strided convolution layer after layer 6_1 and put another residual connection from layer 4_3 to the output. The illustration of SaCNN architecture is given in figure 4.
In the same spirit with SaCNN, Liu et al. (Ming Liu, Jue Jiang, Zhenqei Guo, Zenan Wang, & Yang Liu, 2018) designed a single-column scale-aware CNN by using Feature Pyramid Networks (FPN) (Lin et al., 2017) as the backbone of Fully Convolutional Networks (FCN) (Long, Shelhamer, & Darrell, 2014) that generates a single density map. FPN has several residual connections that is able to bridge information between feature maps with different scale like SaCNN.

To combine the best of both multi-column and single-column scale-aware CNN, there are several research that developed a model containing multiple paths only in several part of the networks. For conciseness, this approach will be defined as semi multi-column scale-aware CNN in this paper. The top example of model in this approach are Feature Pyramid Networks for Crowd Counting (FPNCC), Context-Aware Networks (CAN), and Congested Scene Recognition Networks (CSRNet), which achieve the state-of-the-art performance in at least one of the popular crowd counting datasets.

The idea of semi multi-column CNN is implemented in FPNCC (Cenggoro, Aslamiah, & Yunanto, 2019) by utilizing Feature Pyramid Networks (FPN) that produces multiple outputs in different scale. In this case, three outputs from FPN are used. Each of the FPN output is subjected to a sequential pair of 1x1 convolutional layers to produce several intermediary density maps. This par is the multi-column part of the FPNCC. Finally, all intermediary density maps are then aggregated to generate a single density map by using an aggregator module. This module is a series of 5x5 and 1x1 convolutional layers. Figure 5 depicts the architecture of FPNCC.

In CAN (W. Liu, Salzmann, & Fua, 2019), the semi multi-column architecture is realized by using different block size of Spatial Pyramid Pooling (SPP) (He, Zhang, Ren, & Sun, 2014) on the 10th layer feature maps of VGG16. Each SPP features (scale features) are exposed to a 1x1 conv, then the output is subtracted with the 10th layer feature maps of VGG16 (contrast features). Afterward, the scale features are multiplied with the contrast features to produce weighted features, then the weighted features are concatenated with VGGNet features to serve as the final feature map. The final density map is generated from the final feature map by using a sequence of dilated convolution layers.

Similar to CAN, the multi-column part of CSRNet (Li, Zhang, & Chen, 2018) is attached on top of the 10th layer feature maps of VGG16. However, instead of using SPP, CSRNet use four dilated convolution columns with different dilation rate. Because of the different dilation rate, the output of each column has different size. To aggregate all columns output for final density map, they are resized to uniform size by using bilinear interpolation.

Other than FPNCC, CAN, and CSRNet, there are three other deep learning model that can be included in semi multi-column CNN category: Trellis Encoder Decoder Networks (TEDNet), Aggregated Multicolumn Dilated Convolution Network (AMDCN), and Deformation Aggregation Network (DA-Net).

TEDNet (Jiang et al., 2019) uses a decoding block that takes two inputs from consecutive scaling phase. Each of the outputted density map for different scaling is included in the loss function, but not directly included in the final density map. The information of each intermediary density map is injected to the final density map via the input combination in the decoding block.

Like CSRNet, AMDCN (Deb & Ventura, 2018) use multiple columns of dilated convolution layers with different dilation rate for encoding scale variation. However, the multi-column part is attached before the first single-column part instead of the last layer.

In DA-Net (Zou, Su, Qu, & Zhou, 2018), the multi-column part receive input from the fourth to eighth layer density map of VGG16. The multi-column part that takes fourth to seventh layer feature map use deformable convolution layer (Dai et al., 2017) to generate an intermediary density map. The other column uses a vanilla convolution layer to produce the intermediary density map. Subsequently, all intermediary density maps are summed to produce the final density map.

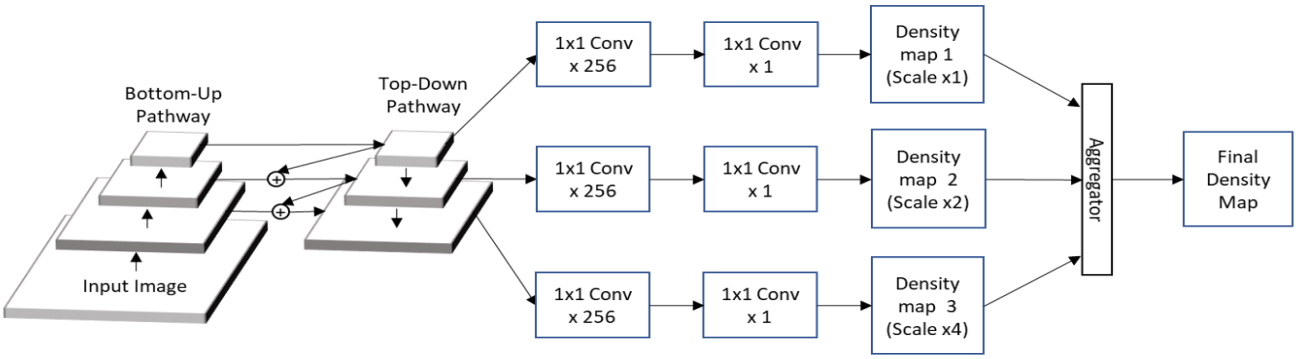

Figure 5 Illustration of FPNCC architecture

2. Multi-Tasking CNN for Crowd Counting

Multi-tasking is one of the prevalent techniques in deep learning that is capable to improve the model performance. This approach let the model to do several related tasks in addition to the main task, that is density map regression in crowd counting case. These additional tasks can serve as regularizers that forces the model to learn a more useful feature representation. The simplest implementation of this approach is to let the model jointly be trained on several tasks without any direct collaboration between tasks. This paradigm is adopted in by the first work that uses multi-tasking CNN for crowd counting by Zhang et al. (C. Zhang et al., 2015). They developed a CNN that jointly regresses the density map as well as the global count. In addition to single-column scale-aware architecture, SaCNN and its improvement (Sang et al., 2019) also employs multi-task learning similar to the work of Zhang et al.

There are two other deep learning model that also uses this simplest paradigm of multi-task learning: Cascaded Multi-Task Learning (CMTL) (Sindagi & Patel, 2017a) and ResNetCrowd (Marsden, McGuinness, Little, & O'Connor, 2017). CMTL is jointly trained for crowd count classification in addition to density map regression. Meanwhile, ResNetCrowd is subjected to multi-task learning of four tasks: behavior recognition, density level classification, count regression, and density map regression.

Later, researchers found that collaboration between tasks can improve the performance of deep learning model for crowd counting, In fact, the most prominent crowd counting models with multi-tasking approach use this paradigm: Perspective-Aware CNN (PACNN) and Attention-injective Deformable Crowd Networks (ADCrowdNet).

PACNN (M. Shi, Yang, Xu, & Chen, 2019) is currently one of the state-of-the-art model for ShanghaiTech Part B and UCSD dataset. In addition to density map regression as its main task, PACNN also regresses to the corresponding perspective map. PACNN produces three density maps with 1:1, 1:2, and 1:4 ratio to the ground truth density map. It also produces two perspective maps with 1:1 and 1:2 ratio. The 1:2 and 1:1 density map are combined adaptively by generating weighting factors from the estimated density map.

Similarly, Attention-injective Deformable Crowd Networks (ADCrowdNet), employs collaborative multi-tasking approach to achieve a state-of-the-art performance on WorldExpo'10 dataset. ADCrowdNet comprises two separate networks with different task: Attention Map Generator (AMG) and Density Map Estimator (DME). Firstly, AMG is trained for classification task of crowd versus not-crowd image. Afterward, DME perform the main task of density map regression with the additional input of attention map generated by AMG.

Other than PACNN and ADCrowdNet, there are three other works that use the collaborative multi-tasking approach: Body Structure Aware Deep Crowd Counting (BSAD) (Huang et al., 2018), Composition Loss (CL) (Idrees et al., 2018), and DecideNet (J. Liu, Gao, Meng, & Hauptmann, 2018).

BSAD use regression of body to body part map and structured density map as its tasks for multi-task learning. The body part map is generated by using CNN for semantic segmentation. The structured density map is generated by using Gaussian distribution modelling on the body part map. The final density map is extracted from the head density map of the whole structured density map.

In CL, four regressors are employed in the network. Three regressors are density map estimators with different parameter to adjust the width of the gaussian kernel in the produced density map. The other regressor directly predicts the absolute count. The loss of each regressor is calculated and averaged for computing the final loss value. For absolute count regressor loss, the predicted count is computed by averaging the summed values of all generated density map and the predicted absolute count itself.

DecideNet employs three CNNs to accomplish three different tasks. The CNNs are named as RegNet, DetNet, and QualityNet. RegNet regresses density map as typical CNN for crowd counting. DetNet detects heads in the style of CNN for object detection. The detected heads are converted into density map by putting Gaussian kernel in the predicted bounding boxes. To produce the final density map, QualityNet generates weighting attention between the density map of RegNet and DenseNet. The authors argue that RegNet is excellent for counting crowded scene while DetNet is better for counting sparse scene. Thus, by combining RegNet and DetNet prediction, DecideNet can achieve a good performance on both crowded and sparse scene.

3. Incorporating Local Context in Crowd Counting CNN

As what is discussed in section V, crowd image typically has different scale across the image. However, the variation in scale changes gradually from upper side of the image to the lower side or from left to right. Therefore, as we take a smaller patch, the variation of scale is also become smaller that to some extent can be neglected. Several research in crowd counting take advantage from the local context to design a deep learning model that is robust to scale variation without any scale-aware architecture. The first CNN model that uses this approach are Contextual Pyramid CNN (CP-CNN) (Sindagi & Patel, 2017b) and Switching-CNN (Sam, Surya, & Babu, 2017).

[CP-CNN]: Use three CNN: Global Context Estimator (GLE), Local Context Estimator (LSE), and Density Map Estimator (DME). The GLE and LSE are trained to classify density level of full image and its patches respectively. The DME is trained as typical CNN for crowd counting. To generate the final density map, the last feature maps of GLE, LSE, and DME are concatenated and fed to a Fusion CNN (F-CNN). The F-CNN is also trained to generate density map like DME.

In Switching-CNN, the input image is split into 9 non-overlapping patches. Afterward, these 9 patches are fed into three-column CNN. Another CNN column is employed to decide which column delivers the best performance given a certain patch.

A year after, the authors of Switching-CNN proposed an improved version of Switching-CNN called as Incrementally Growing CNN (IG-CNN) (Sam, Sajjan, Babu, & Srinivasan, 2018). In IG-CNN, each CNN column is pretrained by using an incrementally growing strategy. Firstly, a base CNN is trained with the whole dataset. Afterward, the base CNN is copied into two new CNN. The training dataset is also split into two subset and fed to the respective new CNN. These steps are repeated until the predetermined maximum tree depth is reached.

The local context approach is one of the promising techniques for future crowd counting model. One of the model in this approach, Deep Recurrent Spatial-Aware Network (DRSAN) (L. Liu, Wang, Li, Ouyang, & Lin, 2018), is able to achieve a state-of-the-art result on ShanghaiTech Part A dataset. The core of DRSAN is a module called as Recurrent Spatial-Aware Refinement (RSAR), which comprises Attentional Region Locator (ARL) module and Local Refinement Networks (LRN) module. First, global features are generated by using a three-column CNN. An initial density map is generated from the global features with a convolutional layer with 1x1 kernels. Subsequently, the initial density map is refined by RSAR, first by locating a local region of interest using ARL. This ARL is constructed by the combination of Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Spatial Transformer Networks (STN) (Jaderberg, Simonyan, Zisserman, & Kavukcuoglu, 2015). The use of LSTM enables RSAR to generates multiple local regions iteratively. For each iteration, the corresponding density sub-map is refined by using LRN, which is a three-column CNN with residual connections.

Recently in 2019, there are two works that also use local context in the model: Recurrent Attentive Zooming Networks (RAZ-Net) (C. Liu, Weng, & Mu, 2019) and Scale-Aware Attention Networks (SAAN) (Hossain et al., 2019).

RAZ-Net is trained in two phases. In the first phase, a main CNN is trained as a typical density map regressor with an extra column to propose a region to zoom. Subsequently, RAZ-Net is trained to recurrently refine the proposed zooming regions.
Meanwhile, SAAN uses three type of CNN: Multi-scale Feature Extractor (MFE), Global Scale Attention (GSA), and Local Scale Attention (LSA). The MFE is a three-column CNN with different kernel size to extract features from different scale. The outputs of MFE is weighted by using global weights from GSA and pixel-wise attention from LSA. The weighted feature maps is fused by a Fusion Network to generate the final density map.

4. Leveraging Ensemble Learning For Crowd Counting With CNN

In machine learning field, ensemble learning technique is currently one of the most prominent technique to improve the performance of a model. In fact, one of the most powerful modern machine learning model, XGBoost (T. Chen & Guestrin, 2016), is essentially an ensemble of decision tree with gradient boosting machine (GBM) paradigm (Friedman, 2001, 2002). Motivated by the impressive performance of GBM. Wallach and Wolf (Walach & Wolf, 2016) designed CNN-Boosting, an ensemble of CNN that use GBM for crowd counting. Although it is not a state-of-the-art model for crowd counting, CNN-Boosting has a competitive performance on UCSD, Mall, and UCF_CC_50.

The use of ensemble learning is not actually popular in crowd counting research. To the best of our knowledge, there are only two works that use ensemble learning in CNN for crowd counting: CNN-Boosting and Decorrelated Convolutional Networks (D-ConvNet) (Z. Shi, Zhang, Liu, et al., 2018). However, the ensemble learning approach is still promising as D-ConvNet is currently one of the state-of-the-art models for WorldExpo'10 dataset.

D-ConvNet infuses ensemble learning concept in CNN by adopting Negative Correlation Learning (NCL) (Y. Liu & Yao, 1999). NCL is originally a method for robust ensemble learning that encourages diversity among models in the ensemble. D-ConvNet use NCL by employing multiple density map generator on top of the VGG16 last features. All generators are treated as an ensemble of models and trained with NCL.

5. Crowd Counting With Generative Adversarial Networks

With the problem formulation of density map regression, crowd counting can be thought as generating density map image given a crowd image. Therefore, this problem can naturally be solved by image generator algorithm such as Generative Adversarial Networks (GAN) (Goodfellow et al., 2014). Despite the straightforward implementation, only three works tried to use GAN for crowd counting to date:

Adversarial Cross-Scale Consistenty Pursuit (ACSCP), Multi-Scale GAN (MS-GAN), and GAN Multiple Target Regression (GAN-MTR).

ACSCP (Shen et al., 2018) uses two GANs, the first one generates density map from full image, the second one generates density map from one fourth of the full image. The author proposed Cross-Scale Consistency Pursuit (CSCP) loss, which minimizes the absolute difference between generated density map from the two GANs employed.

MS-GAN (Yang, Zhou, & Kung, 2018) uses scale-aware CNN architecture for generating density map. The scale-aware architecture is a single column CNN which each layer output is concatenated to produce the density map.

GAN-MTR (Olmschenk, Tang, & Zhu, 2018) incorporates a semi-supervised framework to train a GAN for crowd counting. The supervised loss of the algorithm is the typical crowd counting supervised loss. The unsupervised loss penalized the output if the range is outside of the known output. The density map generated form fake image is punished to zero.

6. Approaches With Semi-Supervised Learning

Compared to other computer vision problem such as image classification and object detection, the size of annotated crowd counting dataset is significantly smaller. Deep learning is known to have a better performance with bigger data, thus the size of crowd counting dataset is a challenge to develop a robust deep learning model. Given the limited size of annotated data, semi-supervised learning algorithms is promising to be used for crowd counting. Currently, there are three semi-supervised CNN that has been developed for crowd counting: Grid Winner-Take-All Autoencoder (GWTA) (Sam, Sajjan, Maurya, & Babu, 2019), Learning to Rank (L2R) (X. Liu, van de Weijer, & Bagdanov, 2018), and the previously explained GAN-MTR.

GWTA is a modified version of Winner-Take-All Autoencoder (WTA) (Makhzani & Frey, 2015) specifically designed for crowd counting. If WTA is straightly applied to crowd counting, it can only be used to select units with largest activation globally. In GWTA, the selection process is done within grids, that are formed by dividing the last feature maps into sub-regions. This allows GWTA to select largest activation in local context, thus introduces a more reliable performance locally. As a typical autoencoder, GWTA can be trained unsupervisedly to reconstruct its input. Only the last two layers are trained in supervisedly.

Meanwhile, L2R uses a semi-supervised framework that can be seen as a multi-task learning with supervised and unsupervised task. The supervised task is the standard density maps regression task. The unsupervised task is to rank between images, which has more crowd count. The ranking data is formed by taking sub-images of a crowd image. The authors assume that the sub-images should have less crowd count than the original image, thus they guide the model to rank the sub-images with less count.

7. Other Approaches For Deep Learning In Crowd Counting

Other than the previous six major approaches, there are five deep learning models for crowd counting that do not fit to the definition of the major approaches: ConvLSTM (Xiong, Shi, & Yeung, 2017), Top-Down Feedback CNN (TDF-CNN) (Sam & Babu, 2018), Deeply Recursive ResNet, (DR-ResNet) (Ding, Lin, He, Wang, & Huang, 2018), NetVLAD for crowd counting (Z. Shi, Zhang, Sun, & Ye, 2018), and Adaptive Counting CNN (A-CCNN) (Amirgholipour, He, Jia, Wang, & Zeibots, 2018).

The main idea of ConvLSTM is to incorporate temporal relationship between images in a video for crowd counting. This temporal information is captured by using Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) combined with convolutional layer to take image as input and generate density map.

In TDF-CNN, two CNN are employed as supporting modules: bottom-up network and top-down network. These two networks are trained separately in two sequential phases. The bottom-up network is trained first as a typical density map regressor. Afterward, the bottom-up network weights are frozen and used to give feedback features to the first layer in the main CNN via top-down networks.

The next model, DR-ResNet, uses ResNet with two residual modules that are stacked to a single CNN. The second module is placed on top of the first module to receive the output from the first module. Subsequently, the output of the second module is passed to the second module once again, thus the second module become a residual recursive module.

Meanwhile, Shi et al. designed a crowd count model that uses NetVLAD (Arandjelovic, Gronat, Torii, Pajdla, & Sivic, 2018). NetVLAD is a trainable version of a popular visual descriptor VLAD (Vector of Locally Aggregated Descriptor) (Jegou, Douze, Schmid, & Perez, 2010).

The last model, A-CCNN, uses fuzzy inference system from Mamdani (Mamdani, 1977) to choose optimal hyperparameter of Counting CNN (CCNN). CCNN is the single column CNN that form the HydraCNN.

8. State-Of-The-Art Methodologies In Crowd Counting

In table 2, we list all methods that achieves a state-of-the-art performance on ShanghaiTech Part A, ShanghaiTech Part B, WorldExpo'10, and UCF_CC_50 dataset. We can conclude from the table that scale-aware method is the most competitive approach for crowd counting. Among the state-of-the-art models with scale-aware method, there is no model that use multi-column architecture. Thus, the introduction of single-column architecture for scale-aware model not only allows faster inference, but also more powerful performance. Multi-tasking approach is also proven to be effective for improving crowd counting CNN, which two models become state-of-the-art models. Other approach such as local context utilization and ensemble learning are also able to produce a state-of-the-art crowd counting model. The performance of all state-of-the-art

models on the popular crowd counting dataset are listed in table 3, 4, 5, and 6

Table 2 State-of-the-art Models in Popular Crowd Counting Datasets

Method	Category
SFCN	Single-column scale-aware CNN
SPN	Single-column scale-aware CNN
CAN	Semi multi-column scale-aware CNN
CSRNet	Semi multi-column scale-aware CNN
FPNCC	Semi multi-column scale-aware CNN
ADCrowdNet	Multi-tasking CNN
PACNN	Multi-tasking CNN
DRSAN	CNN with local context
D-ConvNet	CNN with ensemble learning

Table 3 State-of-the-art Models Performance on ShanghaiTech Part A

Method	MAE	MSE
SFCN	64.8	107.5
SPN	**61.7**	99.5
CAN	62.3	100.0
CSRNet	68.2	115.0
FPNCC	81.2	139.2
ADCrowdNet(AMG-bAttn-DME)	63.2	98.9
ADCrowdNet(AMG-Attn-DME)	-	-
PACNN	66.3	106.4
PACNN + CSRNet	62.4	102.0
DRSAN	69.3	**96.4**
D-ConvNet-v1	73.5	112.3

Table 4 State-of-the-art Models Performance on ShanghaiTech Part B

Method	MAE	MSE
SFCN	**7.6**	13.0
SPN	9.4	14.4
CAN	7.8	12.2
CSRNet	10.6	16.0
FPNCC	**7.6**	12.0
ADCrowdNet(AMG-bAttn-DME)	-	-
ADCrowdNet(AMG-Attn-DME)	-	-
PACNN	8.9	13.5
PACNN + CSRNet	**7.6**	**11.8**
DRSAN	11.1	18.2
D-ConvNet-v1	73.5	112.3

Table 6 State-of-the-art Models Performance on WorldExpo'10

Method	S1	S2	S3	S4	S5	Avg
SFCN	-	-	-	-	-	9.4
SPN	-	-	-	-	-	-
CAN	2.9	12.0	10.0	**7.9**	4.3	7.4
ECAN	2.4	**9.4**	8.8	11.2	4.0	**7.2**
CSRNet	2.9	11.5	**8.6**	16.6	3.4	8.6
FPNCC	1.9	22	12.3	16	4.3	11.3
ADCrowdNet (AMG-bAttn-DME)	1.7	14.4	11.5	**7.9**	3.0	7.7
ADCrowdNet (AMG-Attn-DME)	**1.6**	13.2	8.7	10.6	**2.6**	7.3
PACNN	2.3	12.5	9.1	11.2	3.8	7.8
DRSAN	2.6	11.8	10.3	10.4	3.7	7.8
D-ConvNet-v1	1.9	12.1	20.7	8.3	**2.6**	9.1

Table 5 State-of-the-art Models Performance on UCF_CC_50

Method	S1	S2	S3	S4	S5	Avg
SFCN	-	-	-	-	-	9.4
SPN	-	-	-	-	-	-
CAN	2.9	12.0	10.0	**7.9**	4.3	7.4
ECAN	2.4	**9.4**	8.8	11.2	4.0	**7.2**
CSRNet	2.9	11.5	**8.6**	16.6	3.4	8.6
FPNCC	1.9	22	12.3	16	4.3	11.3
ADCrowdNet (AMG-bAttn-DME)	1.7	14.4	11.5	**7.9**	3.0	7.7
ADCrowdNet (AMG-Attn-DME)	**1.6**	13.2	8.7	10.6	**2.6**	7.3
PACNN	2.3	12.5	9.1	11.2	3.8	7.8
DRSAN	2.6	11.8	10.3	10.4	3.7	7.8
D-ConvNet-v1	1.9	12.1	20.7	8.3	**2.6**	9.1

V. CONCLUSION

In this paper, we extensively review deep learning models that are used for crowd counting. To accelerate the comprehension of recent progress in deep learning for crowd counting, we categorize all models into six groups. With the categorization, we can conclude that scale-aware models are the most successful approach for crowd counting to date. Therefore, the future works of crowd counting might benefit from incorporating scale-aware prior in deep learning models. Other promising approaches are to embrace multi-tasking, local context, and ensemble learning in a deep learning model.

REFERENCES

Amirgholipour, S., He, X., Jia, W., Wang, D., & Zeibots, M. (2018). A-CCNN: Adaptive CCNN for Density Estimation and Crowd Counting. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 948–952). IEEE. https://doi.org/10.1109/ICIP.2018.8451399

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic,

J. (2018). NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6), 1437–1451. https://doi.org/10.1109/TPAMI.2017.2711011

Boominathan, L., Kruthiventi, S. S. S., & Babu, R. V. (2016). CrowdNet: A Deep Convolutional Network for Dense Crowd Counting. In Proceedings of the 2016 ACM on Multimedia Conference - MM '16 (pp. 640–644). New York, New York, USA: ACM Press. https://doi.org/10.1145/2964284.2967300

Cao, X., Wang, Z., Zhao, Y., & Su, F. (2018). Scale Aggregation Network for Accurate and Efficient Crowd Counting. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11209 LNCS, 757–773. https://doi.org/10.1007/978-3-030-01228-1_45

Cenggoro, T. W., Aslamiah, A. H., & Yunanto, A. (2019). Feature Pyramid Networks for Crowd Counting. In To appear: 2019 International Conference on Computer Science and Computational Intelligence. Yogyakarta: Elsevier.

Chan, A. B., Zhang-Sheng John Liang, & Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1–7). IEEE. https://doi.org/10.1109/CVPR.2008.4587569

Chen, K., Loy, C. C., Gong, S., & Xiang, T. (2012). Feature Mining for Localised Crowd Counting. In Procedings of the British Machine Vision Conference 2012 (Vol. 47, pp. 21.1-21.11). British Machine Vision Association. https://doi.org/10.5244/C.26.21

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 (Vol. 19, pp. 785–794). New York, New York, USA: ACM Press. https://doi.org/10.1145/2939672.2939785

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable Convolutional Networks. In 2017 IEEE International Conference on Computer Vision (ICCV) (Vol. 2017-Octob, pp. 764–773). IEEE. https://doi.org/10.1109/ICCV.2017.89

Deb, D., & Ventura, J. (2018). An aggregated multicolumn dilated convolution network for perspective-free counting. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2018-June, 308–317. https://doi.org/10.1109/CVPRW.2018.00057

Ding, X., Lin, Z., He, F., Wang, Y., & Huang, Y. (2018). A Deeply-Recursive Convolutional Network For Crowd Counting. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Vol. 2018-April, pp. 1942–1946). IEEE. https://doi.org/10.1109/ICASSP.2018.8461772

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gao, J., Wang, Q., & Li, X. (2019). PCC Net: Perspective Crowd Counting via Spatial Convolutional Network, 1–13. Retrieved from http://arxiv.org/abs/1905.10085

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems 27, 2672–2680. Retrieved from http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, 1–14. https://doi.org/10.1007/978-3-319-10578-9_23

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 7, pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hossain, M., Hosseinzadeh, M., Chanda, O., & Wang, Y. (2019). Crowd Counting Using Scale-Aware Attention Networks. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1280–1288). IEEE. https://doi.org/10.1109/WACV.2019.00141

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1. https://doi.org/10.1109/TPAMI.2019.2913372

Huang, S., Li, X., Zhang, Z., Wu, F., Gao, S., Ji, R., & Han, J. (2018). Body Structure Aware Deep Crowd Counting. IEEE Transactions on Image Processing, 27(3), 1049–1059. https://doi.org/10.1109/TIP.2017.2740160

Idrees, H., Saleemi, I., Seibert, C., & Shah, M. (2013). Multi-source Multi-scale Counting in Extremely Dense Crowd Images. In 2013 IEEE Conference

on Computer Vision and Pattern Recognition (pp. 2547–2554). IEEE. https://doi.org/10.1109/CVPR.2013.329

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., & Shah, M. (2018). Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11206 LNCS, 544–559. https://doi.org/10.1007/978-3-030-01216-8_33

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial Transformer Networks. Nips, 1–14. https://doi.org/10.1038/nbt.3343

Jegou, H., Douze, M., Schmid, C., & Perez, P. (2010). Aggregating local descriptors into a compact image representation. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 3304–3311). IEEE. https://doi.org/10.1109/CVPR.2010.5540039

Jiang, X., Xiao, Z., Zhang, B., Zhen, X., Cao, X., Doermann, D., & Shao, L. (2019). Crowd Counting and Density Estimation by Trellis Encoder-Decoder Network. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Retrieved from http://arxiv.org/abs/1903.00853

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation, 1(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Lempitsky, V., & Zisserman, A. (2010). Learning To Count Objects in Images. Advances in Neural Information Processing Systems, 1324–1332. https://doi.org/10.1111/1467-9280.03439

Li, Y., Zhang, X., & Chen, D. (2018). CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1091–1100). IEEE. https://doi.org/10.1109/CVPR.2018.00120

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua, 936–944. https://doi.org/10.1109/CVPR.2017.106

Liu, C., Weng, X., & Mu, Y. (2019). Recurrent Attentive Zooming for Joint Crowd Counting and Precise Localization. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Liu, J., Gao, C., Meng, D., & Hauptmann, A. G. (2018). DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5197–5206). IEEE. https://doi.org/10.1109/CVPR.2018.00545

Liu, L., Wang, H., Li, G., Ouyang, W., & Lin, L. (2018). Crowd counting using deep recurrent spatial-aware network. IJCAI International Joint Conference on Artificial Intelligence, 2018-July, 849–855. https://doi.org/arXiv:1807.00601v1

Liu, W., Salzmann, M., & Fua, P. (2019). Context-Aware Crowd Counting. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Retrieved from http://arxiv.org/abs/1811.10452

Liu, X., van de Weijer, J., & Bagdanov, A. D. (2018). Leveraging Unlabeled Data for Crowd Counting by Learning to Rank. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7661–7669). IEEE. https://doi.org/10.1109/CVPR.2018.00799

Liu, Y., & Yao, X. (1999). Ensemble learning via negative correlation. Neural Networks, 12(10), 1399–1404. https://doi.org/10.1016/S0893-6080(99)00073-8

Long, J., Shelhamer, E., & Darrell, T. (2014). Fully Convolutional Networks for Semantic Segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Makhzani, A., & Frey, B. (2015). Winner-take-all Autoencoders. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (pp. 2791–2799). Cambridge, MA, USA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=2969442.2969552

Mamdani, E. H. (1977). Application of fuzzy logic to approximate reasoning using linguistic. Ieee_J_C, C–26(12), 1182--1191. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0017219295&partnerID=tZOtx3y1

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. In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1–7). IEEE. https://doi.org/10.1109/AVSS.2017.8078482

Ming Liu, Jue Jiang, Zhenqei Guo, Zenan Wang, & Yang Liu. (2018). Crowd Counting with Fully Convolutional Neural Network. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 953–957). IEEE. https://

doi.org/10.1109/ICIP.2018.8451787

Olmschenk, G., Tang, H., & Zhu, Z. (2018). Crowd Counting with Minimal Data Using Generative Adversarial Networks for Multiple Target Regression. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (Vol. 2018-Janua, pp. 1151–1159). IEEE. https://doi.org/10.1109/WACV.2018.00131

Oñoro-Rubio, D., & López-Sastre, R. J. (2016). Towards Perspective-Free Object Counting with Deep Learning. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), ECCV 2016: 14th European Conference, Amsterdam, The Netherlands (Vol. 9911, pp. 615–629). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46478-7_38

Sam, D. B., & Babu, R. V. (2018). Top-Down Feedback for Crowd Counting Convolutional Neural Network. Retrieved from http://arxiv.org/abs/1807.08881

Sam, D. B., Sajjan, N. N., Babu, R. V., & Srinivasan, M. (2018). Divide and Grow: Capturing Huge Diversity in Crowd Images with Incrementally Growing CNN. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3618–3626). IEEE. https://doi.org/10.1109/CVPR.2018.00381

Sam, D. B., Sajjan, N. N., Maurya, H., & Babu, R. V. (2019). Almost Unsupervised Learning for Dense Crowd Counting. Aaai. Retrieved from val.serc.iisc.ernet.in/valweb/papers/AAAI_2019_WTACNN.pdf%0A

Sam, D. B., Surya, S., & Babu, R. V. (2017). Switching Convolutional Neural Network for Crowd Counting. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4031–4039). IEEE. https://doi.org/10.1109/CVPR.2017.429

Sang, J., Wu, W., Luo, H., Xiang, H., Zhang, Q., Hu, H., & Xia, X. (2019). Improved Crowd Counting Method Based on Scale-Adaptive Convolutional Neural Network. IEEE Access, 7, 24411–24419. https://doi.org/10.1109/ACCESS.2019.2899939

Shen, Z., Xu, Y., Ni, B., Wang, M., Hu, J., & Yang, X. (2018). Crowd Counting via Adversarial Cross-Scale Consistency Pursuit. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5245–5254). IEEE. https://doi.org/10.1109/CVPR.2018.00550

Shi, M., Yang, Z., Xu, C., & Chen, Q. (2019). Revisiting Perspective Information for Efficient Crowd Counting. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Retrieved from http://arxiv.org/abs/1807.01989

Shi, Z., Zhang, L., Liu, Y., Cao, X., Ye, Y., Cheng, M.-M., & Zheng, G. (2018). Crowd Counting with Deep Negative Correlation Learning. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5382–5390). IEEE. https://doi.org/10.1109/CVPR.2018.00564

Shi, Z., Zhang, L., Sun, Y., & Ye, Y. (2018). Multiscale multitask deep NetVLAD for crowd counting. IEEE Transactions on Industrial Informatics, 14(11), 4953–4962. https://doi.org/10.1109/TII.2018.2852481

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. The International Conference on Learning Representations 2015, 1–14. Retrieved from http://arxiv.org/abs/1409.1556

Sindagi, V. A., & Patel, V. M. (2017a). CNN-Based cascaded multi-task learning of high-level prior and density estimation for crowd counting. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017. https://doi.org/10.1109/AVSS.2017.8078491

Sindagi, V. A., & Patel, V. M. (2017b). Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs. Proceedings of the IEEE International Conference on Computer Vision, 2017-Octob, 1879–1888. https://doi.org/10.1109/ICCV.2017.206

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 07-12-June, pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

Walach, E., & Wolf, L. (2016). Learning to count with CNN boosting. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9906 LNCS, 660–676. https://doi.org/10.1007/978-3-319-46475-6_41

Wang, Q., Gao, J., Lin, W., & Yuan, Y. (2019). Learning from Synthetic Data for Crowd Counting in the Wild. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Retrieved from http://arxiv.org/abs/1903.03303

Wang, Z., Xiao, Z., Xie, K., Qiu, Q., Zhen, X., & Cao, X. (2018). In Defense of Single-column Networks for Crowd Counting. Retrieved from http://arxiv.org/abs/1808.06133

Wu, X., Zheng, Y., Ye, H., Hu, W., Yang, J., & He, L. (2018). Adaptive Scenario Discovery for Crowd Counting. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2382–2386). Retrieved from http://arxiv.org/abs/1812.02393

Xiong, F., Shi, X., & Yeung, D.-Y. (2017). Spatiotemporal

Modeling for Crowd Counting in Videos. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 5161–5169). IEEE. https://doi.org/10.1109/ICCV.2017.551

Yang, J., Zhou, Y., & Kung, S.-Y. (2018). Multi-scale Generative Adversarial Networks for Crowd Counting. In 2018 24th International Conference on Pattern Recognition (ICPR) (Vol. 2018-Augus, pp. 3244–3249). IEEE. https://doi.org/10.1109/ICPR.2018.8545683

Yu, F., & Koltun, V. (2015). Multi-Scale Context Aggregation by Dilated Convolutions. https://doi.org/10.16373/j.cnki.ahr.150049

Zeng, L., Xu, X., Cai, B., Qiu, S., & Zhang, T. (2017). Multi-scale convolutional neural networks for crowd counting. In 2017 IEEE International Conference on Image Processing (ICIP) (Vol. 2017-Septe, pp. 465–469). IEEE. https://doi.org/10.1109/ICIP.2017.8296324

Zhang, C., Li, H., Wang, X., & Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 07-12-June, pp. 833–841). IEEE. https://doi.org/10.1109/CVPR.2015.7298684

Zhang, L., Shi, M., & Chen, Q. (2018). Crowd Counting via Scale-Adaptive Convolutional Neural Network. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (Vol. 2018-Janua, pp. 1113–1121). IEEE. https://doi.org/10.1109/WACV.2018.00127

Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016). Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 589–597. https://doi.org/10.1109/CVPR.2016.70

Zou, Z., Su, X., Qu, X., & Zhou, P. (2018). DA-Net: Learning the fine-grained density distribution with deformation aggregation network. IEEE Access, 6, 60745–60756. https://doi.org/10.1109/ACCESS.2018.2875495