Received February 22, 2021, accepted March 24, 2021, date of publication April 26, 2021, date of current version June 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074797

AVMSN: An Audio-Visual Two Stream Crowd Counting Framework Under Low-Quality Conditions

RUIHAN HU[1,2], QINGLONG MO[1], YUANFEI XIE[3], YONGQIAN XU[1], JIAQI CHEN[1], YALUN YANG[4,5], HONGJIAN ZHOU[6], ZHI-RI TANG[7], AND EDMOND Q. WU[4], (Member, IEEE)

[1]Guangdong Key Laboratory of Modern Control Technology, Guangdong Institute of Intelligent Manufacturing, Guangzhou 523419, China
[2]Xinjiang Production and Construction Corps Key Laboratory of Modern Agricultural Machinery, Shihezi University, Shihezi 832000, China
[3]Electronic Information School, Wuhan University, Wuhan 430072, China
[4]Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China
[5]School of Machinery and Automation, Wuhan University of Science and Technology, Wuhan 430081, China
[6]School of Mechanical and Electrical Engineering, Wuhan Institute of Technology, Wuhan 430070, China
[7]School of Physics and Technology, Wuhan University, Wuhan 430072, China

Corresponding authors: Qinglong Mo (ql.mo@giim.ac.cn), Zhi-Ri Tang (gerintang@163.com), and Edmond Q. Wu (edmondqwu@gmail.com)

This work was supported in part by the Guangdong Academy of Sciences' (GDAS') Project of Science and Technology Development under Grant 2017GDASCX-0115 and Grant 2018GDASCX-0115, in part by the Guangdong Academy of Science for the Special Fund of Introducing Doctoral Talent under Grant 2021GDASYL-20210103087, and in part by the Opening Foundation of Xinjiang Production and Construction Corps Key Laboratory of Modern Agricultural Machinery under Grant BTNJ2021003.

ABSTRACT Crowd counting is considered as the essential computer vision application that uses the convolutional neural network to model the crowd density as the regression task. However, the vision-based models are hard to extract the feature under low-quality conditions. As we know, visual and audio are used widely as media platforms for human beings to touch the physical change of the world. The cross-modal information gives us an alternative method of solving the crowd counting task. In this case, in order to solve this problem, a model named the Audio-Visual Multi-Scale Network (AVMSN) is established to model the unconstrained visual and audio sources for completing the crowd counting task in this paper. Based on the Feature extraction and Multi-modal fusion module, in order to handle the objects of various sizes in the crowd scene, the Sample Convolutional Blocks are adopted by the AVMSN as the multi-scale Vision-end branch in the Feature extraction module to calculate the weighted-visual feature. Besides, the audio, which is the temporal domain transformed into the spectrogram information and the audio feature is learned by the audio-VGG network. Finally, the weighted-visual and audio features are fused by the Multi-modal fusion module, which adopts the cascade fusion architecture to calculate the estimated density map. The experimental results show the proposed AVMSN achieves a lower mean absolute error than other state-of-art crowd counting models under the low-quality conditions.

INDEX TERMS Multi-scale architecture, audio-visual model, cascade fusion, crowd counting.

I. INTRODUCTION

Crowd counting is taken as the computer-vision task, which is used in various fields such as intelligent transportation [1], industrial manufacturing [2] and security systems [3]. Different from the other computer vision tasks such as image classification [4] and scene understanding [5] and so on, the crowd counting models equipped by the convolutional neural network (CNN) should recognize arbitrarily sized people in various situations, including scenes with the extreme conditions such as high-level noise, low-level illumination and high-level occlusion. Consequently, the performance of the vision-driven model can be easily broken and maybe not very appropriate to deal with the crowd counting problem under extreme conditions.

As for investigations in the field of neurobiology, humans mainly depend on ears and eyes to build their perception systems for listening to and looking at significant information such as lip reading [6] and reasoning [7] the world's wild environment. According to the neurobiology phenomenon,

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

for example, the louder sound denotes more people in the scene. The auditory information can be adopted as the auxiliary cue for counting the number of objects in one crowd scene. Meanwhile, the hardware devices such as smartphones, digital cameras and video surveillance equipment can provide a cheaper way to obtain audio-visual information.

In this paper, a simple audio-visual crowd counting model, namely the Audio-Visual Multi-Scale Network (AVMSN) is proposed to count the objects by multi-modal information. Unlike the traditional vision-based learning mechanism, the image and audio sources are incorporated into the Feature extraction module based on the two-stream learning framework, which consists of the Vision-end and the Audio-end branches to extract the features. For the Vision-end branch, in order to model the density map under various scene geometries, the multi-scale CNN formation is constructed to extract the feature from visual information. As for the Audio-end branch, the auditory information is converted into the spectrogram format and then incorporated into the Audio-VGG network to extract the spectrogram feature.

The above merits are attributed to the following three contributions of the AVMSN:

- the two-stream framework, which contains the Vision-end and Audio-end branches of the Feature extraction module, is adopted in the AVMSN to deal with the crowd count tasks.
- the Vision-end branch in the Feature extraction module uses the multi-scale CNN architecture constructed by the different sizes of the average pooling procedure to extract the weighted-visual feature with respect to the visual channel.
- the Multi-modal fusion module, which adopts the cascade fusion architecture, combines the convolution and full connection layers to fuse the weighted-visual and audio features for further calculating the estimated density map.

The rest of this paper is organized as follows: In Sec. II, the methodology of the AVMSN is described. In Sec. III, the configurations of network architecture for the AVMSN, contrast baselines and corresponding benchmarks are elaborated. Sec. IV shows the qualitative and quantitative experiment results when the AVMSN is applied to compare with the other contrast baselines under the low-quality conditions such as high-level noise, low-level illumination and high-level occlusion conditions. Finally, the conclusions and future work are given in Sec. V.

II. RELATED WORK

A. DETECTION-BASED COUNTING

The early works about counting mainly focus on the detection-based mechanisms such as the joint likelihood model [8], scale-invariant feature transform (SIFT) [9] and part-template tree [10] and so on. The approaches mentioned above tend to detect the targets and count them with hand-crafted features following the object detection mechanism. However, the performance of these detection-based methods always has limitations on the highly congested scene which contains the small-size and irregular objects.

B. CNN-BASED COUNTING

With the development of deep learning [11]–[14], the CNN-based methods [15]–[18] can transform the highly congested images into the density-estimation problem. The point-level labeling is adopted to annotate the objects, and the density map is generated following the Gaussian filter [17]. The convolution neural networks are utilized to extract the local features of the pixels and model the ground truth density maps to count the objects in images. For example, the Multi-column CNN (MCNN) [15] was put forward when using the cascade architecture of the CNN to regress the density map. The reverse perspective network (RPN) [16] was designed by using perspective estimators and coordinate transformers through the meta-learning for accomplishing the crowd counting task. The CODA [17] provided the alternative way which uses adversarial learning to match the predicted density map and the ground truth density map. Although the CNN-based methods can yield good performance on the objects that spread uniformly in images by the end-to-end modeling procedure, the performances of these models are severely influenced by some of the extreme conditions such as high noise, occlusion, and low illumination [18].

C. AUDIO-VISUAL MODEL

The audio-visual model is based on the two-stream framework that incorporates the audio and visual inputs respectively to deal with the computer vision tasks. Ephrat [19] used the visual modal as the auxiliary modal to localize the desired sound objects in videos. The PiexelPlayer [20] separated the spatial positions of the sound objects and freely adjusted the volume of each sound source at the pixel level. Gao [21] adopted the deep multi-instance multi-label (MIML) learning framework to construct the audio-visual model when dealing with the multiple labels for each time step to separate the sound sources. As for the recent researches [19]–[21] about the audio-visual modal information, they can provide us with the theoretical inspiration for dealing with the crowd counting problem.

III. METHOD

In this section, the methodology of the AVMSN is presented. Sec. III. A shows the computation framework for the AVMSN; the Sampling Convolutional blocks of the Feature extraction module and the Multi-modal module are elaborated in Sec. III. B and Sec. III. C respectively, while the loss function of the AVMSN is discussed in Sec. III. D.

A. ARCHITECTURE OF THE AVMSN

The two-stream framework [22], [23] is adopted by the Feature extraction module of the AVMSN, and audio-visual sources can be incorporated as the input channel for the AVMSN. The architecture of the AVMSN is depicted

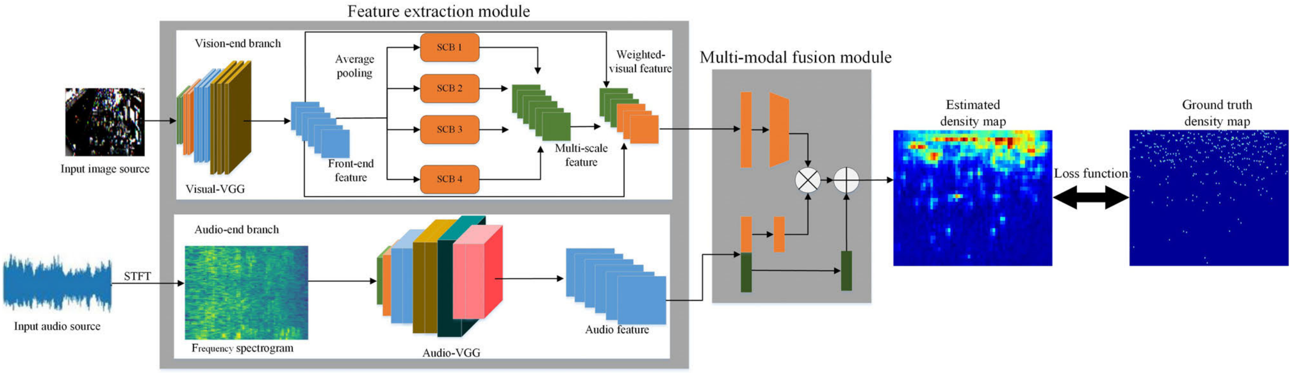

FIGURE 1. The architecture of the AVMSN. The Feature extraction and Multi-modal fusion modules are contained in the AVMSN to calculate the estimated density map. In the feature extraction module of the Vision-end branch, the weighted-visual feature is extracted following the Visual-VGG and Sampling Convolutional Blocks (SCBs) procedures. In the feature extraction module of the audio-end branch, the Audio-VGG net is employed to extract the audio feature. Finally, the weighted-visual feature and the audio feature are incorporated into the Multi-modal module to calculate the predicted density map.

in Fig. 1, from which it can be seen that there are Feature extraction and Multi-modal fusion modules contained in the AVMSN. The Feature extraction module consists of the Vision-end and Audio-end stream branches. As for the Vision-end branch, the front-end feature is extracted by the popular Visual-VGG [24] network. In this paper, the Visual-VGG contains the first 10 convolution layers of the VGG16 network benchmarked on the ImageNet [24] with 3 × 3 kernel sizes. The main motivation to use the VGG16 as the backbone is to limit the network complexity since the small sizes of convolution filters are used in all layers of VGG16. However, as for the architecture of Resnet [25] and Xception [26], the various sizes of the convolution filters such as 1*1, 3*3 and 5*5 are employed in the network, and the computation burden of the network is heavier than that of VGG16. Then, the front-end feature is incorporated into different blocks through several Sampling Convolutional Blocks (SCBs) to calculate the multi-scale feature. In order to retain more background information of the front-end feature, the average pooling layer is adopted. The SCBs are designed to furtherly extract the visual features following the multi-scale architecture from the front-end feature. The responsibility of the multi-scale architecture belonging to the SCBs is applied to deal with the scale, diversity and resolution issues between the different scenes for the crowd counting problem. After being processed by several SCBs, the front-end feature is concatenated and transformed into a multi-scale feature. After that, the multi-scale feature is linked with the front-end feature to calculate the multi-scale weights by the element-wise product operator. The multi-scale weights are added for the purpose of weighting the front-end feature to calculate the weighted-visual feature. The configuration for the Vision-end branch is shown in Tab. 1. Considering the Audio branch, the temporal domain of the audio sources can be transformed into the frequency domain of the audio sources following the Short-Time Fourier Transform (STFT) [27]. The audio feature is extracted following the 16-layer Audio-VGG network [28] that contains some 3 × 3 convolution kernel sizes and 3 full connection layers where hidden units are defined as 4096, 4096 and 128 in the Audio-end branch. The configuration for the Audio-end branch is shown in Tab. 2. In this paper, after being processed by the Audio-end branch, the obtained dimension of each audio feature is 128. Subsequently, the weighted-visual and audio features are incorporated into the Multi-modal fusion module. In the Multi-modal fusion module, the cascade fusion architecture is adopted to fuse the audio-visual feature and then output the estimated density map. Finally, the estimated density map can be optimized following the simple least square loss function.

B. SAMPLING CONVOLUTIONAL BLOCK

In order to make the AVMSN enable to handle the huge variation of the resolutions under the dense crowd scenes, the front-end feature from Visual-VGG [24] is fed to the 4 blocks belonging to the multi-scale architecture of the Sampling Convolutional Blocks that are adopted by the AVMSN. Supposing the visual sources and the front-end feature are defined as I and $\mathrm{f_g}$ respectively in this paper, the front-end feature can be calculated as follows:

$$\mathrm{f_g} = VGG(I) \tag{1}$$

In the actual scenario, one large size of the object that can cover the areas in density maps is equal to the summation of some small sizes of the objects. Hence, the high-density areas in the density maps do not denote more objects in the crowd counting scene. Inspired by the pyramid structure of the network [29], [30], it can be found that the Sampling Convolutional Block (SCB) is proposed in the AVMSN, and the different sizes of the average pooling layers are adopted to model the different sizes of the crowd in the front-end feature. In the SCB, the average pooling layer is adopted for the purpose of retaining more background information of the front-end feature. In this paper, the pooling template sizes of the average pooling layers are predefined as 1, 2, 3 and 6 regarding the 4 blocks of the SCBs. As shown in Fig. 2,

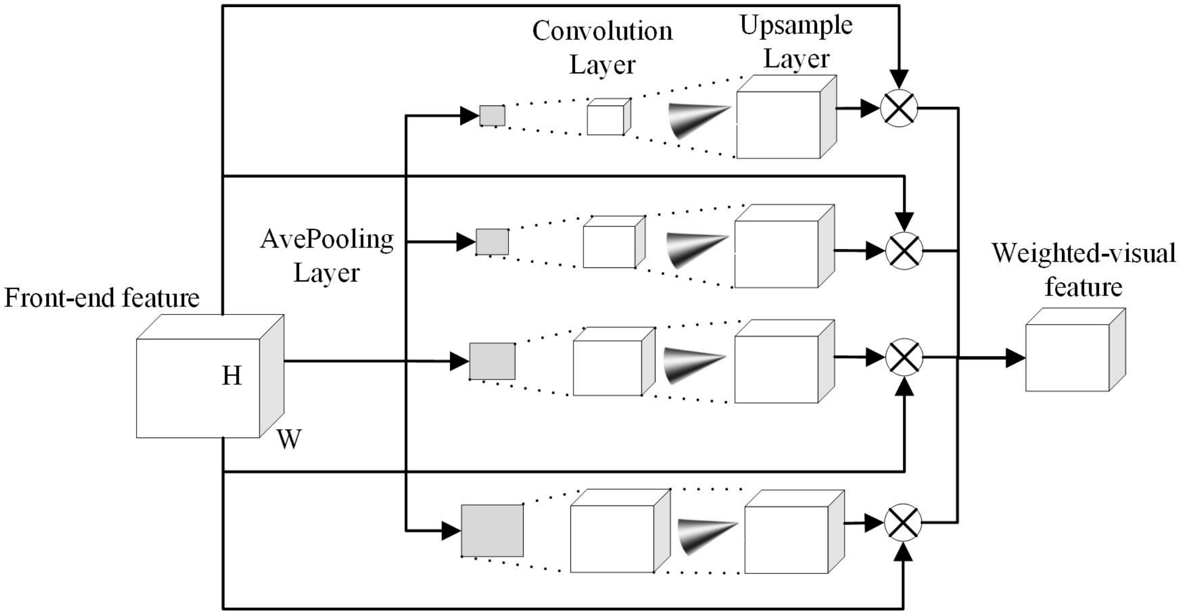

FIGURE 2. The architecture of the sampling convolutional blocks. The front-end feature from the Visual-VGG is processed by the multi-scale architecture which contains the average pooling and convolutional and upsample layers of four blocks. Then the weighting summation procedure is adopted by the SCB to calculate the weighted-visual feature.

TABLE 1. The configuration of the Vision-end branch that contains the Visual-VGG and several SCBs for the AVMSN.

<table>
<tr><th colspan="4">Vision-end branch</th></tr>
<tr><th colspan="4">Visual-VGG</th></tr>
<tr><td colspan="4">Conv2-64-3
Conv2-64-3
Max-Pooling2-2
Conv2-128-3
Conv2-128-3
Max-Pooling2-2
Conv2-256-3
Conv2-256-3
Conv2-256-3
Max-Pooling2-2
Conv2-512-3
Conv2-512-3
Conv2-512-3</td></tr>
<tr><th colspan="4">SCB(four blocks)</th></tr>
<tr><td>Ave-Pooling2-1</td><td>Ave-Pooling2-2</td><td>Ave-Pooling2-3</td><td>Ave-Pooling2-6</td></tr>
<tr><td>Conv2-512-1</td><td>Conv2-512-1</td><td>Conv2-512-1</td><td>Conv2-512-1</td></tr>
<tr><td colspan="4">Conv2-512-1
Conv2-512-1</td></tr>
</table>

TABLE 2. The configuration of the Audio-end branch that contains the Audio-VGG and full connection layers for the AVMSN.

Audio-end branch
Audio-VGG
Conv2-64-1
Max-Pooling2-2
Conv2-128-1
Max-Pooling2-2
Conv2-256-1
Conv2-256-1
Max-Pooling2-2
Conv2-512-1
Conv2-512-1
Max-Pooling2-2
Full connection layers
FC-4096
FC-4096
FC-128

it is supposed that the height and the width dimension of the front-end feature f_g are H × W and the pooling size of the average pooling layers L_{avp} for the i_{th} block of the SCB is p_i. After being processed by the L_{avp} in different SCBs, the height and width of the features become 1×1, 2×2, 3×3 and 6×6. The convolution layer L_{cov} with the convolutional size θ_i is applied to calculate these features following the L_{avp} to build the multi-scale architecture of the SCB. Then, the Upsample Layer L_{up} is adopted in the multi-scale architecture to restore the multi-scale feature f_i, which is kept as the same size as the front-end feature f_g. In this paper, the bilinear interpolation mechanism is adopted in the L_{up}. Hence, the f_i for different blocks of the SCBs can be denoted as follows:

$$f_i = L_{\text{up}}(L_{cov}(L_{avp}(f_g, p_i), \theta_i)) \tag{2}$$

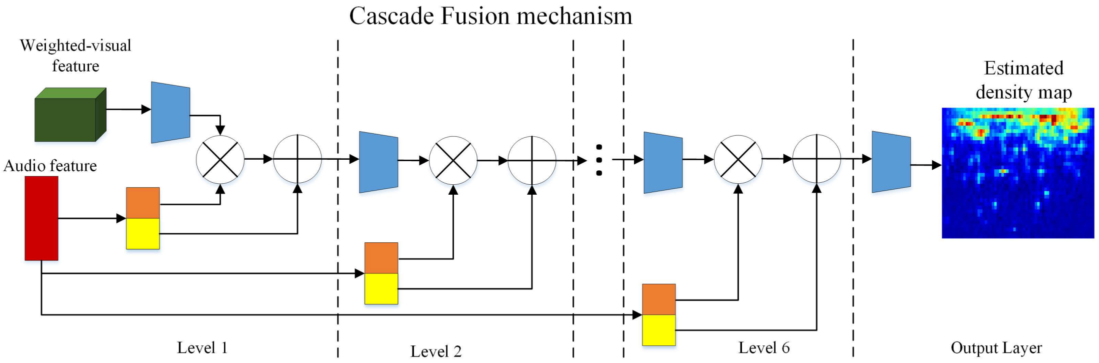

FIGURE 3. **The architecture of the Multi-modal fusion module. The architecture of the Multi-modal fusion module is constructed by the six levels of the cascade fusion architecture which contains the convolution and full connection layers for each level. The mechanism is applied to link the convolution and full connection layers to calculate the fusion feature. Finally, fusion feature is processed by the convolution layer to calculate the Estimated density map.**

Then the multi-scale weight $\mathrm{f_w}$ is calculated to model the difference between the position and their neighbors of the objects for the visual sources. In order to link the $\mathrm{f_g}$ and the f_i, the f_i from different blocks of SCBs is concatenated to calculate the $\mathrm{f_w}$ as shown below:

$$\mathrm{f}_{\mathrm{w}}^{(i)} = \mathrm{Sigmoid}(L_{\mathrm{cov}}(\mathrm{f}_i \otimes f_g, \theta_i)) \tag{3}$$

From (3), it can be seen that the multi-scale weight is obtained after the element-wise product operator $\otimes$, the convolutional layer and the Sigmoid operator. The Sigmoid operator is adopted to score the importance of different SCB channels. Then the multi-scale weights are applied to calculate the weighted-visual feature f as follows:

$$\mathrm{f} = \sum_{\mathrm{i}=1}^{4} \mathrm{f}_{\mathrm{w}}^{(i)} \mathrm{f}_i \tag{4}$$

As described in Fig. 2, compared with the 2D dimension of the density map, the dimension of f is 3D, which is hard to fuse the weighted-visual and audio feature. In order to transform the 3D dimension of f to the same size of the density map, the f is incorporated into the multi-modal fusion module in the AVMSN. In Section 2.3, the multi-modal fusion module of the SCB is described.

C. MULTI-MODAL FUSION MODULE

The architecture of the Multi-modal fusion module is drawn in Fig. 3. As shown in Fig. 3, the multi-modal fusion module adopts the cascade fusion architecture that contains 6 fusion levels of the convolutional and full connection. In Fig. 3, the convolution and full connection layers in six fusion levels are denoted as $\mathrm{L}_{conv}^{(1)}, \mathrm{L}_{conv}^{(2)}, \ldots, \mathrm{L}_{conv}^{(6)}$ and $FC^{(1)}$, $FC^{(2)}, \ldots, FC^{(6)}$ respectively. Supposing the hidden units for the i_{th} full connection layer are denoted as $\mathrm{h}^{(i)}$. For the first layer of the Multi-modal fusion module, the convolutional layer is applied to filtering the weighted-visual feature f and the full connection layer is applied to the audio feature $\mathrm{f_a}$ to calculate the fusion feature $\mathrm{f_{at}}$ as shown below:

$$\mathrm{f}_{at}^{(1)} = \mathrm{L}_{conv}^{(1)}(f) FC^{(1)}_{1:\frac{h^{(1)}}{2}}(\Gamma(f_a)) + FC^{(1)}_{\frac{h^{(1)}}{2}:h^{(1)}}(\Gamma(f_a)) \tag{5}$$

Operator Γ shows the unsqueeze and extension operator, which keeps the $\mathrm{f_a}$ as the same size as the f. According to Eq. 5, when the level stage is greater than 1, the $\mathrm{f_{at}}$ is conducted as the input of the next level stage and concatenated with the audio feature $\mathrm{f_a}$. The number of fusion levels is randomly predefined, and besides, it could be set as other numbers of the fusion level. Eq. (5) shows that number of units in the convolution layer is defined as the half size for full connection for each layer. Through several levels of the cascade architecture, the fusion feature f_{at} finally passes through the convolution layer as the output layer to calculate the predicted density map D^{pred}.

D. LOSS FUNCTION OF THE AVMSN

In this paper, the loss function for the AVMSN uses the least square loss function to learn the ground truth density map D^{gt} as follows:

$$\mathrm{loss} = \left\| D^{gt} - D^{\mathrm{pred}} \right\|_2^2 \tag{6}$$

To minimize the loss function of the AVMSN, the Adam optimization [31] method with batch size 32 for the training dataset is adopted. In order to obtain the D^{gt}, the point-wise method is used to calibrate the $2D$ positions of the counting objects in the images. The D^{gt} is generated following the fixed Gaussian Kernel [32], and the sum of the density maps equals the crowd count. Supposing the total $2D$ points are C and the computation concerning the D^{gt} of the image I equals the summation of the normal distribution of calibration points, then there is:

$$D^{gt}(p|I) = \sum_{c=1}^{C} N(p|\mu = P_{(c)}, \sigma^2) \tag{7}$$

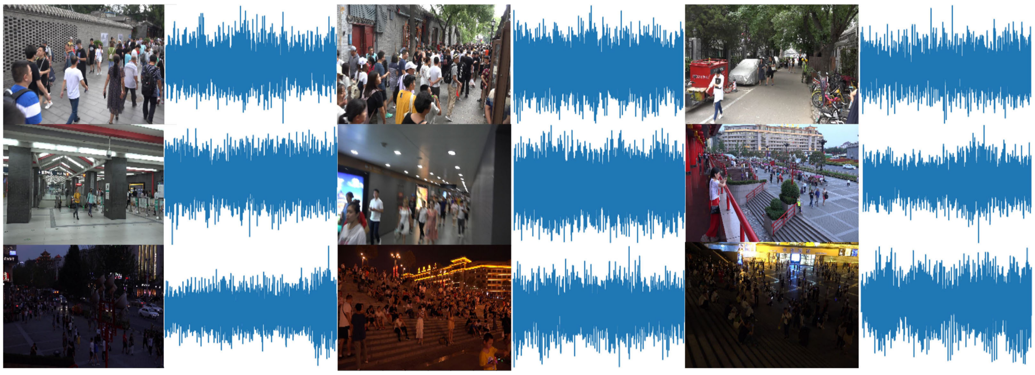

FIGURE 4. Some examples from DISCO. The left panes show the original images and the right panes show the sound waves for the left images.

In this paper, the mean absolute error (MAE) is adopted as the computing metrics as follows:

$$MAE = \frac{1}{N}\sum_{\mathrm{i}=1}^{N}\left|P_i - \hat{P}_i\right| \tag{8}$$

where N denotes the number of the images, while P_i and $\hat{P}_i$ are the number of the estimated and ground-truth crowd objects, respectively.

IV. EXPERIMENT EXPERIMENTAL SETUP

A. TRAINING DETAILS AND DATASETS

In terms of the audio source for the AVMSN, the hot length and sliding window size for the STFT are defined as 256 and 1022 for the frequency domain. In addition to that, as for the visual source for the AVMSN, each static image is cropped into the 224 × 224 size from different inputs. The image enhancement operation is executed following the flipping, brightness, and color enhancement. Besides, the learning rates for sub-modules such as Audio-VGG, Visual-VGG and multi-modal fusion modules are defined as 0.0005, 0.0005 and 0.002 accordingly. The maximum training epoch is set as 500. For each epoch, the several trained models are evaluated on the validation dataset, and one of the best performances is left behind. The AVMSN is implemented by the PyTorch platform on the TITAN-X GPU processor under Windows Operating System.

In order to evaluate the performance of the model under the high-noise, low-illumination and high-occlusion conditions, the AVMSN is applied to the DISCO [29] dataset. Some examples for the DISCO dataset are shown in Fig. 4. The visual sources are drawn in the left pane, and corresponding audio sources are drawn in the right panel.

The visual sources for the DISCO dataset are collected by the HDR-CX900E of the Sony camera, which is used as the benchmark in other works [29]. The resolution of each image is 1920 × 1080. The audio sources are subsampled by 48000Hz, and the frame number equals 25. The DISCO dataset consists of 8095 images and corresponding audios. The average, minimum and maximum number of the crowd objects in DISCO are 87.99, 1 and 709 accordingly. When compared with the famous Crowd counting benchmarks such as SHHA, UCF-QNRF and SHHB [30], [31], the statistics about illumination and crowd degree are calculated in Tab. 3 that shows the range of the counting numbers is [15], [33] more widespread than other datasets for DISCO datasets.

TABLE 3. Dataset specifications.

Dataset	No. of Images	Resolution	Min	Max	Ave
DISCO	8095	1920X1080	1	709	86.99
SHHA	482	Varied	33	3139	501
UCF_QNRF	50	Varied	94	4543	1279
SHHB	716	768X1024	9	578	123

B. CONTRASTS

In order to evaluate the audio-visual model for further modeling the crowd scene, several state-of-art vision-based counting models are used as contrast baselines for the AVMSN:

MCNN [15]: Different from the AVMSN, the MCNN only adopts the images as the visual source of the network. One image is incorporated into the 'L, M, S' column branches by different sizes of the convolutional kernel sizes. For the L column, the convolutional kernel sizes are set as 9 × 9, 7 × 7, 7 × 7 and 7 × 7. For the M column, the convolutional kernel sizes are set as 7 × 7, 5 × 5, 5 × 5 and 5 × 5. For column S, the convolutional kernel sizes are set as 5 × 5, 3 × 3, 3 × 3 and 3 × 3. Then, the output concerning the last layers of the three branches is merged into the feature maps.

CAN [17]: In the CAN model, the multi-scale contextual framework that contains the front-end network, multi-scale network and back-end decoder is proposed. Different from the AVMSN, the perspective map information for the crowd can also be modelled as an addition branch to enhance the performance and then be inserted into the CAN model.

CSRNet [34]: The CSRNet uses the Visual-VGG as the front-end network to extract the visual feature and dilated convolution layer as the back-end network to extract the

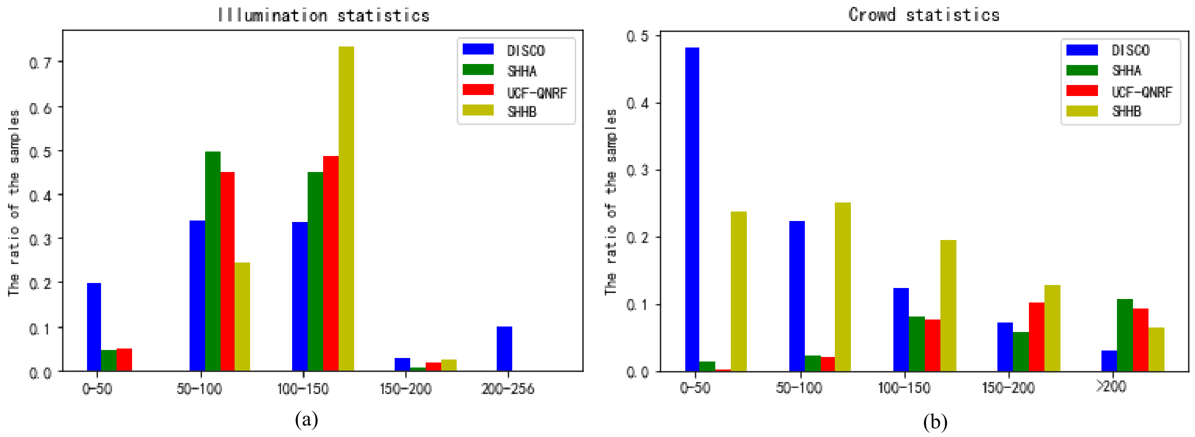

FIGURE 5. The illumination and crow statistics. Fig. 5(a) draws the illumination statistics of DISCO, SHHA, UCF-QNRF and SHHB. Fig. 5(b) draws the crowd statistics of DISCO, SHHA, UCFQNRF and SHHB.

saliency information of the density maps. The dilated convolution layer helps the CSRNet to model the highly congested scenes.

C. EXPERIMENTAL RESULTS

In order to evaluate the performance of the AVMSN under low-quality conditions such as the images corrupted from the Gaussian noise, shrinking of the brightness and shrinking of the occlusion. In this paper, the hyperparameters rate σ, r, c are adopted to denote the decay rates of the noise, illumination and occlusion, respectively.

1) THE IMPORTANCE OF AUDIO-END BRANCH

Theoretically speaking, the Audio-Visual Multi-Scale Network (AVMSN) introduces the audio processing branch against extreme conditions because the audio can be used as the auxiliary modal of visual images to present the crowd scene when images are decayed by the extreme conditions. In other words, the audio feature extracted by the Audio-VGG is kept the same under the different levels of extreme conditions. According to the content of Section III. A of the revised manuscript, the predicted density map D^{pred} is calculated from the multi-scale feature f_i and audio feature f_a. Therefore, the quality of the D^{pred} can be employed to explicate how the AVMSN can be against extreme conditions. In order to demonstrate our description, the Structural SIMilarity metric (SSIM) [35] is utilized to measure the similarity between the predicted and ground truth density maps (D^{pred} and D^{gt}) of the crowd scene.

$$\mathbf{SSIM} = \left[l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma}\right] \tag{9}$$

where the $l(x, y)$, $c(x, y)$ and $s(x, y)$ denote the luminance, contrast and structure degrees respectively, while α, β and γ are the parameters that control the mean and variance.

Fig. 6 shows the visual image decayed by different levels of noise and illumination, respectively. Besides, the (AVMSN-V) denotes the AVMSN only dominated by the multi-scale architecture of the Sampling Convolutional Blocks and without the audio processing branch. From Fig. 6(a) and (b), It can be seen that the similarity between the D^{pred} and D^{gt} that achieved by the AVMSN is better than that achieved by the AVMSN-V.

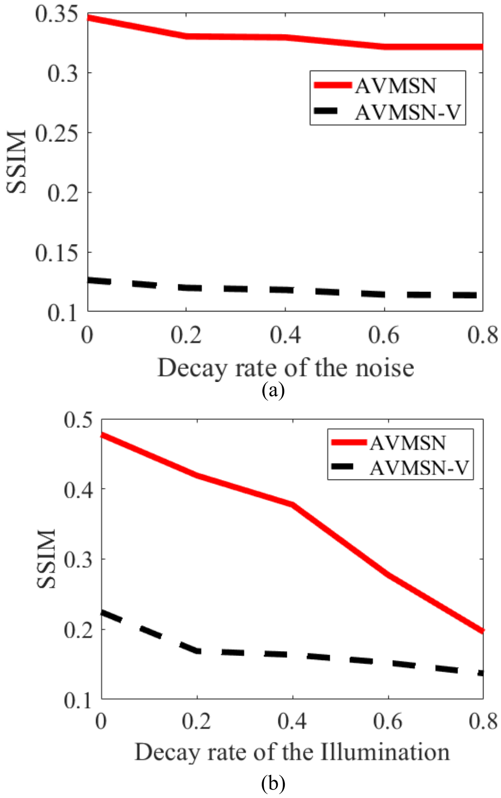

FIGURE 6. The comparison between the AVMSN and AVMSN-V with the decay rate of the noise 6(a) and illumination 6(b).

2) THE QUANTITATIVE AND QUALITATIVE ANALYZATION UNDER THE NOISY CONDITION

In this section, the performance of the AVMSN is evaluated under the high-level noise condition. Fig. 7 shows the quantitative performance of the five example images selected from the DISCO, and the performances after 200 training epochs

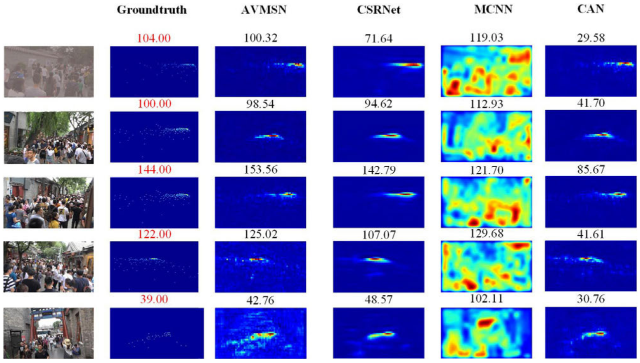

FIGURE 7. The comparisons about crowd counting among four models. The performances for the AVMSN, CSRNet, MCNN, and CAN under the decay rate 0.2 of the Gaussian noise.

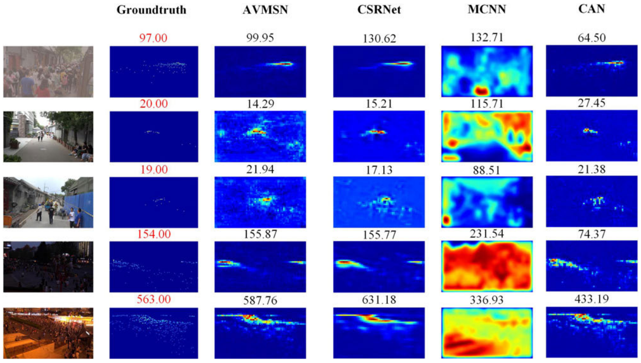

FIGURE 8. The quantitative visualization about crowd counting among four models. The performances for the AVMSN, CSRNet, MCNN, and CAN under the decay rate 0.8 of the illumination.

TABLE 4. The comparisons about crowd counting among four models. The performances for the AVMSN, CSRNet, MCNN, and CAN with the decay rate of the noise from 0-0.8.

Methods	DISCO				
	σ=0	σ=0.2	σ=0.4	σ=0.6	σ=0.8
CSRNet	5.8	6.1	6.7	7.5	8.1
MCNN	79.3	82.7	83.5	84.2	85.71
CAN	38.2	38.5	39.2	41.7	43
AVMSN	**7.0**	**7.2**	**7.3**	**7.3**	**7.3**

for the AVMSN, CSRNet, MCNN and CAN are depicted in the third, fourth, fifth and sixth columns. Furthermore, the decay rate σ of the Gaussian noise is predefined as 0.2. The noisy images are depicted in the second column.

The numbers with red color on the Ground truth density maps denote the real crowd counts, while those with black color are the estimated crowd counts. For example, as shown in the first row of Fig. 7, it can be seen that the AVMSN predicts 100.32 numbers in the crowd scene, and it is the value closest to the ground truth number 104.00. Besides, estimated crowd counts for the CSRNet, MCNN and CAN are 71.64, 119.03, 29.58, respectively. From the estimated density maps of Fig. 7, it can be seen that the performance of the AVMSN is better than that of other contrast baselines under the noisy condition.

In order to make a qualitative analysis on the performance of the AVMSN, the images are learned by models under the decay rate σ of the noise, which ranges from 0-0.8. The σ equals 0 denoting the image without suffering any noise. The performances for the AVMSN, CSRNet, MCNN and CAN are shown in Tab. 4. The MAE is adopted as the metric and the performance of the models after the 200 training epochs. In addition to that, it can be seen that the performance of the four models is shrunk with an increase in noise decay rate.

From Tab. 4, the performances of the AVMSN and CSRNet are better to achieve better results when compared with the CAN and MCNN models for the MAE metric. At the σ equals to 0, the performance of the AVMSN and CSRNet is better than that of the AVMSN. When the decay rate of the noise increased, the performance (from 7.0 to 7.3) of the AVMSN is not decayed obviously. According to the results from Tab. 4, the AVMSN is robust to the noisy image. In this paper, it is believed that the success of the AVMSN's performance under the high-level noise results from the multi-scale architecture of the Sampling Convolutional Blocks adopted by the AVMSN. The learning effect for the visual channel of the Vision-end branch is not easily decayed by the noise.

3) THE QUANTITATIVE AND QUALITATIVE ANALYSIS UNDER THE LOW ILLUMINATION CONDITION

In this section, the performance of the AVMSN is evaluated under the low-level illumination condition. Fig. 8 shows the five example images selected from the DISCO, and the performance after 200 training epochs for AVMSN, CSRNet, MCNN and CAN are depicted in the third, fourth, fifth and sixth columns. Furthermore, the decay rate r of the illumination for each image is predefined as 0.8.

The numbers with red color on the Ground truth density maps are the real crowd counts, while those with black color on the estimated density maps are the estimated crowd counts. For example, from the first row of Fig. 8, it can be seen that the AVMSN predicts 99.95 numbers in the crowd scene, which is the value closest to the ground truth number 97.00. Besides, estimated crowd counts for the CSRNet, MCNN and CAN are 130.62, 132.71, 64.50, respectively. From the estimated density maps of Fig. 8, it can be seen that the performance of the AVMSN is better than that of other contrasts under the low illumination condition.

In order to perform quantitative analysis on the AVMSN, the images are learned by models under the decay rate r of the illumination control, which ranges from 0-0.8. The r is equal to 0, which refers to the image without suffering any illumination decay. The performances for the AVMSN, CSRNet, MCNN and CAN are shown in Tab. 5. In addition to that, the MAE is adopted as the metric or the models after 200 training epochs. Furthermore, it can be seen that the performance of the four models is shrunk with the illumination decay rate increase.

TABLE 5. The comparisons about crowd counting among four models. The performances for the AVMSN, CSRNet, MCNN, and CAN with the decay rate of the illumination from 0-0.8.

Methods	DISCO				
	r=0	r=0.2	r=0.4	r=0.6	r=0.8
CSRNet	17.78	23	29.23	30.07	31.34
MCNN	84.9	104	118.6	129.5	151.8
CAN	41.48	52.7	68.35	74.56	89.11
AVMSN	**13.1**	**17.88**	**21.83**	**25.38**	**30.23**

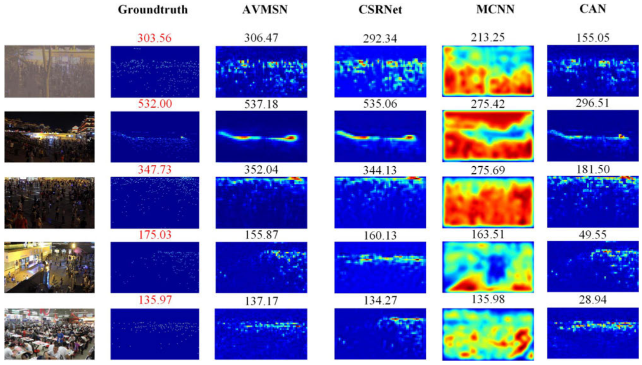

FIGURE 9. The quantitative visualization about crowd counting among four models. The performances for the AVMSN, CSRNet, MCNN, and CAN under the decay rate 0.2 of the occlusion.

From Tab. 5, the performances of the AVMSN and CSRNet are better when compared with the CAN and MCNN models for the MAE metric. The results are kept the same with the quantitative visualization in Fig. 8. As for the comparison between the AVMSN and CSRNet, when the decay rate is equal to 0.2 of the illumination for the visual sources, the performance of the AVMSN is worse than that of the CSRNet. However, when the decay rate increased (decay rate equals to 0.8), the performance gap of the AVMSN and CSRNet becomes smaller than that of the decay rate that is 0.2. It is thought in this study that the success of the AVMSN's performance is due to the Audio-end branch is added in the AVMSN as the auxiliary procedure, and the low-illumination condition cannot have a severe impact on the audio channel.

4) THE QUANTITATIVE AND QUALITATIVE ANALYSIS UNDER THE HIGH OCCLUSION CONDITION

In this section, the performance of the AVMSN is evaluated under the high-level occlusion condition. Fig. 9 shows the five example images selected from the DISCO, and the performances after 200 training epochs for the AVMSN, CSRNet, MCNN and CAN are depicted in the third, fourth, fifth and sixth columns. Besides, the decay rate c of the occlusion for each image is predefined as 0.2 for Fig. 9.

The numbers with red color on the Ground truth density maps are the real crowd counts, while those with black color on the estimated density maps are the estimated crowd counts. For example, from the first row of Fig. 9, it can be seen that the AVMSN predicts 306.47 numbers in the crowd scene, which is the closest value to the ground truth number 303.56. Furthermore, estimated crowd counts for the CSRNet, MCNN and CAN are 292.34, 213.25, 155.05, respectively. From the estimated density maps of Fig. 9, it can be found that the performance of the AVMSN is better than other contrast baselines under the high-level occlusion condition.

TABLE 6. The comparisons about crowd counting among four models. The performances for the AVMSN, CSRNet, MCNN, and CAN with the decay rate of the occlusion from 0-0.8.

Methods	DISCO				
	c=0	c=0.2	c=0.4	c=0.6	c=0.8
CSRNet	17.78	23	29.23	30.07	31.34
MCNN	84.9	104	118.6	129.5	151.8
CAN	41.48	52.7	68.35	74.56	89.11
AVMSN	**13.1**	**17.88**	**21.83**	**25.38**	**30.23**

In order to make a quantitative analysis on the performance of the AVMSN, the images are learned by models under the decay rate c of the occlusion of the images, ranging from 0-0.8. The c is equal to 0 which means the image without suffering the occlusion. The performances for the AVMSN, CSRNet, MCNN and CAN are shown in Tab. 6, and the MAE is adopted as the metric.

From Tab. 6, it can be seen that the better performances of the AVMSN and CSRNet are achieved when compared with the CAN and MCNN models for the MAE metric. As c equals 0, the comparable performances between the AVMSN and CSRNet are achieved. The performance of the AVMSN varies from 13.1 to 30.23 when c ranges from 0 to 0.8. However, when the decay rate of the occlusion increases, the performance of the CSRNet is decayed severely.

V. CONCLUSION

In this paper, as the audio-visual model, the AVMSN is proposed to solve the crowd counting task following the two-stream learning framework by adopting the visual and audio sources as the input channel. The Feature extraction module in the AVMSN fuses the Vision-end and Audio-end branches to learn the input sources and extract the weighted-visual features and audio features for them, respectively. As for the Vision-end branch, ten layers of the VGG network are adopted to extract the front-end visual feature. In order to handle the variation of the resolutions under the dense crowd scenes, the average pooling operator is adopted as the multi-scale architecture to model front-end features by four Sampling Convolutional Blocks (SCBs) to calculate the visual feature. For SCBs, the front-end feature is processed through the convolution, average pooling and upsample layers for further calculating the multi-scale features. Then the multi-scale features from different SCBs are adapted to the front-end feature by the element-wise product operator to obtain the multi-scale weights from the VGG network. Finally, the front-end feature is weighted by the multi-scale weights to compute the weighted-visual feature.

The size of the weighted-visual feature is 3D-dimensional, and hard to keep the same dimension with the estimated density map. After the Feature extraction module, the Multi-modal fusion is put forward. The weighted-visual feature passes through the six levels of the cascade architecture belonging to the Multi-modal fusion module. After processing through the convolution layer, the estimated density map is calculated and optimized continuously according to the ground truth density map.

In this paper, the advantage of introducing the audio-end branch against high-noise and low-illumination is explored theoretically. The experiment results from crowd counting tasks show that the AVMSN can achieve better performance on the mean absolute error (MAE) metric than other state-of-the-art works. Furthermore, the experiment results demonstrate that the AVMSN is characterized by a more accurate prediction for the crowd numbers than other state-of-art models under low-quality images.

Overall, the experiment results show that the AVMSN can adaptively model the content of the audio and visual sources under low-quality conditions such as high-level noise, low-level illumination and high-level occlusion conditions. It is expected that our work can give inspiration to the crowd counting task for unconstrained videos when introducing the audio modal as the axillary information to enhance the performance.

Furthermore, the performance of the AVMSN will be enhanced when the audio modal is fused into the visual modal. However, for the generation of the density map, the fixed size of the Gaussian kernel is adopted. Are the different sizes of kernels suitable for the particular crowd scene? Our future work will focus on the crowd counting learning tasks under the adaptive density map generation for crowd counting tasks.

REFERENCES

[1] Q. Zhou, J. Zhang, L. Che, H. Shan, and J. Z. Wang, "Crowd counting with limited labeling through submodular frame selection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1728–1738, May 2019.

[2] J. Fu, H. Yang, P. Liu, and Y. Hu, "A CNN-RNN neural network join long short-term memory for crowd counting and density estimation," in *Proc. IEEE Int. Conf. Adv. Manuf. (ICAM)*, Nov. 2018, pp. 471–474, doi: 10.1109/AMCON.2018.8614939.

[3] Y. Hu, H. Chang, F. Nian, Y. Wang, and T. Li, "Dense crowd counting from still images with convolutional neural networks," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 530–539, Jul. 2016.

[4] R. Hu, S. Chang, H. Wang, J. He, and Q. Huang, "Efficient multispike learning for spiking neural networks using probability-modulated timing method," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 1984–1997, Jul. 2019.

[5] R. Hu, S. Zhou, Z. R. Tang, S. Chang, Q. Huang, Y. Liu, W. Han, and E. Q. Wu, "DMMAN: A two-stage audio–visual fusion framework for sound separation and event localization," *Neural Netw.*, vol. 133, pp. 229–239, Jan. 2021.

[6] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 21, 2019, doi: 10.1109/TPAMI.2018.2889052.

[7] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1086–1099, May 2018.

[8] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 90–97.

[9] X. Wang, B. Wang, and L. Zhang, "Airport detection in remote sensing images based on visual attention," in *Proc. Int. Conf. Neural Inf. Process.*, 2011, pp. 475–484,

[10] Z. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 604–618, Apr. 2010.

[11] R. Hu, Q. Huang, S. Chang, H. Wang, and J. He, "The MBPEP: A deep ensemble pruning algorithm providing high quality uncertainty prediction," *Int. J. Speech Technol.*, vol. 49, no. 8, pp. 2942–2955, Aug. 2019.

[12] R. Hu, Z.-R. Tang, X. Song, J. Luo, E. Q. Wu, and S. Chang, "Ensemble echo network with deep architecture for time-series modeling," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 4997–5010, May 2021, doi: 10.1007/s00521-020-05286-8.

[13] R. Hu, Q. Huang, H. Wang, J. He, and S. Chang, "Monitor-based spiking recurrent network for the representation of complex dynamic patterns," *Int. J. Neural Syst.*, vol. 29, no. 8, Oct. 2019, Art. no. 1950006.

[14] Z. Tang, R. Zhu, R. Hu, Y. Chen, E. Q. Wu, H. Wang, J. He, Q. Huang, and S. Chang, "A multilayer neural network merging image preprocessing and pattern recognition by integrating diffusion and drift memristors," *IEEE Trans. Cognit. Develop. Syst.*, early access, Jun. 18, 2020, doi: 10.1109/TCDS.2020.3003377.

[15] 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.

[16] 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.

[17] N. Ilyas, A. Ahmad, and K. Kim, "CASA-Crowd: A context-aware scale aggregation CNN-based crowd counting technique," *IEEE Access* vol. 7, pp. 182050–182059, 2019.

[18] W. Li, L. Yongbo, and X. Xiangyang, "CODA: Counting objects via scale-aware adversarial density adaption," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 193–198.

[19] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *Sepcial Interest Group Comput.*, vol. 37, no. 4, pp. 1–11, 2018.

[20] Z. Hang, G. Chuang, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 587–604.

[21] G. R. H. Feris and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 36–54.

[22] M. Tong, M. Zhao, Y. Chen, and H. Wang, "D3-LND: A two-stream framework with discriminant deep descriptor, linear CMDT and nonlinear KCMDT descriptors for action recognition," *Neurocomputing*, vol. 325, pp. 90–100, Jan. 2019.

[23] Z. Tan, Y. Chen, S. Ye, R. Hu, H. Wang, J. He, Q. Huang, and S. Chang, "Fully memristive spiking-neuron learning framework and its applications on pattern recognition and edge detection," *Neurocomputing*, vol. 403, pp. 80–87, Aug. 2020, doi: 10.1016/j.neucom.2020.04.012.

[24] S. Karen and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, Jul. 2017, pp. 1800–1807.

[27] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[28] L. Shi, K. Du, C. Zhang, H. Ma, and W. Yan, "Lung sound recognition algorithm based on VGGish-BiGRU," *IEEE Access*, vol. 7, pp. 139438–139449, 2019.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[30] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[31] R. K. Yadav and Anubhav, "PSO-GA based hybrid with Adam optimization for ANN training with application in medical diagnosis," *Cognit. Syst. Res.*, vol. 64, pp. 191–199, Dec. 2020.

[32] M.-S. Yang and H.-S. Tsai, "A Gaussian kernel-based fuzzy C-means algorithm with a spatial bias correction," *Pattern Recognit. Lett.*, vol. 29, no. 12, pp. 1713–1725, Sep. 2008.

[33] 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. Eur. Conf. Comput. Vis.*, 2018, pp. 544–559.

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

[35] U. Brandes and J. Lerner, "Structural similarity: Spectral methods for relaxed blockmodeling," *J. Classification*, vol. 27, no. 3, pp. 279–306, Nov. 2010.

RUIHAN HU received the Ph.D. degree in microelectronics from the School of Physics and Technology, Wuhan University, in 2019. He is currently working with the Guangdong Key Laboratory of Modern Control Technology, Guangdong Institute of Intelligent Manufacturing, Guangzhou, China. His research interests mainly include spiking neural networks, computer vision, deep learning, and some related applications.

QINGLONG MO received the B.Sc. degree from the Guangdong University of Technology, Guangzhou, China, in 2007. He is currently working with the Guangdong Key Laboratory of Modern Control Technology, Guangdong Institute of Intelligent Manufacturing, Guangzhou. His research interests include neural computing, machine learning, and image processing.

YUANFEI XIE received the B.Sc. degree in mechanical design and theory from the School of Electrical Engineering and Automation, Wuhan University, in 2020, where she is currently pursuing the M.Eng. degree. Her research interests mainly include crowd counting and crowd tracking.

YONGQIAN XU received the M.Eng. degree from the South China University of Technology, Guangzhou, China, in 2008. He is currently a Professor with the Guangdong Key Laboratory of Modern Control Technology, Guangdong Institute of Intelligent Manufacturing, Guangzhou. His research interests include neural computing, machine learning, and image processing.

JIAQI CHEN received the B.Sc. degree from the Guangdong University of Technology, Guangzhou, China, in 2005. She is currently working with Guangdong Key Laboratory of Modern Control Technology, Guangdong Institute of Intelligent Manufacturing, Guangzhou. Her research interests include neural computing, machine learning, and image processing.

YALUN YANG received the M.Eng. degree in mechanical engineering from the Wuhan University of Science and Technology, Wuhan, China, in 2018, where she is currently pursuing the Ph.D. degree in computer vision. Her research interests include feature identification and image preprocessing.

HONGJIAN ZHOU received the Ph.D. degree in mechanical design and theory from the School of Power and Mechanical Engineering, Wuhan University, Wuhan, China, in 2020. He is currently working with the School of Mechanical and Electrical Engineering, Wuhan Institute of Technology, Wuhan. His research interest mainly includes machine learning and some related applications.

ZHI-RI TANG received the B.Sc. and M.Eng. degrees from Wuhan University, in 2017 and 2019, respectively. His main research interests include cognitive science, biomedical informatics, and machine learning.

EDMOND Q. WU (Member, IEEE) received the Ph.D. degree in control theory and application from Southeast University, Nanjing, China, in 2009. He is currently an Associate Professor with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China. He is also with the Science and Technology on Avionics Integration Laboratory, China National Aeronautical Radio Electronics Research Institute, Shanghai. His research interests include deep learning, cognitive modeling, and industrial information processing.