IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 23, 2021, accepted January 13, 2022, date of publication January 20, 2022, date of current version January 28, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3144607

FSC-Set: Counting, Localization of Football Supporters Crowd in the Stadiums

OMAR ELHARROUSS[1], (Member, IEEE), NOOR ALMAADEED[1], KHALID ABUALSAUD[1], (Senior Member, IEEE), SOMAYA AL-MAADEED[1], (Senior Member, IEEE), ALI AL-ALI[2], AND AMR MOHAMED[1], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, Qatar University, Doha, Qatar
[2]Supreme Committee for Delivery and Legacy, Doha, Qatar

Corresponding author: Omar Elharrouss (elharrouss.omar@gmail.com)

This work was supported by research grant from Supreme Committee for Delivery and Legacy (SC), Qatar, under Grant QUEX-CENG-SCDL-19/20-1.

ABSTRACT Counting the number of people in a crowd has gained attention in the last decade. Due to its benefit to many applications such as crowd behavior analysis, crowd management, and video surveillance systems, etc. Counting crowded scenes, like stadiums, represents a challenging task due to the inherent occlusions and density of the crowd inside and outside the stadiums. Finding a pattern to control thousands of people and counting them is a challenging task. With the introduction of Convolutional Neural Networks (CNN), enables performing this task with acceptable performance. The accuracy of a CNN-based method is related to the size of data used for training. The availability of the dataset is sparse. In particular, there is no dataset in the literature that can be used for training applications for crowd scene. This paper proposes two main contributions including a new dataset for crowd counting, and a CNN-based method for counting the number of people and generating the crowd density maps. The proposed dataset for Football Supporters Crowd (FSC-Set) is composed of 6000 annotated images (manually) of different types of scenes that contain thousands of people gathering in or around the stadiums. FSC-Set contains more than 1.5 Million individuals. The collected images are captured under varying Fields of Views (FOV), illuminations, resolutions, and scales. The proposed dataset can also be utilized for other applications, such as individual's localization and face detection as well as team recognition from supporter images. Further, we propose a CNN-based method named FSCNet for crowd counting exploiting context-aware attention, spatial-wise attention, and channel-wise attention modules. The proposed method is evaluated on our established FSC-Set and other existing datasets then compared to state-of-the-art methods. The obtained results show satisfactory performances on all the datasets. The dataset is made publicly available and can be requested using the following link: https://sites.google.com/view/fscrowd-dataset/

INDEX TERMS Crowd counting, football supporters crowd, density map, crowd management.

I. INTRODUCTION

The analysis of data is a challenging task due to vast growth of the amount of data in the majority of domains, especially when considering the analysis of data related to video technologies which is typically associated with requiring large communication, computation, storage, and transmission [1]. Also, the achieved development in video surveillance techniques makes the analysis of the data stored difficult [2]. Storing lengthy videos without useful meta-data that describes the context and semantics of these videos may limit their utilization in modern applications [3]. Thus, the extraction of meaningful and interesting information from videos represents a key major task [4]. The extracted features can be used to interpret surveilled scenes. Researchers have carried a substantial amount of work to detect the pertinent information, according to the purpose and the analyzed situations, from the visualized video.

The associate editor coordinating the review of this manuscript and approving it for publication was Juntao Fei.

FIGURE 1. Images from FSC dataset.

Automatic analysis of surveillance scenes can help security agents to overcome many critical cases. The growth of gatherings on different occasions such as sports events, religious events, and festivals, increases the importance and complexity of their surveillance task [5], [6]. For that, crowd analysis becomes a hot computer vision topic and a challenging task. The management of the crowd behavior, as well as the recognition and the prediction of the abnormal event using an automated system, can help the security agent efficiently manage the crowd, which is an integral part of public security [7]. With the new technologies including smart cameras and sensors, the analysis of the crowd becomes easier [8]. Using deep learning techniques, the learning from large-scale datasets increased the performance of such a system [9]. In addition, it enables covering various aspects such as abnormal activity detection and recognition, crowd motion analysis, crowd counting, and crowd activity learning.

For sports activity, the analysis of the crowd is essential for ensuring the security of people and for easy management. Also, the sports events are the most occasions where the people gather In the stadiums, the estimation of a number of people, detection of the existing face, recognition of team supporters can help the stadium managers. In the literature, there is no crowd dataset for sports fans that can be used by researchers. For that, this work proposes a dataset of the football supporters' crowds as well as a technique of counting this crowd using a deep learning method. The dataset is composed of thousands of images collected and then annotated. In addition, the images are classified by team, which makes it usable for recognizing the team from the image of supporters. It contains more than 1.5 M annotations represent the number of people in all the images. Some images from the dataset are shown in Figure 1. This paper also proposes a CNN-based architecture for crowd counting. Also, some existing methods have been trained on the present dataset as well as on the existing datasets in the literature. The summarization of the contributions in this work is described as following:

- Large-scale Football Supporters Crowd dataset (FSC-Set) is collected and annotated. FSC-set contains 6000 images including empty scenes (no people and no crowd), and more than 1.5 million annotated instances. FSC-Set images are collected by team and can be used for teams recognition from supporter images.
- FSCNet crowd counting method is proposed. FSCNet consists of a CNN-based architecture which is based on context-aware attention, spatial-wise attention and channel-wise attention modules.
- Ten existing methods have been trained and tested on FSC-Set. The obtained results have been compared with the FSCNet results on the same dataset.
- The proposed crowd counting method has been tested on other existing datasets including shngaiTech_(A, B), UCF_QNRF, and UCF_CC_50.

The paper sections are organized as follows. The existing dataset and related works are presented in section 2. Section 3 describes the collected dataset. While the proposed crowd counting method is presented in section 4. The obtained results and discussion of it are provided in Section 5. A conclusion is presented in section 6.

TABLE 1. CNN-based methods for face anti-spoofing detection.

Type	Dataset	Year	# images	Scenes
Survaillance	UCF CC50 [54]	2013	50	Public
	WorldExpo10 [55]	2015	3 980	Street
	Shanghai Tech Part A [49]	2016	482	Events
	Shanghai Tech Part B [49]	2016	716	Streets
	UCF-QNRF [56]	2018	1 535	Religious event
	Crowd surveillance [57]	2019	13 945	Variant
	NWPU-Crowd [58]	2020	5 109	Diffrent Events
Drone	DroneCrowd [60]	2019	33 600	Streets
	GCC [29]	2019	15 212	Synthetic scenes
	VisDrone-CC [61]	2020	3367	Steeets

II. RELATED WORKS

Crowd Analysis such as crowd counting and management is a real issue for many computer vision applications including sports crowd surveillance and management, face detection in the crowd, and supporter behavior analysis. For that, Crowd analysis becomes one of the hot topics in computer vision. In this section, recent methods for crowd counting will be presented as well as the popular datasets.

A. CROWD COUNTING METHODS

Crowd counting is the operation of estimating the number of people or objects in a surveillance scene. For people counting in the crowd, many works have been proposed for estimating crowd mass. Which can be divided into many categories such as regression-based methods, density estimation-based methods, detection-based methods, and deep-learning-based methods. Comparing the accuracy of each one of these categories the CNN-based methods are the most effective methods. For that, we focus on this category for presenting the existing approaches.

The introduction of deep learning techniques makes the computer vision tasks more effective and the Convolutional Neural Network (CNN) improves the performance accuracy of each task specialty those used large-scale datasets. On crowd counting, the use of deep learning techniques makes the estimation of crowd density more accurate comparing with the traditional and sequential method in terms of accuracy and the computational cost [10], [11]. Also, CNN-based methods can estimate density maps and localize the pedestrians in the scene, unlike the regression-based methods. However, crowd density estimation in complex scenes still a challenging task due to the variations of scale, shape, and location of people.

Many methods attempted to handle scale variations for an accurate estimation of crowd density. To do that, Zou *et al.* [12] based on contextual dependencies for re-calibrate multiple scale-associated information. In [13] the authors used a fusion-based technique by the analysis of band-pass and roiling guidance stages for handling the scale variation to count the crowded mass. In the same context, some researchers implemented a CNN-based model exploiting VGG-16 backbone for features extraction followed by a crowd density estimation block [14]. While the authors in [15] proposed a multi-task method for crowd counting as well as localization of the position of each person in the scene using bi-branch CNN model. For [16], [17], [22], the algorithms consist of segmenting the crowd region then estimating the crowd density. For estimating the crowd density maps, but this time on Drone images, the authors in [18] proposed a CNN-based method that used warp features on warped images to estimate the density maps. For an accurate estimation, the labeling deviations should be handled. For that, a dilated-based model has proposed two networks named Density Attention Network (DANet) and Attention Scaling Network (ASNet) to estimate density maps. While the authors in [20] used a scale-attention-based model to estimate the crowd density after segmenting the crowd regions in the image.

Backbones and interconnection between the parts of a network have an impact on the accuracy of a CNN model. Different backbone including VGG-16, VGG-19, ResNet-50, and others has been used in different crowd counting models, but the most used backbone for crowd counting is VGG-16. The use of these backbones can increase the computational cost, especially on large-scale datasets. In order to reduce the number of parameters and the size of a network, the authors in [21] proposed a lightweight generation network method named Structured Knowledge Transfer (SKT) using two modules: teacher module that used Intra-Layer Pattern Transfer and student exploited Inert-Layer Relation Transfer. In another research paper, the authors used MobileNetV2 backbone to reduce the FLOPs and implemented a Light-weight encoder-decoder crowd counting model [23].

In addition to the scale variation, we can find that the object intensity, as well as the density of the crowd, can affect the performance of the proposed methods. To handle this, the authors in [24] proposed a crowd counting method named DENet composed of two-stage networks: detection network DNet and estimation network ENet. Detection network DNet count the people in each region and the estimation network ENet work on the complex and crowded regions in the image. DENet used VGG-16 as backbone for the feature extraction stage. Using the same backbone another method has been proposed named CANNet for estimating the crowd density map [25]. Also in [26] the authors based on the contextual and spatial information of the image to propose the crowd counting model. The method named SCAR consists of a Spatial-wise attention module and a Channel-wise Attention module before combining the results of each module for the final estimation. Using another version of VGG family which is VGG-19, the authors in [28] proposed a method based on density probabilities construction and Bayesian loss function to estimate the crowd density maps. In the same context, the authors in [29] proposed a crowd counting dataset as well as a crowd counting method based on a special FCN model. The proposed crowd counting methods achieved good results using different deep learning architectures, but for the complex scenes, the performance of these methods needs more improvement.

In order to handle the scale variations problem that represents a challenge for crowd counting methods, the

TABLE 2. FSC dataset comparing with the existing datasets.

Dataset	Size	Resolution	Annotations per images			Crowding degree
			Min	Max	Total	
UCF_CC-50 [54] 2013	50	2101×2888	94	4,543	63,974	Congested
WorldExpo10 [55] 2015	3 980	576×20	1	253	199,923	Medium
ShanTech Part_A [49] 2016	482	589×868	33	501	241,677	Congested
ShanTech Part_B [49] 2016	716	768×1024	9	578	88,488	Medium
UCF-QNRF [56] 2018	1 535	2013×2902	49	12,865	1,251,642	Congested
Crowd surveillance [57] 2019	13 945	1342×840	-	-	386,513	Congested
NWPU-Crowd [58] 2020	5 109	2191×3209	0	20,033	2,133,375	Congested
FSC-Set 2020	**6 000**	**660 × 340 to 4106 × 2727**	**0**	**10,200**	**+1.5M**	**Congested**

authors in [31] proposed a scale-driven-CNN-based method (SD-CNN) that consists of detecting the heads with different scales based on a scale map. The scale map is developed by annotating the heads in the images then mapping the head sizes. In the same context, the authors in [32] proposed two architecture for detecting the heads in an image including sparse-scale-CNN that detects the heads then dense-scale-CNN that generates the scale-map. Also for handling the scale variations for crowd counting, the authors in [33] proposed a multi-scale convolutional module and self-attention residual network that are fused for generating the crowd density map.

The crowd counting methods counted the number of people in a scene by generating the density maps using annotated datasets for training. In order to use this dataset, for reconstructing the density map using the image generation method, the authors in [34] proposed a domain-adaptive crowd counting (DACC) which is an image translation and density map reconstruction method. Another method proposed in [35] consists of using a small part of the dataset while the model have been trained on the density map (localization-level annotations) and a part of the dataset while just the number of people in the images is used for training the model (count-level annotations, like the images classification labeled by the number of people in the images). The method proposed two models while the first one is weak-supervised and the second is full-supervised. Image translation by density map generation is used also in [36]. the authors proposed a neuron linear transformation (NLT) network to predict the density map then estimate the number of dots in the map for estimating the crowd number.

B. CROWD COUNTING DATASETS

Detecting and understanding a specific object in a specific region using visual analysis is more difficult when the objects are away from the camera when the scene is crowded. This is caused by many factors like occlusion between objects in the scene, specific objects that can be represented with a few pixels, the variations of poses and appearances of the objects, the clothing, and the orientation of the camera. For crowd analysis, the same challenges are considered, due to the high density of the crowd where some human bodies are occluded partially that produce a miss-classification. Because of that, the researchers use the faces for crowd analysis (face detection in the crowd, crowd counting), according to the fact that the faces are usually been captured by high altitude surveillance cameras.

Crowd counting is an essential part of crowd analysis, the traditional methods find it difficult to count the people gathering in a surveilled scene. With the introduction of deep learning techniques, people counting in the crowd becomes easier and the methods become more accurate due to the learning method using neural networks. The main need of any good training is the dataset, which should generally be large-scale. Many datasets have been proposed for this purpose. In this section, the existing dataset will be presented by describing each dataset. Also a summarization of each dataset characteristics is presented in Table 2.

In the literature, the existing crowd counting datasets have different categories like real-world or synthetic with many datasets for people or vehicle counting. We can find also datasets taken from surveillance cameras and others taken using drones. In this paper, we will focus on the frequently used real-world datasets for people counting captured from surveillance cameras as well as those taken using drones.

One of the oldest and challenging crowd counting datasets is the public **UCF_CC-50 dataset [54]**. UCF_CC-50 dataset contains many scenes with different densities. The dataset composed of just 50 annotated images which makes the counting and the learning from it very difficult, which demonstrate the inability of most methods to estimate the number of the crowd. The Mean Absolute Error (MAE) and Mean Squared Error (MSE), obtained using these methods are very far from the optimal.

In the same context, **WorldExpo'10 [55]** is a large-scale dataset captured in different scenes during Shanghai 2010 WorldExpo. The dataset is collected using 106 surveillance cameras and contains 1132 videos with annotations which represents 199,923 persons. It is composed of 3920 frames with a resolution of 576 × 720.

Another famous dataset that widely used for crowd counting is **Shanghai-Tech [49]**. It is composed of two parts Part_A and Part_B. Part_A represents the images that contain crowded scenes with different distributions collected from the Internet. While Part_B contains the image captured by cameras in a street in Shanghai. The total number of images in the two parts is 1198 images with 330,165 annotations while Part_A contains more images than Part_B. The number of images in each part is not enough for training the deep learning method which obligates the data augmentation process.

Another crowd counting dataset which is a large-scale dataset named **UCF-QNRF [56]** is composed of 1535 images of many scenes including Hajj, which is a religious event where a very big mass of people gather. The dataset contains annotated images of more than 1 million annotations while the images captured from different angles with different resolutions. The UCF-QNRF is a good dataset for deep learning methods which needs high-performance machines for training.

A surveillance-based dataset for crowd counting is proposed in [57] which is a large-scale dataset of 13,945 images of more that 300K of annotated persons with high resolution images. Also, the authors provide an block annotations with the number of people in the specific regions in the images.

NWPU-Crowd [58] is the most recent dataset for crowd counting. The dataset is composed of 5109 images with about 2 million people annotated. The dataset also contains negative samples as well as different image resolutions and large appearance variations.

For the crowd counting datasets captured using a drones, the content of images can be different from the surveillance dataset [59]. The differences can be in terms of the depth of the objects in the images and the angle of view which is usually from high altitude for the drone data.

One of drone datasets for crowd counting in literature is named **DroneCrowd** [60]. The dataset consists of 112 videos collected using multiple drone devices with 33 600 frames with different image resolutions which are on average 1920×1080. Along with that it contains more than 4 million annotated persons. In addition to crowd counting, this dataset consists of crowd localization.

Another dataset is GTA5 Crowd Counting **(GCC) [29]** is a drone-based crowd counting dataset. The frames in this dataset are collected from an electronic game, Grand Theft Auto V (GTA5) and is composed of 15,212 images with 7,625,843 annotated persons. The images' resolution is 1080×1920 which represents a larger data volume.

VisDrone2020 Dataset [61] is acrowd dataset while the images are captured using drones. The dataset composed 3390 images of 113 video each video sequence contains 30 images. the resolution of the images is 1920×1080. In VisDrone2020, 2430 images of 82 video sequences are annotated, and the rest of the images without annotations.

III. FOOTBALL SUPPORTERS CROWD (FSC-SET) DATASET DESCRIPTION

In this section, we present the proposed Football Supporters crowd dataset (FSC-Set). A detailed description will be performed, including data collection, annotation, split in evaluation, the challenges as well as the computer vision tasks that can use the proposed dataset.

A. DATA COLLECTION

The FSC-Set images are collected from the Internet, Facebook Ultras pages, Google, and several football teams web sites. All the images are taken from public groups and websites. The collected images represent the most famous football teams in the world from a global perspective. Each team has more than 50 images. The dataset is composed of 20 football team supporters which can be used also for football team recognition from the images of the supporters. These images represent the typical crowded scenes in the stadiums, where the supporters hold many objects like flags and are with different appearances like colored clothes and painted faces. It also contains images from different points of view with different densities. All these characteristics make counting a challenging task.

B. DATA ANNOTATION

In order to annotate the crowd images, the head point has been used by taking the coordinate of each head in each image. Using Matlab, the points for each image are saved in a mat file (.mat). For that, and in order to have all the existing heads in the images including the depth ones with different scales, we zoomed in/out all regions of image during annotations process.

C. DATASET CHARACTERISTICS

Football Supporters Crowd (FSC) dataset is the only dataset available for sports supporters inside and outside stadiums. FSC-Set consists of 6000 images of different sizes and is composed of more than 1.5M annotated persons. Compared with the crowd counting dataset, the proposed dataset contains images from different angles of view, variant scales, various depth from the images as well as the number of person in an image. In addition to the annotated crowd in the scenes, the FSC-Set dataset contains 200 empty images without any person in the scene. The empty scenes can help the model to learn from the empty scene which can be similar to crowd scene in terms of texture features.

In terms of image resolutions, FSC-Set is composed of different image sizes starting from 660×340 to 4106×2727. Various images size can be a challenge for crowd counting model that allows to estimate the crowd from any scene. In terms of the number of people in an image, it ranges from 2 to 4000 person in the scene without considering the empty scenes. This allows high variations of the appearances within the dataset images. In addition, in every crowd image, we can find some person represented by 4 pixels, while in some other images we can find a head presented by more than half of the region of the image.

D. FSC-SET CHALLENGES

The performance of such crowd counting methods even deep-learning-based approaches can be affected by many factors that represent a crucial challenge. These factors can include scale variations related to the small targets, the fast motion of some objects, weather changes like cloud and rain, and complex background while the image contains the crowd of people and many other objects.

In addition to the cited challenges, the propped dataset has other challenges that we can't be found in the other crowd

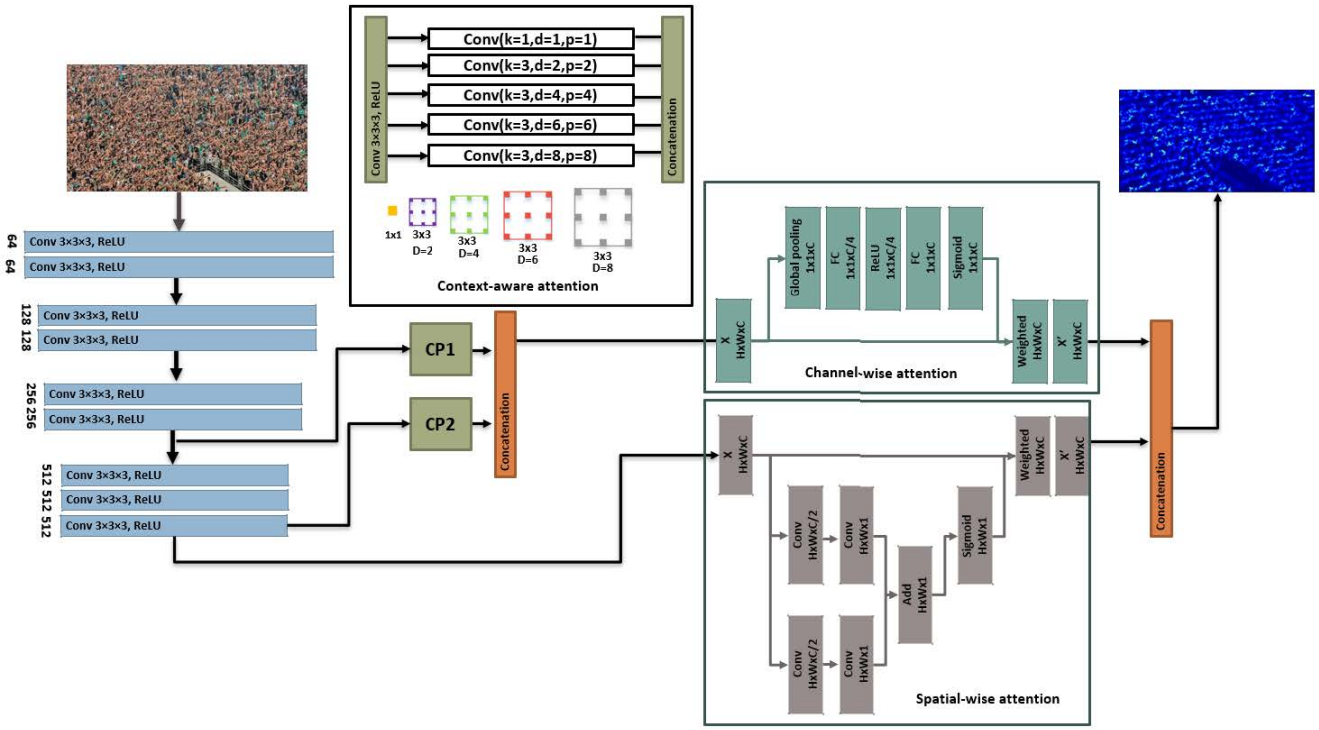

FIGURE 2. Proposed architecture model.

counting datasets (the existing dataset contains images of the crowd with different scales but similar in the other aspects). In this dataset, additional information and objects can be found, unlike the other dataset, which can also represent a challenge using FSC-Set dataset:

- **Flares**: the burning of flares can produce an immense amount of smoke that can fog the monitored scenes outside or inside the stadiums. consequently, reduce the number of people estimated using a crowd counting method.
- **Flying flag**: the flying flags in a monitored scene can produce a miss estimation of many people in it. Also, some flags can contain many objects that can be estimated as a person which can affect the real number of people in the crowd.
- **Skin color**: the skin color of supporters is another type of added information in this dataset and not in the other dataset. It can affect the performance of a method due to the type of colors contained in the training images as represented in the first images in figure 1.
- **Type and color of worn clothes**: FSC-Set contains supporters from all countries around the world. So, the variations of types and colors of clothes can also affect the estimation accuracy.
- **Painted faces**: All crowd counting datasets are annotated using faces and/or heads. The variations of the face color can prevent an algorithm or a method to find the pattern. Painted faces also present a challenge for the crowd counting method to detect these faces for estimating the density map due to the colors of paint that can be the same as the color of the clothes or the flags.

The presented challenges can encourage researchers to proposes some techniques to overcome it for accurate crowd counting.

E. OTHER USES OF THE DATASET

During the collection of the images of the dataset, we attempted to collect the images by teams. More than 20 teams of different leagues From Europe, South America, Africa, and Asia. The national football teams in the datasets include FC Barcelona, Real Madrid CF, Boca Juniors CA, Borussia Dortmund, Liverpool FC, Everton FC, Manchester City FC, AC Milan, Arsenal FC, Celtic FC, Inter Milan, Juventus FC, Manchester United FC, Olympique Marsilya, Paris Saint-Germain FC, AS Monaco, AS Roma, Raja CA, Wydad AC, Al-Ahli Saudi FC. For that, the dataset can be used for teams recognition from supporter's images. In addition, the dataset can be used for testing the face detection method due to the fact that the current version of the dataset does not include face annotations. In addition, we collect the others supporter's images in another file that includes different images from international teams during the big leagues like the world cup, Euro-cup, Copa America, and African cup.

IV. PROPOSED CROWD COUNTING NETWORK

Counting the people in the crowd can be difficult with respect to the scale, the shape variations, and the depth of desired

TABLE 3. Implementation details.

Method	Hardware	Processing speed	Language	Optimizer	Lr	Momentum	Decay	Epochs
FSCNet	Google Colab Pro	25 GB RAM	PyTorch	SGD	1e-7	0.95	5×10^{-3}	50

objects [37], [38]. Many works attempted to handle these problems by proposing different techniques.

To work on crowd images that contain different scales, shapes, and depth levels, each proposed method performs one of these challenges to extract pertinent features for an effective estimation of the crowd. In this paper, we attempted to work on low-level, high-level, shape, and scale of the data using three modules including Context-aware attention, channel-wise attention (high-level feature maps), and spatial-wise attention (low-level feature maps) modules. For that, VGG-16 backbone is used to extract the initial features from the original images to guide the entire network to learn from many features. The scale problem has been handled through the use of a context-aware attention module by taking the output of different blocks of the backbone (VGG-16 last block for CP2 and before the last VGG-16 block for CP1). Also, each CP module is composed of adopted Atrous convolutions with a variation of dilation rate and a pending parameter that takes into consideration the different shapes and scales in the image. The global network is presented in detail in Figure 2.

A. CONTEXT-AWARE ATTENTION MODULE

Analyzing the context is the way to extract, recognize segment or classify the content of an image or a video. To do that, the existing deep learning architectures combine multiple convolutional and pooling layers to extract features for good learning like in [11]. For crowd counting, the people in the scene can be under the variations of shape and scale. The combination of the convolutional and pooling layers cannot be effective for handling all these variations. The authors in [51] attempted to implement a module inspired by SIFT feature extraction method [52] that can extract the features based on the shape, scale, and locations of the objects. For that atrous convolution [53] is used to extract the features of the same scale. For multi-scale features, the use of the output of each convolutional-pooling block can be taken. The same strategy is used for implementing the context-aware pyramid (CP) module shown in Figure 2. Two CP blocks are used for scale-shape-based extraction. The first one takes the output of the third VGG-16 block as input of the first CP1 module while the second one CP2 takes the output of the last block of the backbone. Each CP module is composed of adopted atrous convolutions with a variation of dilatation rate and pending parameters. The outputs of each atrous convolution are concatenated. Then the two CP modules are combined to get features maps of Context-aware pyramid module then used as input of the next channel-wise attention module.

B. CHANNEL-WISE ATTENTION MODULE

A channel-wise attention module is a channel-based attention module for fully convolutional neural networks. The purpose of the channel is to extract the important features of the input image with a feature detector that corresponds to each channel in the feature map.

The spatial dimension of the input function map was compressed to measure the channel attention efficiently. As shown in Figure 2, First, by using global pooling, space information of a map was aggregated and this generated different spatial context descriptors to obtain a channel-wise feature vector v. To capture channel-wise dependencies two fully connected layers (FCs) are used. in order to limit the complexity of the module, we encode the channel-wise feature vector by forming a bottleneck with two FC layers around the non-linearity. Then, through using sigmoid operation, we take the normalization processing measures to the encoded channel-wise feature vector mapped to [0,1]. The channel-wise (CA) is expressed as follows:

$$CA = \sigma(f_c(r(f_c(v, W_0), W_1))) \tag{1}$$

where σ indicates the sigmoid function. W refers to parameters in channel-wise attention block, f_c denotes FC layer, and r is ReLU activation function

C. SPATIAL-WISE ATTENTION MODULE

A Spatial-wise Attention Module is a spatial attention module for fully conventional neural networks. It produces a spatial care map by the use of the inter-space function relationship. Unlike the attention of the channel, the focus of spatial attention is where an information component complements the attention of the channel. We first apply the average pooling and max pooling operations along the channel axis and concatenate them to produce an effective characteristic descriptor for calculating spatial attention. In order of get global information as well as increasing receptive field, two convolutions with two kernels $1\times k$ and $k\times 1$, then, using sigmoid.

$$C_1 = conv_2(conv_1(f, W_0), W_1) \tag{2}$$
$$C_2 = conv_1(conv_2(f, W_0), W_1) \tag{3}$$
$$SA = \sigma(C_1 + C_2) \tag{4}$$

where σ denotes a function sigmoid, $conv_1$ and $conv_2$ refers to $1\times k\times C$ and $k\times 1\times 1$ convolution layer respectively, and W refers to parameters in spatial-wise attention block.

D. TRAINING DETAILS

In order to train the proposed dataset as well as reducing the size of data all images for all datasets used for training the proposed method are resized to 768×1024 pixels. Also, the

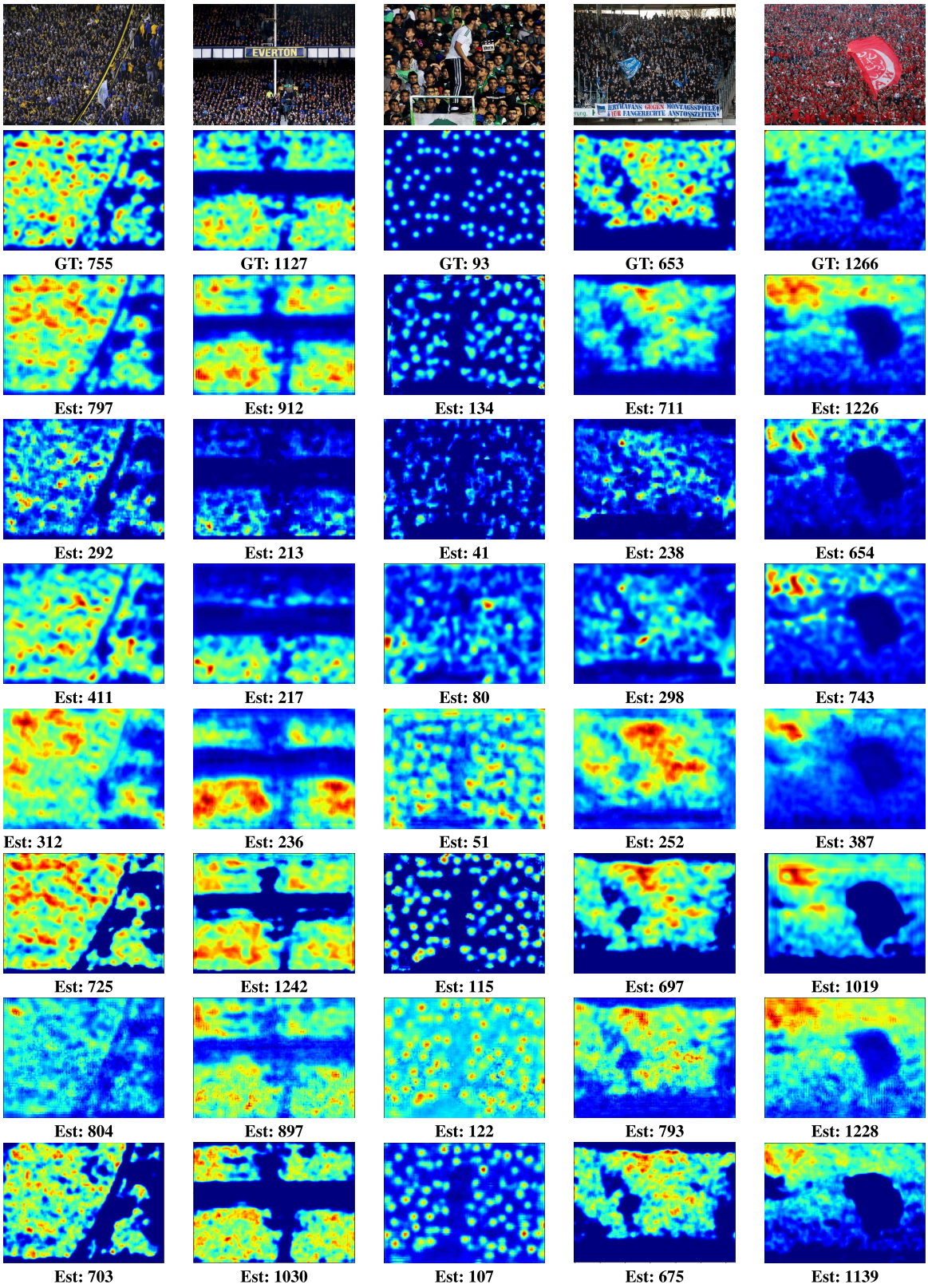

FIGURE 3. Comparisons of estimated density maps and estimated crowd number between the proposed method and the other methods. (first row) input image. (second row) groundtruth. (third row) CSRNet. (fourth) MCNN. (fifth row) DENet. (sixth row) SKT. (seventh row) SCAR. (eighth row) CANNet. (last row) FSCNet.

TABLE 4. Estimation errors for each method on FSC dataset. The **bold** and underline fonts respectively represent the **first** and second and the third place.

Method	Backbone	MAE	MSE
MCNN [49](2016)	-	85.719	131.001
SANet [50](2018)	MCNN	105.634	236.957
CSRNet [10] (2018)	VGG-16	69.635	164.067
SPN [11] (2019)	VGG-16	125.296	270.835
SKT [21](2020)	FS	88.348	193.662
MobileCount [23] (2020)	MobileNetV2	106.026	242.765
DENet [24] (2020)	VGG-16	89.374	134.593
ASNet [20] (2020)	Xception	79.398	127.134
CANNet [25](2019)	VGG-16	51.251	75.082
SCAR [26] (2019)	VGG-16	53.274	79.158
FSC-Net baseline	VGG-16	**47.309**	**59.140**

TABLE 5. Quality of density map on FSC dataset. PSNR and SSIM metrics before and after smoothing operation are shown for each method. The **bold** and underline fonts respectively represent the **first** and second place.

	Original		After smoothing	
Method	PSNR	SSIM	PSNR	SSIM
CSRNet [10] (2018)	17.501612	0.440301	18.050703	0.441067
SPN [11] (2019)	15.490336	0.271204	15.740604	0.261996
SKT [21](2020)	15.699434	0.175745	15.895486	0.166666
MCNN [49](2019)	15.151169	0.552638	15.306079	0.559340
MobileCount [23] (2020)	14.426194	0.106462	14.684645	0.100336
DENet [24] (2020)	17.337937	0.257780	17.787568	0.245685
CANNet [25](2019)	17.146045	0.493663	17.005465	0.387988
SCAR [26] (2019)	18.843281	**0.588591**	19.096763	**0.587974**
FSCNet	**19.17281**	0.573461	**19.372466**	0.561179

data are split into a training set of 80% of data, a validation set of 10%, and a testing set of 10% of data for each dataset. The FSC-Net method is implemented and tested using Google Colab Pro which uses Python and PyTorch, with 25 GB of RAM, and High-GPU option. The implementation and training details are illustrated in Table 3.

In order to compute the distance between the ground-truth and the estimated density maps, we used Euclidean loss used in many methods [10], [11], [40] and defined by:

$$L(\Theta) = \frac{1}{2N}\sum_{i=1}^{N} ||Z(X_i;\Theta) - Z_i^{GT}||_2^2 \tag{5}$$

where X_i is the input image. Θ denote the learning parameters used in the proposed model. Z_i^{GT} is the density map ground truth of X_i. $Z(X_i;\Theta)$ refers to the estimated density map of X_i. L denotes the loss between ground truth and the estimated density and. N is the total number of images.

V. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method on the proposed dataset as well as on the existing datasets, we compared the obtained results with the state-of-the-art methods while the code is available. The list of methods used in the comparison are: CSRNET [10], SPN [11], ASNet [20], MCNN [49], SANet [50], CANNet [25], SCAR [26], MobileCount [23], SKT [21], and DENet [24]. The comparison is performed using quantitative and qualitative results of all methods including the results of the proposed FSCNet.

A. EVALUATION METRICS

In order to measure the effectiveness of each method including the proposed method, Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics are used. The two metrics are defined in [10] by the following expressions:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |z_i - z_i^{gt}| \tag{6}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |z_i - z_i^{gt}|^2} \tag{7}$$

where N is total number of images used in testing, z_i^{gt} represents the real crowd count, and z^i denotes the estimated number of people in the crowd.

B. EVALAUTION AND DISCUSSION

The performance of each method on the proposed dataset is performed using both quantitative and qualitative results are presented in this section. For presenting the quantitative results, MAE and MSE metrics are used. For the qualitative results, we exploited PSNR and SSIM metrics for each method including the proposed architecture. A visualization of some examples, from the FSC dataset, is presented with the estimated number of the crowd as well as the density maps.

C. QUANTITATIVE RESULTS

To demonstrate the quality of crowd numbers obtained by each method, MAE and MSE are used. Table 4 present these metrics for all methods on the FSC dataset. Also, in Figure 3, we illustrated some examples from the FSC dataset by presenting the estimated crowd number as well as the density map for each method. From the presented table and figure, we can observe that CANNet [25], SCAR [26] and the proposed methods FSCNet gives the most accurate results. The obtained MAE values are 51.251, 53.274, and 47.309 for CANNet, SCAR, and FSCNet respectively. Also, FSCNet results outperform MobileCount method by 59 points and CSRNet by 22 points for MAE. For the other method like CSRNet, it can be considered among the methods that give convincing results in terms of MAE and MSE by 69.635 and 164.067 respectively. In addition, these methods are the best for counting the very crowded scenes as well as for the less crowded scenes like in Figure 3 (third column) or (fifth column). We can observe that the most used backbone for crowd counting methods is VGG-16 and the obtained results using these methods are the most accurate ones. For MobileCount method that used MobileNetV2 as a backbone for features extraction, the number of parameters is less than the other method also the computational cost can be less also, but the accuracy of density map estimation is not robust comparing with the other methods.

TABLE 6. The performance of each method on the existing crowd counting dataset. The **bold** and underline fonts respectively represent the **first** and second place.

Method	ShanTech_A		ShanTech_B		UCF_QNRF		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CSRNet [10] (2018)	68.2	115.0	10.6	16.0	-	-	266.1	397.5
SPN [11] (2019)	61.7	99.5	9.4	14.4	-	-	259.2	335.9
MCNN [49](2019)	110.2	173.2	26.4	41.3	277	426	377.6	509.1
CANNet [25](2019)	62.3	100.0	**7.8**	**12.2**	107	183	212.2	243.7
SCAR [26] (2019)	66.3	114.1	9.5	15.2	-	-	259.0	374.0
SS-CNN+SD-CNN [31] (2019)	-	-	-	-	-	-	235.7	345.6
SKT [21](2020)	62.73	102.33	7.98	13.13	96.24	156.82	-	-
MobileCount [23] (2020)	84.8	135.1	8.6	13.8	127.7	216.5	284.5	421.2
DENet [24] (2020)	65.5	101.2	9.6	15.4	-	-	241.9	345.4
SS-CNN [32] (2021)	-	-	-	-	115.2	175.7	229.4	325.6
Liu et al. [33] (2021)	64.4	100.2	8.4	13.4	-	-	242.3	320.4
DACC [34] (2021)	112.4	176.9	13.1	19.4	203.5	343.5	-	-
MATT [35] (2021)	80.1	129.4	11.7	17.5	-	-	355.0	550.2
NLT [36] (2021)	91.4	153.4s	10.4	18.8	165.8	279.7	-	-
FSCNet	**55.6**	**92.753**	8.25	17.79	**90.8**	**131.5**	**194.3**	**226.6**

The visualized results in Figure 3, demonstrate the obtained accuracies in Table 4 and confirm that the method with fewer MAE values has the most similar density maps to the ground-truth and the min errors in the visualized examples. For example, as demonstrated in the table, we can find from the visualized results that SCAR and proposed method results are the most similar to the ground-truth, but For SCAR method, the number of people estimation is more than the real number like in the second row, third row, and fourth row. For the CANNet method, even the MAE values in the Table are good but the estimated density maps are not similar well to the ground-truth. For the other methods, as we can see, the MobileCount method is the worst in terms of estimated density map quality. Some methods are tested in terms of MAE and MSE, but their density maps are not good enough and so it's not illustrated in Figure 3.

D. QUALITATIVE RESULTS

In order to evaluate the quality of the counting results produced by the density map, PSNR, and SSIM metrics have been used. To the fact that a good density map means a good estimated number of the crowd, a smoothing operation is exploited for enhancing the generated density maps which implicate an enhancement of the estimated number of people in the crowd. Table 5 presents the obtained PSNR and SSIM results for each method as well as the proposed method. From the table, we can find that the best results in terms of PSNR and SSIM confirm the results obtained in Table 4 which make the SCAR and FSCNet the best methods that can estimate the crowd number with convincing results and with a density map most similar to the ground-truth compared with other methods. The density maps generated using CANNet are not good enough comparing to the obtained MAE and MSE. While CSRNet gives good results but with less precision.

From the table also, we can observe that the estimated results are convincing but not perfect according to the complexity of the dataset. In addition, the number of people in each image varies from 2 to more than 4000 people. Also, the collected images are taken from different points of view with also different resolutions that make the estimation of crowd numbers with perfect precision difficult. This can be shown in Figure 4 which illustrates many scenes with a different number of people in each scene. From the figure, we can notice that for some images like the second row the right image contains 2471 person and the people are far from the camera which make the estimation difficult. Even this, FSCNet the number of estimated people for this images reached 2310 which is a good estimation according to the small scale of the people in this images. Another example, the third image in the first row contains 106 persons which are few comparing with the other images, but the proposed method succeeds to estimate the number of people in this image which has an error of 2 people. Also, the quality of the estimated density map is similar to the ground-truth. From all these results we can conclude that FSCNet can estimate the density of a crowd with a minimum error even with different scales and shape also if the images are very crowded or less crowded.

E. EVALUATION ON EXISTING DATASETS

Besides the evaluation of the proposed method FSCNet on the proposed dataset FSC-Set, an evaluation on the existing datasets is performed to demonstrate the effectiveness of FSCNet compared with the state-of-the-art methods such as MCNN, CSRNet, SNP, MobileCount, SKT, DENet, CANNet, and SCAR using MAE and MSE metrics. The existing datasets used in this evaluation are:ShanTech_Part_A [49], ShanTech_Part_B [49], UCF_QNRF [56], and UCF_CC_50 [54].

The obtained results using MAE and MSE metrics are presented in Table 6. From this table, we can observe that many methods succeed to estimate the number of

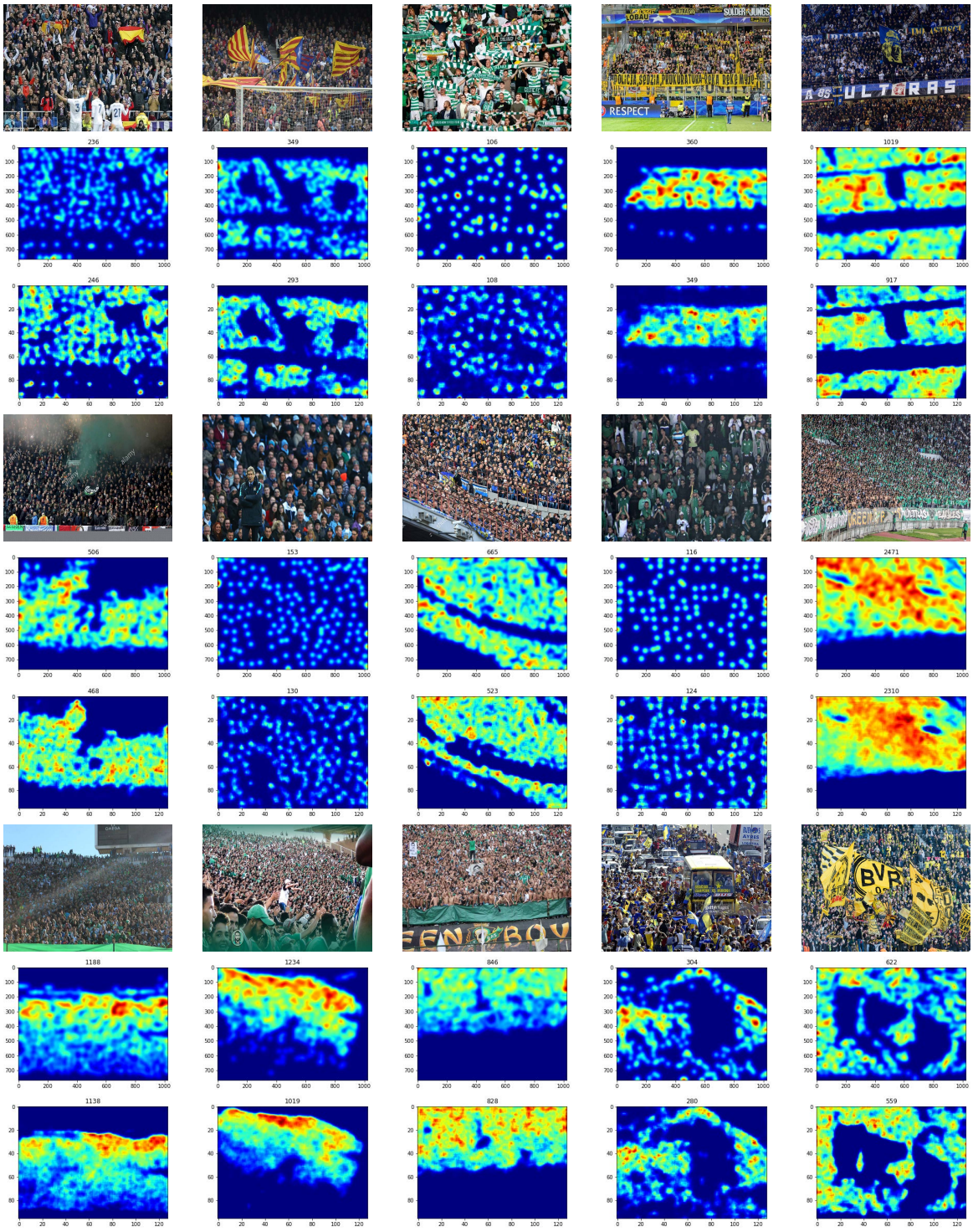

FIGURE 4. Some example of the estimated density map Using FSCNet method on FSC-Set. First row represent the original image. Second row represent the groudtruth. Third row represent the estimated density map.

people in the crowd with promising results especially for ShanTech_Part_B dataset due to simple crowd density in this dataset and all the images contains the same depth of the crowd and the same distribution of the people in the scenes. We can find also that the ShanTech_Part_A comes in second place in terms of MAE reached due to the same reasons of ShanTech_Part_B but here the images are more crowded than the images in ShanTech_Part_B. For the other datasets including UCF_QNRF, and UCF_CC_50, the images are more crowded which can reach 4000 people per image which

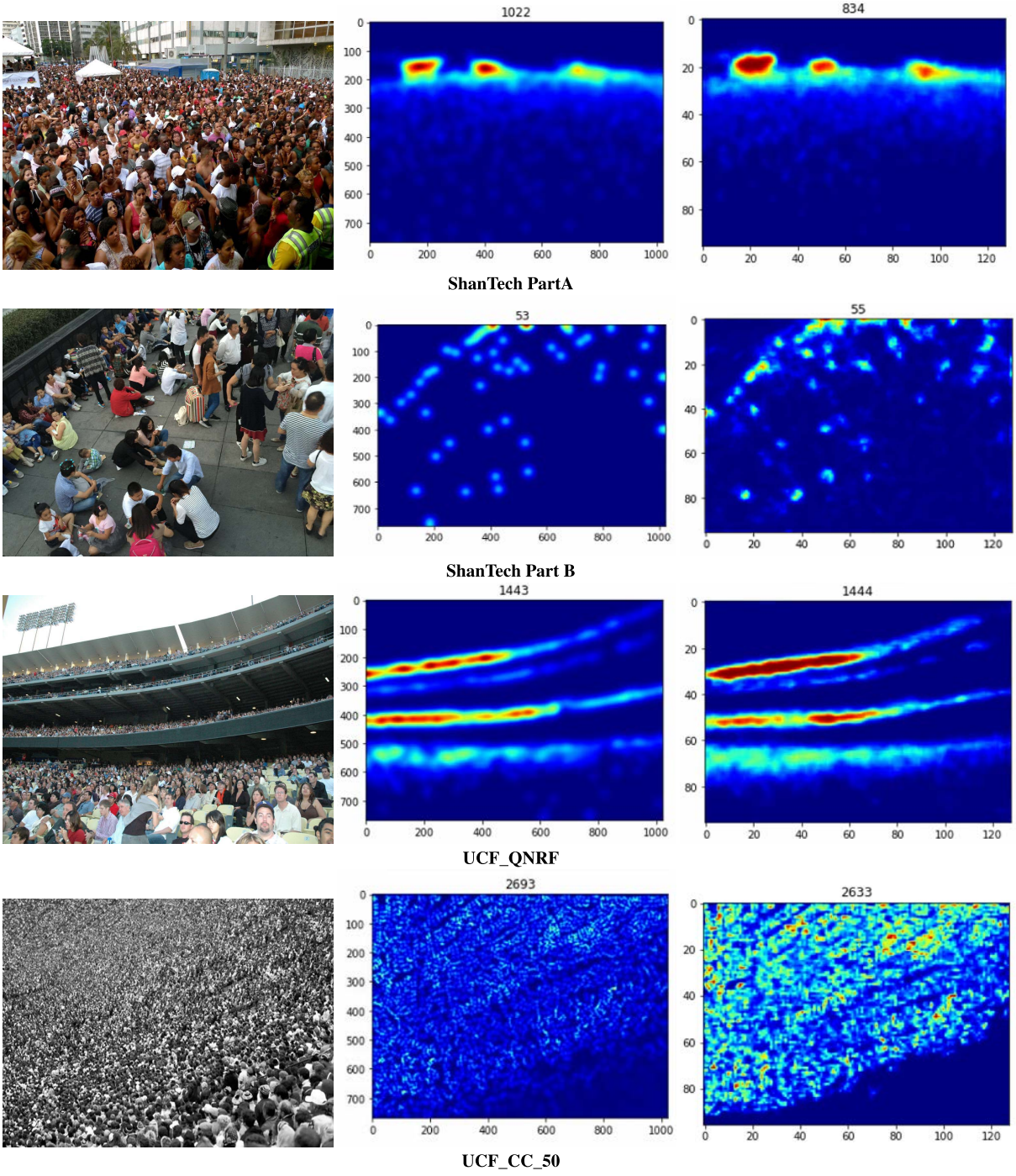

FIGURE 5. Comparisons of estimated density maps using the proposed method FSCNet on different datasets.

makes the estimation of the density maps more complex. Also, the scale and shape variations in these datasets affect the performance of each method.

For the obtained MAE and MSE of each method, the results in the table show that each method reaches good results in a dataset better than the others. And this comes from the treatment used for each method, for example, some methods are working on the scale variation while others used segmentation of the crowd region before estimating the crowd density. For the proposed FSCNet method, we can see that it outperform the others method in three datasets including ShanTech_Part_A, UCF_QNRF, and UCF_CC_50 with an MAE of 55.6 on ShanTech_Part_A and less by 6 points than the SPN method which comes in the second place. While we can find that the SPN, SKT CANNet method reached close results of MAE values. On UCF_QNRF, FSCNet

TABLE 7. Estimation errors for each method on FSC-Set dataset. The bold and underline fonts respectively represent the first and second and the third place.

Method	MAE	MSE
FSC-Net context-aware+channel-wise	70.833	95.691
FSC-Net: channel-wise+spatial-wise	71.400	101.377
FSC-Net: context-aware+spatial-wise	<u>65.729</u>	<u>87.385</u>
FSC-Net all	**47.309**	**59.140**

reached 90.8 as mean error rate better than SKT method of a difference of 6 points. Also For UCF_CC_50 dataset, FSC-Net achieved the min MAE results better of CANNet with 18 points. For these results, we can conclude that the proposed method is more effective compared with the other methods and this is due to the set of challenges that the proposed model tries to handle including scale-and shape variations with the use of spatial-wise and channel-wise as well as context-aware attention modules in the proposed architecture.

The obtained results are presented also by the crowd density maps for four datasets in Figure 5. From the figure, we can find that the proposed method can estimate the crowd number with a good quality of density maps.

F. ABLATION ANALYSIS

The fusion of different blocks, while each block performs a specific analysis, in a CNN-based model can improve the performance of a crowd counting method. The context-wise, spatial-wise, and channel-wise attention modules are used in the proposed crowd counting model. the use of just one of these modules is not like the use of all of them in terms of performance. This section, attempts to present the impact of using different modules. Table 7 represents the results of using each combination. From the obtained results, we can find that the use of all three modules outperforms the other results by a difference of more than 20 points. While the obtained MAE and MSE results using context-aware+channel-wise and channel-wise+spatial-wise are close with an MAE of 70.833 and 71.400 respectively. The results using context-aware+spatial-wise are more improved than the two previous by a difference of 5 points.

VI. CONCLUSION

In this paper, A Football supporters crowd FSC-Set dataset is proposed. The dataset is collected annotated as well as classified into classes that represent teams supporter. The collected images are with different resolutions, illuminations, appearances variation, various scales, and from different fields of view. In addition, we proposed a CNN-based model for crowd counting exploiting the VGG-16 and three attention modules. The proposed method is trained and tested on different datasets including the proposed dataset FSC-Set. The obtained results are improved for almost all datasets and the quality of density maps is satisfying compared with the state-of-the-art methods.

REFERENCES

[1] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," *J. Vis. Commun. Image Represent.*, vol. 77, May 2021, Art. no. 103116, doi: 10.1016/j.jvcir.2021.103116.

[2] O. Elharrouss, N. Al-Maadeed, and S. Al-Maadeed, "Video summarization based on motion detection for surveillance systems," in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jun. 2019, pp. 366–371.

[3] Y. Akbari, N. Almaadeed, S. Al-Maadeed, and O. Elharrouss, "Applications, databases and open computer vision research from drone videos and images: A survey," *Artif. Intell. Rev.*, vol. 54, no. 5, pp. 3887–3938, Jun. 2021.

[4] A. Abbad, O. Elharrouss, K. Abbad, and H. Tairi, "Application of MEEMD in post-processing of dimensionality reduction methods for face recognition," *IET Biometrics*, vol. 8, no. 1, pp. 59–68, Jan. 2019.

[5] B. Sheng, C. Shen, G. Lin, J. Li, W. Yang, and C. Sun, "Crowd counting via weighted VLAD on a dense attribute feature map," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1788–1797, Aug. 2016.

[6] K. Abualsaud, T. M. Elfouly, T. Khattab, E. Yaacoub, L. S. Ismail, M. H. Ahmed, and M. Guizani, "A survey on mobile crowd-sensing and its applications in the IoT era," *IEEE Access*, vol. 7, pp. 3855–3881, 2019.

[7] U. Sajid, H. Sajid, H. Wang, and G. Wang, "ZoomCount: A zooming mechanism for crowd counting in static images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3499–3512, Oct. 2020.

[8] H. Gayathri, P. M. Aparna, and A. Verma, "A review of studies on understanding crowd dynamics in the context of crowd safety in mass religious gatherings," *Int. J. Disaster Risk Reduction*, vol. 25, pp. 82–91, Oct. 2017.

[9] C. Celes, A. Boukerche, and A. A. F. Loureiro, "Crowd management: A new challenge for urban big data analytics," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 20–25, Apr. 2019.

[10] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[11] X. Chen, Y. Bin, N. Sang, and C. Gao, "Scale pyramid network for crowd counting," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1941–1950.

[12] Z. Zou, Y. Liu, S. Xu, W. Wei, S. Wen, and P. Zhou, "Crowd counting via hierarchical scale recalibration network," 2020, *arXiv:2003.03545*.

[13] Y. Chen, C. Gao, Z. Su, X. He, and N. Liu, "Scale-aware rolling fusion network for crowd counting," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[14] M. K. K. Reddy, M. A. Hossain, M. Rochan, and Y. Wang, "Few-shot scene adaptive crowd counting using meta-learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2814–2823.

[15] Y. Hou, C. Li, F. Yang, C. Ma, L. Zhu, Y. Li, H. Jia, and X. Xie, "BBA-NET: A bi-branch attention network for crowd counting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4072–4076.

[16] X. Kong, M. Zhao, H. Zhou, and C. Zhang, "Weakly supervised crowd-wise attention for robust crowd counting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2722–2726.

[17] X. Pan, H. Mo, Z. Zhou, and W. Wu, "Attention guided region division for crowd counting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2568–2572.

[18] Y. Yang, G. Li, Z. Wu, L. Su, Q. Huang, and N. Sebe, "Reverse perspective network for perspective-aware object counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4374–4383.

[19] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, "Adaptive dilated network with self-correction supervision for counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4594–4603.

[20] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, and Y. Pang, "Attention scaling for crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4706–4715.

[21] L. Liu, J. Chen, H. Wu, T. Chen, G. Li, and L. Lin, "Efficient crowd counting via structured knowledge transfer," 2020, *arXiv:2003.10120*.

[22] J. Gao, Q. Wang, and X. Li, "PCC Net: Perspective crowd counting via spatial convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3486–3498, Oct. 2020.

[23] P. Wang, C. Gao, Y. Wang, H. Li, and Y. Gao, "MobileCount: An efficient encoder-decoder framework for real-time crowd counting," *Neurocomputing*, vol. 407, pp. 292–299, Sep. 2020.

[24] L. Liu, J. Jiang, W. Jia, S. Amirgholipour, Y. Wang, M. Zeibots, and X. He, "DENet: A universal network for counting crowd with varying densities and scales," *IEEE Trans. Multimedia*, vol. 23, pp. 1060–1068, 2021.

[25] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5099–5108.
[26] J. Gao, Q. Wang, and Y. Yuan, "SCAR: Spatial-/channel-wise attention regression networks for crowd counting," *Neurocomputing*, vol. 363, pp. 1–8, Oct. 2019.
[27] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[28] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6142–6151.
[29] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8198–8207.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] S. Basalamah, S. D. Khan, and H. Ullah, "Scale driven convolutional neural network model for people counting and localization in crowd scenes," *IEEE Access*, vol. 7, pp. 71576–71584, 2019.
[32] S. D. Khan and S. Basalamah, "Sparse to dense scale prediction for crowd couting in high density crowds," *Arabian J. Sci. Eng.*, vol. 46, no. 4, pp. 3051–3065, Apr. 2021.
[33] Y.-B. Liu, R.-S. Jia, Q.-M. Liu, X.-L. Zhang, and H.-M. Sun, "Crowd counting method based on the self-attention residual network," *Appl. Intell.*, vol. 51, no. 1, pp. 427–440, 2021.
[34] J. Gao, T. Han, Y. Yuan, and Q. Wang, "Domain-adaptive crowd counting via high-quality image translation and density reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 15, 2021, doi: 10.1109/TNNLS.2021.3124272.
[35] Y. Lei, Y. Liu, P. Zhang, and L. Liu, "Towards using count-level weak supervision for crowd counting," *Pattern Recognit.*, vol. 109, Jan. 2021, Art. no. 107616.
[36] Q. Wang, T. Han, J. Gao, and Y. Yuan, "Neuron linear transformation: Modeling the domain shift for crowd counting," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 27, 2021, doi: 10.1109/TNNLS.2021.3051371.
[37] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. CVPR*, Jun. 2016, pp. 5525–5533.
[38] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 650–657.
[39] S. Lamba, N. Nain, and H. Chahar, "A robust multi-model approach for face detection in crowd," in *Proc. 12th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, 2016, pp. 96–103.
[40] B. Kneis, "Face detection for crowd analysis using deep convolutional neural networks," in *Proc. Int. Conf. Eng. Appl. Neural Netw.* Cham, Switzerland: Springer, Sep. 2018, pp. 71–80.
[41] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 643–650.
[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[43] S. Zhang, L. Wen, H. Shi, Z. Lei, S. Lyu, and S. Z. Li, "Single-shot scale-aware network for real-time face detection," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 537–559, Jun. 2019.
[44] L. Wang, X. Yu, T. Bourlai, and D. N. Metaxas, "A coupled encoder–decoder network for joint face detection and landmark localization," *Image Vis. Comput.*, vol. 87, pp. 37–46, Jul. 2019.
[45] C. E. Bencheriet, "New face features to detect multiple faces in complex background," *Evolving Syst.*, vol. 10, no. 2, pp. 79–95, Jun. 2019.
[46] O. Elharrouss, D. Moujahid, and H. Tairi, "Motion detection based on the combining of the background subtraction and the structure–texture decomposition," *Optik, Int. J. Light Electron Opt.*, vol. 126, no. 24, pp. 5992–5997, Dec. 2015.
[47] J.-M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and A. Dev, "Color and scale: The spatial structure of color images," in *Proc. 6th Eur. Conf. Comput. Vis.*, vol. 1, 2000, pp. 331–341.
[48] H. Zhou, Y. Chen, and R. Feng, "A novel background subtraction method based on color invariants," *Comput. Vis. Image Understand.*, vol. 117, no. 11, pp. 1589–1597, Nov. 2013.
[49] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 589–597.
[50] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.
[51] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.
[52] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[53] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[54] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. CVPR*, Jun. 2013, pp. 2547–2554.
[55] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. CVPR*, Jun. 2015, pp. 833–841.
[56] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. ECCV*, 2018, pp. 532–546.
[57] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding, "Perspective-guided convolution networks for crowd counting," in *Proc. ICCV*, Oct. 2019, pp. 952–961.
[58] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: A large-scale benchmark for crowd counting and localization," 2020, *arXiv:2001.03360*.
[59] O. Elharrouss, N. Almaadeed, K. Abualsaud, A. Al-Ali, A. Mohamed, T. Khattab, and S. Al-Maadeed, "Drone-SCNet: Scaled cascade network for crowd counting on drone images," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 6, pp. 3988–4001, Dec. 2021, doi: 10.1109/TAES.2021.3087821.
[60] L. Wen, D. Du, P. Zhu, Q. Hu, Q. Wang, L. Bo, and S. Lyu, "Drone-based joint density map estimation, localization and tracking with space-time multi-scale attention network," 2019, *arXiv:1912.01811*.
[61] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, and H. Ling, "Vision meets drones: Past, present and future," *CoRR*, vol. abs/2001.06303, pp. 1–20, Jan. 2020.
[62] V. Iglovikov, S. Mushinskiy, and V. Osin, "Satellite imagery feature detection using deep convolutional neural network: A kaggle competition," 2017, *arXiv:1706.06169*.
[63] V. Iglovikov, S. Seferbekov, A. Buslaev, and A. Shvets, "TernausNetV2: Fully convolutional network for instance segmentation," in *Proc. CVPR Workshops*, Jun. 2018, p. 237.
[64] S. R. Bulo, L. Porzi, and P. Kontschieder, "In-place activated BatchNorm for memory-optimized training of DNNs," 2017, *arXiv:1712.02616*.

OMAR ELHARROUSS (Member, IEEE) received the master's degree from the Faculty of Sciences, Dhar El Mehraz, Fez, Morocco, in 2013, and the Ph.D. degree from the LIIAN Laboratory, USMBA-Fez University, in 2017. His research interests include pattern recognition, image processing, and computer vision.

NOOR ALMAADEED received the bachelor's degree in computer science from Qatar University, in 2000, the M.Sc. degree in computer and information sciences from the City, University of London, U.K., in 2005, and the Ph.D. degree from Brunel University London, U.K., in 2014. She is currently an Assistant Professor with the Department of Computer Science and Engineering, Qatar University. Her areas of research interests include speech signal detection, speaker identification, and audio/visual speaker recognition. She was awarded the Qatar Education Excellence Day Platinum Award-New Ph.D. Holders Category, from 2014 to 2015.

KHALID ABUALSAUD (Senior Member, IEEE) is currently with the Department of Computer Science and Engineering, Qatar University, Qatar. He has more than 25 years of professional experience in information technology. He teaches courses in hardware and software systems. He is active in getting research funding from different sources, including the Qatar National Research Foundation, the Supreme Committee for Delivery and Legacy (FIFA'2022), and some other organizations in Qatar. His research work has been presented in international conferences and journals. He participated actively in organizing several IEEE international conferences in Qatar, namely, ICIoT2020, IEEE WCNC'2016, PLM'2015, AICCSA'2014, RelMiCS'2011, and AICCSA'2008. He is also a LPI of research project which achieved significant outcomes. His research interests include health systems, wireless sensors for IoT applications, cybersecurity, cloud computing, and computer network protocols. He has served as a technical program committee (TPC) member and the chair for various reputable IEEE conferences. He received several awards from different local and international organizations. Recently, he served as a Guest Editor in Connected Healthcare Special Issue for IEEE NETWORK. He is also an Associate Editor of *IET Quantum Communication* journal.

SOMAYA AL-MAADEED (Senior Member, IEEE) received the Ph.D. degree in computer science from Nottingham, U.K., in 2004. She is currently the Head of the Computer Science Department, Qatar University. She is also a Coordinator of the Computer Vision and AI Research Group, Qatar University. She enjoys excellent collaboration with national and international institutions and industry. She is a Principal Investigator of several funded research projects generating approximately five million. She has published extensively in pattern recognition and delivered workshops on teaching programming for undergraduate students. She attended workshops related to higher education strategy, assessment methods, and interactive teaching. In 2015, she was elected as the IEEE Chair for the Qatar Section.

ALI AL-ALI received the master's degree in business administration and management (general from executive master of business administration) from the Scandinavian Business School, Paris, France. He is currently the Deputy Executive Director of security at Supreme Committee for Delivery and Legacy. He has an experienced officer with a demonstrated history of working in the law enforcement industry. He has skilled in international relations, management, government, strategic planning, and police.

AMR MOHAMED (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electrical and computer engineering from The University of British Columbia, Vancouver, BC, Canada, in 2001 and 2006, respectively. He has worked as an Advisory IT Specialist at IBM Innovation Centre, Vancouver, from 1998 to 2007, taking a leadership role in systems development for vertical industries. He is currently a Professor with the College of Engineering, Qatar University. He has over 25 years of experience in IoT, edge computing, pervasive AI, and wireless networking research and industrial systems development. He has authored or coauthored over 200 refereed journals and conference papers, textbooks, and book chapters in reputable international journals and conferences. His research interests include wireless networking and edge computing for IoT applications. He holds three awards from IBM Canada for his achievements and leadership, and four best paper awards from IEEE conferences. He is serving as a technical editor for two international journals. He has served as a technical program committee (TPC) co-chair for many IEEE conferences and workshops.