Improving Crowd Density Estimation by Fusing Aerial Images and Radio Signals

KAI-WEI YANG, National Yang Ming Chiao Tung University, Taiwan
YEN-YUN HUANG, National Yang Ming Chiao Tung University, Taiwan
JEN-WEI HUANG, National Yang Ming Chiao Tung University, Taiwan
YA-ROU HSU, National Yang Ming Chiao Tung University, Taiwan
CHANG-LIN WAN, National Yang Ming Chiao Tung University, Taiwan
HONG-HAN SHUAI, National Yang Ming Chiao Tung University, Taiwan
LI-CHUN WANG, National Yang Ming Chiao Tung University, Taiwan
WEN-HUANG CHENG, National Yang Ming Chiao Tung University and National Chung Hsing University, Taiwan

A recent line of research focuses on crowd density estimation from RGB images due to a variety of applications, e.g., surveillance, traffic flow control. However, the performance drops dramatically for low-quality images, such as occlusion, or poor light conditions. On the other hand, people are equipped with various wireless devices, allowing the received signals to be easily-collected at the base station. As such, another line of research utilizes received signals for crowd counting. Nevertheless, received signals only offer the information of the number of people, while the accurate density map cannot be derived. As UAVs are now treated as flying base stations and equipped with cameras, we make the first attempt to leverage both RGB images and received signals for crowd density estimation on UAVs. Specifically, we propose a novel network to effectively fuse the RGB images and RSS information. Moreover, we design a new loss function that considers the uncertainty from RSS and makes the prediction consistent with the received signals. Experimental results manifest that the proposed method successfully helps break the limit of traditional crowd density estimation methods and achieves state-of-the-art performance. The proposed dataset is also released as a public download for future research.

CCS Concepts: • **Computing methodologies** → **Neural networks**.

Additional Key Words and Phrases: crowd density estimation, unmanned aerial vehicles, data fusion, datasets

ACM Reference Format:
Kai-Wei Yang, Yen-Yun Huang, Jen-Wei Huang, Ya-Rou Hsu, Chang-Lin Wan, Hong-Han Shuai, Li-Chun Wang, and Wen-Huang Cheng. 2021. Improving Crowd Density Estimation by Fusing Aerial Images and Radio Signals. 1, 1 (November 2021), 22 pages. https://doi.org/10.1145/1122445.1122456

Authors' addresses: Kai-Wei Yang, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu, Taiwan, andyst.eed06@nctu.edu.tw; Yen-Yun Huang, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu, Taiwan, milu0970488651.eed06@nctu.edu.tw; Jen-Wei Huang, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu, Taiwan, admsd.eed06@nctu.edu.tw; Ya-Rou Hsu, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu, Taiwan, st000320.eed06@g2.nctu.edu.tw; Chang-Lin Wan, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu, Taiwan, wanchunglin.eed06@g2.nctu.edu.tw; Hong-Han Shuai, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu, Taiwan, hhshuai@nctu.edu.tw; Li-Chun Wang, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu, Taiwan, lichun@cc.nctu.edu.tw; Wen-Huang Cheng, National Yang Ming Chiao Tung University and National Chung Hsing University, 1001 University Road, Hsinchu, Taiwan, whcheng@nctu.edu.tw.

1 INTRODUCTION

5G network aims to provide a communication service with higher data rate, ultra-low latency, and massive user capacity [1]. However, the widely-used fixed base stations (BSs) cannot effectively facilitate on-demand coverage or adapt to the dynamic environments, especially for crowded scenes or complex terrains [49]. One of the promising solutions is to deploy unmanned aerial vehicles (UAVs) as flying BSs to provide line-of-sight (LoS) connections for ground user equipments (UE) due to their characteristics in terms of low-cost, high-agility and easy-to-deploy, which can augment the coverage and capacity of existing cellular networks [23, 55]. To precisely allocate UAVs for enhancing the communication services, an accurate crowd density map is required to identify the communication demand of each area. One simple approach is to make use of previous works that utilize the received signal strength (RSS) to count the crowd [10, 37, 42]. Nevertheless, the crowd estimation from RSS only provides the clue for people counting while the crowd density map cannot be obtained since two signals with same strength can be transmitted from any directions.

On the other hand, with the advance of deep learning technology, a recent line of research studies the problem of crowd density estimation from RGB images [7, 8, 17, 25, 30, 47, 51], which can facilitate a variety of applications, e.g., activity surveillance, traffic arrangement, security facilities planning. For example, Li *et al.* [26] propose CSRNet by using the 2D dilated convolutional layer to capture the multi-scale features without significantly increasing the parameters. Moreover, Dai *et al.* [29] propose the Dense Dilated Convolution Block (DDCB) to consider the continuously varied scale features. Since flying BSs are usually equipped with cameras, it is promising to combine the crowd density estimation from RGB images and crowd counting from RSS to improve the performance of crowd density estimation. In fact, RGB images and RSS are complementary to each other since i) RSS only provides the information of the distance between UEs and flying BSs but does not contain the information about directions, while RGB images provide the estimation of the crowd density map, and ii) people who are occluded by objects, e.g., trees, cars, cannot be observed from RGB images, while received signals still contain the information of occluded people.

Therefore, in this paper, we make the first attempt to integrate the information from RGB images and received signals. However, several challenges arise for integrating RGB images and RSS. 1) *Unavailability of labeled datasets.* To learn how to estimate the density map from both the RSS and RGB data, one of the challenges is to find a suitable dataset. Even though various existing sources provide image data for crowd counting, the unavailability of corresponding signal data is the primary problem. 2) *Data heterogeneity.* Aerial images and RSS are with different dimensions (2D vs. 1D) and contain heterogeneous information. As such, it is difficult to directly train the model in an end-to-end manner. How to effectively combine the two kinds of information so that one can benefit from the other requires a careful design. 3) *Different receptive fields between RSS and RGB.* Even when the RGB and RSS data are transformed to represent the density map, the receptive fields of cameras and transmitted signals are not perfectly matched (rectangle vs. round). Therefore, it is necessary to deal with the unmatched receptive fields for a better prediction.

To address the first challenge, inspired by [50], we collect our own dataset containing both synthetic aerial images and RSS data with the game engine from Grand Theft Auto V. Moreover, to address the second challenge, we derive the relationship between RSS and aerial images by introducing the communication channel model to convert RSS into RSS Density Map (RDM), i.e. the synthetically-aligned density map derived according to RSS. As for the structure of our network, one intuitive design will be to separately use RGB images and RSS for estimating the crowd density map and ensemble the results in the final stage. However, without the guidance from RGB images, it is difficult to estimate the density map only from RSS since RSS does not contain the location information of UEs. Therefore, we propose a novel network architecture, named RDM-Image Fusion Network (RIFNet), to effectively integrate RDM with RGB

image feature maps. Finally, to address the third challenge, we further propose a new Ring Loss function by considering the uncertainty to differentiate the importance of each pixel. Experimental results show that the proposed method for fusing RGB images with signal data can further improve the image-based baselines by at least 11.2% in terms of Patch Mean Absolute Error (PMAE).

The contributions of this paper are summarized as follows.

- Today flying BSs are usually allocated by heuristics or based on the crowd counting from received signals. To accurately estimate the crowd density map, we propose an approach, new to the current practice of crowd density estimation, by integrating the RGB aerial images and complementary information from RSS. To the best of our knowledge, this is the first work attempting to combine these two complementary information for crowd density estimation.
- We propose the RIFNet structure to fuse the image data with corresponding RSS, while the proposed RIFNet can be appended to existing CNN-based crowd density estimation methods. Moreover, the Ring Loss function further takes the uncertainty of the signal information into consideration and assists the network in making the connection between two input data.
- Experimental results show that the proposed method for fusing RGB images with signal data together with the Ring Loss function outperforms the baselines. Moreover, the dataset containing RGB images and RSS is released as a public download [1] for future research.

The remainder of this paper is structured as follows. Section 2 presents the related works. In Section 3, we describe the relation of RSS and RGB images and propose a novel model for integrating the RGB aerial images and RSS. We introduce our experimental setup and present our experimental results in Section 4. Finally, Section 5 concludes this research.

2 RELATED WORK

Crowd counting plays an important role for many applications, *e.g.,* video surveillance [57], automatic driving technologies [2, 24, 52]. Existing approaches of crowd counting or crowd density estimation can be divided into two main categories, image-based and non-image-based, by their sources. The image-based approach can be further split into the following aspects: detection-based, regression-based, and CNN-based methods.

2.1 Image-based Approach

2.1.1 Detection-based Methods. Previous works estimate the crowd count by first detecting and locating the head, parts of the body, or the full body for each person based on the hand-crafted and low-level features, and then counting the number of detected people [9, 11, 12, 19, 45], which can be further integrated with tracking [22, 44]. For example, Dalal and Triggs [9] propose a method which trains a classifier using features extracted from the full body. Moreover, Dollár *et al.* [11] analyze the statistics of pedestrian scale, occlusion, and location of multiple datasets in pedestrian detection and measure the performance in relation to these statistical information. These methods may be successful in low-density crowd scenes. However, they are inapplicable to extremely-congested crowd scenes due to the occlusion of facial and body parts.

2.1.2 Regression-based Methods. Due to the failure of detection-based methods in highly congested scenes, a line of research aims to directly estimate the total number of the people from images by regressing the image features to the

[1] Link to the dataset and some experimental results have been made available at: **https://github.com/RIFNet/RIFNet**

crowd count [5, 18, 36, 39, 54]. Typically, the process contains two main steps: i) feature extraction, including edge features and texture features, and ii) people count regression, which regress the corresponding image features to the total count. For example, Ma and Chan [36] propose an integer-programming method to estimate the number of people in a set of overlapping sliding windows through mapping local features to the people count. Moreover, Ryan *et al.* [39] split a single person into multiple foreground blobs, then use the local features to count the people in each blob segment. However, regression-based methods only output the count number instead of the crowd density map and neglect the spatial information, which makes it unsuitable to be applied on UAV allocation.

2.1.3 CNN-based Methods. Due to the success of deep CNN in the computer vision community, a recent line of research mostly focuses on CNN-based approaches to predict density maps [4, 6, 8, 14, 26, 32, 33, 35, 40, 43, 47, 51, 56]. For example, Zhang *et al.* [56] assume that different sizes of kernels can extract different sizes of features, which can alleviate the scale-variation problem, and propose a Multi-column CNN (MCNN) for crowd counting. Later on, to deal with different scales of crowd separately, Sam *et al.* [40] propose a Switch-CNN to classify each input image patch to choose a suitable column for estimating the crowd. Following the similar idea, Li *et al.* [26] propose to use dilated convolutional layers for reducing the required parameters of modeling different scales and improving the performance. Besides multi-column approaches, other strategies like multi-layer regression [47] and deformable convolution [14] are also applied to crowd estimation models to improve the accuracy of prediction. To further improve the performance, attention mechanism has been widely used [6, 20, 28] and introduced to enhance the extracted features. For instance, Jiang *et al.* [20] integrate the Density Attention Network and the Attention Scaling Network to provide attention masks related to regions of different density levels. Also, Chen *et al.* [6] propose the Crowd Attention Convolutional Neural Network, which can assess the importance of a human head at each pixel location by automatically encoding a confidence map. Using this guidance of the confidence map, the network focuses more on the position of human head in estimated density map, avoiding misjudgements effectively. Meanwhile, Lian *et al.* [27] propose to integrate the depth map and RGB images for crowd counting and localization by using the depth-aware kernels and anchors. Even though these approaches improve the performance of crowd density estimation, it is noticeable that image-based crowd density estimation approaches still have their limitations, including the lack of robustness against bad environmental conditions (such as poorly-lit spaces or bad weathers), and the incapability to deal with people behind objects.

2.2 Non-Image-based Approach

Without utilizing cameras to capture the images, some studies focus on crowd analysis using WiFi signals [15, 37, 58]. For example, Zhou *et al.* [58] try to count the number of people in a room by identifying a set of differential WiFi Channel State Information (CSI) measurements. Another research, Ooi *et al.* [37] focus on the periodic transmission of probe-request frames and study the correlation of these WiFi frames with the actual number of people present in a crowd. However, these research can only be applied to indoor environments or scenes without crowded people. Moreover, since public WiFi is usually used in touristic areas only and may be sparse at less popular places, it is difficult to estimate the density map directly. On the other hand, as mobile phones have become ubiquitous, the cellular-based solutions have become popular [42, 53], which are able to perform people counting in any location, from indoor environments (via picocells) to vast outdoor areas (via macrocells). However, WiFi-based and cellular-based approaches only output the count number instead of the density map. To the best of our knowledge, the proposed RIFNet is the first work combining RSS and aerial images for crowd density map estimation.

3 PROPOSED METHOD

Given the aerial images captured by UAVs and the RSS on UAVs (flying base stations), the goal of this paper is to estimate the crowd density map by combining aerial images with RSS. Therefore, to address the challenge of data heterogeneity, we first model the relationship between RSS and aerial images by introducing the communication channel models to convert RSS into RSS Density Map (RDM), i.e., the aligned density map derived from RSS. Equipped with RDM, we then propose RDM-Image Fusion Network (RIFNet) that effectively integrates RGB aerial images and RDM through middle-layer and late-layer fusion. Finally, to address the challenge of uncertainty between RSS and RGB images, we design a new loss that considers the uncertainty to differentiate the importance of each pixel.

3.1 RSS Density Map

To model the relationship between RSS and aerial images for alignment, one simple solution is to use a data-driven approach, which directly feeds both raw RSS and RGB images into the model and uses the crowd count as the supervision for learning how to fuse the data. Nevertheless, this simple approach requires a large amount of data since the two inputs have different dimensionality and lack direct connections. Therefore, we propose to first transform the RSS information into RDM, which provides the clue of density map from the RSS information. Specifically, due to the attenuation through propagation, the RSS is related to the distance between the transmitters (UEs of people) and the receivers (UAVs). Therefore, we can evaluate how far a person is from the UAV according to RSS.

To obtain a transformation from the RSS to the distance of the crowd, we introduce a general channel model for radio signals. An emitted signal propagates through space in all directions, resembling an inflating sphere. Based on the conservation of energy, every single point on the sphere surface should have the same energy. In addition, since the surface area is proportional to the square of distance, the signal power should be inversely proportional to the square of its propagating distance theoretically. Generally, according to previous research [46], the average path loss $\overline{L}_{path}(d)$ of a received power can be indicated by the n-th power of the distance d, i.e., $\overline{L}_{path}(d) \propto (\frac{d}{d_0})^n$, where d_0 is the close-in distance, implying a reference point for radio field strength measurements near the transmitter. Within this close-in distance, the behavior of the signal attenuation becomes unpredictable. n is the Path Loss Exponent (PLE), which is an environmental dependent parameter, representing the rate of path dissipation. Therefore, $\overline{L}_{path}(d)$ can be measured in terms of decibel as follows.

$$\overline{L}_{path}(d)_{dB} = \overline{L}_{path}(d_0) + 10n \times log_{10}(\frac{d}{d_0}). \tag{1}$$

By using Eq.1, RSS can be transformed into the distance between UAV and UEs. It is worth noting that the effect of multipath fading is not considered here since the crowd density estimation is usually deployed in open places, e.g., parades on squares. For indoor environments, Rician fading can be considered to build a more precise communication model.

Meanwhile, to take maximum advantage of the signal data, the range of the received signal should be aligned with the image as much as possible. Therefore, we assume that the direction of the camera is set to -90 degrees, facing directly downward.[2] There are two major benefits: i) the projection coordinate of the UAV is at the center of the whole image, which maximizes the overlap between the range of received signals and aerial images; and ii) the setting minimizes the perspective distortion when aligning the information with the aerial images.

However, when it comes to the real-world environment, the path loss with a fixed transmission distance should be regarded as a random variable because the transmitted signal is affected by external factors such as geographical

[2]Later in the experiments, we will show that the proposed approach is error-tolerant with the variation of the direction of the camera.

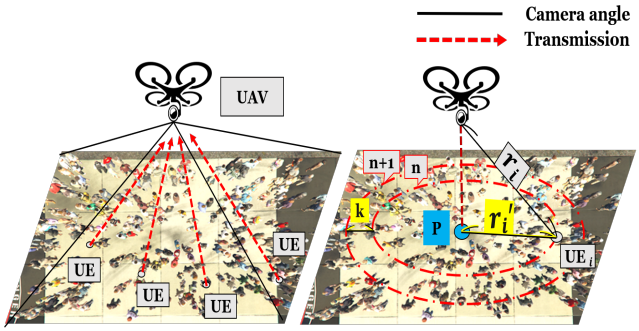


Fig. 1. Illustrative example of the flying base station for crowd density map estimation.

terrains or other existing objects. A common approach to build a practical model of path loss is to utilize the statistical regression to estimate the parameter n that best describe the measurements.[3] Nevertheless, the measurement error still exists in the regression model. To deal with the measurement error, instead of using the exact transformed distance from RSS, we calculate the histogram of the distances with the bin width equal to k meters for uncertainty tolerance.

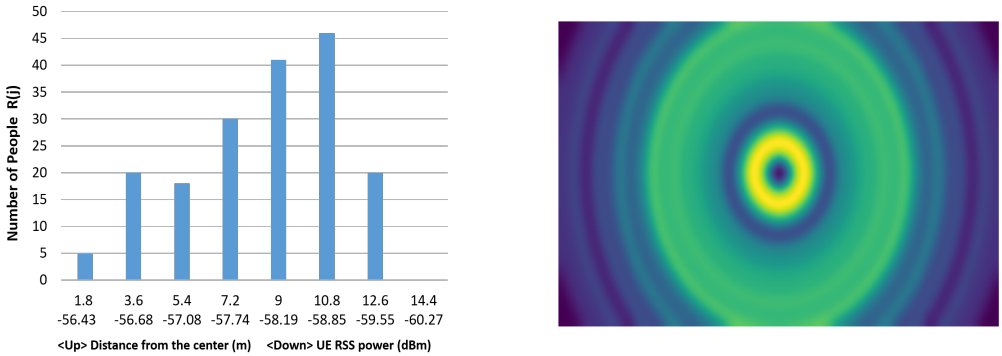
Take Fig. 1 as an example. The camera on the UAV faces down and captures the aerial images, while the UEs request the communication from the UAV (flying base station). Let P denote the projection point of the UAV on the horizontal plane and r_i is the distance between UAV and the i-th UE (UE_i) calculated from the channel model. Moreover, r_i' is defined as the projection of r_i on the horizontal plane, indicating the distance between a UE_i and P. By using Eq.1, we can transform RSS into the distance between UAV and UEs. Let $R(j)$ denote the number of UEs in the j-th bin of the distance histogram, which is calculated by adding up the number of r_i' that falls in a ring area, defined as $(j-1)k < r_i' \leq jk, j \in \mathbb{Z}^+$. As such, the transformation from RSS to the distance histogram is illustrated as Fig. 2(a).

After transforming RSS into the distance histogram, the one-dimensional feature is still not aligned to the aerial images. To facilitate the fusion of RSS and images, we convert the distance histogram shown in Fig. 2(a) into a density map of the same size as the images, referred to as *RSS Density Map (RDM)*. To do so, we first calculate the magnification ratio of the images. Consider a camera with focal length f, size of the photosensitive element $W \times L$, located at altitude h (can be derived by altimeter sensors), and taking pictures of $w \times l$ pixels vertically downward. The ratio of the actual size of the object to its size in the image is denoted as q/p, where q is the object distance and p is the image distance. From the Thin Lens Equation $1/p + 1/q = 1/f$, the scale ratio q/p is approximately q/f when photographing at a long distance such that $q \gg f$. While f remains a constant, the scale ratio q/f is directly proportional to the object distance q. All pixels are assumed to have the same scale ratio as the center of the image[4], which is h/f because q equals h. That is,

$$scale \;\; ratio = \frac{q}{p} \approx \frac{q}{f} = \frac{h}{f}. \tag{2}$$

[3]In this paper, we also apply the regression method to generate a path loss model that best fits the reality. The regression process will be explained in detail later in Section 4.1.2.

[4]It is worth noting that objects appearing at different positions in the image, such as at the middle and near the boundaries, have distinct object distance q, which causes some minor perspective distortion but can be neglected in our scenario (h is large).



(a) Distance Histogram

(b) RDM

Fig. 2. An illustrative example of generating RDM. (a) The distance histogram that shows the number of people in each ring. Every bin in the graph represents one ring, with width of the rings k equals 1.8 meters. (b) An illustration of RSS Density Map (RDM). Every pixel value represents the expected number of people at that specific area.

Based on the scale ratio, the number of pixels in the image that corresponds to 1 meter of width and length in the real world are expressed by,

$$M_{width} : \frac{w}{W \times h/f}(pixel/m),$$
$$M_{length} : \frac{l}{L \times h/f}(pixel/m). \tag{3}$$

Generally speaking, under most circumstances, the magnification ratios of the width and the length are identical. Thus, we define the magnification ratio $M = M_{width} = M_{length}$. Using this magnification ratio, in addition to Bresenham's circle algorithm [3] with every increment being k' pixels on the radius, and the center of the image being the center of the circle, we generate concentric circles until the diameter exceeds the diagonal length of the image. Let k' denote the number of pixels converted from k meters, where $k' = k \times M$. Consequently, the spaces between each concentric circles are the results of mapping the ring areas from the real-world to the image. Afterward, dividing the number of people on individual rings with their corresponding area leads to the average crowd density in the ring. We eventually establish a matrix with the same size as the input image, assign the value of each entry with its corresponding crowd density, and then pass it through a Gaussian filter to blur the boundaries of adjacent rings to alleviate the quantization error. Fig.2(b) illustrates an example of the output of RDM from the RSS shown in Fig. 2(a). To sum up, the original received RSS is now transformed into RDM which has the same size and dimension as the input images. It simultaneously contains information concerning distance and the number of people, greatly reducing the difficulty of merging the two kinds of raw data.

3.2 RIFNet

Equipped with aligned RGB aerial images and RDM, the next goal is to effectively integrate RDM with RGB images. One naive way is to concatenate the image and RDM into a new tensor and use neural networks to process the tensor. Nevertheless, the large-sized tensor requires a large amount of data to train the neural network for fusing. Another

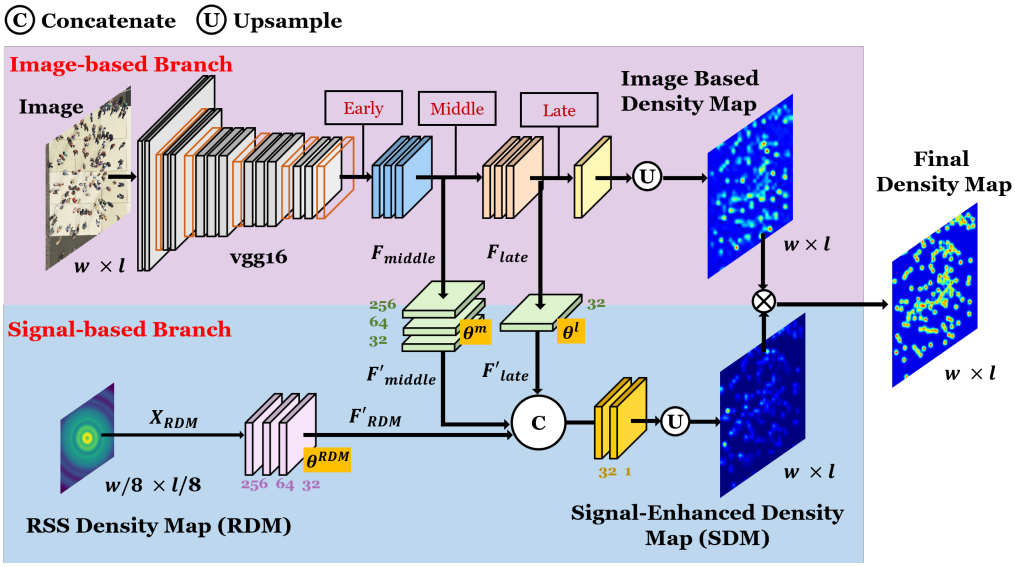


Fig. 3. Structure of the RIFNet, composed of the image-based Branch (the upper part), which takes an RGB image as input, and the signal-based Branch (the lower part), which takes an RDM as input. Their outputs are the Image-based Density Map and Signal-Enhanced Density map, respectively. In this illustrative figure, CSRNet is implemented as the imaged-based branch but, in fact, any CNN-based state-of-the-art approach can be implemented as the image-based branch. For the signal-based branch, it is composed of an RDM encoder and the output layer. The two branches are connected by the middle-layer CNN and the late-layer CNN. Numbers next to the feature maps represent their channels. The kernel size of all CNNs are 3×3. Dilation rate of the RSS encoder in the signal-based branch is 2.

approach is to use existing fusion approaches to fuse the RGB aerial images and RDM, e.g., element-wise (Hadamard) product [41], but the values in the two data represent completely different meanings (i.e. raw RGB color code and average crowd density), which leads to an inferior performance.

Therefore, to address the second challenge of data heterogeneity, we propose a novel RDM-Image Fusion Network (RIFNet) to fuse the information. The network architecture of RIFNet is illustrated in Fig. 3, which contains the image-based branch (upper part) and signal-based branch (lower part). The key idea is to fully exploit the information from RDM to complement the information from RGB images. The image-based approaches generate the image-based density map. Meanwhile, the image-based branch also provides the signal-based branch with additional spatial guidance to predict the Signal-Enhanced Density Map (SDM) as shown in bottom-right side of Fig. 3.

Specifically, for the image-based branch, given the RGB aerial images, we use state-of-the-art image-based density map estimation to generate a density map as the output, referred to as the *image-based density map*. On the other hand, for the signal-based branch, the goal is to efficiently make use of the image-based branch and extract the complementary information from the RDM. Since the pixel values in both density maps indicate the expected number of people in the image, the extracted features close to the output layer in the image-based branch should be more similar to the features extracted from the signal-based branch. Therefore, we conduct a multi-layered feature fusion, which takes the extracted features from the middle layer and late layer of the image-based branch, denoted by F_{middle} and F_{late}, respectively,

and then uses CNNs to transform F_{middle} and F_{late} into F'_{middle} and F'_{late} for the signal-based branch, i.e.,

$$F'_{middle} = g_m(F_{middle}; \theta^m), \tag{4}$$

$$F'_{late} = g_l(F_{late}; \theta^l), \tag{5}$$

where g_m and g_l are the transformation function composed by CNNs with θ^m and θ^l as learnable parameters. Since the middle-layer and late-layer information is fused afterward, the number of layers for θ^m is greater than that of θ^l to make the information compatible. Furthermore, let X_{RDM} denote the RDM. We use another transformation function g_{RDM} to extract features from X_{RDM}, denoted as F'_{RDM}, for matching F'_{middle} and F'_{late}, i.e.,

$$F'_{RDM} = g_{RDM}(X_{RDM}; \theta^{RDM}), \tag{6}$$

where θ^{RDM} represents the learnable parameters of the CNN encoder. Finally, F'_{middle}, F'_{late} and F'_{RDM} are concatenated and fed into the decoder to derive the Signal-enhanced Density Map (SDM). Using SDM as the auxiliary information, a Hadamard product is applied on the image-based density map and the SDM to generate the final density map. In summary, RIFNet fully leverages the image-based approach to extract the complementary information from RDM to achieve a better estimation.

3.3 Loss Function

The mean square error (MSE) loss is a standard loss that computes the pixel-wise Euclidean distance between ground truth density map and the predicted density map. Let $X(i)$ and $X_{RDM}(i)$ denote the i-th RGB image and RDM, respectively. The formula for pixel-wise MSE loss L_{MSE} can be expressed by

$$L_{MSE} = \sum_i ||Z(X(i), X_{RDM}(i); \Theta) - Z^{GT}(i)||_2^2, \tag{7}$$

where $Z(X(i), X_{RDM}(i); \Theta)$ is the predicted density map of i-th training sample by RIFNet, Θ represents all learnable parameters in the RIFNet, and $Z^{GT}(i)$ is the ground truth density map of i-th training sample. Another commonly-used loss is the difference between the predicted count and the ground truth count, denoted by L_c. Let $C(i)$ and $C^{GT}(i)$ denote the predicted count and ground truth count of the i-th training sample, respectively.

$$L_c = \sum_i ||C(i) - C^{GT}(i)||. \tag{8}$$

However, the errors of different regions in the predicted density map are not equally important due to the third challenge of uncertainty between RSS and RGB. For example, when the ring region converted from j-th bin of the distance histogram is fully inside the image, the predicted count should be the same as $R(j)$. If the ring region is only partially inside the image due to the rectangularity of image, the predicted count may have a larger error caused by the uncertainty.

Therefore, we propose a novel Ring Loss, denoted by L_{ring}. Let $P^{estimate}(i, j)$ and $P^{gt}(i, j)$ denote the estimated count and ground truth count of the (partial) ring region converted from $R(j)$ in the i-th training sample, respectively. The Ring Loss function is defined as follows.

$$L_{ring} = \sum_i \sum_{j=1}^{n_r} \frac{A_j^{in}}{A_j} \times ||P^{estimate}(i, j) - P^{gt}(i, j)||, \tag{9}$$

where n_r is the number of ring regions. A_j^{in} and A_j are respectively the ring area inside the image and the area of the whole ring area converted from $R(j)$. In other words, A_j^{in}/A_j represents the certainty of region $R(j)$. Finally, the total loss function can be written as,

$$L_{total} = L_{MSE} + \lambda \times L_{ring}, \tag{10}$$

where λ is a hyper-parameter controlling the importance of the ring loss.[5]

3.4 Training Strategy

The training process consists of three phases:

- **Phase 1:** Load the pretrained weights (if any) and train a traditional image-based crowd density estimation model as the image-based branch.
- **Phase 2:** Load the pre-trained image-based branch, freeze the pre-trained parameters, and train the middle-layer CNN, the late-layer CNN, and the signal-based branch.
- **Phase 3:** Unfreeze all the parameters and train the whole model (end-to-end).

To facilitate the learning from the signal-based branch, we randomly block a 50×50 (pixel) square in the input RGB image in phase 2 and phase 3. This makes the model acquire more information from the signal-based branch since it does not have any clue about the blocked areas from the RGB image. It is worth noting that one alternative solution is to train the architecture in an end-to-end manner, *i.e.*, directly taking RSS and RGB as inputs and using CNNs to fit the groundtruth. However, training from scratch makes the model focus on the RGB branch and ignore the RSS branch since it is easier to fit RGB to density map. Moreover, as the model initially does not know how to use RSS input, it may interfere the RGB branch and leads to an inferior performance. In contrast, by first pretraining the image-based branch and fixing the parameters, the network is equipped with basic abilities and focuses on using RSS to further improve the results.

4 EXPERIMENT

4.1 Experimental Setup

4.1.1 UAV-GCC Dataset. To train and evaluate the performance of the proposed RIFNet, it requires a dataset that contains both RSS and RGB images. Inspired by previous work [50], we build such dataset by a similar manner, i.e., generating the synthetic crowd scenes from game engines, together with the corresponding RSS. Specifically, we create a new dataset, *UAV-GCC dataset*, using the game engine from Grand Theft Auto V (GTA5) to collect synthetic images with the camera facing straightly downward and locating at 15 or 20 meters above the ground. There are 3, 125 images in total with the resolution of 225×400. To increase the diversity of the crowd density, the number of people in our dataset ranges from 20 people to 300 people, with an average of 115 people. Since the game engine is unable to create synthetic RSS data, we generate hypothetical signals from the ground truth using a regression channel model (details in Section 4.1.2). It is worth noting that the proposed synthetic dataset is further transformed by applying the SE Cycle GAN proposed in [50]. Fig. 4 demonstrates two examples of the translated synthetic images. We partition the proposed dataset into 5 folds and perform the 5-fold cross validation.

4.1.2 Channel Model Generation. Since the GTA5 game engine cannot generate synthetic RSS data, we use real-world RSS data for regressing the channel model. Table 1 summarizes the experimental equipment. Specifically, we use Asus

[5]The effect with different values of λ will be further discussed in Section 4.4.

Fig. 4. Illustration of the translated synthetic images generated by SE Cycle GAN.

Table 1. Parameters of Radio Signal Strength (RSS) Measurement for Regression Model

Radio Signal	Wi-fi (802.11)
Frequency	5.8 GHz
Transmitted Power	25 dBm
User Equipment (UE)	Asus ZS630KL

ZS630KL as the UE and communicate with the flying base station through Wi-fi (802.11) with the transmit power of $25dB$. We choose an outdoor open space as the experiment environment to be consistent with the scenarios in the UAV-GCC dataset. Therefore, the path loss model in [21] previously mentioned as Eq.1 can be applied, i.e.,

$$\overline{L}_{path}(d)_{dB} = \overline{L}_{path}(d_0) + 10n \times log_{10}(\frac{d}{d_0}) + X. \tag{11}$$

The close-in distance d_0 is set to 1 meter in our experiment and the additional random variable X represents the uncertainties caused by the environment. We utilize the 5.8 GHz WiFi signal to simulate the lower frequency 5G signals (sub $6GHz$) since there is an overlap on their frequency band. Meanwhile, since it is also less likely to experience any interference, we use a UAV installed with a WiFi access point as the transmitter. Then, considering the projection point of the UAV to be the center point, we collect the real-world RSS at 4 different locations on 20 concentric circles, where each with its radius 1 meter larger than the previous one. Thus, in a circular area with a radius of 20 meters, RSS at 80 different locations are measured. An illustrative image of our experiment can be seen in Fig 5. Through the regression of the parameters of the channel model, PLE(n), $\overline{L}_{path}(d_0)$[dB], and X[dB] are set to 2.76744, 47.7103, and 1.0875, respectively. Finally, the path loss model is derived as,

$$\overline{L}_{path}(d)_{dB} = 47.7103 + 27.6744 \times log_{10}(d) + 1.0875. \tag{12}$$

From Eq.12 and the magnification ratio mentioned in Section 3.1, we can calculate the RSS of each people in an image to generate a hypothetical RSS data for the UAV-GCC dataset.

4.1.3 Ground Truth Generation. Similar to current CNN-based crowd counting methods [40, 56], we generate the ground-truth density map Z^{GT} by convolving the ground truth dot map with a Gaussian Kernel as follows.

$$Z^{GT}(x) = \sum_{j=1}^{N_p} \delta(x - x_j) \times G_\sigma(x), \tag{13}$$

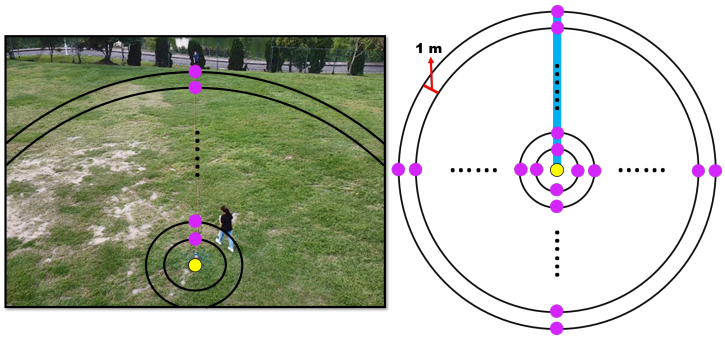


Fig. 5. An illustrative example of collecting real-world RSS data for channel model regression. The black curve in the photo on the left represents the concentric circles. The purple dots and the yellow dot are respectively the locations we measure the RSS data and the center of concentric circles. There are 40 circles in total and four RSS data are measured on each circle at different locations.

where x is the position of an arbitrary pixel, x_i is the position of the pixel where a person is present, N_p indicates the total number of people in the image, δ is the Dirac Delta Function, and $G_\sigma(x)$ is the Gaussian kernel with standard deviation $\sigma = 3.5$.

4.1.4 Baselines and Implementation Details. Since this is the first work combining RSS and RGB aerial images, we first evaluate the performance of the proposed RIFNet by using the following five CNN-based crowd density estimation approaches for the image-based branch.

- **CSRNet [26].** The first 10 layers of the VGGNet is used as a front-end pretrained model in order to effectively extract features of input images. Afterward, a dilated convolutional neural network is added to not only increase the receptive field without using any pooling layer but also ensure that the output feature map size is the same as the input size.
- **CAN [31].** This model first extracts basic features from an image by using a pre-trained VGG16 front-end, and the VGG features are fed to a Spatial Pyramid Pooling [16] to extract multi-scale context information. The network can thus learn the importance of each such feature at every location in an image.
- **MCNN [56].** The network structure of MCNN can be divided into three separate columns. Each column comprises CNN filters with different sizes of receptive fields and is adaptive to variations in people/head size. By combining the outputs of all CNN with learnable weights, MCNN produces the final density map based on geometry-adaptive kernels without the requirement of knowing the perspective map of the input image.
- **SFCN [50].** The proposed network is first pre-trained on synthetic data generated by an automatic data collector and labeler and fine-tuned by using real data. They also proposed the SFCN structure, which is a domain adaptive method by combining the advantage of Fully Convolutional Network (FCN) [34] and the Spatial Encoder [38].
- **SCAR [13].** The network structure is composed of two branches, the Spatial-wise Attention Model (SAM) and the Channel-wise Attention Model (CAM). SAM is able to encode the pixel-wise context of the input images, while SAM extracts more discriminative information to help the model pay attention to the head regions.

Most of these existing models adopt VGG-16 as part of the structure. To increase the diversity of our experiment, we implement SFCN with ResNet-101 as its backbone. The trained models then serve as the image-based branch in the RIFNet structure. Please note that the random-blocking is also used for baselines during training to guarantee the

fairness of comparison. We use Adam as the optimizer with the batch size as 1 and the weight decay as 5×10^{-4}. The hyperparameter λ is set to 0.005 for RIFNet.

4.1.5 Evaluation Metrics. Let D_{test} denote the testing data. The Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) are common evaluation criteria of former researches on crowd density estimation, which can be represented as follows.

$$MAE = \frac{1}{|D_{test}|} \sum_{i=1}^{|D_{test}|} |C(i) - C^{GT}(i)|, \tag{14}$$

$$RMSE = \sqrt{\frac{1}{|D_{test}|} \sum_{i=1}^{|D_{test}|} (C(i) - C^{GT}(i))^2}, \tag{15}$$

where $|D_{test}|$ is the size of the testing set, $C^{GT}(i)$ and $C(i)$ are respectively the number of people in the ground truth and the estimated number of people in the i-th image, which are obtained by the sum of all pixels in the estimated/ground truth density maps. However, MAE and RMSE only consider the error in the whole image, lacking the consideration of the correctness in each local area. In other words, when the model overestimates the number of people in some regions and underestimates the number of people in some other regions, MAE or RMSE may still be very small. Therefore, following previous work [48], we adopt the Patch Mean Absolute Error (PMAE) and the Patch Root Mean Squared Error (PRMSE) as the primary criteria.

$$PMAE = \frac{1}{n_p \times |D_{test}|} \sum_{i=1}^{|D_{test}|} \sum_{j=1}^{n_p} |C(i,j) - C^{GT}(i,j)|, \tag{16}$$

$$PRMSE = \sqrt{\frac{1}{n_p \times |D_{test}|} \sum_{i=1}^{|D_{test}|} \sum_{j=1}^{n_p} (C(i,j) - C^{GT}(i,j))^2}, \tag{17}$$

where n_p is the number of patches in one image, $C(i,j)$ and $C^{GT}(i,j)$ are respectively the estimated number and the ground truth number of people in the j-th patch of i-th image. By computing MAE and RMSE of each patch, PMAE and PRMSE reflect the local performance of each sub-area. In the following sections, we evaluate the experiment results by partitioning every image into 12 patches, i.e., $n_p = 12$.

4.2 Evaluations on Density Map Quality

Table 2 compares the performance of CSRNet, CAN, MCNN, SFCN and SCAR models with and without RIFNet. The results manifest that, in terms of all evaluation metrics, the performance of all models improves with the Ring Loss and the signal-based branch appended. For example, the proposed approach respectively improves CSRNet, CAN, MCNN, SFCN and SCAR by 15.9%, 11.2%, 18.3%, 21.7% and 21.2% in terms of Patch Mea Absolute Error (PMAE). The improved density maps show not only a better local performance (PMAE and PRMSE) but also a more accurate people count (MAE and RMSE), proving the usefulness of the signal-based branch and the compatibility of the proposed RIFNet with existing approaches. Moreover, we also conduct the experiments of directly fusing the density maps from RGB images and the RSS (RDM) based on Hadamard product. The experimental results (with direct fusion) suggest that if the inputs are not aligned with a proper method, the additional RSS data might destroy the original features of the RGB images, making the performance even worse than that of using RGB-only.

Table 2. The experimental results of each state-of-the-art image-based crowd density estimation method appended with the signal-based branch, together with the baselines (direct fusion) of using Hadamard product for fusing the RGB-based and RSS-based density maps.

	PMAE	PRMSE	MAE	RMSE
CSRNet	0.6465	0.9576	4.55	6.9686
CSRNet with direct fusion	0.8364	1.2898	5.80	7.6851
CSRNet+RIFNet (ours)	0.5437	0.7939	3.20	4.7457
CAN	0.6286	0.9577	4.85	7.3317
CAN with direct fusion	0.8313	1.2847	7.02	11.9908
CAN+RIFNet (ours)	0.5582	0.8289	3.80	5.6308
MCNN	1.0607	1.5125	9.70	13.0963
MCNN with direct fusion	1.2463	1.8306	10.23	14.0749
MCNN+RIFNet (ours)	0.8669	1.2497	6.66	8.9281
SFCN	0.7430	1.1122	6.12	8.8649
SFCN with direct fusion	1.0905	1.7614	8.07	11.1076
SFCN+RIFNet (ours)	0.5816	0.9089	3.88	6.3093
SCAR	0.6377	0.9020	5.32	6.9174
CAR with direct fusion	1.2618	1.9394	8.70	11.2314
SCAR+RIFNet (ours)	0.5034	0.7407	2.83	4.5372

Table 3. The computational efficiency of the baselines and RIFNet with different state-of-the-art image-based model as the image-based branch, measured in terms of frames per second (fps). The result shows that RIFNet is able to achieve real-time performance.

Baseline	117 fps	93 fps	141 fps	26 fps	110 fps
RIFNet	+CSRNet	+CAN	+MCNN	+SFCN	+SCAR
	66 fps	60 fps	70 fps	26 fps	67 fps

Fig. 6 shows the qualitative results of the resulting density maps, where the first column shows the input RGB images and the ground truth density map, and the remaining columns show the density map generated by baselines and baselines with adding the signal-based branch (+). The number at the bottom right corner of each density map represents the total count of people. The results show that the output density maps of the baselines ($1st$, $3rd$, $5th$ and $7th$ rows) appear to be blurry at the areas with a higher crowd density. This is in line with the theory of Jiang *et al.* [20] that traditional image-based crowd density estimation methods seem to overestimate areas with clustered crowd. In contrast, the proposed RIFNet successfully separates each individual in the output density maps ($2nd$, $4th$, $6th$ and $8th$ rows). This is because the original uncertainty is clarified by the RSS information. Hence, with additional knowledge of the number of people given by RSS, combined with spatial information given by the RGB image, individuals in crowded areas can be distinctly identified. This indicates that the proposed model effectively leverages the additional enhancement of RSS information, and is able to estimate density maps with more explicit details.

4.3 Evaluations on Running Time

The computational efficiency is of crucial importance for providing low latency services of 5G cellular network. To evaluate the running time, the CSRNet [26], CAN [31], MCNN [56], SFCN [50], and SCAR [13] are respectively used as the image-based branch of the proposed RIFNet. The running time is tested on one NVIDIA Tesla V100 32GB GPU. Recall that the size of the images in the UAV-GCC dataset used for training and testing is 225×400. The results are shown in Table 3, which implies that RIFNet is able to achieve real-time performance for most approaches. SFCN-based

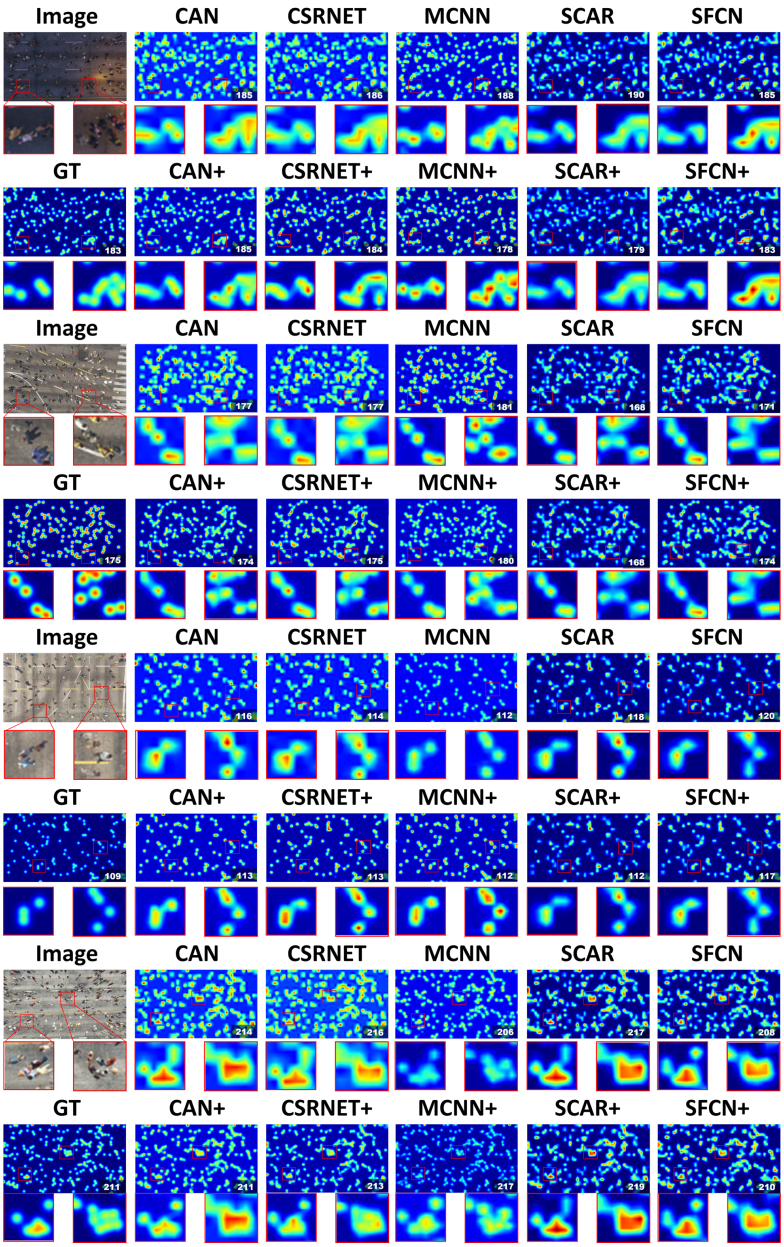


Fig. 6. Comparison results of the ground truth, traditional image-based approach, and the RIFNet structure using CAN, CSRNET, MCNN, SCAR and SFCN as the image-based branch.

RIFNet shows a relatively low computational efficiency as compared to the others because it has a much larger number of parameters involved in its network structure.

Table 4. Diagnostic experiments of the RIFNet structure using CSRNet and CAN as image-based branch.

		PMAE	PRMSE	MAE	RMSE
CSRNet	Purely image-based	0.6465	0.9576	4.55	6.9686
	SBB w/o Ring Loss	0.6098	0.8918	4.19	6.2539
	Ring Loss w/o SBB	0.6052	0.9214	4.51	7.0752
	SBB+Ring Loss	**0.5437**	**0.7939**	**3.20**	**4.7457**
CAN	Purely image-based	0.6286	0.9577	4.85	7.3317
	SBB w/o Ring Loss	0.5907	0.9148	3.97	6.3632
	Ring Loss w/o SBB	0.6022	0.8942	4.70	7.0131
	SBB+Ring Loss	**0.5582**	**0.8289**	**3.80**	**5.6308**

Table 5. Performance of the model with feature extraction from different stages in the CSRNet and CAN image-based branch. (SBB refers to the signal-based branch)

	CSRNet as image-based branch				CAN as image-based branch			
	PMAE	PRMSE	MAE	RMSE	PMAE	PRMSE	MAE	RMSE
w/o SBB	0.6465	0.9576	4.55	6.9686	0.6286	0.9577	4.85	7.3317
early	0.6624	1.0214	5.18	8.2789	0.6211	0.9154	4.35	6.6632
middle	0.6424	0.9530	4.46	6.8464	0.6086	0.9132	4.13	6.4677
late	0.6449	0.9709	4.90	7.4721	0.6109	0.9249	4.62	7.0869
early+middle	0.6270	0.9737	5.10	7.9178	0.6088	0.9024	4.06	**6.2200**
early+late	0.6250	0.9383	4.39	6.8590	0.6073	0.9090	4.04	6.4004
middle+late	**0.6098**	**0.8918**	**4.19**	**6.2539**	**0.5907**	0.9148	**3.97**	6.3632
early+middle+late	0.6225	0.9354	4.41	6.9018	0.6077	**0.8934**	4.31	6.7247

4.4 Ablation Study

Here, we conduct the ablation study to show the contributions of the signal-based branch and the proposed Ring Loss. Specifically, we compare several variants with CSRNet and CAN serving as the image-based branch: i) Purely image-based branch, which simply evaluates the performance without any RSS information, ii) SBB w/o Ring Loss, which appends the signal-based branch (SBB) to the image-based branch but without using the Ring Loss, iii) Ring Loss w/o SBB, which uses the Ring Loss as the additional constraint in the training process without appending the signal-based branch to the image-based branch, and iv) SBB+Ring Loss, which provides complete information of RSS to assist the crowd estimation by appending the signal-based branch and using the Ring loss in the training process. As shown in Table 4, both the Ring Loss and the signal-based branch can effectively provide RSS information to assist the original image-based approaches in generating a more accurate density map.

Moreover, since the architecture of RIFNet involves extracting features from the image-based branch and fusing them with information of RSS, we examine the fusion at different stages to justify the designed architecture. Hence, we apply the pre-trained model of CSRNet and CAN as the image-based branch and extract features from the early layer, the middle layer, and the late layer. The early layer is defined as the output of the VGG16 backbone, while the middle and late layers are defined as the center and the output of the model appended to the VGG16 backbone as shown in Fig. 3. The extracted features are fed into the signal-based branch for further fusion. Table 5 shows the results of fusion at different stages for CSRNet and CAN, respectively. The results indicate that the model performs better if the features are extracted from multiple layers, instead of a single layer, then fused with the RDM, since the multi-level fusion can help the model learn to identify objects of different sizes. On the other hand, the result also suggests that fusion with early-stage features results in a less accurate prediction. This is because the input of the signal-based branch is the

Table 6. Performance of the model with different weight of the ring loss while using CSRNet and CAN as image-based branch

	CSRNet-based RIFNet				CAN-based RIFNet			
	PMAE	PRMSE	MAE	RMSE	PMAE	PRMSE	MAE	RMSE
λ=0.05	0.6360	0.9237	4.95	7.2118	0.6244	0.9300	4.43	7.1705
λ=0.03	0.6279	0.9576	4.63	7.2588	0.6172	0.9486	4.76	7.4942
λ=0.01	0.6151	0.9355	4.66	7.3483	0.6102	0.9541	4.73	7.6196
λ=0.005	**0.6052**	**0.9214**	**4.51**	**7.0752**	**0.6022**	**0.8942**	4.70	7.0131
λ=0.003	0.6059	0.9407	4.97	7.6147	0.6058	0.8991	**4.24**	**6.4988**
λ=0.001	0.6185	0.9367	4.82	7.3625	0.6076	0.9262	4.57	7.0962
λ=0.0005	0.6296	0.9723	5.02	7.8198	0.6192	0.9207	4.80	7.6667

Table 7. The experimental results of each baseline methods appended with the signal-based branch under angle variation.

	PMAE	PRMSE	MAE	RMSE
CSRNet	0.6408	0.9431	4.52	6.9542
CSRNet + RIFNet	0.5453	0.8005	3.25	4.9861
CAN	0.6073	0.9158	4.79	7.1496
CAN + RIFNet	0.5601	0.8342	3.92	5.7185
MCNN	1.0739	1.5208	9.75	13.1287
MCNN + RIFNet	0.8812	1.2643	6.81	9.0224
SFCN	0.7736	1.1238	6.26	8.9641
SFCN + RIFNet	0.6139	0.9187	3.94	6.4261
SCAR	0.6324	0.8619	5.18	6.9006
SCAR + RIFNet	0.5146	0.7647	3.02	4.9372

RDM which possesses more similar information with the image-based density map than with the raw image data (RGB color code). However, the features extracted from the early layer of the image-based branch is more related to the raw image data. Consequently, it is more difficult for the model to learn the connections between them.

To analyze and determine the hyper-parameter λ in the loss function shown in Eq.10, we evaluate the performance of the proposed method for λ ranging in $(0.0005, 0.05)$ with CSRNet and CAN. As shown in Table 6, the optimal value is found to be 0.005. However, the model is rather stable with $\lambda \in (0.01, 0.003)$.

Finally, we examine the robustness of the proposed model with different camera view angles. In Section 3.1, we assume that the UAV's angle of view is -90 degrees, i.e., facing directly downward. However, in the real world environment, there might be some external factors causing the variation of the angle of view. We therefore conduct an experiment to assess the influence of this variation on our performance. Specifically, we further collect 897 testing images by randomly selecting the angle of view between -80 and -88 degree. The result is shown in Table 7, which manifests that the proposed RIFNet still helps the baselines to better estimate the crowd density map. Compared with Table. 2, the performance of purely image-based baselines without appending RIFNet has no significant difference, while the baselines with the RIFNet is slightly worse than the results of no angle variation, due to the mismatch between the image view and the signal sensing range. It is worth noting, though, that the proposed RIFNet still respectively improves CSRNet, CAN, MCNN, SFCN and SCAR by 14.9%, 7.8%, 17.9% 15.5%, and 18.6% in terms of PMAE.

Table 8. Evaluation of the RIFNet structure using existing CNN-based models as image-based Branch with unedited testing images (w/o RC) and testing images with random cropping (w/ RC).

	PMAE			PRMSE		
	w/o RC	w/ RC	error↑	w/o RC	w/ RC	error↑
CSRNet	0.6465	0.9785	51%	0.9576	1.7740	85%
CSRNet+RIFNet	0.5437	0.7008	29%	0.7939	1.1370	43%
CAN	0.6286	0.8978	43%	0.9577	1.6818	76%
CAN+RIFNet	0.5582	0.6458	16%	0.8289	0.9965	20%
MCNN	1.0607	1.3556	28%	1.5125	2.1312	41%
MCNN+RIFNet	0.8669	1.1755	36%	1.2497	1.9332	55%
SFCN	0.7430	1.0648	43%	1.1122	1.8805	69%
SFCN+RIFNet	0.5816	0.9426	62%	0.9089	1.6994	87%
SCAR	0.6377	0.9251	45%	0.9020	1.6183	79%
SCAR+RIFNet	0.5034	0.6388	27%	0.7407	0.9733	31%

Table 9. Comparison between the performance of the baseline and the RIFNet structure under poor light situation.

	PMAE	PRMSE	MAE	RMSE
CSRNet	0.8648	1.3251	8.05	11.6487
CSRNet + RIFNet	0.7025	1.0295	5.30	7.1532
CAN	1.0055	1.6101	10.40	15.0653
CAN + RIFNet	0.8514	1.3956	7.76	12.2706
MCNN	1.2552	1.7168	11.38	15.2219
MCNN +RIFNet	1.0383	1.4370	8.41	10.6979
SFCN	1.7975	2.9411	17.96	23.7208
SFCN + RIFNet	1.5747	2.7008	15.53	20.0638
SCAR	1.2576	2.2014	12.49	15.1542
SCAR + RIFNet	1.2440	2.1160	8.86	12.4611

4.5 Case Study

Evaluations on Random Cropped Images. To examine whether the additional RSS data can serve as a complementary information under the circumstances where the crowd is partially occluded in the RGB images, we randomly block a section of 50×50 pixels in every image throughout the testing process to simulate such situation. Furthermore, during each testing procedure, the seed for generating the random sections are identical to guarantee the fairness of evaluation. Table 8 shows the results of RIFNet incorporating different image-based approaches with and without random cropping. In comparison to the testing results without random cropping the images, several conclusions can be drawn. i) Since the signal-based branch in the RIFNet structure merely serves as a enhancement for the original image-based approach, the output still depends on the RGB image input. Thus, despite the supplementary RSS information, there appears to be an increase in all evaluation criteria. ii) For most models, the increment of error is smaller when incorporating with RIFNet, proving that the fusion with signal information enhances the robustness of the model against occluded crowds. iii) For the case in which SFCN and MCNN is used as the image-based branch, the increment in error rate appears to be higher with addition information of RSS, which might be caused by the excessive number of learnable parameters and fusing too many information from different branches. However, the error of SFCN with RIFNet is still obviously smaller than that of SFCN without RIFNet, proving the usefulness of RSS information.

Table 10. Comparison between the performance of the baseline and direct fusion.

	PMAE	PRMSE	MAE	RMSE
CSRNet	0.6465	0.9576	4.55	6.9686
CSRNet Hadamard	0.8364	1.2898	5.80	7.6851
CAN	0.6286	0.9577	4.85	7.3317
CAN Hadamard	0.8313	1.2847	7.02	11.9908
MCNN	1.0607	1.5125	9.70	13.0963
MCNN Hadamard	1.2463	1.8306	10.23	14.0749
SFCN	0.7430	1.1122	6.12	8.8649
SFCN Hadamard	1.0905	1.7614	8.07	11.1076
SCAR	0.6377	0.9020	5.32	6.9174
SCAR Hadamard	1.2618	1.9394	8.70	11.2314

Table 11. Comparison results between baseline and RIFNet on VisDrone dataset

	PMAE	PRMSE	MAE	RMSE
SCAR	1.8966	3.3494	12.97	20.5141
SCAR + RIFNet	1.7102	2.9241	11.11	14.9240

Evaluation on Poor Light Situation. To prove the robustness of the RIFNet with low-quality images, we selected 130 images from the UAV-GCC dataset with poor-light situations. The first row of Fig. 6 shows an example of such poor-lighted images. The experimental results are shown in Table 9. Compared with Table 2, the data indicates that the performance of the baselines drops when the input image is with very little light. On the other hand, equipped with the RIFNet structure, most evaluation metrics suggest that the performance is less affected.

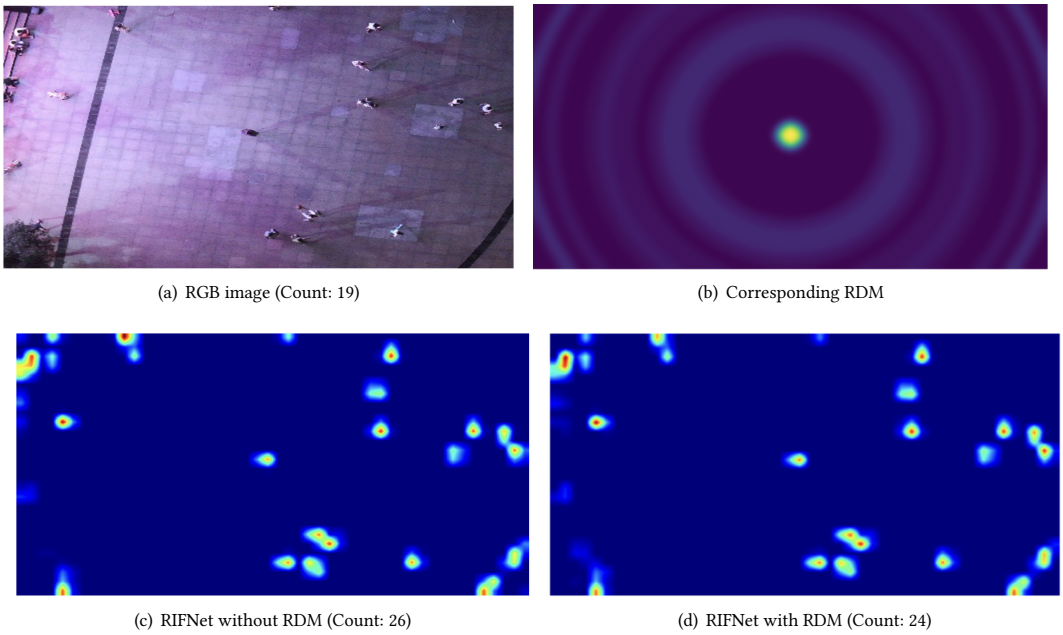
Evaluations on Direct Fusion We also try to align the two inputs of our network by means of doing Hadamard product with raw RGB images and the RDM directly. The experiment results are shown in table 10.This suggests that if the input are not aligned with a proper method, the additional RSS data might somehow destroy the original features of the RGB images, making the performance even worse than that of the baseline.

4.6 Evaluations on Real Datasets

We further conducted experiments on a real UAV datasets, *i.e.*, VisDrone dataset [59], with the model trained on our synthetic datasets and SCAR as the image branch. Specifically, we first manually estimate the distance between the UAV and crowd on 100 randomly-selected images, and then use the real RSS data with the same distance as the corresponding RSS. Table 11 indicates that the proposed approach still achieves an obviously better result than the approach using only RGB images. Figure 7 further illustrates one qualitative result with the original image, corresponding RDM, and the results with and without RDM, which shows that the proposed approach can be directly applied to the real scene without any modifications.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose the novel structure of RDM-Image Fusion Network (RIFNet) to combine the information from RGB images and the RSS to generate a more accurate crowd density map. Specifically, we first align RSS with RGB images by using the communication model and propose RIFNet to effectively fuse the information. Moreover, a new loss function is proposed to consider the uncertainty from RSS and makes the prediction consistent with the received



(a) RGB image (Count: 19) (b) Corresponding RDM

(c) RIFNet without RDM (Count: 26) (d) RIFNet with RDM (Count: 24)

Fig. 7. Qualitative results on VisDrone dataset.

signals. Experimental results confirm that RIFNet can improve the overall performance. The proposed UAV-GCC dataset containing RGB-images and their corresponding RSS data is released as a public download. In the future, since the performance of crowd density estimation varies with different scenes, especially for signal-based methods, we plan to leverage the concept of meta learning for training RIFNet to quickly fit a particular scene. Moreover, we plan to study the problem of joint density map estimation by integrating the results from multiple UAVs for a large-scale density map estimation.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude for the kind support from Shuo Tang. We are also grateful to the National Center for High-performance Computing for computer time and facilities. This work is supported in part by the Ministry of Science and Technology (MOST) of Taiwan under the grants MOST-109-2223-E-009-002-MY3, MOST-110-2634-F-007-015, MOST-110-2634-F-009-021, MOST-109-2218-E-002-015, MOST-109-2221-E-009-114-MY3, MOST-110-2218-E-A49-018, MOST-109-2221-E-009-097 and MOST-109-2221-E-001-015.

REFERENCES

[1] Jeffrey G. Andrews, Stefano Buzzi, Wan Choi, Stephen Hanly, Angel Lozano, Anthony C. K. Soong, and Jianzhong Charlie Zhang. 2014. What Will 5G be? *IEEE Journal on Selected Areas in Communications* 32, 6 (2014), 1065–1082.

[2] Anas Basalamah. 2016. Automatic update of crowd and traffic data using device monitoring. US Patent 9,401,086.

[3] Jack Bresenham. 1965. Algorithm for Computer Control of a Digital Plotter. *IBM Systems Journal* 4, 1 (1965), 25–30.

[4] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, , and Fei Su. 2020. Scale Aggregation Network for Accurate and Efficient Crowd Counting. In *Proceedings of the European Conference on Computer Vision (ECCV).*

[5] Antoni B. Chan and Nuno Vasconcelos. 2012. Counting People With Low-Level Features and Bayesian Regression. *IEEE Transactions on Image Processing* 21, 4 (April 2012), 2160–2177.

[6] Jiwei Chen, Wen Su, and Zengfu Wang. 2020. Crowd counting with crowd attention convolutional neural network. *Neurocomputing* 382 (2020), 210 – 220. https://doi.org/10.1016/j.neucom.2019.11.064

[7] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G. Hauptmann. 2019. Learning Spatial Awareness to Improve Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision (ICCV)*. 6151–6160.

[8] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, Jun-Yan He, and Alexander G. Hauptmann. 2019. Improving the Learning of Multi-Column Convolutional Neural Network for Crowd Counting. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. 1897–1906.

[9] Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 886–893.

[10] Simone Di Domenico, Mauro De Sanctis, Ernestina Cianca, and Giuseppe Bianchi. 2016. A Trained-Once Crowd Counting Method Using Differential WiFi Channel State Information. In *Proceedings of the 3rd International on Workshop on Physical Analytics (WPA)*. 37–42.

[11] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2012. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 34, 4 (April 2012), 743–761.

[12] Huiyuan Fu, Huadong Ma, and Hongtian Xiao. 2014. Crowd Counting via Head Detection and Motion Flow Estimation. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM)*. 877–880.

[13] Junyu Gao, Qi Wang, and Yuan Yuan. 2019. SCAR: Spatial-/Channel-Wise Attention Regression Networks for Crowd Counting. *Neurocomputing* 363 (October 2019), 1–8.

[14] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. 2019. DADNet: Dilated-Attention-Deformable ConvNet for Crowd Counting. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. 1823–1832.

[15] Marcus Handte, Muhammad Umer Iqbal, Stephan Wagner, Wolfgang Apolinarski, Pedro José Marrón, Eva María Muñoz Navarro, Santiago Martinez, Sara Izquierdo Barthelemy, and Mario González Fernández. 2014. Crowd Density Estimation for Public Transport Vehicles. In *Proceedings of the International Conference on Extending Database Technology/International Conference on Database Theory (EDBT/ICDT)*. 315–322.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2014. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *Lecture Notes in Computer Science* (2014), 346–361. https://doi.org/10.1007/978-3-319-10578-9_23

[17] Yaocong Hu, Huan Chang, Fudong Nian, Yan Wang, and Teng Li. 2016. Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation* 38 (2016), 530–539. https://doi.org/10.1016/j.jvcir.2016.03.021

[18] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. 2013. Multi-source Multi-scale Counting in Extremely Dense Crowd Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2547–2554.

[19] Haroon Idrees, Khurram Soomro, and Mubarak Shah. 2015. Detecting Humans in Dense Crowds Using Locally-Consistent Scale Prior and Global Occlusion Reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 37, 10 (2015), 1986–1998.

[20] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, and Y. Pang. 2020. Attention Scaling for Crowd Counting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4705–4714.

[21] Wahab Khawaja, Ismail Guvenc, David Matolak, Uwe-Carsten Fiebig, and Nicolas Schneckenburger. 2019. A Survey of Air-to-Ground Propagation Channel Modeling for Unmanned Aerial Vehicles. *IEEE Communications Surveys Tutorials* 21, 3 (2019), 2361–2391.

[22] Mehmet Kemal Kocamaz, Jian Gong, and Bernardo R. Pires. 2016. Vision-based counting of pedestrians and cyclists. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–8. https://doi.org/10.1109/WACV.2016.7477685

[23] C. Lai, L. Wang, and Z. Han. 2019. Data-Driven 3D Placement of UAV Base Stations for Arbitrarily Distributed Crowds. In *2019 IEEE Global Communications Conference (GLOBECOM)*. 1–6. https://doi.org/10.1109/GLOBECOM38437.2019.9014210

[24] Wei-Cheng Lai, Zi-Xiang Xia, Hao-Siang Lin, Lien-Feng Hsu, Hong-Han Shuai, I-Hong Jhuo, and Wen-Huang Cheng. 2020. Trajectory Prediction in Heterogeneous Environment via Attended Ecology Embedding. In *Proceedings of the ACM International Conference on Multimedia*.

[25] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. 2015. Crowded Scene Analysis: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* 25, 3 (2015), 367–386. https://doi.org/10.1109/TCSVT.2014.2358029

[26] Yuhong Li, Xiaofan Zhang, and Deming Chen. 2018. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1091–1100.

[27] Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao. 2019. Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Chuanbin Liu, Hongtao Xie, Zhengjun Zha, Lingyun Yu, Zhineng Chen, and Yongdong Zhang. 2020. Bidirectional Attention-Recognition Model for Fine-Grained Object Classification. *IEEE Transactions on Multimedia* 22, 7 (2020), 1785–1795. https://doi.org/10.1109/TMM.2019.2954747

[29] Qinghui liu, Michael Kampffmeyer, Robert Jenssen, and Arnt-Børre Salberg. 2019. Dense Dilated Convolutions Merging Network for Semantic Mapping of Remote Sensing Images. In *Proceedings of Joint Urban Remote Sensing Event (JURSE)*. 1–4.

[30] Weizhe Liu, Krzysztof Maciej Lis, Mathieu Salzmann, and Pascal Fua. 2019. Geometric and Physical Constraints for Drone-Based Head Plane Crowd Density Estimation. 244–249.

[31] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. 2019. Context-Aware Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5094–5103.

[32] Xiyang Liu, Jie Yang, and Wenrui Ding. 2020. Adaptive Mixture Regression Network with Local Counting Map for Crowd Counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[33] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. 2020. Semi-Supervised Crowd Counting via Self-Training on Surrogate Tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation. arXiv:1411.4038 [cs.CV]

[35] Yu-Jen Ma, Hong-Han Shuai, and Wen-Huang Cheng. 2021. Spatiotemporal Dilated Convolution with Uncertain Matching for Video-based Crowd Estimation. *IEEE Transactions on Multimedia* (2021), 1–1. https://doi.org/10.1109/TMM.2021.3050059

[36] Zheng Ma and Antoni B. Chan. 2013. Crossing the Line: Crowd Counting by Integer Programming with Local Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2539–2546.

[37] Yaik Ooi, Kong Zan Wai, Ian Tan, and Ooi Boon Sheng. 2016. Measuring the Accuracy of Crowd Counting Using WiFi Probe-Request-Frame Counting Technique. *Journal of Telecommunication, Electronic and Computer Engineering* 8 (January 2016), 79–81.

[38] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2017. Spatial As Deep: Spatial CNN for Traffic Scene Understanding. arXiv:1712.06080 [cs.CV]

[39] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. 2009. Crowd Counting Using Multiple Local Features. In *Proceedings of the Digital Image Computing: Techniques and Applications (DICTA)*. 81–88.

[40] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. 2017. Switching Convolutional Neural Network for Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4031–4039.

[41] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. 2019. Revisiting Perspective Information for Efficient Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7279–7288.

[42] Kyosuke Shibata and Hiroshi Yamamoto. 2019. People Crowd Density Estimation System Using Deep Learning for Radio Wave Sensing of Cellular Communication. In *Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. 143–148.

[43] Vishwanath A. Sindagi, Rajeev Yasarla, Deepak Sam Babu, R. Venkatesh Babu, and Vishal M. Patel. 2020. Learning to Count in the Crowd from Limited Labeled Data. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[44] Chon Hou Sio, Yu-Jen Ma, Hong-Han Shuai, Jun-Cheng Chen, and Wen-Huang Cheng. 2020. S2SiamFC: Self-supervised Fully Convolutional Siamese Network for Visual Tracking. In *Proceedings of the ACM International Conference on Multimedia*.

[45] Russell Stewart, Mykhaylo Andriluka, and Andrew Yan-Tak Ng. 2016. End-to-End People Detection in Crowded Scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[46] Gordon L. Stüber. 2017. *Principles of Mobile Communication* (4th ed.). Springer, Cham.

[47] Xin Tan, Chun Tao, Tongwei Ren, Jinhui Tang, and Gangshan Wu. 2019. Crowd Counting via Multi-Layer Regression. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. 1907–1915.

[48] Yukun Tian, Yiming Lei, Junping Zhang, and James Ze Wang. 2019. PaDNet: Pan-Density Crowd Counting. *IEEE Transactions on Image Processing* 29 (November 2019), 2714–2727.

[49] Haijun Wang, Haitao Zhao, Weiyu Wu, Jun Xiong, Dongtang Ma, and Jibo Wei. 2019. Deployment Algorithms of Flying Base Stations: 5G and Beyond With UAVs. In *IEEE Internet of Things Journal*, Vol. 6. 10009–10027.

[50] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. 2019. Learning from Synthetic Data for Crowd Counting in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8198–8207.

[51] Shuheng Wang, Hanli Wang, and Qinyu Li. 2019. Multi-Dilation Network for Crowd Counting. In *Proceedings of the ACM Multimedia Asia (MMAsia)*. 1–6.

[52] Zi-Xiang Xia, Wei-Cheng Lai, Li-Wu Tsao, Lien-Feng Hsu, Chih-Chia Hu Yu, Hong-Han Shuai, and Wen-Huang Cheng. 2020. Human-Like Traffic Scene Understanding System: A Survey. *IEEE Industrial Electronics Magazine* (2020).

[53] Peng Yu, Wenjing Li, Fanqin Zhou, Lei Feng ans Mengjun Yin, Shaoyong Guo, Zhipeng Gao, and Xuesong Qiu. 2018. Capacity Enhancement for 5G Networks Using MmWave Aerial Base Stations: Self-Organizing Architecture and Approach. *IEEE Wireless Communications* 25, 4 (August 2018), 58–64.

[54] Anran Zhang, Xiaolong Jiang, and ad Xianbin Cao Baochang Zhang. 2020. Multi-Scale Supervised Attentive Encoder-Decoder Network for Crowd Counting. *ACM Transactions on Multimedia Computing, Communications, and Applications Article 28* 16, 1 (April 2020).

[55] H. Zhang, L. Song, and Z. Han. 2020. *Unmanned aerial vehicle applications over cellular networks for 5G and beyond.* Springer.

[56] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 589–597.

[57] Zhaoxiang Zhang, Mo Wang, and Xin Geng. 2015. Crowd counting in public video surveillance by label distribution learning. *Neurocomputing* 166 (Oct. 2015), 151–163.

[58] Rui Zhou, Xiang Lu, Yang Fu, and Mingjie Tang. 2020. Device-free crowd counting with WiFi channel state information and deep neural networks. *Wireless Networks* 26 (02 2020). https://doi.org/10.1007/s11276-020-02274-7

[59] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling. 2020. Vision Meets Drones: Past, Present and Future. arXiv:2001.06303