# Crowd Counting for Real Monitoring Scene

LI Yiming[1], LI Weihua[2], SHEN Zan[3], NI Bingbing[1]

(1. Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;
2. Video Production Line, ZTE Corporation, Chongqing 401121, China;
3. Institute of Technology, Ping An Technology (Shenzhen) Co., Ltd., Shanghai 200120, China)

DOI: 10.12142/ZTECOM.202002009

http://kns.cnki.net/kcms/detail/34.1294.TN.20191210.1718.002.html, published online December 11, 2019

Manuscript received: 2019-01-17

**Abstract**: Crowd counting is a challenging task in computer vision as realistic scenes are always filled with unfavourable factors such as severe occlusions, perspective distortions and diverse distributions. Recent state-of-the-art methods based on convolutional neural network (CNN) weaken these factors via multi-scale feature fusion or optimal feature selection through a front switch-net. L2 regression is used to regress the density map of the crowd, which is known to lead to an average and blurry result, and affects the accuracy of crowd count and position distribution. To tackle these problems, we take full advantage of the application of generative adversarial networks (GANs) in image generation and propose a novel crowd counting model based on conditional GANs to predict high-quality density maps from crowd images. Furthermore, we innovatively put forward a new regularizer so as to help boost the accuracy of processing extremely crowded scenes. Extensive experiments on four major crowd counting datasets are conducted to demonstrate the better performance of the proposed approach compared with recent state-of-the-art methods.

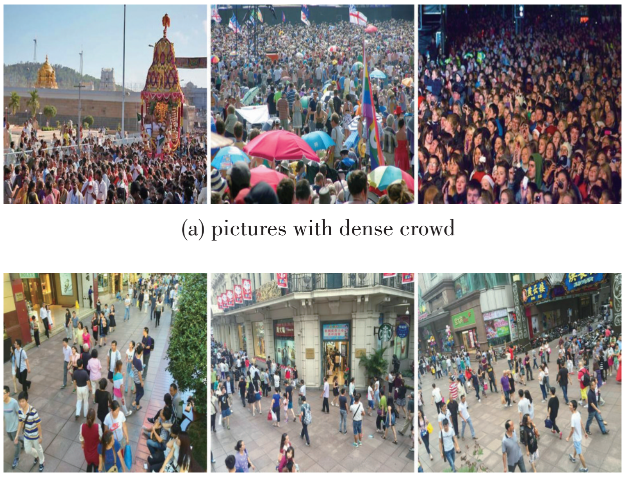
**Keywords**: crowd counting; density; generative adversarial network

**Citation** (IEEE Format): Y. M. Li, W. H. Li, Z. Shen, et al., "Crowd counting for real monitoring scene," *ZTE Communications*, vol. 18, no. 2, pp. 74 – 82, Jun. 2020. doi: 10.12142/ZTECOM.202002009.

## 1 Introduction

With the population density of major cities increasing in recent years, crowd scene analysis has already become an important safety index in the field of video surveillance, especially the crowd count and high-quality density map which has a wide range of applications in public safety, traffic monitoring, scene understanding and flow monitoring. However, predicting accurate crowd count while ensuring high-quality density map is a really challenging task, because complex crowd scenes are always accompanied with severe occlusions, perspective distortions and diverse distributions, and also put forward a great challenge to the algorithm model. Several typical still crowd images from the ShanghaiTech dataset[1] are shown in **Fig. 1**.

In order to solve these problems in computer vision field, a great many algorithms have been proposed, which can be mainly divided into two categories, namely, the hand-crafted feature based regression and the convolutional neural network (CNN) based regression. Recent works[1 - 3] indicate that the CNN based regression has a more excellent performance. Such methods obtain the number of people from a still image by mapping the image to its density map through a CNN architecture. They have achieved significant improvements on


(a) pictures with dense crowd

(b) pictures with relatively sparse crowd

▲**Figure 1. Examples of crowd scene from the ShanghaiTech dataset[1].**

**This work was supported by ZTE Industry-University-Institute Cooperation Funds.**

count estimates, whereas the quality of their estimated density map is unfortunately poor due to the throng scene and self-defect of Euclidean loss.

In the past two years, generative adversarial networks (GANs)[4] have become the most popular frameworks in all relevant fields of image generation. Some of its derivatives such as conditional GANs (cGANs)[5] and information maximizing generative adversarial nets (InfoGANs)[6] can generate extremely realistic images. Therefore, the key point is whether we can draw the advantages of GANs to generate high-quality and high-resolution density maps. Inspired by this, we propose a novel crowd counting model based on cGANs called Crowd Counting Network for Real Monitoring Scene.

The initial inspiration of Crowd Counting Network for Real Monitoring Scene derives from Ref. [7] which uses cGANs to realize pixel-to-pixel translation. Usually, most existing CNN-based approaches on crowd counting add several max-pooling layers in their networks, forcing them to regress on down-sampled density maps, and traditional Euclidean loss is employed to optimize their network parameters which will eventually lead to a relatively blurry result. While in our proposed approach, the generator of cGANs is designed to generate density maps having the same size as input images through a U-net[8] structure with the same amount of convolutional and deconvolutional layers. In other words, it executes a pixel-wise translation from a crowd image to its estimated density map. Thanks to the combination of pixel-wise Euclidean loss, perceptual loss, inter-frame loss and the adversarial training loss provided by GANs, the density map predicted by the generator overcomes blurry results obtained by optimizing only over Euclidean loss and achieves higher quality than that of the previous methods. Besides, we innovatively propose a novel regularizer which provides a very strong regularization constraint on the consistency of parent-child-relationship density maps between different scales to excavate multi-scale consistent information. Unlike using different sizes of filters to extract multi-scale features, we care more about local and overall interrelation between adjacent image patches.

Contributions of this paper are summarized as follows.

- We propose a novel crowd counting framework based on cGANs, called Crowd Counting Network for Real Monitoring Scene. It implements end-to-end training. The use of adversarial training loss helps generate high-quality crowd density map.
- A novel regularizer is introduced to help solve perspective distortions and diverse distributions problems in crowd scenes by providing a very strong constraint on the consistency of parent-child-relationship patches to excavate multi-scale consistent information.
- An inter-frame loss is denoted for the crowd counting in video stream, which can improve the continuity of detection by constraining the number of people calculated by density map between adjacent frames. The loss can also enhance the stability of the network in predicting the density map of video information.
- Our method obtains state-of-the-art performance on four major crowd counting datasets involving the ShanghaiTech dataset, WorldExpo'10 dataset, UCF_CC_50 dataset and UCSD dataset.

## 2 Related Work

A large number of algorithms have been proposed to tackle crowd counting task in computer vision. Early works estimate the number of pedestrians via head or body detection[9 – 11]. Such detection based methods are limited by severe occlusions in extremely dense crowd scenes. Methods in Refs. [12 – 15] use regressors trained with low-level features to predict global counts, and Ref. [16] makes a fusion of hand-crafted features from multiple sources, including the histogram of oriented gradients (HOG), scale-invariant feature transform (SIFT), Fourier analysis, and detections. These methods cannot provide the distribution of crowd, and such low-level features are outperformed by features extracted from CNN which have better and deeper representations.

Several works focus on crowd counting in videos by trajectory-clustering. RABAUD et al.[17] utilized a highly parallelized version of the Kanade-Lucas-Tomasi Tracking (KLT) tracker to extract a set of feature trajectories from videos. Fragmentation of trajectories is restrained by conditioning the trajectories spatially and temporally. BROSTOW et al.[18] proposed an unsupervised data driven Bayesian clustering algorithm, which uses space-time proximity and trajectory for clustering. However, such tracking based methods are limited in crowd counting from arbitrary still image for lack of temporal information.

In recent years, crowd counting has entered the era of CNN. WANG et al.[19] trained a classic Alexnet style CNN model to predict crowd counts. Regrettably, this model has limitation in crowd analysis as it does not provide the estimation of crowd distribution. ZHANG et al.[3] proposed a deep convolutional neural network for crowd counting which is alternatively regressed with two related learning objectives: the crowd count and the density map. Such switchable objective-learning helps improve the performance of both objectives. However, the application of this method is limited as it requires perspective maps which are not easily available in practice during the process of both training and testing. Multi-column CNN is employed by Refs. [1] and [20]. Different CNN columns with varied receptive fields are designed to capture scale variations and perspectives, and then features from these columns are fused together by a 1×1 convolutional layer to regress crowd density. Switch-CNN[2] based on the multi-column convolutional neural network (MCNN)[1] is a patch-based switching architecture before the crowd patches go into multi-column regressors. The switch-net is trained as a classifier to choose the most appropriate regressor for a particular input patch, which takes advantage of patch-wise variations in density within a single image. These methods have made great contributions to the progress of crowd counting by deep learning; at the same time, they add max-pooling layers in their networks and use L2 loss to optimize the whole model. Namely, they pay more attention to the accuracy of predicted crowd count, and neglect the quality of the regressed density map. The latest proposed contextual pyramid CNN (CP-CNN)[21] is a contextual Pyramid CNNs for incorporating global and local contexts which are obtained by learning various density levels. This contextual information is fused with high dimensional feature maps extracted from a multi-column CNN[1] by a fusion-CNN consisting of a set of convolutional and fractionally-strided layers. Adversarial loss is used to help generate high-quality density maps in the last fusion-CNN. Up to now, this approach acquires the lowest counting error on three major crowd datasets in addition to generating high-quality density maps.

The above methods utilize multi-scale features fusion or optimum feature selection to deal with crowd in varied scales, but to some extent they only consider crowd in different scales having different sensitivities to diverse convolutional kernel, which is a relatively local consideration. The latest one incorporates contextual information by classifying images or patches into five density levels independently, while ignoring the correlation between adjacent patches. In other words, none of them research on the statistical consistency of the crowd counts in multi-scale joint patches; for example, a patch is supposed to be equally divided into four sub-patches and the estimated crowd count of the patch ought to be equal to the sum of the estimated crowd counts of these four sub-patches. Such multi-scale consistency offers an effective and strong regularization constraint for crowd count and density estimation. Unfortunately, these methods do not take it into consideration.

## 3 Our Approach

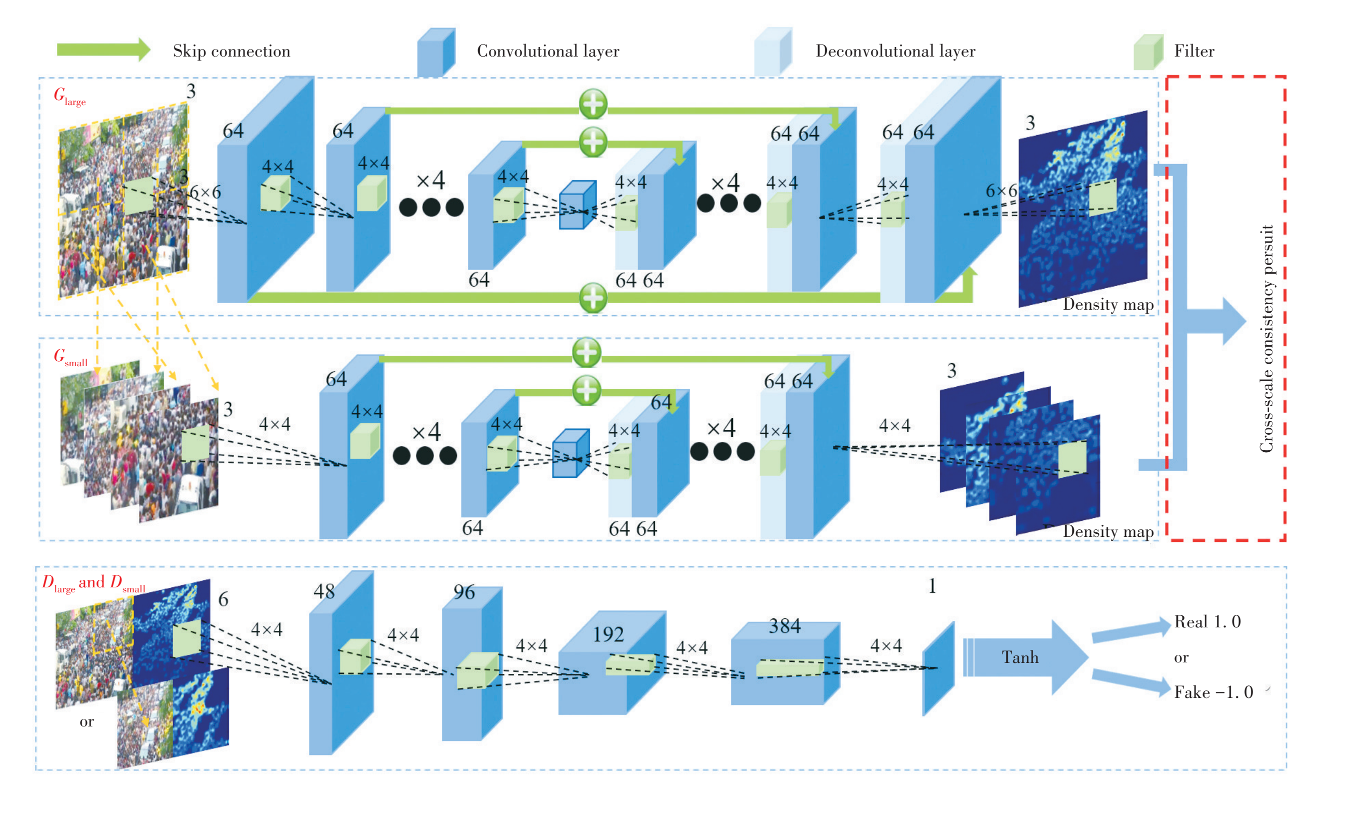
We proposed a novel GANs-based crowd counting framework called Real Monitoring Scene Network (RMSN) for Crowd Counting. Many of the previous state-of-the-art methods[1–2] choose L2 loss to regress density map, which is widely acknowledged to result in low-quality and blurry results especially for image reconstruction tasks[7], [22]. To overcome this flaw and generate high-quality and high-resolution density maps, we design a weighted combination of loss including: adversarial training loss, perceptual loss and pixel-wise Euclidean loss, and a new regularizer is proposed in our GANs-based model to excavate multi-scale consistent information. After generating the density map, we will get the density matrix information between −1 and 1 and then normalize it. The count number from density map can be obtained by summing the normalized matrix divided by a certain coefficient 0.12.

### 3.1 Architecture

RMSN is based on the idea of pixel-to-pixel translation, and in order to leverage the proposed regularizer, our network architecture consists of two complementary conditional GANs: $\mathrm{GAN}_{\mathrm{large}}$ and $\mathrm{GAN}_{\mathrm{small}}$. A classic GAN architecture usually contains two models: a generator $G$ trained to produce outputs and a discriminator $D$ trained to distinguish the real target and fake outputs from $G$. In our method, the generator $G$ learns an end-to-end mapping from input crowd image to its density map. **Fig. 2** shows the integral architecture of RMSN. The general structures of the two GANs are quite similar. Specific details are discussed below.

A general problem of pixel-to-pixel translation is the difficulty to efficiently map a high resolution input image to a high resolution output image. Fortunately, previous works[7], [23–24] have provided an excellent solution by using an encoder-decoder network[25]. In RMSN, a U-net[8] structure is introduced to the generator $G$ as an encoder-decoder. Let us start with the large GANs in our proposed architecture. In the generator $G_{\mathrm{large}}$, eight convolutional layers along with batch normalization layers and LeakyReLU activation layers are stacked in the encoder part which serves as a feature extractor. Then, eight de-convolutional layers along with batch normalization layers and ReLU activation layers (except for the last one) are added in the decoder part, followed by a tanh function. Note that the de-convolutional layers are a mirrored version of the foregone convolutional layers. We set a stride of 2 in all layers, which means convolutions in the encoder down-sample by a factor of 2, whereas deconvolutions upsample by a factor of 2. In addition, three dropout layers are added behind the first three de-convolutional layers with dropout ratio set to 0.5 in order to alleviate over-fitting. Skip connections are also added between mirror-symmetry convolutional and de-convolutional layers to help improve the performance and efficiency, similar to Ref. [7]. The architecture of $G_{\mathrm{large}}$ can be depicted as: C(64,6)-C(64,4)-C(64,4)-C(64,4)-C(64,4)-C(64,4)-C(64,4)-C(64,4)-DCD(64,4)-DCD(64,4)-DCD(64,4)-DC(64,4)-DC(64,4)-DC(64,4)-DC(64,4)-DC(3,6)-Tanh, where C is a Conv-BN-LReLU layer, DCD is a deConv-BN-Dropout-ReLU layer, DC is a deConv-BN-ReLU layer and the first number in every parenthesis represents the number of filters while the second number represents filter size.

The generator $G_{\mathrm{small}}$ which is similar to $G_{\mathrm{large}}$ contains 7 convolutional layers and 7 deconvolutional layers. 4 × 4 filters are used in all layers with a stride of 2. The architecture of generator $G_{\mathrm{small}}$ can be depicted as: C(64,4)-C(64,4)-C(64,4)-C(64,4)-C(64,4)-C(64,4)-C(64,4)-DCD(64,4)-DCD(64,4)-DC(64,4)-DC(64,4)-DC(64,4)-DC(64,4)-DC(3,4)-Tanh. The inputs of generator $G_{\mathrm{large}}$ are 240×240×3 sized crowd patches, and the inputs of generator $G_{\mathrm{small}}$ are 120×120×3 sized crowd patches equationally cropped from the input of the generator $G_{\mathrm{large}}$ without overlapping, as shown in the upper left corner of Fig. 2 Their outputs are of the same size as their inputs. That means the



▲**Figure 2. Architecture of the proposed Crowd Counting Network for Real Monitoring Scene (RMSN): The top level is the structure of generator $G_{large}$, the middle part is the structure of generator $G_{small}$, and the bottom part is the discriminators $D_{large}$ and $D_{small}$ that have the same structure.**

density maps generated from our RMSN contain more details and have better characterization capabilities than previous density-map-based works[1 - 3] as their density maps are always much smaller than the origin images.

The discriminators $D_{large}$ and $D_{small}$ have the same structure, displayed at the bottom of Fig. 2. Five convolutional layers along with batch normalization layers and LeakyReLU activation layers (except for the last one) act as a feature extractor. A tanh function is stacked at the end of these convolutional layers to regress a probabilistic score ranges from −1.0 to 1.0. The architecture of discriminators $D_{large}$ and $D_{small}$ can be depicted as: C(48,4)-C(96,4)-C(192,4)-C(384,4)-C(1,4)-Tanh. The inputs of the discriminators $D_{large}$ and $D_{small}$ are 240×240×6 and 120×120×6 sized concatenated pairs of crowd patch and density map, respectively. The values of the output matrix indicate whether the input is real (close to 1.0) or fake (close to −1.0).

### 3.2 Loss Function

In our problem, motivated by recent success of GANs, we propose an adversarial loss of generating crowd density map from image patch. The adversarial loss involves a discriminator $D$ and a generator $G$ playing a two-player minimax game: $D$ is trained to distinguish synthetic images from ground truth while $G$ is trained to generate images to fool $D$. The adversarial loss is denoted as:

$$L_A(G,D) = \mathbb{E}_{x,y\sim \mathrm{Pdata}(x,y)}\left[\log D(x,y)\right] + \mathbb{E}_{x\sim \mathrm{Pdata}(x)}\left[\log\left(1 - D(x,G(x))\right)\right], \tag{1}$$

where $x$ denotes a training patch and $y$ denotes corresponding ground-truth density map. $G$ tries to minimize this objective, whereas $D$ tries to maximize it.

Due to the lack of direct constraint from ground truth, just using an adversarial loss may sometimes lead to aberrant spatial structure. Thus, we include two conventional losses to smooth and improve the solution, which is denoted as follows.

In our problem, $l_2$ loss $L_E(G)$ can force the generated estimated density map to fool $D$ and be close to the ground truth in an L2 sense.

$$L_E(G) = \frac{1}{C}\sum_{c=1}^{C}\left\|p^G(c) - p^{GT}(c)\right\|_2^2, \tag{2}$$

where $p^G(c)$ represents the pixels in generated density map and $p^{GT}(c)$ represents the pixels in ground-truth density map, with $c$=3.

Perceptual loss is first introduced by JOHNSON et al.[24] for image transformation and super resolution task. By minimizing the perceptual differences between the two images, the synthetic image can be more semantically similar to the objective image. The perceptual loss is defined as:

$$L_P(G) = \frac{1}{C}\sum_{c=1}^{C}\left\|f^G(c) - f^{GT}(c)\right\|_2^2, \tag{3}$$

where $f^G(c)$ represents the pixels in high level perceptual features of generated density map and $f^{GT}(c)$ represents the pixels in high level perceptual features of ground-truth density map, with $c$=128.

Therefore, the integrated loss is expressed as:

$$L_{\mathrm{I}} = \arg\min_G \max_D L_A(G,D) + \lambda_e L_E(G) + \lambda_p L_P(G), \tag{4}$$

where $\lambda_e$ and $\lambda_p$ are predefined weights for Euclidean loss and perceptual loss. Suggested by previous works[26], we set $\lambda_e = \lambda_p = 150$.

In our problem, we propose a new inter-frame loss for the prediction in video stream, which can improve the continuity of detection by constraining the number of people between adjacent frames and enhance the stability of the network in predicting the density map of video information. The loss is defined as the distance between two adjacent frames of generated density maps, which is denoted as:

$$L_i(G) = \frac{1}{N_{\mathrm{pix}}}\left\|n^G(c) - n^{*G}(c)\right\|_2^2, \tag{5}$$

where $N_{\mathrm{pix}}$ represents the whole numbers of pixels in generated density maps, $n^G(c)$ represents the number of pedestrians calculated from the current frame in generated density map, and $n^{*G}(c)$ represents the number of pedestrians calculated from the previous frame.

Therefore, for video stream information, the integrated loss $L_{\mathrm{I}}$ should be denoted as:

$$L_{\mathrm{I}} = \arg\min_G \max_D L_A(G,D) + \lambda_e L_E(G) + \lambda_p L_P(G) + \lambda_i L_i(G), \tag{6}$$

where $\lambda_i$=150 is predefined weights for inter-frame loss.

To restrain the cross-scale consistency of parent-child-relationship density maps, we propose a Cross-Scale Consistency Pursuit loss[27] defined as the discrepancy/distance between $P_{\mathrm{concat}}$ and $P_{\mathrm{parent}}$. The CSCP loss of a W×H density map with channels is defined as:

$$L_C(G) = \frac{1}{C}\sum_{c=1}^{C}\left\|p^{\mathrm{prt}}(c) - p^{\mathrm{cnt}}(c)\right\|_2^2, \tag{7}$$

where $p^{\mathrm{prt}}(c)$ represents the pixels in density map $P_{\mathrm{parent}}$ and $p^{\mathrm{cnt}}(c)$ represents the pixels in density map $P_{\mathrm{concat}}$, with $c$=3.

As pointed out above, the four loss functions are weightedly combined to a final objective,

$$L_{\mathrm{II}} = L_{\mathrm{I}} + \lambda_c L_C(G), \tag{8}$$

where $\lambda_c$=10 is the predefined weight for cross-scale consistency pursuit loss.

### 3.3 Training Details

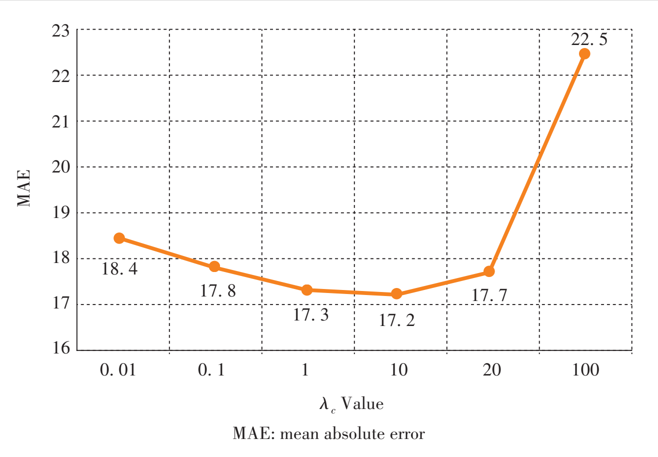
During training, the input is an image pair consisting of a crowd patch and its corresponding density map. Such an image pair is first input to the large-scale subnet $G_{\mathrm{large}}$, and then evenly divided into four equidistant image pairs without overlapping and finally input to the small-scale subnet $G_{\mathrm{small}}$. Both subnets are jointly trained. The RMS prop optimizer has a learning rate set to 0.00005 and is used to update the parameters of the network. We follow the update rule: in each iteration, $G_{\mathrm{small}}$'s four updates are followed by a $G_{\mathrm{large}}$.

To increase the training data, one of the general methods is to resize the input image pair to a larger size and randomly crop the image pair of a particular size. However, such data increases are not appropriate in our crowd counting tasks because image interpolation algorithms such as recent and bilinear algorithms inevitably change the number of people in the density map. Therefore, in our experiments, we use filled and flipped images to replace image size adjustments with a probability of 50% for data enhancement.

Our model requires approximately 300 periods of training to converge. In order to balance the training of the two sub-networks, in the first 100 periods, the predefined weight $\lambda_c$ in Eq. (6) is set to 0, then it is adjusted to 10 and the training process is continued. Finally, the well-trained generator $G_{\mathrm{large}}$ is used to predict the density map of the test image. Training and testing of the proposed network is implemented on the Torch7 framework.

### 3.4 Parameter $\lambda_c$ Study

We did comparative experiments performed on Part_B of the ShanghaiTech dataset to choose the optimum value of $\lambda_c$. As shown in **Fig. 3**, mean absolute error (MAE) decreases



▲ **Figure 3. Comparisons of MAE for different $\lambda_c$ values on ShanghaiTech Part_B.**

when the value of $\lambda_c$ increases. The lowest MAE value is obtained at $\lambda_c$=10. After that, when the value of $\lambda_c$ increases, the error rises rapidly, because the comparison of the weight of cross-scale consistency loss and $L_1$ loss becomes too significant. Therefore, we finally assign 10 to $\lambda_c$.

## 4 Experiments

We evaluate our method in four major crowd counting datasets, including the ShanghaiTech dataset, WorldExpo'10 dataset, UCF CC 50 dataset and UCSD dataset. Compared with the state-of-the-art methods, our method gains a superior or at least competitive performance in all datasets used for evaluation. Training and testing of the proposed network are implemented on Torch7 framework.

We use MAE and mean squared error (MSE) to evaluate the performance of our method on existing works.

Adversarial pursuit seeks to exploit adversarial loss, perceptual loss and U-net structured generator to improve the quality of generated density maps. It is worth noting that our predicted density map is better distributed than the MCNN population, with less blur and noise. In addition, comparative experiments were performed on the ShanghaiTech[1] and WorldExpo'10[3] datasets in **Table 1** above. It can be observed that training with additional adversarial loss and perceptual loss (i.e. LI) results in far less errors than training with Euclidean loss only.

### 4.1 ShanghaiTech

The ShanghaiTech dataset is created by ZHANG et al.[1], which that consists of 1 198 annotated images. The dataset is divided into two parts. Part A contains 482 images downloaded from the Internet with extremely dense crowd, and Part B contains 716 images taken from the busy street in Shanghai with normal flow of crowd. Our model is trained and tested on the training and testing set split by author respectively. To augment the training data, we resize all the images to 720× 720 and cropped patches from each image. Each patch is 1 size of origin image and is cropped from different locations. Ground-truth density maps are generated by geometry-adaptive Gaussian kernels. At the test time, a window of size 240× 240 slides on the test image to crop patches with 50% overlapping as inputs of the well trained generator. Then, outputs from the generator are integrated to a weight-balanced density map which has the same size of the test image. Finally, the estimated crowd count of the image can be calculated by the sum of the density map. The proposed method is compared with four current state-of-the-art CNN-based approaches: a switchable objective-learning CNN[3], MCNN[1], Switch-CNN[2] and CP-CNN[21]. ZHANG et al.[3] proposed a switchable objective-learning CNN which is alternatively regressed with two related learning objectives: crowd count and density map. This method is highly dependent on the perspective maps during training and testing. ZHANG et al.[1] employed a MCNN to extract multi-scale features and to fuse them to get a better representation. Switch-CNN[2] trained a prepositive switch-net to intelligently choose the optimal regressor instead of multi-column feature fusion. CP-CNN[21] incorporated global and local contextual information with fused multi-column features, and is trained in an end-to-end fashion using a combination of adversarial loss and pixel-level Euclidean loss. From **Table 2** we can see, on Part B of which images are closer to the real monitoring screens, the proposed approach obtains appreciable improvement in contrast to the best model CP-CNN at the time. On Part A, besides CP-CNN, our method has also achieved the best results, compared with the other three ones. In order to fairly evaluate the quality of the generated density map, we choose the same set of test images published in MCNN[1] paper along with ground-truth and predicted density maps, shown in **Fig. 4**. It can be intuitively seen that our predicted density maps conform to the distribution of crowd much better than MCNN's with noticeable blur and noise, which benefits from our GANs-based architecture and new regularizer.

### 4.2 WorldExpo'10 Dataset

The WorldExpo'10 dataset is created by ZHANG et al.[3] with 1 132 annotated video sequences captured by 108 surveillance cameras from Shanghai 2010 World Expo. A total of 199 923 pedestrians in 3 980 frames are labeled at the centers of their heads. In these frames, 3 380 frames are treated as the training set; the rest 600 frames are used as the test set, which are sampled from five different scenes, each containing 120 frames. The pedestrian number in the test scene ranges from 1 – 220. This dataset also provides perspective maps, the value of which represents the number

**▼Table 1. Comparisons of errors for training with different losses**

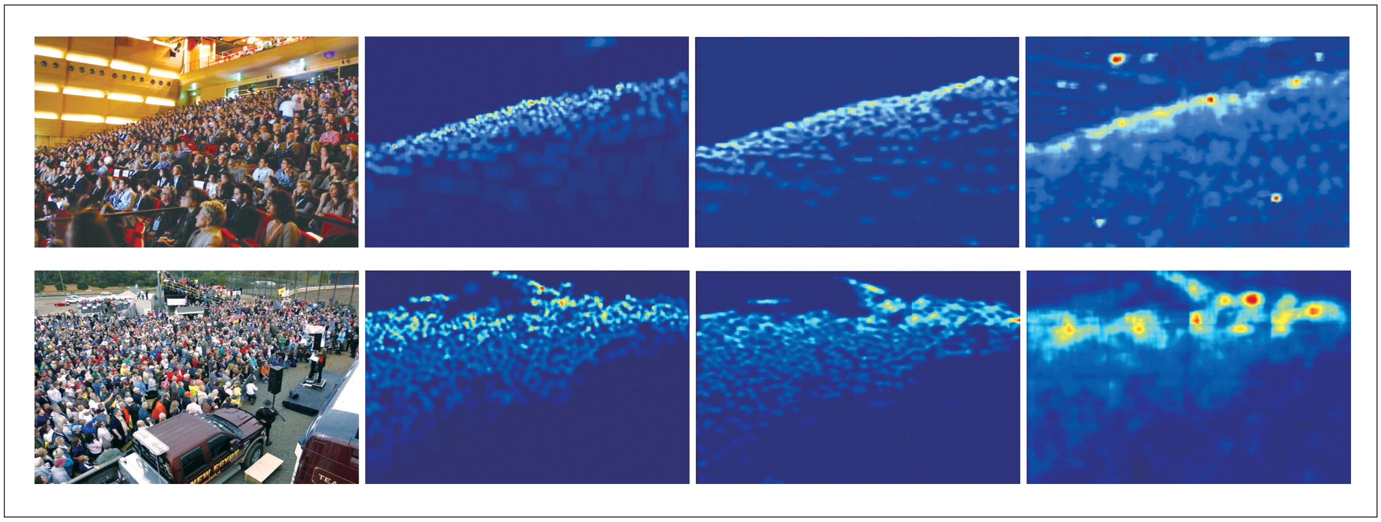
| | Part A | | Part B | | WorldExpo'10 |
|---|---|---|---|---|---|
| Objective | MAE | MSE | MAE | MSE | AMAE |
| $L_E$ | 95.8 | 149.4 | 24.1 | 36.4 | 9.95 |
| $L_I$ | 83.2 | 131.3 | 18.4 | 28.8 | 8.48 |
| $L_{II}$ | **75.7** | **102.7** | **17.2** | **27.4** | **7.5** |

AMAE: average mean absolute error MAE: mean absolute error MSE: mean squared error

**▼Table 2. Comparison of RMSN with other three state-of-the-art CNN-based methods on ShanghaiTech dataset**

| | Part A | | Part B | |
|---|---|---|---|---|
| Methods | MAE | MSE | MAE | MSE |
| The approach in Ref. [3] | 181.8 | 277.7 | 32.0 | 49.8 |
| MCNN[1] | 110.2 | 173.2 | 26.4 | 41.3 |
| Switch-CNN[2] | 90.4 | **135.0** | 21.6 | 33.4 |
| The proposed RMSN | **86.2** | 145.4 | **17.2** | **27.4** |

MAE: mean absolute error
MCNN: multi-column convolutional neural network
MSE: mean squared error
RMSN: real monitoring scene network



**▲Figure 4. Two test images sampled from the ShanghaiTech Part A dataset (From left to right, the four columns successively denote test images, ground-truth density maps, our estimated density maps and the multi-column convolutional neural network (MCNN)'s[1] respectively).**

of pixels in the image covering one square meter at real location. For fair comparison, we choose the crowd density distribution kernel introduced by Ref. [3], which contains two terms: a normalized Gaussian kernel as a head part and a bivariate normalized distribution as a body part, to generate density maps with perspective information. To follow the previous methods, only the crowd in region of interest (ROI) are taken into consideration. So we multiply predicted density map by specifing ROI mask, which means that the area out of ROI is set to zero. MAE is suggested by ZHANG et al.[3] to evaluate the performance of crowd counting model on this dataset.

**Table 3**, in which MAE is used to evaluate the performance on each scene and the average result across scenes, reports the performance of our method on five different test scenes in comparison to other four state-of-the-art methods. Our method refreshes the scores of three scenes: Scene2, Scene3 and Scene5, while achieving comparable performance on the rest two scenes, and outperforms the leader CP-CNN[21] by a margin of 0.41 points in terms of average MAE across scenes.

### 4.3 UCF_CC_50 Dataset

The UCF_CC_50 dataset, which is a very challenging dataset composed of 50 annotated crowd images with a large variance in crowd counts and scenes, is firstly introduced by IDREES et al.[28]. The crowd counts range from 94 to 4 543. We follow Ref. [28] and use 5-fold cross-validation to evaluate the proposed method.

We compare our method with five existing methods on UCF_CC_50 dataset using MAE and MSE as metrics in **Table 4**. IDREES et al.[28] proposed to use multi-source features like head detections, Fourier analysis and texture features. Our approach acquires the best MAE and comparable MSE among existing approaches.

**▼Table 3. Comparison of RMSN with other four state-of-the-art CNN-based methods on the WorldExpo'10 dataset**

| Methods | Scene 1 | Scene 2 | Scene 3 | Scene 4 | Scene 5 | Average |
|---|---|---|---|---|---|---|
| The approach in Ref. [3] | 9.8 | 14.1 | 14.3 | 22.2 | 3.7 | 12.9 |
| MCNN[1] | 3.4 | 20.6 | 12.9 | 13.0 | 8.1 | 11.6 |
| Switch-CNN[2] | 4.4 | 15.7 | 10.0 | 11.0 | 5.9 | 9.4 |
| CP-CNN[21] | **2.9** | 14.7 | 10.5 | **10.4** | 5.8 | 8.9 |
| The proposed RMSN | 4.1 | **14.05** | **9.6** | 11.8 | **2.9** | **8.49** |

CP-CNN: contextual pyramid convolutional neural network
MCNN: multi-column convolutional neural network
RMSN: real monitoring scene network

**▼Table 4. Comparative results on the UCF_CC_50 dataset**

| Methods | MAE | MSE |
|---|---|---|
| The approach in Ref. [28] | 419.5 | 541.6 |
| The approach in Ref. [3] | 467.0 | 498.5 |
| MCNN[1] | 377.6 | 509.1 |
| Switch-CNN[2] | 318.1 | 439.2 |
| CP-CNN[21] | 295.8 | 320.9 |
| The proposed RMSN | **291.0** | **404.6** |

MAE: mean absolute error
MCNN: multi-column convolutional neural network
MSE: mean squared error
RMSN: real monitoring scene network

### 4.4 UCSD Dataset

We also evaluate our method on the single-scene UCSD dataset with video stream. This dataset consists of 2 000 labeled frames with size of 158×238. Ground truth is labeled at the center of every pedestrian and the largest number of people is under 46. The ROI and perspective map are provided as well. In order to cover the pedestrian contour, we choose a bivariate normalized distribution kernel shaped ellipse to generate density maps. We follow the same train-test setting in Ref. [13]. The 800 frames from 601 to 1 400 are treated as training set and the rest 1 200 frames as test set. At the test time, MAE and MSE are used as evaluation metrics.

**Table 5** exhibits the comparison of our method with other state-of-the-art methods on UCSD dataset. Crowd count is calculated within the given ROI. The first two methods[12], [14] adopts hand-crafted features, while the rest three are CNN-based. All their results are relatively close due to the comparatively simple scene with low variation of crowd density. Nevertheless, our method outperforms most of the methods, which shows that our approach is also applicable in relatively sparse and single crowd scene.

**Fig. 5** shows the application of our method under video information from UCSD dataset. In practical applications, we calculate the pedestrian flow and retention based on the density map. In the velocity map, we can see the small arrows around pedestrian area which represents the direction of pedestrian movement. In the retention map, we use the chromatic area of different colors near head to indicate the length of retention of the corresponding pedestrian, based on the residence time of the pedestrian in a certain place.

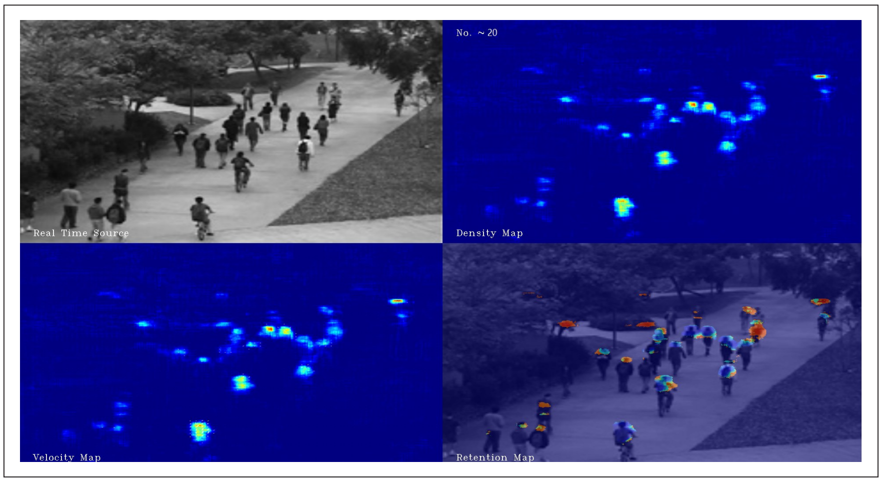
## 5 Conclusions

In this paper, we propose a GANs-based crowd counting network which takes full advantage of excellent performance of GANs in image generation. To better reduce errors caused by different scales of the crowd, we propose a novel regularizer which provides a strong regularization constraint on multi-scale crowd density estimation. Extensive experiments indicate that our method achieves the state-of-the-art performance on major crowd counting datasets used for evaluation.

▼**Table 5. Comparative results on the UCSD dataset**

| Methods | MAE | MSE |
|---|---|---|
| Kernel Ridge Regression[12] | 2.16 | 7.45 |
| Cumulative Attribute Regression[14] | 2.07 | 6.86 |
| The approach in Ref. [3] | 1.60 | 3.31 |
| Switch-CNN[2] | 1.62 | 2.10 |
| The proposed RMSN | **1.47** | **1.98** |

CNN: convolutional neural network
MAE: mean absolute error
MSE: mean squared error
RMSN: real monitoring scene network



▲**Figure 5. One test video information sampled from the UCSD dataset (from left to right and top to bottom, the four images successively denote real time source, density map, velocity map and retention map respectively).**

**References**

[1] ZHANG Y Y, ZHOU D S, CHEN S Q, et al. Single-image crowd counting via multi-column convolutional neural network [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 589 – 597. DOI: 10.1109/cvpr.2016.70

[2] SAM D B, SURYA S, BABU R V. Switching convolutional neural network for crowd counting [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017. DOI:10.1109/cvpr.2017.429

[3] ZHANG C, LI H, WANG X, et al. Cross-scene crowd counting via deep convolutional neural networks [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA, 2015: 833 – 841. DOI: 10.1109/cvpr.2015.7298684

[4] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Advances in neural information processing systems, 2014 (3): 2672 – 2680

[5] MIRZA M, OSINDERO S. Conditional generative adversarial nets [EB/OL]. (2014-11-06) [2018-10-12]. https://arxiv.org/abs/1411.1784

[6] CHEN X, DUAN Y, HOUTHOOFT R, et al. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets [C]//Conference and Workshop on Neural Information Processing Systems. Barcelona, Spain, 2016

[7] ISOLA P, ZHU J-Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks [EB/OL]. (2016-09-21) [2018-10-12]. https://arxiv.org/abs/1611.07004

[8] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany, 2015: 234 – 241. DOI: 10.1007/978-3-319-24574-4_28

[9] LIN Z L, DAVIS L S. Shape-based human detection and segmentation via hierarchical part-template matching [J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 32(4): 604 – 618. DOI: 10.1109/tpami.2009.204

[10] WANG M, WANG X. Automatic adaptation of a generic pedestrian detector to a specific traffic scene [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA, 2011. DOI: 10.1109/cvpr.2011.5995698

[11] WU B, NEVATIA R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors [C]//Tenth IEEE International Conference on Computer Vision (ICCV’05). Beijing, China, 2005: 90 – 97. DOI: 10.1109/iccv.2005.74

[12] AN S J, LIU W Q, VENKATESH S. Face recognition using kernel ridge regression [C]//IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1110 - 1116. DOI: 10.1109/cvpr.2007.383105

[13] CHANA B, LIANG Z-S J, VASCONCELOS N. Privacy preserving crowd monitoring: counting people without people models or tracking [C]//IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1 – 7. DOI: 10.1109/cvpr.2008.4587569

[14] CHEN K, LOY C C, GONG S, et al. Feature mining for localised crowd counting [C]//British Machine Vision Conference. Surrey, UK, 2012. DOI: 10.5244/c.26.21

[15] KONG D, GRAY D, TAO H. A viewpoint invariant approach for crowd counting [C]//International Conference on Pattern Recognition. Hong Kong, China, 2006. DOI: 10.1109/icpr.2006.197

[16] BANSAL A, VENKATESH K S. People counting in high density crowds from still images [EB/OL]. (2015 - 07 - 30) [2018 - 10 - 12]. https://arxiv. org/abs/1507.08445v1

[17] RABAUD V, BELONGIE S J. Counting crowded moving objects [C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). New York, USA, 2006: 705 – 711. DOI: 10.1109/cvpr.2006.92

[18] BROSTOW G J, CIPOLLA R. Unsupervised bayesian detection of independent motion in crowds [C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, USA, 2006: 594 – 601. DOI: 10.1109/cvpr.2006.320

[19] WANG C, ZHANG H, YANG L, et al. Deep people counting in extremely dense crowds [C]//ACM International Conference on Multimedia. Brisbane, Australia, 2015. DOI: 10.1145/2733373.2806337

[20] BOOMINATHAN L, KRUTHIVENTI S S, BABU R V. CrowdNet: a deep convolutional network for dense crowd counting [C]//ACM Conference on Multimedia. Vienna, Austria, 2016. DOI: 10.1145/2964284.2967300

[21] SINDAGI V A, PATEL V M. Generating high-quality crowd density maps using contextual pyramid CNNs [C]//IEEE International Conference on Computer Vision. Venice, Italy, 2017

[22] LI C, WAND M. Precomputed real-time texture synthesis with markovian generative adversarial networks [C]//European Conference on Computer Vision. Amsterdam, Netherland, 2016: 702 – 716. DOI: 10.1007/978-3-319-46487-9_43

[23] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: feature learning by inpainting [C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2536 – 2544. DOI: 10.1109/cvpr.2016.278

[24] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution [C]//European Conference on Computer Vision. Amsterdam, Netherland, 2016: 694 – 711. DOI: 10.1007/978-3-319-46475-6_43

[25] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504 – 507. DOI: 10.1126/science.1127647

[26] ZHANG H, SINDAGI V, PATEL V M. Image de-raining using a conditional generative adversarial network [EB/OL]. (2017-01-21) [2018-10-12]. https://arxiv.org/abs/1701.05957

[27] SHEN Z, XU Y, NI B B, et al. Crowd counting via adversarial cross scale consistency pursuit [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, 2018: 5245 – 5254. DOI: 10.1109/cvpr.2018.00550

[28] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images [C]//IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2547 – 2554. DOI: 10.1109/cvpr.2013.329

## Biographies

**LI Yiming** received the B.S. degree in information engineering from Shanghai Jiao Tong University, China in 2018. From 2018 to the present, he is pursuing his M.S. degree at the Institute of Image Communications and Network Engineering of Shanghai Jiao Tong University. His research interests include crowd counting in dense scenes and the image enhancement, segmentation and texture recognition technologies in materials science.

**LI Weihua** received the B.S. degree in information engineering from Southwest University, China in 1996. He is currently responsible for the VSS product planning at ZTE Corporation.

**SHEN Zan** received the B.S. and M.S. degrees in electronics and information engineering from Shanghai Jiao Tong University, China in 2016 and 2019 respectively. He once participated in the internship of Tencent Youtu Lab in 2018. After graduation, he works at Ping An Technology (Shenzhen) Co, Ltd. His research interests include but not limited to deep learning, computer vision, and machine learning. He has published one technical paper in IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

**NI Bingbing** (nibingbing@sjtu.edu.cn) received the B.S. degree in electronic engineering of Shanghai Jiao Tong University, China in 2005, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore in 2011. He is currently a special researcher, long-hired associate professor and doctoral supervisor at the Department of Electronics of Shanghai Jiao Tong University. His main research areas are computer vision, machine learning and multimedia computing, specializing in face recognition, video understanding, intelligent interactive creative media generation and intelligent medical treatment. He has published more than 100 papers in top international journals and conferences in the field of artificial intelligence/computer vision and more than 40 papers in CVPR, ICCV and other international top computer vision conferences.

⬅ From Page 56

[27] PAN C H, REN H, DENG Y S, et al. Joint blocklength and location optimization for URLLC-enabled UAV relay systems [J]. IEEE communications letters, 2019, 23(3): 498−501. DOI:10.1109/lcomm.2019.2894696

[28] BOYD S, VANDENBERGHE L. Convex optimization [M]. Cambridge, U.K: Cambridge University Press, 2004. DOI:10.1017/cbo9780511804441

[29] BOYD S. Convex optimization II [EB/OL]. (2013-09-11)[2019-10-20]. http://www.stanford.edu/class/ee364b/lectures.html

## Biographies

**ZHANG Pengyu** received the B.E. degree from Guangdong University of Technology, China in 2017. He is pursuing his master degree in the School of Information Engineering, Guangdong University of Technology. His research interests include UAV communications, mobile edge computing, and ultra-reliable and low-latency communications.

**XIE Lifeng** received the B.E. degree from Guangdong University of Technology,China in 2016. He is currently a Ph.D. candidate in the School of Information Engineering, Guangdong University of Technology. His research interests include energy harvesting in wireless communications, wireless information and power transfer, and UAV communications.

**XU Jie** (xujie@cuhk.edu.cn) received the B.E. and Ph.D. degrees from University of Science and Technology of China in 2007 and 2012 respectively. From 2012 to 2014, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore. From 2015 to 2016, he was a post-doctoral Research Fellow with the Engineering Systems and Design Pillar, Singapore University of Technology and Design. From 2016 to 2019, he was a professor with the School of Information Engineering, Guangdong University of Technology, China. He is currently an associate professor with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China. His research interests include energy efficiency and energy harvesting in wireless communications, wireless information and power transfer, UAV communications, and mobile edge computing and learning.