Highlights

Collaborative Algorithms that Combining AI with IoT Towards Monitoring and Control System

Tao Zhang,Yan Zhao,Wenjing Jia,Mu-Yen Chen

- Our proposed system can provide effective public danger prediction
- Conditional Random Fields can promote the feature fusion result remarkably
- Density-adaptive Gaussian kernel can elevate the quality of density maps

Collaborative Algorithms that Combining AI with IoT Towards Monitoring and Control System

Tao Zhang[a,1], Yan Zhao[a,*,2], Wenjing Jia[b,3] and Mu-Yen Chen[c,4]

[a]School of Artificial Intelligence and Computer, Jiangnan University, Lihu Revenue 1800, Wuxi, 214122, Jiangsu, China
[b]Faculty of Engineering and Information Technology, University of Technology, Broadway, Sydney, 123, NSW, Australia
[c]Department of Engineering Science, National Cheng Kung University, Broadway, Tainan, 70101, Taiwan, China

ARTICLE INFO

Keywords:
Collaborative Algorithm
Artificial Intelligence of Things
Crowd counting
Monitoring and control system

ABSTRACT

Recently, a new IoT structure known as the Artificial Intelligence of Things (AIoT) comes into play. Crowd counting is a promising field in data analysis of AIoT, however, due to poor transparency and high data security risks, developing a novel network architecture that can precisely elevate the counting of heavy crowd is extremely difficult. In addition, the fusion of IoT and AI also poses several challenges. The focus of this work is on the effective design of IoT framework and deep learning algorithm towards security of smart city. The system can be used to estimate the crowd traffic in public places, and can prevent the occurrence of congestion, stampede and other accidents, such as stations, airports, large-scale exhibitions, tourist attractions and other places. The constructed system contains video collection, upload and display as well as data analysis and early warning operation at the embedded device end, and automatically tracks densely crowd areas by controlling the video monitoring device. Moreover, the cloud platform can be controlled through the network. Our proposed algorithms are composed of two main aspects, i.e., division and focus. Firstly, we propose a novel density-adaptive Gaussian kernel to elevate the quality of density maps. Then, we propose a module based on conditional random fields for feature fusion. Finally, we propose a block segmentation module to predict our segmentation results and extract the context-aware information in segmentation stage. Experiments on our captured data, the Shanghai Tech, UCF_CC_50 and UCF_QNRF datasets demonstrate that our solution has obtained better performance and lower count errors over the state of the art.

1. INTRODUCTION

The tremendous expansion of the global population and urbanization has resulted in frequent crowd gathering in public places. In the scenarios where the crowd density becomes too high, stampedes and crushes may cause emergency circumstances and can be really dangerous. Such incidents have occurred from time to time in recent years. Consequently, it is essential to find a reliable solution to obtain the density estimation and object counting in densely crowded scenes with real-time images captured by smart devices. The rapid development in communication devices and software has facilitated the spread of Internet of Things (IoT) security system. The Internet of Things (IoT) refers to a network comprised of physical objects capable of gathering and sharing electronic information. The Internet of Things includes a wide variety of smart devices, from industrial machines that transmit data about the production process to sensors that track information about the human body. In the context of the rapid development of Internet of Things technology, we should utilize advanced Internet of Things devices, such as smart cameras, to carry out context-aware public security data analysis, which is also known as IoT Security. As a hot topic in the research of distributed computing and IoT, collaborative learning has been a hot topic and popular with researchers recently. Combining AI algorithm and the data we received by IoT technology, we can figure out extremely difficult problems in real world based on deep learning. Crowd counting, for instance, is a promising field in data analysis of IoT data. However, due to the occlusion, background clutters, large scale and perspective variations, as illustrated in Fig 1, developing a novel network architecture that can precisely elevate the counting of heavy crowd is extremely difficult.

Overall, the main contributions of our paper are summarized as follows:

*Corresponding author
taozhang@jiangnan.edu.cn (T. Zhang); zhaoyan_jlu@163.com (Y. Zhao); Wenjing.Jia@uts.edu.au (W. Jia); mychen119@gs.ncku.edu.tw (M. Chen)
ORCID(s):

(a) Background Clutters (b) Scale Variation

Figure 1: Common scenes in crowd counting. In (a) on the left, the background scene includes building, lights and cars, which is prone to be recognized as crowd and interferes the estimation accuracy. In (b), the sizes of human heads vary severely with an image owing to different distance from camera.

- Our work focuses on the robust construction of Internet-based monitoring and control systems, this system is intelligent integration of hardware and software. The main aim of our constructed system is to output the crowd information with less noise in the surveillance scene through our designed smart devices. The designed smart device uses a modified camera to capture and further process the captured video information, which can be transmitted through remote or wireless devices.
- Different from the previous approaches using Gaussian kernels with invariant hyper parameters and radius, we construct a new Gaussian kernel which parameters vary with the change of local density and distance of the annotation points. Therefore, the density maps created by this method is obviously more robust and have better performance in the training process.
- Inspired by Shen, Xu, Ni, Wang and Yang (2018), we propose a supervised learning network which provides both the global targets count and the spatial distribution of targets by constructing divided density maps. And we propose a novel convolutional framework which uses conditional random fields (CRFs) to fuse and reconstruct the features in various scales. Extensive experiments on benchmarks show that the feature fusion module we proposed can promote the performance of encoder-decoder network.

The rest of the paper is organized as follows. Sect. 2 focuses on the Background and Related Works. Sect. 3 and 4 detail the mentioned modules of our proposed algorithm and detailed experiment setting. Sect. 5 presents experimental results where performance is compared with the state-of-the-art approaches. We conclude the paper and discuss the direction of our future work in Sect. 6.

2. BACKGROUND AND RELATED WORKS

Designing an appropriate system detection framework for IoT is difficult because IoT is heterogeneous Guo and Shen (2017),Yang, Zheng and Tang (2017),Zhang, Li, Jia, Sun and Yang (2017b),Zhang, Yue, Shen, Zhu, Zhen, Cao and Shao (2019), concurrent by nature, and sensitive to timing. Some difficulties happen all the time due to the difficult fusion problem of hardware and software when IoT uses some pattern recognition algorithms for computation Cheng, Wang, Jiang, Hua, Feng, Zhang and Zhou (2018),Zhou, Hu, Wang, Lu and Zhao (2013),Ghayvat, Mukhopadhyay, Gui and Suryadevara (2015),Zhang, Jia, Gong, Sun and Song (2017a). Generally speaking, the AI field does not focus on the security and real-time performance of hardware and software but concentrate on how to use robust algorithms to solve specific difficult problems Niu, Lin and Ke (2018),Zou, Dong and Wu (2018),Shen, Lee, Shu and Guo (2016). Due to the lack of consideration of algorithm complexity, there are few opportunities for practical application Li, Chen, Tang and Yan (2018a),Chen, Guo and Bao (2016),Shi, Cao, Zhang, Li and Xu (2016). Combining AI algorithm with

IoT system, the main problem is to predict the expected performance of the algorithm used to solve the AI problem, the performance in the worst case and the high cost of planning AI task framework.

IoT technology is a new decentralized data architecture and distributed computing paradigm, with the characteristics of transparency, decentralization, high credibility, non-tampering, traceability and high encryption security. In the IoT architecture, consensus mechanisms such as workload proof are used to realize non-tampering and non-forgery functions of transaction data Cheng, Jiang, Wang, Hua, Feng, Guo and Wu (2019),Wang, Zhang, Lin, Ge and Han (2018). The data security mechanism is used to implement hash calculation and asymmetric data encryption for the content of transaction data. Hu and Ni (2018) provides methods to detect vehicles rapidly and reduce the data volume by storing data as the detection results. Ding, Wu, Zhang, Lin, Tsiftsis and Yao (2018) and Kim and Ben-Othman (2018) offer effective ways to accomplish amateur smart city surveillance with the help of UAVs. The communication subject is P2P network, and each node has equal status and is interconnected and interactive in a flat topology structure. There is no centralized special node or hierarchical structure. Each node can undertake network routing, verifying block data, disseminating block data, discovering new nodes and other functions. Decentralized systems based on the blockchain are designed for IoT data exchange and storage Novo (2018),Huang, Su, Zhang, Shi, Zhang and Xie (2017),Kumar, Murugan, Muruganantham and Sriman (2020) and that makes the whole framework more trustworthy and robust. This paper concentrates on developing a context-aware IoT system with the help of AI algorithm, so in this part we give some detailed introductions about crowd counting methods.

Detection-based and regression-based methods Ryan, Denman, Sridharan and Fookes (2015),Sindagi and Patel (2017b),Kang, Ma and Chan (2017),Wang, Zhang, Yang, Liu and Cao (2015) are the dominating solutions at the early stage. However, these methods are often unreliable and weak generation, especially when the crowd becomes very dense. As far as we know, the detection-based methods work better in sparse scenes whereas the regression-based methods are on the contrary. Liu, Gao, Meng and Hauptmann (2018a) generated detection-based and regression-based density maps respectively.

With the rapid development of deep learning, CNN-based methods have been striving to address the counting problem in crowded scenes. These methods have achieved great improvements while being compared with classical solutions. Zhang, Zhou, Chen, Gao and Ma (2016),Ooro-Rubio and Lopez-Sastre (2016),Sam, Surya and Babu (2017),Ranjan, Le and Hoai (2018) have achieved remarkable performance using multi-scale and multi-branch networks for addressing the scale variation. They apply convolutional neural networks with various receptive fields to extract features of different head sizes.

Therefore, three obvious drawbacks in current works are summarized as follows: First, at present, there is no specific technology to alleviate the large density difference caused by perspective distortion. Secondly, the segmentation feature maps generated during training implies plenty of spatial information and semantic information. However, existing methods have neglected these features. Thus, we construct binary segmentation maps and use them to co-generate density maps with an accompanying segmentation loss function. Thirdly, because of the scale and density variation in different image regions, we propose a division strategy to analyze and evaluate the predicted density maps.

In this section, we will give a brief introduction on the most related works of crowd counting and review the corresponding surveys Ryan et al. (2015),Sindagi and Patel (2017b),Kang et al. (2017) for reference. We roughly group existing crowd counting approaches into several major categories and review the most representative works in each category as below.

2.1. Counting by CNNs

The main idea of this approach is to learn one target at a time. But in recent years, more works choose to focus on multi-task training because of its success in other computer vision fields. Multi-task learning networks Ranjan et al. (2018),Sam, Sajjan, Babu and Srinivasan (2018) mostly are made up with several subnets, which have different aims to learn. Gao, Wang and Li (2019) takes perspective change into consideration, which is usually ignored by many other works. Its whole framework consists of three main parts, i.e., a Density Map Estimation, a Random High-Level Density Classification and Fore/Background Segmentation. Besides, a DULR module is embedded in PCCNet to extract the perspective features. Yang, Li, Wu, Su, Huang and Sebe (2020) also utilized a perspective method to wrap the density maps and images to make the scale variation more acceptable. In our work, a novel scheme supervised by binary segmentation is proposed to improve the network structure through effectively discriminating the foreground and background.

Most of related works offer their own ways to handle the multi-scale features in the task. Luo, Yang, Li, Nie, Jiao, Zhou and Cheng (2020) present HyGnn to interweave the multi-scale features and localization together and reason

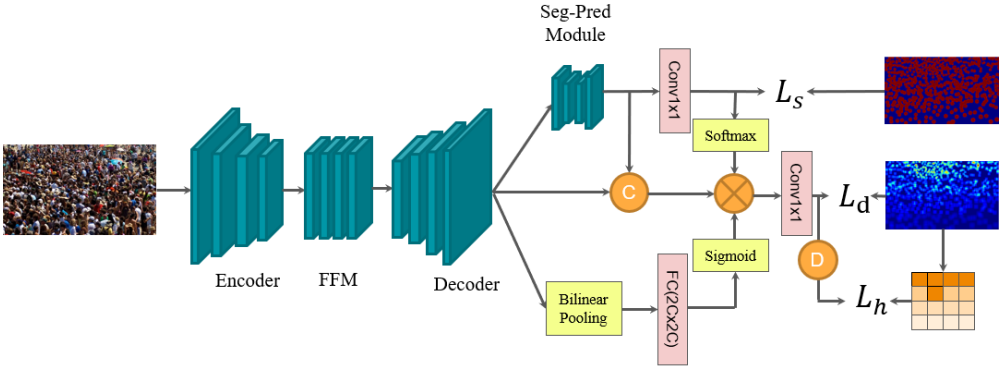


Figure 2: Overview of our network. The Encoder is composed of a few convolutional layers and the Decoder is composed of a few deconvolutional layers. The details of FFM is in Sect 3.3. The C denotes concatation operation and the D denotes dividing operation. Seg_Pred Module is designed for segmentation of the foreground and background.

jointly with a graph structure. Liu, Qiu, Li, Liu, Ouyang and Lin (2019) used CRFs to recalculate the features in different scales.

Due to the huge cost of time and manpower on the annotation task, un/semi-supervised learning approaches are exploited to train neural networks. Olmschenk, Chen, Tang and Zhu (2019) proposes a semi-supervised structure that depends on GAN to seek the number of individuals in both real and fake images and strives to discriminate them. Liu et al. Liu, van de Weijer and Bagdanov (2018b),Liu, van de Weijer and Bagdanov (2019) exploited a way to pre-train unlabeled datasets, based on the fact that counts in cropped images are always less than those in original images. They tried to train the network to sort images of different sizes. Wang, Li and Xue (2019) algorithm offered a method to train with a scale-aware adversarial density adaption.

Owing to the fact that the density of crowds varies extremely drastically, some studies have been concentrating on different regions of images to fix this issue Xiong, Lu, Liu, Liu, Cao and Shen (2019),Shen et al. (2018),Sajid, Sajid, Wang and Wang (2020). Sajid et al. (2020) cropped images into patches and fed them into multi-way classification modules with labels representing their different density level without using any density maps. Shen et al. (2018) design a novel scale-consistency module that links the count of local patches and the count of their region union. Sajid (2019) proposed a Decision Module to decide to use zoom-in or zoom-out mode to resolve scale variation. Liu, Weng and Mu (2019) intended to zoom in the high-density regions and improve the resolution for a re-inspection and eventually elevate the performance of localization.

2.2. Counting by CRFs

Conditional random fields have been exploited in many fields of computer vision to enhance and refine vision features. Benefited from the message passing mechanism, feature maps enable feature fusion and optimization enhancement in different ranges. For instance, Zheng, Jayasumana, Romera-Paredes, Vineet, Su, Du, Huang and Torr (2015) proposed a network to reconstruct the pixel-wise relationship and used CRFs to refine segmentation maps. Xu, Ouyang, Alameda-Pineda, Ricci, Wang and Sebe (2017) proposed an Attention-Gated CRFs Module to reproduce rich representations and fuse features in different scales. Liu et al. Liu et al. (2019) utilized CRFs to reconstruct multiple features in different scales of crowd counting for the first time and they confirmed that CRFs was effective in many computer vision tasks.

In our research, based on idea of CRFs, we design a module called FFM and it is placed between encoder and decoder in the encoder-decoder backbone network. Multiscale features can be refined with the message fusion mechanism in this module.

3. PROPOSED SOLUTION

The focus of this section is on the details of our proposed architecture. The whole structure of our network is depicted in Fig.2. An original module is firstly fed into our proposed Encoder-Decoder structure, we train our network with a multi-task loss function that is based on density and dividing. We firstly introduce a novel method to generate high-quality ground-truth density maps, and then give descriptions of the fundamental parts of our network framework. Finally, a brief representation of the combinatorial loss function is introduced in this paper.

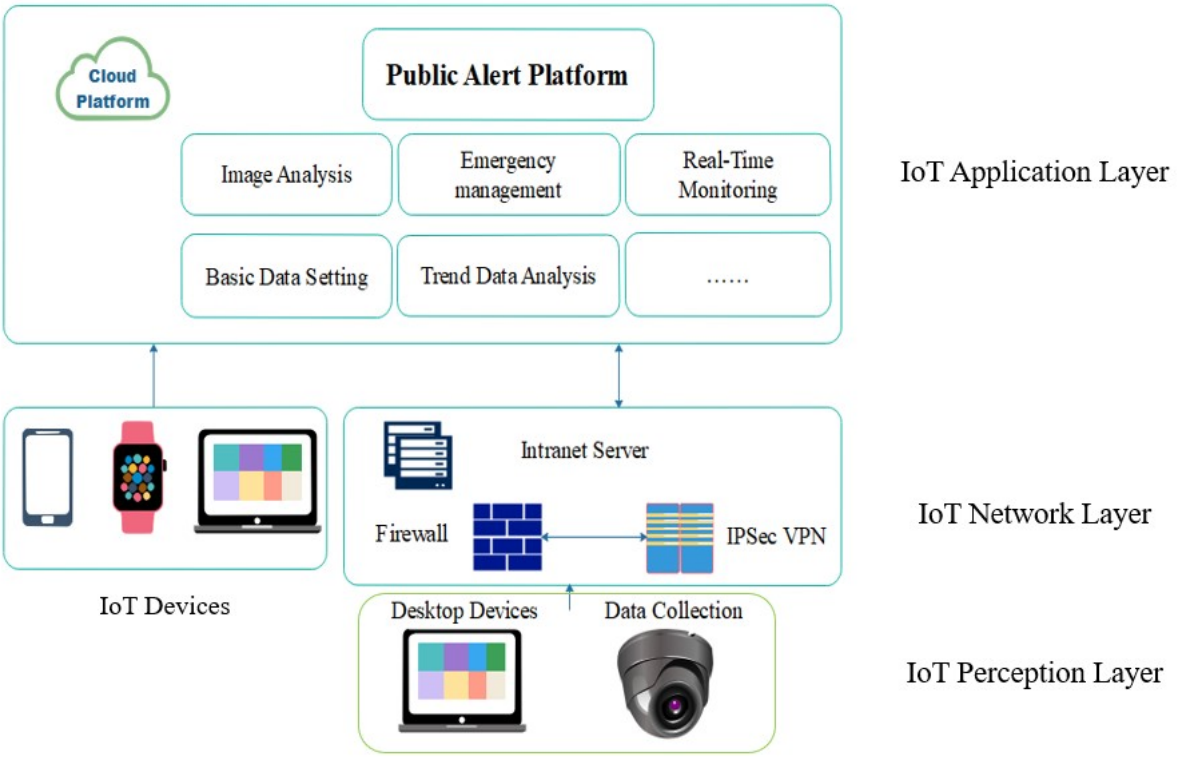


Figure 3: The overall framework of the proposed system.

3.1. Context-aware IoT system modeling

Illustrated in Fig.3, the framework we proposed is made up with four main parts, including Public Alert Platform(PAP), mobile and desktops devices connected to the alert platform, a data collection part with desktop devices and smart cameras and an intranet server for safe data transmission from the collection part. Data transmission and storage should be built in a decentralized way which is based on blockchain Novo (2018),Huang et al. (2017),Kumar et al. (2020). The PAP consists of image analysis, emergency management, real-time monitoring, basic data setting, trend data analysis and some other necessary functions. With the assistance of 5G and distributed computing, our framework is capable of accomplishing a series of public danger prediction and problem management.

The embedded development board collects the video through the camera, uploads the collected video to the server through the network module, and displays the collected video and the analyzed chart through the capacitor screen. At the same time, it receives the data from the server, and then controls the cloud platform to realize the tracking and monitoring of the camera. In order to make the web applications run better and make full use of hardware resources, multithreading technology is used in both the embedded device side and server side. The embedded development board of the system contains four CPU cores, and the hardware resources of the embedded development board can be fully utilized by using multithreading technology. We created a separate thread for TCP video transmission, with the main thread and the child threads executing in parallel on different CPUs without affecting video capture, display, and statistics, making the application run more smoothly. We created two threads on the server side. One thread is used

for image analysis, and one thread is used for video storage. Qt signal and slot mechanism is used to realize memory sharing between threads, so that each thread executes sequentially and concurrently.

The implementation of the whole camera tracking algorithm is divided into two parts, the server side and the embedded device side. The server side realizes the mass center calculation of crowd density map. The best monitoring effect can be obtained by centering the camera on the centroid area of the density map. After the embedded device receives the position information from the server, it is converted to the angle that needs to be adjusted by the steering gear, and the conversion relationship needs to be adjusted according to the camera imaging angle and the assembly relationship between the camera and the steering gear. After converting the angle, the PWM signal needed by the steering gear is calculated. Finally, the PWM signal is sent to the steering gear through the steering gear drive, and the steering gear makes actions according to the signal to realize the tracking function of the camera.

The overall framework of proposed context-aware IoT system consists of several function modes. Function modes are system structures to be scheduled and exploited on the hardware system. Each function module describes a state, implemented by one or more machine learning algorithms. Mode changes characterize the relationship of each functional module of the system, and several approaches were explored to solve deadlines misses problems Xu, Hammadeh, Kroller, Ernst and Quinton (2015),Pazzaglia, Mandrioli, Maggio and Cervin (2019). The proposed context-aware IoT system in this paper uses those function modes to keep the security, robustness and reliability of designed system; in addition, the proposed context-aware IoT system switches to stop mode immediately to prevent damage to the system in the event of a failure or unexpected event.

The proposed context-aware IoT system in Fig.4 mainly contain 5 function modes:

- stop
- preprocess
- train
- detect
- count

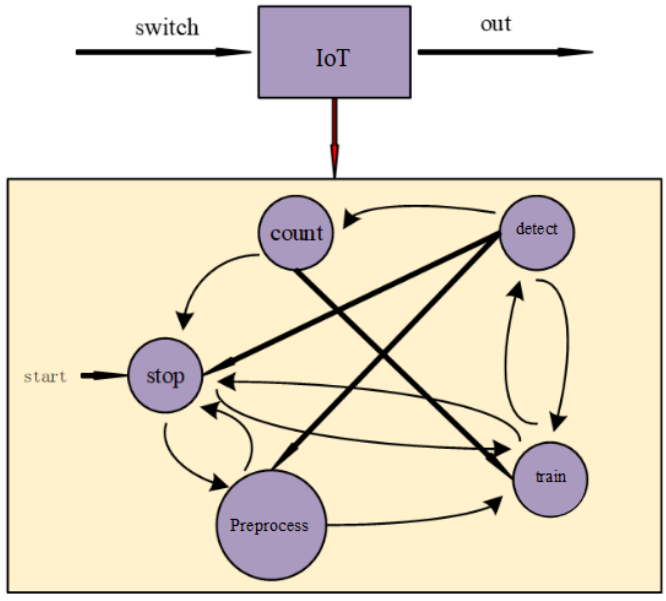


Figure 4: Proposed model for different modes of the system.

Fig.4 depicts the whole framework of the proposed system. It describes the five different function modes. The proposed context-aware IoT system begins at the stop mode. When the system is running, the preprocess mode is first

activated, then it can change to the train mode, and the training process operates on a set of crowd data samples to be tested, immediately it changes to the detect mode, starting a crowd detection process through some machine learning algorithms. When the crowd density map is generated in the detect module, the system switches to the count mode, where a self-learning generative adversarial networks model is proposed to count the crowd accurately. If it fails, the system will return to the detect module, otherwise switches to the preprocess and train mode, starting to seek the optimal training model.

When the crowd is no more detected in the video stream of the camera, the system will switch to the stop mode, or reconfirm this case in the preprocess and detect mode. The proposed system can switch to the stop mode at any time when something unexpected happens.

3.2. Density-Adaptive Kernel Estimation

We formulate the crowd counting problem as a density map estimation problem. Our work takes N training images $I_1, I_2, \cdots, I_N$, and an annotation set P_i corresponding to each image I_i, as input samples. Given annotated positions of every person, we create a density map D_i accordingly by convolving the head points with a Gaussian kernel $G(x)$, and generate the corresponding density map based on annotation points as shown in the following equation:

$$D_i(p) = \sum_{P \in P_i} N\left(p|\mu = P, \sigma_P^2\right), \tag{1}$$

where p denotes a single point annotation and N denotes the normalized Gaussian filter with a mean P and an isotropic covariance σ_p.However, this method only works well under the circumstance that the persons in the image are uniformly distributed.

Some studies have managed to eliminate the dramatic scale distortion by warping images and the ground-truth density maps according to perspective factors Gao et al. (2019),Yang et al. (2020). In order to overcome this conspicuous flaw, we create a density-adaptive kernel for congested scenarios to dynamically cover the image. We calculate our filter based on the density around persons, which makes the filter tune much more flexible. We define a hyper parameter N to divide the whole image into $N \times N$ parts. We also make a comparison between different N in Table 5. And based on the density in each sub regions, we can define a Density-Adaptive kernel to create the density maps in a more reasonable and elegant way.

3.3. Feature Fusion Module

Conditional Random Fields (CRFs) have been applied widely in the tasks of nature language processing and computer vision in recent years. As a conditional probability distribution model, CRFs take a set of high-dimension features as input and output a set of predicted features with a specific feature function. Multi-scale features in the set can fine-tune features in other scales and complement each other with a message transmission mechanism. Considering the fact that features in crowd counting task also vary in scales, we believe that constructing a module through fusing CRFs scheme can boost the performance of the network. As we can see in Fig.2, the Feature Fusion Module locates between the encoder and the decoder to transport fined features into the decoder.

The structure of FFM is showed clearly in Fig.5. With K hidden states, we update one state at one time with the other $K-1$ states with feature fusion operation. We also propose second version of FFM, where the hidden states are convolved into $W \times H \times C$ instead of concatenation. We will compare the effect of the two different version of FFM in Table 2. In the encoder of our backbone network, we employ both normal convolution and dilated convolution to extract features. After nine levels of convolution, we integrate features of levels 4, 5, 7 and 8 into FFM. The reason why we abandon features of level 6 is that the convolution layer in level 6 has a larger dilation rate and it is more likely to cause gridding artefacts consequently.

We adjust the four feature maps to have the same channel number 96 in order to avoid over-fitting. The four feature maps are passed into FFM to enhance the features instead of being passed into decoder directly, which makes features robust to the huge scale diversity of head size. Experiment results presented in the next section demonstrate that our proposed FFM improves the robustness of our net significantly. The decoder part consists of three deconvolution layers which have 4×4 kernel size and a stride of 2 respectively to recover high resolution. And we place three convolution layers around deconvolution layers to refrain from checkerboard artefact problems. As we can see from Fig.6, the network trained with FFM performs more stably and have lower MAE and RMSE apparently.

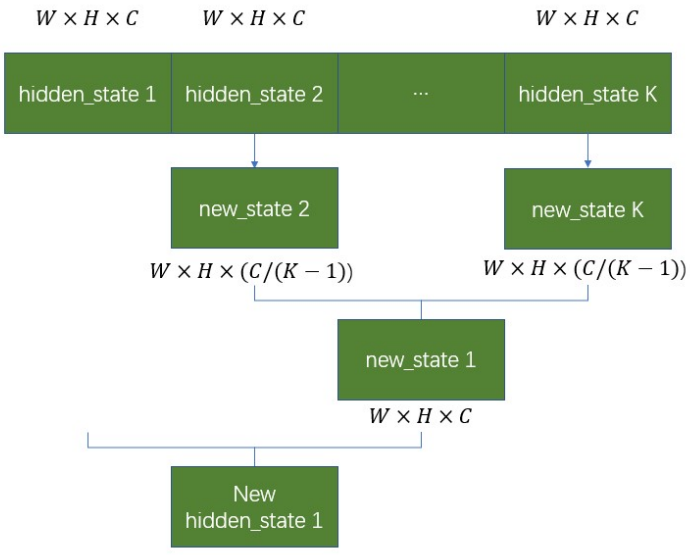
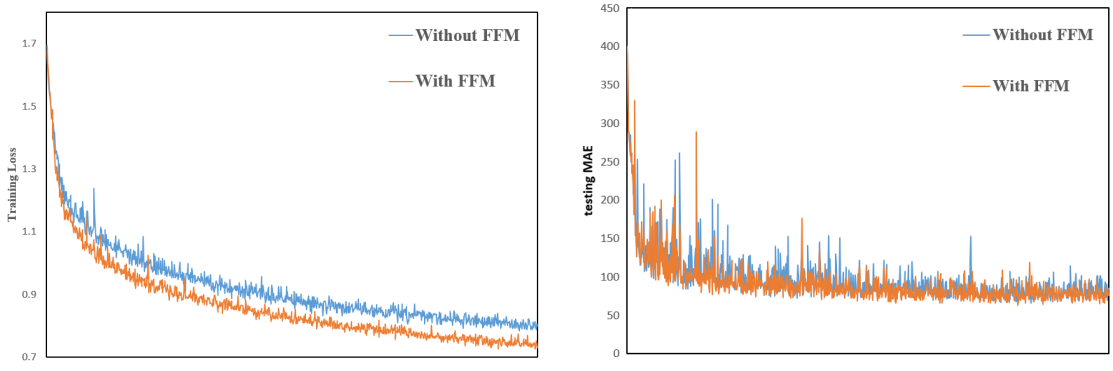


Figure 5: Structure of proposed FFM. W, H and C denote the width, height and channel of the K feature maps respectively. We take the first iteration for an example. We choose $K-1$ feature maps from 2 to K at a time and convolve them with same kernels and generate *new_state 1* with a concatation operation. Finally, we generate the new *hidden_state 1* with *hidden_state 1* and *new_state 1* through convolution operations. We can generate a group of new hidden states in one iteration, and finally we could get our feature-fusioned maps for decoding after T iterations.



(a) Training Loss (b) Testing MAE

Figure 6: These two pictures illustrate training loss and testing MAE on ShanghaiTech Part A dataset with and without FFM respectively.

3.4. Loss Function

The details of the loss functions applied in this work have been introduced and explained in earlier sections. In this section, we will summarise the loss functions used in this research.

We have three losses in total, including Local Density Loss (L_h), Density Map Estimation Loss (L_d) and Segmentation Loss (L_s). And L_h and L_s are constructed as Eq. (3.4) and Eq. (3.4). L_d is for pixel-wise density map estimation, constructed by combining $L1$ and $L2$ loss in order to elevate robustness to outliers:

$$L_d = \frac{1}{2}\left\|Den_p - Den_g\right\|_2^2 + \left\|Den_p - Den_g\right\|_1. \tag{2}$$

Table 1
Seg-Pred Module. The 'shape' column represents the size and amount of convolutional kernels and 'op' means the operations after convolution. $1 \times 1 \times 16$ denotes a filter with size 1×1 and 16 channels. *bn* denotes batch-normalization and *relu* is Rectified Linear Unit (ReLU).

Shape	*Op*	
$1 \times 1 \times 16$ conv	*bn*	*relu*
$3 \times 3 \times 32$ conv	*bn*	*relu*
$3 \times 3 \times 64$ conv	*bn*	*relu*
$3 \times 3 \times 32$ deconv	*bn*	*relu*
$3 \times 3 \times 16$ deconv	*bn*	*relu*
$1 \times 1 \times 2$ conv	\	\

$$L_s\left(V,\theta_s\right) = \sum_{l\in\{0,1\}} -\alpha^l S^l \left(1-\rho\left(F_s\left(V,\theta_s\right)\right)\right)^\gamma \log\left(\rho\left(F_s\left(V,\theta_s\right)\right)\right), \tag{3}$$

$$L_h\left(D_g, D_p\right) = \sum_{h\in(0,H)}^{H-1} -D_g^{h,*}\left(1-D_p^{h,*}\right)^\gamma \log\left(1-D_p^{h,*}\right), \tag{4}$$

In this equation, Den_p represents the predicted density map and Den_g denotes the generated density map. The multi-level loss of our work is made up with the three loss functions as follows:

$$L_{total} = \lambda_h L_h + \lambda_d L_d + \lambda_s L_s. \tag{5}$$

In this work, $\left(\lambda_h, \lambda_d, \lambda_s\right)$ are set as (1,1,10) to keep the balance among them, since the weight on segmentation is always lower than others.

4. EXPERIMENTS

4.1. Evaluation Metrics

We use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to evaluate the gap between the ground-truth counts and the estimated counts as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|D_i - \hat{D}_i|, \tag{6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(D_i - \hat{D}_i\right)^2}, \tag{7}$$

where N denotes the number of images in a testing dataset, D_i and $\hat{D}_i$ are ground-truth and estimated counting numbers of the i_{th} image, respectively.

4.2. Experiments Settings

4.2.1. Utilization

During the utilization of the input images, we normalize the pixel values of images by dividing the whole image and segmentation map by 255 and multiplying the density map by 100. Then, we randomly crop them into 128×128 patches for data augmentation in training phase.

4.2.2. Training Details

We carry out our method with TensorFlow 1.14.0 on Ubuntu with a GTX2080Ti GPU. The total training epoch is set as 1000 and our network is trained with Adam Optimizer. We set β_1 to 0.9 and β_2 to 0.999 as default. The initial learning rate is set to 0.0001. We set 16 as the batch size of our research and 1e-4 as learning rate. We set a stepwise β based on d_a for kernel estimation. When $d_a \leq 0.3d$, we propose β equals to 0.5; while when $d_a \geq 0.7d$, β is set to 0.1

Table 2
Comparison of different results on ShanghaiTech dataset,UCF_CC_50 and UCF_QNRF

Method	PartA		PartB		UCF_CC_50		UCF_QNRF	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN Zhang et al. (2016)	110.2	173.2	26.4	41.3	377.6	509.1	277.0	426.0
Switching-CNNSam et al. (2017)	90.4	135.0	21.6	33.4	318.1	439.2	-	-
CP-CNN Sindagi and Patel (2017a)	73.6	106.4	20.1	30.1	295.8	**320.9**	-	-
ACSCP Shen et al. (2018)	75.7	102.7	17.2	27.4	291.0	404.6	-	-
CSRNet Li, Zhang and Chen (2018b)	68.2	115.0	10.6	16.0	-	-	-	-
SANet Cao, Wang, Zhao and Su (2018)	67.0	104.5	8.4	13.6	258.4	334.9	-	-
CFF Shi, Mettes and Snoek (2019)	65.2	109.4	7.2	12.2	266.1	397.5	93.8	146.5
TEDNet Jiang, Xiao, Zhang, Zhen, Cao, Doermann and Shao (2019)	64.2	109.1	8.2	12.8	**249.4**	354.5	113	188
Ours (FFM Ver.2)	67.2	110.7	8.1	12.9	283.2	386.1	95.4	150.1
Ours (FFM Ver.1)	**63.2**	**102.5**	**7.7**	**10.2**	283.2	350.5	**87.8**	**132.1**

resulting from the sparse crowd and 0.3 in other cases. Note that, in most of other works, β is just a constant parameter, which is fixed to 0.3. Furthermore, in order to examine the effect of the hyper parameter N, we set N to 1,2,4 and 8 to verify our opinion.

The development board processor uses Exynos 4412 processor with ARM Cortex-A9 kernel, equipped with 1G of memory, 4GB of EMMC storage, and runs in the embedded Linux system. Other hardware modules of the system include a USB camera, a 7-inch capacitive screen and a WiFi module. After joining the cloud platform, the system can automatically track densely populated areas, expand the scope of monitoring and improve the quality of monitoring.

4.2.3. Datasets

In our research, four benchmark datasets including ShanghaiTech, UCF_CC_50 and UCF_QNRF are used to evaluate the performance of our DFNet on crowd counting. The details of these datasets are presented as follows.

- **ShanghaiTech** dataset Zhang et al. (2016) consists of 1,198 images taken from streets in Shanghai and 330,165 people. The whole dataset is divided into two parts, *i.e.*, Part A and Part B, depending on the density and range of images. Crowds in images of ShanghaiTech Part A are much more congested than those in ShanghaiTech Part B. Part A has 482 images and Part B has 716 images. Each part is divided into training and testing subsets.

- **UCF_CC_50** Idrees, Saleemi, Seibert and Shah (2013) consists of 50 images with different perspectives and huge density diversity. The numbers of annotation points in each image range from 94 to 4,543 with an average of 1280 people, which makes the dataset a difficult one.

- **UCF_QNRF** Idrees, Tayyab, Athrey, Zhang, Al-Máadeed, Rajpoot and Shah (2018) is one of the most challenging datasets considering that it consists of 1535 high-resolution images with extremely congested crowds ranging from 49 to 12,865 people in each image.

- **Our captured data**. We set up cameras in the hallway of a teaching building on the Campus of Shanghai Jiao Tong University, and recorded 12 video clips, each one contains 2 minutes clip. Students and teachers are coming in and out of the building, sometimes carrying a schoolbag, sometimes in groups. This dataset is made up of 28800 frames, and a cycle is set to 300 frames.

5. RESULTS AND ANALYSIS

In this section, we analyze the effect of our proposed Feature Fusion Module and division mechanism in DAK. We also perform ablation experiments on ShanghaiTech Part A dataset to compare the classical geometry-adaptive kernel and the Density-Adaptive Kernel proposed in this work. In this paper, we compare our approach with the-state-of-art algorithms to demonstrate that our solution achieves remarkable improvement over the previous works. Our method achieves a good effect on all benchmarks and as we can see in Fig.6, several typical examples in Shanghai TechA are showed to confirm the effectiveness of our work. Fig.6 depicts that our method is capable of predicting density maps in spite of scale variation, complicated scenes and variable noises.

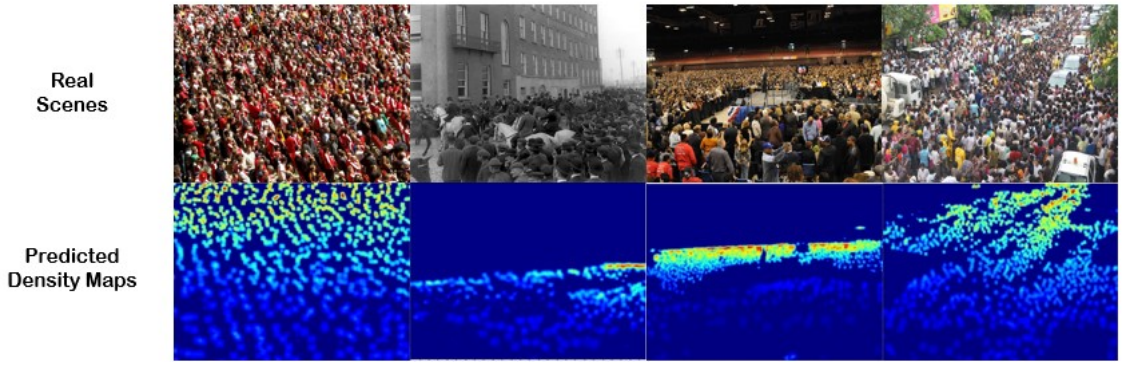


Figure 7: Experiment Results on real scenes in Shanghai Tech

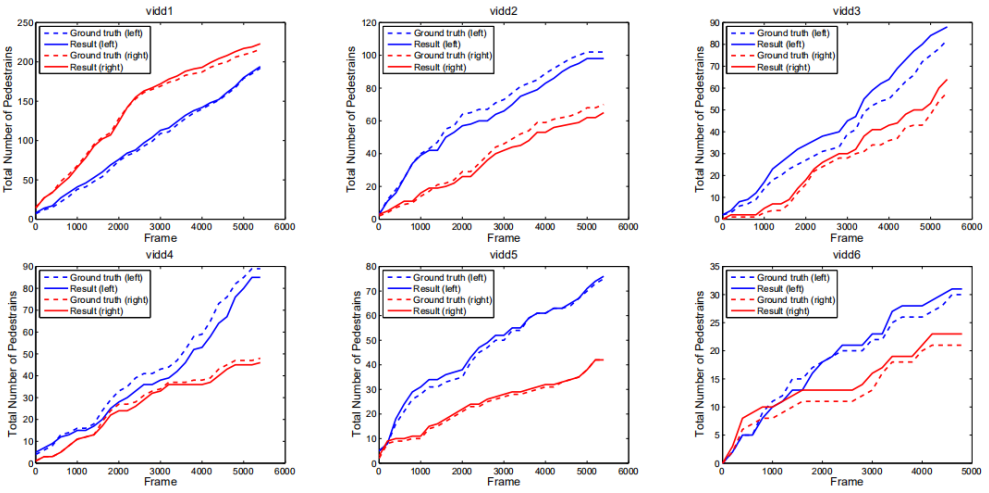


Figure 8: Comparison on our constructed dataset

5.1. Performance Evaluation

Firstly, we analyse the performance improvement brought be our proposed DFNet on the UCF_CC_50 dataset as illustrated in Table 3. Specifically, we perform 5-fold cross-validation as recommended in Idrees et al. (2013) to evaluate the effect of our method. As shown in the table, with our proposed approach, MAE is reduced to 251.3 and RMSE is reduced to 350.5 from those achieved by other recent approaches.

On ShanghaiTech Part A dataset, we have also achieved consistent improvement over other existing works, as shown in Table 3. From this table, it can be seen that, our method outperforms most of the recent works on ShanghaiTech Part B in both MAE and RMSE. In particularly, compared with CSRNet Li et al. (2018b), our method improves the MAE performance by 7.3% and RMSE by 10.9%. Also, our performance is 3.0% better in MAE and 6.3% in RMSE compared to CFF Shi et al. (2019).

5.2. Ablation Study

5.2.1. Feature Fusion Module

We further conduct experiments to evaluate the performance improvement offered by the FFM module, which is shown in Table 3.

As it can be seen from this table, without the FFM, our base network is able to reach 65.4 in MAE and 110.0 in RMSE on ShanghaiTech Part A. With FFM, the MAE and RMSE are reduced to 63.2 (4.4% improvement) and 102.5 (8.3% improvement). This demonstrate the significant ability of our proposed FFM to fuse multilevel features and improve the robustness of the network. However, FFM fails to improve the performance of our network on

Table 3
The impact of FFM on MAE

Dataset	Without FFM	With FFM
ShanghaiTech Part A	65.4	**63.2**
ShanghaiTech Part B	**7.7**	7.8
UCF_CC_50	305.2	**251.3**

Table 4
Comparison of GAK and DAK on ShanghaiTech Part A dataset

Methods	GAK	DAK
MCNN Zhang et al. (2016)	110.2	104.3
CSRNet Li et al. (2018b)	68.2	66.2
Ours	68.7	**63.2**

Table 5
Comparison of different N on MAE

	N			
Dataset	1	2	4	8
ShanghaiTech Part A	67.6	67.4	63.2	66.8
ShanghaiTech Part B	11.2	10.9	7.7	9.8

ShanghaiTech Part B, which may owe to the sparseness of crowds in Part B compared with the congested scenes in Part A. As we can see, FFM performs impressively well in UCF_CC_50 as well. Apparently, with the ability of reconstructing and fusing multi-scale features, FFM performs better in more congested images.

As we mentioned above, we proposed two versions of FFM. In order to test the effect of these two versions, we set experiment respectively based them. And the experiment results are showed in the last two columns of Table 2. And the result depicts that the Version 2 performs better almost on every benchmark.

5.3. Density-Adaptive Kernel

We also make a comparison between the Density-Adaptive Kernel and the traditional Geometry-Adaptive Kernel on ShanghaiTech Part A dataset. As we can see from Table 4 with taking both density and annotation allocations into consideration, MAE apparently is reduced compared with GAK-based methods. This may be due to the changeable radius of the kernel, which suits the variation of feature scales.

In this work, we propose a dividing mechanism to divide density maps into $N \times N$ subregions to focus on different parts with various density. To understand the impact of this dividing mechanism, in our experiment, N is set to 2,4 and 8 to examine the effect of dividing. We also set up a situation that N equals to 1, which means no dividing is employed on density maps. The experiment results are shown in Table 5. As we can see, our network performs best when N is set to 4.

6. CONCLUSION

In our work, we propose density-adaptive Gaussian kernel to generation high-quality density maps. And we construct a module called Feature Fusion Module to address the scale variation problem. And based on the model we constructed, we proposed a context-aware IoT system to solve all kinds of problems in urban intelligent security.

Although the advent of Artificial Intelligence of Things has spawned a large number of new technologies and applications, the fusion of IoT and AI also poses several emerging research challenges. In this paper, our constructed system contains a variety of practical functions, such as dense crowd quantity detection, crowd warning, visualization of statistical results, preservation of statistical results, automatic crowd tracking and other functions. What's more important is that the system puts the parts with large amount of computation and complex computation on the server side, which coincides with the idea of intelligent interconnection and realizes the intelligent control on the embedded device side. However, the crowd counting system is still far from perfect. Computational efficiency is essential in this distributed system. In future work, we are planning to focus on the following issues: 1) utilizing the developed system

for analyzing crowd flow in a continuous mode from a video sequence. 2) focusing on the optimization of distributing computing based on AI to better elevate the performance of IoT networks. 3) elevating its ability to discriminate the noises in images and analyzing the crowds in more crowded scenes. 4) improving the existing crowd counting algorithm continually.

With the construction of 5G network, the super bandwidth, ultra-low latency and ultra-high rate of 5G network are used for video streaming data transmission, which can make data processing on the cloud server as convenient and fast as local data processing. The cloud server will then feedback the data processing results to the embedded system. It applies intelligence to the edge and gives devices the ability to understand the data, observe the environment around them, and decide what to do best.

7. ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NO. 61702226); the 111 Project (B12018); the Natural Science Foundation of Jiangsu Province (NO. BK20170200); the Fundamental Research Funds for the Central Universities (NO. JUSRP11854,NO. JUSRP11851).

CRediT authorship contribution statement

Tao Zhang: Investigation, Methodology, Writing- original draft, Supervision. **Yan Zhao:** Writing- Reviewing and Editing, Methodology, Visualization. **Wenjing Jia:** Validation, Visualization, Data curation. **Mu-Yen Chen:** Supervision,Validation.

References

Cao, X., Wang, Z., Zhao, Y., Su, F., 2018. Scale aggregation network for accurate and efficient crowd counting, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 757–773.

Chen, I.R., Guo, J., Bao, F., 2016. Trust management for soa-based iot and its application to service composition. IEEE Transactions on Services Computing 9, 482–495.

Cheng, Y., Jiang, H., Wang, F., Hua, Y., Feng, D., Guo, W., Wu, Y., 2019. Using high-bandwidth networks efficiently for fast graph computation. IEEE Transactions on Parallel and Distributed Systems 30, 1170–1183.

Cheng, Y., Wang, F., Jiang, H., Hua, Y., Feng, D., Zhang, L., Zhou, J., 2018. A communication-reduced and computation-balanced framework for fast graph computation. Frontiers of Computer Science in China 12, 887–907.

Ding, G., Wu, Q., Zhang, L., Lin, Y., Tsiftsis, T.A., Yao, Y.D., 2018. An amateur drone surveillance system based on the cognitive internet of things. IEEE Communications Magazine 56, 29–35.

Gao, J., Wang, Q., Li, X., 2019. Pcc net: Perspective crowd counting via spatial convolutional network. IEEE Transactions on Circuits and Systems for Video Technology , 1–1.

Ghayvat, H., Mukhopadhyay, S., Gui, X., Suryadevara, N.K., 2015. Wsn- and iot-based smart homes and their extension to smart buildings. Sensors 15, 10350–10379.

Guo, L., Shen, H., 2017. Efficient approximation algorithms for the bounded flexible scheduling problem in clouds. IEEE Transactions on Parallel and Distributed Systems 28, 3511–3520.

Hu, L., Ni, Q., 2018. Iot-driven automated object detection algorithm for urban surveillance systems in smart cities. IEEE Internet of Things Journal 5, 747–754.

Huang, Z., Su, X., Zhang, Y., Shi, C., Zhang, H., Xie, L., 2017. A decentralized solution for iot data trusted exchange based-on blockchain, in: 2017 3rd IEEE International Conference on Computer and Communications (ICCC).

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.

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Máadeed, S., Rajpoot, N.M., Shah, M., 2018. Composition loss for counting, density map estimation and localization in dense crowds, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 544–559.

Jiang, X., Xiao, Z., Zhang, B., Zhen, X., Cao, X., Doermann, D., Shao, L., 2019. Crowd counting and density estimation by trellis encoder-decoder networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6133–6142.

Kang, D., Ma, Z., Chan, A.B., 2017. Beyond counting: Comparisons of density maps for crowd analysis tasks - counting, detection, and tracking. IEEE Transactions on Circuits Systems for Video Technology PP, 1–1.

Kim, H., Ben-Othman, J., 2018. A collision-free surveillance system using smart uavs in multi domain iot. IEEE Communications Letters 22, 2587–2590.

Kumar, S.G., Murugan, A., Muruganantham, B., Sriman, B., 2020. Iot–smart contracts in data trusted exchange supplied chain based on block chain. International Journal of Electrical and Computer Engineering 10, 438–446.

Li, T., Chen, W., Tang, Y., Yan, H., 2018a. A homomorphic network coding signature scheme for multiple sources and its application in iot. Security and Communication Networks 2018, 1–6.

Li, Y., Zhang, X., Chen, D., 2018b. 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.

Liu, C., Weng, X., Mu, Y., 2019. Recurrent attentive zooming for joint crowd counting and precise localization, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1217–1226.

Liu, J., Gao, C., Meng, D., Hauptmann, A.G., 2018a. 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.

Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L., 2019. Crowd counting with deep structured scale integration network .

Liu, X., van de Weijer, J., Bagdanov, A.D., 2018b. Leveraging unlabeled data for crowd counting by learning to rank, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7661–7669.

Liu, X., van de Weijer, J., Bagdanov, A.D., 2019. Exploiting unlabeled data in cnns by self-supervised learning to rank. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 1862–1878.

Luo, A., Yang, F., Li, X., Nie, D., Jiao, Z., Zhou, S., Cheng, H., 2020. Hybrid graph neural networks for crowd counting, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11693–11700.

Niu, Y., Lin, W., Ke, X., 2018. Cf-based optimisation for saliency detection. Iet Computer Vision 12, 365–376.

Novo, O., 2018. Blockchain meets iot: An architecture for scalable access management in iot. IEEE Internet of Things Journal 5, 1184–1195.

Olmschenk, G., Chen, J., Tang, H., Zhu, Z., 2019. Dense crowd counting convolutional neural networks with minimal data using semi-supervised dual-goal generative adversarial networks., in: CVPR Workshops.

Ooro-Rubio, D., Lopez-Sastre, R.J., 2016. Towards perspective-free object counting with deep learning, in: European Conference on Computer Vision (ECCV).

Pazzaglia, P., Mandrioli, C., Maggio, M., Cervin, A., 2019. Dmac: Deadline-miss-aware control, in: Leibniz International Proceedings in Informatics (LIPIcs); 133, pp 1-24 (2019), p. 24.

Ranjan, V., Le, H., Hoai, M., 2018. Iterative crowd counting, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 270–285.

Ryan, D., Denman, S., Sridharan, S., Fookes, C., 2015. An evaluation of crowd counting methods, features and regression models. Computer Vision Image Understanding 130.

Sajid, U., 2019. ZiZoNet: A Zoom-In and Zoom-Out Mechanism for Crowd Counting in Static Images. Ph.D. thesis. University of Kansas.

Sajid, U., Sajid, H., Wang, H., Wang, G., 2020. Zoomcount: A zooming mechanism for crowd counting in static images. IEEE Transactions on Circuits and Systems for Video Technology 30, 3499–3512.

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.

Sam, D.B., Surya, S., Babu, R.V., 2017. Switching convolutional neural network for crowd counting .

Shen, Z., Lee, P.P., Shu, J., Guo, W., 2016. Encoding-aware data placement for efficient degraded reads in xor-coded storage systems, in: 2016 IEEE 35th Symposium on Reliable Distributed Systems (SRDS), pp. 239–248.

Shen, Z., Xu, Y., Ni, B., Wang, M., Yang, X., 2018. Crowd counting via adversarial cross-scale consistency pursuit, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L., 2016. Edge computing: Vision and challenges. IEEE Internet of Things Journal 3, 637–646.

Shi, Z., Mettes, P., Snoek, C., 2019. Counting with focus for free, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4199–4208.

Sindagi, V.A., Patel, V.M., 2017a. Generating high-quality crowd density maps using contextual pyramid cnns .

Sindagi, V.A., Patel, V.M., 2017b. A survey of recent advances in cnn-based single image crowd counting and density estimation. Pattern Recognition Letters .

Wang, C., Zhang, H., Yang, L., Liu, S., Cao, X., 2015. Deep people counting in extremely dense crowds .

Wang, J., Zhang, X.M., Lin, Y., Ge, X., Han, Q.L., 2018. Event-triggered dissipative control for networked stochastic systems under non-uniform sampling. Information Sciences 447, 216–228.

Wang, L., Li, Y., Xue, X., 2019. Coda: Counting objects via scale-aware adversarial density adaption .

Xiong, H., Lu, H., Liu, C., Liu, L., Cao, Z., Shen, C., 2019. From open set to closed set: Counting objects by spatial divide-and-conquer, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8362–8371.

Xu, D., Ouyang, W., Alameda-Pineda, X., Ricci, E., Wang, X., Sebe, N., 2017. Learning deep structured multi-scale features using attention-gated crfs for contour prediction, in: Advances in Neural Information Processing Systems, pp. 3961–3970.

Xu, W., Hammadeh, Z.A., Kroller, A., Ernst, R., Quinton, S., 2015. Improved deadline miss models for real-time systems using typical worst-case analysis, in: 2015 27th Euromicro Conference on Real-Time Systems, pp. 247–256.

Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., Sebe, N., 2020. Reverse perspective network for perspective-aware object counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4374–4383.

Yang, Y., Zheng, X., Tang, C., 2017. Lightweight distributed secure data management system for health internet of things. Journal of Network and Computer Applications 89, 26–37.

Zhang, A., Yue, L., Shen, J., Zhu, F., Zhen, X., Cao, X., Shao, L., 2019. Attentional neural fields for crowd counting, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5713–5722.

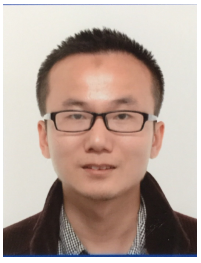
Zhang, T., Jia, W., Gong, C., Sun, J., Song, X., 2017a. Semi-supervised dictionary learning via local sparse constraints for violence detection. Pattern Recognition Letters 107, 98–104.

Zhang, T., Li, J., Jia, W., Sun, J., Yang, H., 2017b. Fast and robust occluded face detection in atm surveillance. Pattern Recognition Letters 107, 33–40.

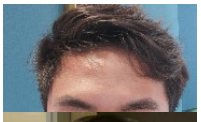
Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y., 2016. Single-image crowd counting via multi-column convolutional neural network, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S., 2015. Conditional random fields as recurrent neural networks, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1529–1537.

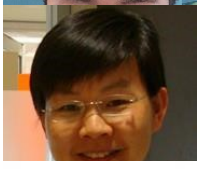
Zhou, J., Hu, L., Wang, F., Lu, H., Zhao, K., 2013. An efficient multidimensional fusion algorithm for iot data based on partitioning. Tsinghua Science Technology 18, 369–378.
Zou, J., Dong, L., Wu, W., 2018. New algorithms for the unbalanced generalised birthday problem. Iet Information Security 12, 527–533.



Tao Zhang received the bachelor's degree from Henan Polytechnic University, Jiaozuo, China, in 2008, and the Ph.D.degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2016.He is currently an associate professor with the Jiangsu Provincial Engineering Laboratory for Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi, China. He has led many research projects (e.g., the National Science Foundation and the National Joint Fund), He has authored over thirty quality journal articles and conference papers. His current research interests include data mining, information systems, wireless network, artificial intelligence, IoT and security, medical data analysis, visual surveillance, scene understanding, behavior analysis, object detection, and pattern analysis.



Yan Zhao received the bachelor's degree from Jilin University, Changchun, China in 2019, and studying for master degree in Jiangnan University, Wuxi, China. His research interests are target detection, crowd counting and artificial intelligence.



Wenjing Jia is currently a Lecturer at the School of Computing and Communications, UTS. She received her PhD degree in Computing Sciences from University of Technology, Sydney (UTS) in 2007. Her research interests are mainly in the areas of image processing/analysis, computer vision and pattern recognition, particularly car license plate detection and text detection.

Dr.Mu-Yen Chen is an Associate Professor of Engineering Science at National Cheng Kung University, Taiwan. He received his PhD in Information Management from National Chiao-Tung University, Taiwan. His current research interests include artificial intelligence, soft computing, bio-inspired computing, data mining, deep learning, context-awareness, and machine learning, with more than 100 publications in these areas. He has served as Editor in Chief and Associate Editor of international journals (e.g. International Journal of Big Data and Analytics in Healthcare, IEEE Access, Applied Soft Computing, Human-centric Computing and Information Sciences, Journal of Information Processing Systems, International Journal of Social and Humanistic Computing) and currently serves an editorial board member on several SCI journals.