IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 11, 2021, accepted August 31, 2021, date of publication September 13, 2021, date of current version September 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3112174

U-ASD Net: Supervised Crowd Counting Based on Semantic Segmentation and Adaptive Scenario Discovery

ADEL HAFEEZALLAH[1], **AHLAM AL-DHAMARI**[2,3], **AND SYED ABD RAHMAN ABU-BAKAR**[2], **(Senior Member, IEEE)**

[1]Department of Electrical Engineering, Taibah University, Madinah, Saudi Arabia
[2]Department of Electronics and Computer Engineering, School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, Johor 81310, Malaysia
[3]Department of Computer Engineering, Hodeidah University, Hodeidah, Yemen

Corresponding authors: Adel Hafeezallah (ahafiz@taibahu.edu.sa), Ahlam Al-Dhamari (kmahlam@utm.my), and Syed Abd Rahman Abu-Bakar (e-syed@utm.my)

This work was supported in part by the Deputyship of Research and Innovation, Ministry of Education, Saudi Arabia, under Project 758, and in part by the Research Management Center, Universiti Teknologi Malaysia, under Grant 4C396.

ABSTRACT Crowd counting considers one of the most significant and challenging issues in computer vision and deep learning communities, whose applications are being utilized for various tasks. While this issue is well studied, it remains an open challenge to manage perspective distortions and scale variations. How well these problems are resolved has a huge impact on predicting a high-quality crowd density map. In this study, a hybrid and modified deep neural network (U-ASD Net), based on U-Net and adaptive scenario discovery (ASD), is proposed to get precise and effective crowd counting. The U part is produced by replacing the nearest upsampling in the encoder of U-Net with max-unpooling. This modification provides a better crowd counting performance by capturing more spatial information. The max-unpooling layers upsample the feature maps based on the max locations held from the downsampling process. The ASD part is constructed with three light pathways, two of which have been learned to reflect various densities of the crowd and define the appropriate geometric configuration employing various sizes of the receptive field. The third pathway is an adaptation path, which implicitly discovers and models complex scenarios to recalibrate pathway-wise responses adaptively. ASD has no additional branches to avoid increasing the complexity. The designed model is end-to-end trainable. This integration provides an effective model to count crowds in both dense and sparse datasets. It also predicts an elevated quality density map with a high structural similarity index and a high peak signal-to-noise ratio. Several comprehensive experiments on four popular datasets for crowd counting have been carried out to demonstrate the proposed method's promising performance compared to other state-of-the-art approaches. Furthermore, a new dataset with its manual annotations, called Haramain with three different scenes and different densities, is introduced and used for evaluating the U-ASD Net.

INDEX TERMS Computer vision, deep learning, crowd counting, density map estimation, U-Net, adaptive scenario discovery.

I. INTRODUCTION

In situations involving crowd movements such as religious gatherings, sporting events, and public protests, crowd analysis and management are critical and have supreme significance in avoiding stampedes and saving lives. Crowd analysis can be a powerful tool in these situations for early prediction of crowding and selecting appropriate necessary measures for crowd control and management. Thus, avoiding any disaster that is about to happen. The variety of crowd management applications has prompted and inspired researchers from different disciplines to propose innovative and efficient methods for crowd analysis and relevant tasks, including counting [1], [2], behavior analysis [3], tracking [4], density estimation [5], [6], anomaly detection [3], [7], [8], scene understanding [9], segmentation [10]–[12], and mobile crowd sensing [13], [14]. Among these, density estimation and crowd counting are critical elements that serve as the foundation for various

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

purposes. Crowd surveillance and analysis are not trivial problems and bring along different obstructions, like, occlusion, background noise, changes in lighting, scale, people distribution, and perspective. Researchers in this area have come a long way to tackle some of these issues. The current crowd scene analysis methods range from straightforward crowd counting, predicting the total number of individuals in a scene, to density map estimating, which shows crowd distribution characteristics. The density map assists in obtaining more precise and intensive details, which may be crucial in making appropriate decisions, especially in risky scenarios. Notwithstanding, producing precise distribution models is very challenging. One significant trouble stems from the estimation way. Because the produced values of the density are based on pixel-by-pixel estimation, the generated density maps should have spatial coherence to demonstrate the smooth transition amongst adjacent pixels. This is challenging because of the wide range of crowd density values. As shown in Fig. 1, some of the samples consist of hundreds of pedestrians, while other samples containing only a few. This issue can be difficult for a single CNN to deal with the full range of crowd densities. To tackle this challenge, multi-column CNN architectures were introduced widely in the literature. Such architectures can have different parallel CNN branches with various sizes of the receptive field. In this kind of architecture, a branch of a network with smaller receptive fields can effectively address the high crowd density images. In contrast, a network branch with larger receptive fields can address low crowd density images well [15]. In addition, the task would be hard to complete due to the variety of views, which include infrequent crowd clusters and multiple camera viewpoints, particularly when using conventional methods without deep neural networks.

The proposed U-ASD model is inspired by U-Net [10] with an additional adaptive scenario discovery (ASD) [16]. The model on an encoder-decoder layout with three light parallel branches is built. The encoder part of the U-Net is supplanted by VGG16-bn [17]. In addition, the output of the U-Net encoder has been used as a backbone to the last branches that represent the adaptive scenario discovery. Adding the ASD as a binary classifier improves the model's crowd counting efficiency. The ground truth attention map is fed into the adaptive scenario discovery branches, and the output is combined with U-Net using a combined loss.

To sum up, the following contributions are made:

- A hybrid and modified network structure capable of predicting precise density map half the size or resolution of the input is proposed.
- A modified U-Net is produced by replacing the nearest upsampling with max-unpooling. The upsampling using max-unpooling for U-Net is proposed to extract more spatial information through the complex max-unpooling layers. Thus, a better crowd counting accuracy is achieved. To the best of our knowledge, max-unpooling has not yet been utilized in the literature for U-Net architecture for crowd counting. In this study, a comparison in terms of the counting accuracy, the number of parameters, and training runtime between utilizing the nearest upsampling and max-unpooling are presented in Section V.
- A new dataset, dubbed Haramain, with its manual annotations is presented. The Haramain dataset consists of three various scenes and densities.
- The efficacy of the proposed U-ASD Net is tested on four challenging datasets for crowd counting. Interestingly, it surpasses state-of-the-art approaches, according to our findings.

The other sections of this article are arranged as follows. Section II highlights some critical and timely relevant research. The proposed model architecture is presented in Section III. In Section IV, the evaluation metrics, experimental setup, and qualitative and quantitative findings are presented. Section V presents discussions and analyses of the findings. The proposed work is concluded in Section VI.

II. RELATED WORK

Significant CNN-based crowd counting methods and related density map prediction methods are being demonstrated in this section. Furthermore, since the proposed U-ASD Net uses segmentation and spatial CNN to address the crowd counting task, related research studies about those methods are briefly reviewed.

A. PATCH-BASED AND IMAGE-BASED METHODS

Because of its effectiveness in capturing local features and generating a huge amount of training samples, patch-based methods have been utilized in many methods [4], [18]. Patch-based methods train a model and estimate sliding windows through the testing stage by cropping the images of various sizes. Convolutional Neural Networks (CNN) have been utilized in several methods for crowd counting purposes [19]–[21]. Zhang *et al.* [19] developed a deep-qualified CNN for crowd counting and estimating the level of crowd density. Li *et al.* [20] suggested using the VGG16 encoder as well as dilated convolutional layers as a decoder to assemble contextual features on a variety of scales. Cao *et al.* [21] introduced a scale aggregation network to extract multi-scale features utilizing the encoder that uses scale aggregation modules and estimate high-resolution density maps using the decoder that uses a collection of transposed convolutions. Fu *et al.* [18] suggested categorizing the image into five diverse classes, where each class represents different intensities of the density, rather than estimating density maps. Layered boosting and selective sampling procedures were presented by Walach and Wolf [22]. The layered boosting means that CNN layers are added to the model in an iterative manner so that each new layer is learned to predict the residual error of the previous estimation. Kumagai *et al.* [23] introduced the Mixture of CNNs (MoCNN) model that comprises a combination of expert CNNs as well as a gating CNN. The gating CNN specifies

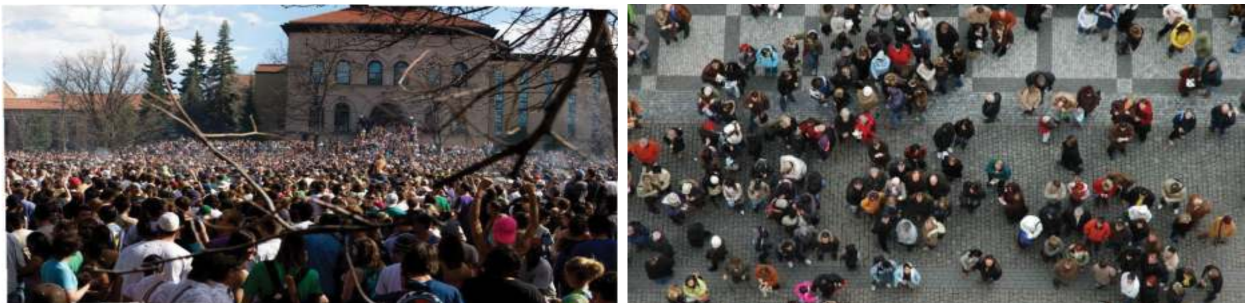


FIGURE 1. A CNN-based density estimation can be utilized to pose the crowd counting problem, but because of the large fluctuation in densities across pixels in the various samples, this issue can be challenging and difficult for a single CNN. This figure presents two images from the Shanghaitech dataset with substantially varying crowd densities.

a probability to each expert CNN layer, and the expert CNNs estimate the crowd count. The weighted average count of all the expert CNNs is the final output crowd count. According to Sam *et al.* [24], the better output is determined by training regressors with a specific group of training patches and exploiting variation in crowd density. Moreover, Sam *et al.* put forward a switching CNN that intelligently determines the best regressor suited for each input patch. Because the patch-based methods seem unable to represent the global contextual data, the whole image-based methods had been concentrated by the proposed works [5], [25], [26]. Zhang *et al.*. [19] put forward a multi-column CNN that processes the input image with adjustable resolution and uses every column to comply with various scales. Sheng *et al.* [24] introduced a novel image representation that integrates semantic attributes and spatial cues to enhance the discriminative power of feature representations. Marsden *et al.* [27] introduced incorporating scale into the models with fewer model parameters and proposed a single column fully convolutional network (FCN) to estimate density map. A cascaded CNN architecture (Cascaded-MTL) was proposed by Sindagi and Patel [26]. The Cascaded-MTL integrates learning of a high level prior to lifting the performance of the density estimation.

B. SEGMENTATION AND SPATIAL CNN

Utilizing pixel-wise regression, density estimation-based techniques can predict a density estimation map and thus localize the crowd. Then, the process of crowd counting is performed by computing the integral image of the density map [28]. To create density maps with keeping the spatial size as the inputs or half the inputs, encoder-decoder architectures are considerably used [10]–[12], [29], [30]. In 2015, U-Net was developed by Ronneberger *et al.* [10] for biomedical image segmentation and was then extensively employed for image segmentation in many other fields with different encoders such as ResNet, Inception, and DenseNet modules. By applying skip connections between the corresponding encoding and decoding path blocks, U-Net designed a symmetric network structure in which convolutional features are stacked from activations of the encoder to the decoder parts. In [31], Shen *et al.* made use of a U-Net structure with an adversarial loss to produce high-quality density maps. Huynh *et al.* [32] put forward an inception U-Net-based multi-task learning for crowd counting, density map-generating, and density level classification. In [30], the authors proposed U-Net-like architecture called W-Net, which applies a reinforcement branch to improve the crowd counting accuracy and converge quicker.

III. PROPOSED METHOD

The workflow of the proposed method is shown in Fig. 2. The estimated maps from the modified U-Net and ASD Net are multiplied to generate the final crowd density map. The overall U-ASD network architecture is illustrated in Fig. 3. The complete design of the U-ASD model is described in detail in subsection A, followed by the specifics of its implementation.

A. U-ASD OVERALL ARCHITECTURE

The proposed U-ASD model is built on U-Net [10] and ASD [16]. The model on an encoder-decoder layout with three light parallel branches is built. The U-Net architecture comprises both an "encoder part" to capture context and a ''symmetric decoder part'' to provide precise localization and estimate the density map. To extract multi-scale visual features from input image sequences, the backbone is utilized. Following [30], a pre-trained VGG16-bn model, which is a variant of VGG16 with batch normalization, is utilized. The first VGG16-bn layers with four-max-pooling are used as the backbone, and it replaces the encoder block of the U-Net. Table 1 presents the configuration of the proposed U-ASD model. The ASD part is used to assist the network in converging faster and providing better performance with low errors. The ASD Net's output density map is 1/16 of the input image. To fuse the U-Net's output with the ASD Net's output, a Nearest-neighbor upsample layer (US) was introduced as shown in Fig. 3 to upsample the output.

B. DENSITY MAP ESTIMATION

Semantic segmentation and density map estimation are classification and pixel-wise regression issues, respectively. Accordingly, numerous studies in crowd counting address the concepts and hypotheses in semantic segmentation. Ronneberger *et al.* [10] designed U-Net (looks like a U letter) to concentrate on the pixel-wise classification of an image sequence. U-Net can focus on low-level abstract features (extracted from the first convolutional layers of the

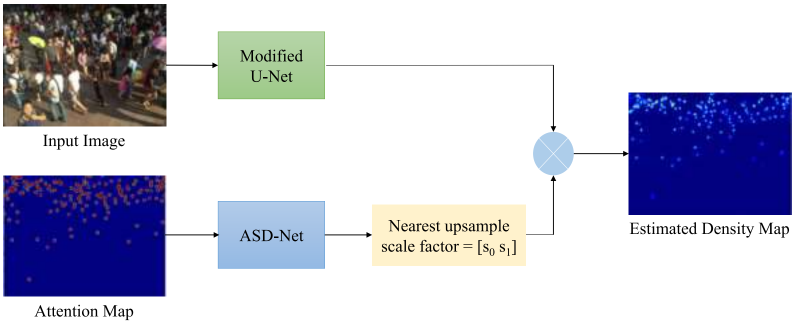


FIGURE 2. The overall training workflow of the proposed method. The final density map is obtained by multiply the generated density maps of the modified U-Net and ASD Net.

encoder part) and high-level semantic abstraction features (extracted from the decoder part's final layers). In the proposed U-Net, the max-unpooling operations utilizing the memorized max-pooling indices from the relating encoder layer replace the nearest upsampling in the U-Net structure. Further details about the U-Net are in subsection E.

Kang *et al.* [4] investigated the generated maps produced by density estimation approaches for crowd analysis applications such as detection, counting, and tracking. They investigated the performance of those applications in great detail when employing full-resolution density maps. Their findings revealed that full resolution density maps enhanced the effectiveness of localization tasks, including tracking and detection. Furthermore, they mentioned that good counting accuracy does not always necessitate full-resolution density maps, and adopting reduced-resolution maps can speed up the predictions while maintaining good counting performance as in [19] and [25]. Because of downsampling strides in the convolution layers and the pooling layers, most existing CNN algorithms normally create density maps with a resolution lower than the source images.

C. CLASSIFICATION VS. REGRESSION FOR COUNTING

As is well known, the network output of the CNN-based classification models is a vector of the same size as the number of classes. The input image's confidence score belongs to the i-th class is expressed by the i-element in the vector. The final classification result has been chosen according to the index that has the highest confidence score during the testing stage. For most classification problems, softmax loss is extensively utilized [33]. Taking the human count number as the class index, on the other hand, is not appropriate for crowd counting problems. The difference between the ground-truth map and the predicted map can be better retained in the proposed U-ASD model while determining the estimation error. Such information is extremely useful for more precisely optimizing the CNN weights during the back-propagation stage. To allow the entire model to implicitly detect all crowd scenarios and respond to varied crowd images in a precise way, two types of architectures in our model are used: U-Net and ADSNet, which will be well explained in the next subsections. For each architecture, a different loss function is defined. For the modified U-Net, the 2-D pixel-wise mean square error (MSE) loss is utilized for the density regression task, which can be defined as in Equation 1 below:

$$L_{mse}(d^g, d^p) = \frac{1}{n}\sum_{i=1}^{n}\left|d_i^g - d_i^p\right|^2 \tag{1}$$

where d^g and d^p refer to the ground-truth map and the predicted density map, respectively, and n is the total number of pixels each.

In the proposed U-ASD model, the main aim of training the ASD is to minimize the binary cross-entropy (BCE) loss, which is used for measuring the error of a reconstruction, which is defined as follows [30], [34], [35]:

$$L_{bce}(g, p) = -\frac{1}{m}\sum_{i=1}^{m}(g_i log(p_i) + (1 - g_i) log(1 - p_i)) \tag{2}$$

where $g_i \in \{0, 1\}$ is the ground-truth attention map, where there are two classes (background '0') and (foreground '1'), p_i is the predicted attention map, and m is the total number of pixels. To put it another way, the predicted foreground mask is compared to the ground truth map using a binary cross-entropy error function, and the low value of the $L_{bce}(g, p)$ means better accuracy.

D. POOLING AND UNPOOLING

To lower the size of the representation and make it more manageable, the pooling layer is employed. It processes the input and downsamples it without affecting the depth [36]. In other words, the process is done spatially. Thus, the input and output depths stay the same. Unpooling is used to achieve upsampling in the network. For density map estimation, precise pixel prediction is required to acquire accurate counting. If max-unpooling is utilized, the feature map will be heterogeneous because of the loss of spatial information produced by max-pooling from the low-resolution image. Nonetheless, after max-pooling, there is no data regarding the locations of the feature vector in the low receptive field. When max-unpooling is performed, the maxima locations inside each pooling zone can be captured in a set of switch variables stored in a continuous array after applying the max-pooling. These switches are utilized in the related max-unpooling to set the signal from the present feature map into the up-sampled feature map's relevant locations. Therefore,

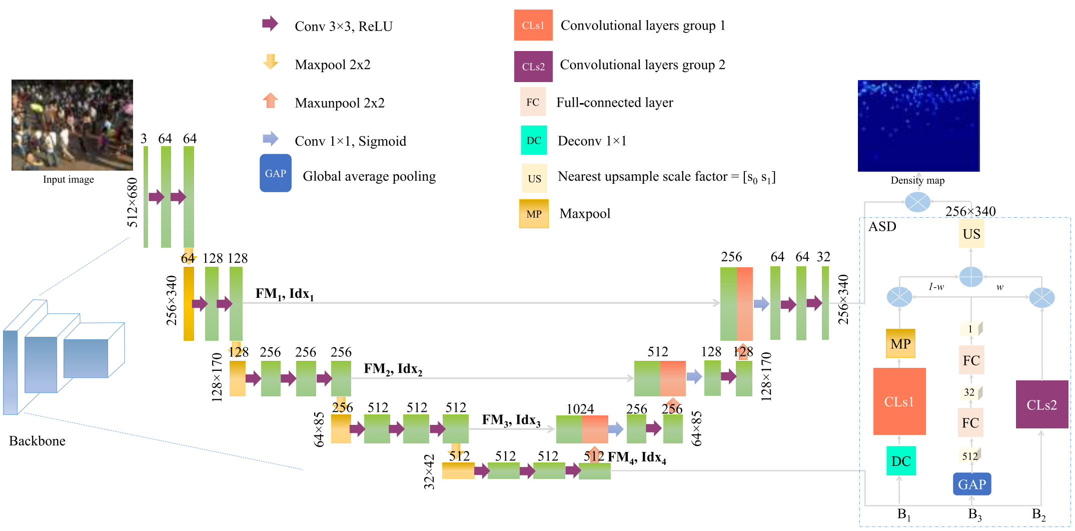


FIGURE 3. The architecture of U-ASD Net. The parameters of the convolutional layer are indicated as (Conv kernel size). 'Conv' refers to a convolution layer, and 'Deconv' refers to a deconvolution layer.

more fine detail can be recovered and preserved the spatial information lost during max-pooling.

E. U-NET

1) ENCODER PART

Only the first layers from the pre-trained VGG16-bn network are used as an encoder for U-Net to generate the multi-scale feature maps, and the fully connected layers for the classification process are excluded. Following the U-Net structure [10], the feature maps (FMs) resulting from the encoder part (FM_1, FM_2, FM_3, and FM_4 shown in Fig. 3) are employed as inputs to the decoder part.

2) DECODER PART

The decoder part is illustrated in Fig. 4. First, the FM_4 output and its index Idx_4 are used to upscale the input using Max-unpooling, and then the output of FM_3 is concatenated with it. After that, this concatenated input is passed to Block 1 demonstrated in Fig. 4. Block 1 includes 2×max-unpool and two convolutional layers with $1 \times 1 \times 256$ and $3 \times 3 \times 256$, respectively, followed by batch normalization (BN) and rectified linear unit (ReLU). The output of this block is upscale using 2×max-unpool and concatenated with the output of FM_2. Similarly, the process of increasing the size is reiterated prior to feeding Block 2 (with the same architecture as in Block 1 but with a different channel size). At last, another upscaling and concatenation from Block 2 are performed, and Block 3 generates the final feature map. At the training phase, the loss function of U-Net is the 2-D MSE loss, which is defined previously in Equation 1.

F. ASD NET

Recent papers in [30] and [37]–[39] have been utilized VGG16-bn for crowd counting, and the proposed models of these papers achieved high performance. Therefore, following these studies, the VGG16-bn has been used as a backbone for our model instead of VGG16 that was used in the original ASD Net. After the backbone, the ASD part incorporates three light parallel paths as in Fig. 3. The first part, B_1, is intended to address the sparse crowds. It contains a deconvolutional layer (DC), which upscales the inputs. After the DC, there are five convolutional layers with larger receptive fields followed by max-pooling. Fig. 5 presents the details about the structure of the convolutional layers group 1 (CLs1) in the B_1 pathway. The second pathway, B_2, is

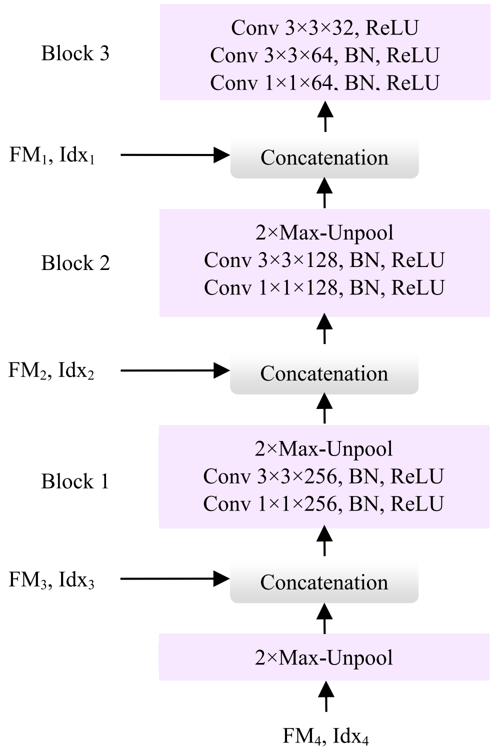


FIGURE 4. Decoder architecture for U-ASD Net. Batch normalization layers "BN" are added between each "Conv" and "ReLU" layer.

TABLE 1. Configuration of the proposed U-ASD network. The parameters in the convolutional layers "Conv2d" and the deconvolutional layer "ConvTranspose2d" are denoted as "kernel size, number of filters, stride, dilation". Unless otherwise stated, the stride and dilation are set to 1 by default, the padding is set to 0 by default. Maxpooling "MaxPool2d" and maxunpooling "Maxunpool2d" layers are denoted as "kernel size, stride". Global average pooling "GAP" is denoted as "dimensions". Fully connected layers (FCs) are referred to as "size of the input" and "size of output," respectively.

	Encoder part		
	Layer name	**Output Size**	**Configuration**
	Conv2d-1	512×680	3×3, 64
	Conv2d-2	512×680	3×3, 64
	MaxPool2d-1	256×340	2, *stride* 2
	Conv2d-3	256×340	3×3, 128
	Conv2d-4	256×340	3×3, 128
	MaxPool2d-2	128×170	2, *stride* 2
	Conv2d-5	128×170	3×3, 256
	Conv2d-6	128×170	3×3, 256
	Conv2d-7	128×170	3×3, 256
	MaxPool2d-3	64×85	2, *stride* 2
	Conv2d-8	64×85	3×3, 512
	Conv2d-9	64×85	3×3, 512
	Conv2d-10	64×85	3×3, 512
	MaxPool2d-4	32×42	2, *stride* 2
	Conv2d-11	32×42	3×3, 512
	Conv2d-12	32×42	3×3, 512
	Conv2d-13	32×42	3×3, 512
	Decoder part		
	Layer name	**Output Size**	**Configuration**
	Maxunpool2d-1	64×85	2, *stride* 2
	Conv2d-1	64×85	1×1, 256
	Conv2d-2	64×85	3×3, 256
	Maxunpool2d-2	128×170	2, *stride* 2
	Conv2d-3	128×170	1×1, 128
	Conv2d-4	128×170	3×3, 128
	Maxunpool2d-3	256×340	2, *stride* 2
	Conv2d-5	256×340	1×1, 64
	Conv2d-6	256×340	3×3, 64
	Conv2d-7	256×340	3×3, 32
	ASD part		
	Layer name	**Output Size**	**Configuration**
B1	ConvTranspose2d-1	64×84	2×2, 512, *stride* = 2
	Conv2d-1	64×84	1×1, 512
	Conv2d-2	64×84	9×9, 256, *padding* = 4
	Conv2d-3	64×84	7×7, 128, *padding* = 3
	Conv2d-4	64×84	7×7, 64, *padding* = 3
	Conv2d-5	64×84	3×3, 1, padding = 1
	MaxPool2d-1	32×42	2×2, *stride* 2
B2	Conv2d-6	32×42	3×3, 256, *padding* = 1
	Conv2d-7	32×42	3×3, 128, *padding* = 1
	Conv2d-8	32×42	3×3, 64, *padding* = 1
	Conv2d-9	32×42	3×3, 1, *padding* = 1
B3	GAP	512	(2,3)
	FC-1	32	512, 32
	FC-2	1	32, 1

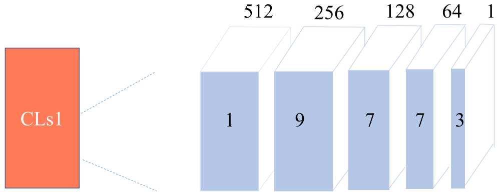


FIGURE 5. Structure of the convolutional layers group 1 (CLs1), where (512, 256, 128, 64, 1) are the number of filters, and (1, 9, 7, 7, 3) are the kernel size.

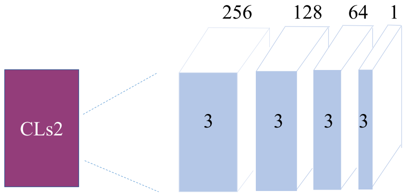


FIGURE 6. Structure of the convolutional layers group 2 (CLs2), where (256, 128, 64, 1) are the number of filters, and (3, 3, 3, 3) are the kernel size.

intended for the dense crowd. The structure of the convolutional layers group 2 (CLs2) in B_2 is shown in Fig. 6. Both B_1 and B_2 pathways are relative and can estimate a density map. For fusing the density maps, a dynamic weighting method named adaption discovery is used. Adaptation discovery is a process that permits the network to carry out feature recalibration for the weight of the B_1 and B_2 branches, through which it can learn to utilize global information to emphasize informative features while suppressing less helpful ones selectively. B_3 in Fig. 3 presents the details of this process. It contains a global average pooling (GAP) and two fully connected layers (FCs) followed by ReLU and Sigmoid-Normalization. The GAP can be used to calculate the global average value of each channel. The multi-layer feature map M that is extracted by the convolutional layers of the U-Net encoder is considered as an input to the GAP. Its dimensions are $h \times w \times c$, where h denoting height, w denoting width, and c are the number of channels. $M_c(i, j)$ is the element at location (i, j) in the $c - th$ channel (i, j). The output is $1 \times 1 \times c$. To capture the interdependencies between channels, two fully connected layers followed by an activation function of Sigmod, which are not shown in Fig. 3, have been added after the GAP. The first FC layer minimizes the dimension from c to $c/16$, and the second FC reduces the dimension from $c/16$ to $c/32$. An initial response w after the sigmoid function is obtained, the w adaptively recalibrates the weight of the B_1 and B_2 pathways. Thus, w is normalized into the interval of [0, 0.5] [16], [40], the output of the B_1 and B_2 paths can be computed as follows:

$$output_{B_1, B_2} = (1 - w)B_1 + wB_2 \tag{3}$$

IV. EXPERIMENTS

The evaluation metrics and experimental details are initially addressed in this section. The results of the proposed U-ASD Net are then reported and analyzed on five challenging crowd counting datasets.

A. EVALUATION METRICS

The counting accuracy of the CNN-based crowd counting networks can be measured by mean absolute error (MAE), mean squared error (MSE), and the resolution of the density map [33]. Further details are explained in the following paragraphs.

- The most well-known evaluation metrics in the scope of evaluating crowd counting methods are the MAE

TABLE 2. **Datasets details. Num: number of images/frames, Avg: average number of count, Min: minimal crowd count, Max: maximum crowd count, Total: total number of annotations.**

Dataset	Type	Scenes	Resolution	Num	Avg	Min	Max	Total
ShanghaiTech Part A	Image	482	Varied	482	501	33	3,139	241,677
ShanghaiTech Part B	Image	716	768×1024	716	123	9	578	88,488
UCF CC 50	Image	50	Varied	50	1279	96	4,633	63,974
UCSD	Video	1	158×238	2,000	24.9	11	46	49,885
Mall	Video	1	640×480	2,000	31	13	53	62,315
Haramain H1	Video	1	576×720	70	44	39	51	3,119
Haramain H2	Video	1	576×720	60	444	408	502	26,640
Haramain H3	Video	1	1280×720	60	523	465	572	31,388

and MSE, respectively, which can be described as below [41]–[43]:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|c_i - \hat{c}_i\right| \tag{4}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left|c_i - \hat{c}_i\right|^2 \tag{5}$$

where N represents the total number of patterns in the test set, c_i is the count label, and $\hat{c}_i$ is the predicted count value for the i-th test pattern. The MAE metric represents the precision of the estimated count, and the MSE metric is a measure of the robustness of counting.

- Peak signal-to-noise ratio (PSNR), the method of computing the mean square error between the predicted density map and its ground truth of all pixels, is preferred for determining the accuracy of the predicted map. Mathieu *et al.* [44] argued that the PSNR is a better metric for assessing quality. The PSNR is defined as follows [45], [46]:

$$PSNR(M, I) = 10log_{10}\frac{max_I^2}{(1/N)\sum_{j=0}^{N}(M_i - I_j)^2} \tag{6}$$

where M refers to the density map image, max_I is the highest possible value of image intensities, and N denotes the total number of pixels in the map image. Generally, a higher PSNR value shows a higher image quality.

- Structural Similarity Index (SSIM) is frequently utilized to assess the quality of the estimated density map [47]. The SSIM estimates the image similarity based on contrast, structure, and luminance, which can be calculated by multiplying the three terms described. The SSIM value is in the $[-1,\ 1]$ range. The higher the SSIM value, the lower the distortion has been. The SSIM formula is defined as follows [48]:

$$SSIM(g, p) = [l(g, p)]^{\alpha}.[c(g, p)]^{\beta}.[s(g, p)]^{\gamma} \tag{7}$$

where:

$$l(g, p) = \frac{2\mu_g\mu_p + C_1}{\mu_g^2 + \mu_p^2 + C_1},$$

$$c(g, p) = \frac{2\sigma_g\sigma_p + C_2}{\sigma_g^2 + \sigma_p^2 + C_2},$$

$$s(g, p) = \frac{2\sigma_{gp} + C_3}{\sigma_g\sigma_p + C_3}$$

μ_g, μ_p, σ_g, σ_p, σ_{gp} are the local means, standard deviations, and cross-covariance for both the ground-truth density (g) and predicted density (p) maps, respectively. If $\alpha = \beta = \gamma = 1$, and $C_3 = C_2/2$ the SSIM can be written as:

$$SSIM(g, p) = \frac{(2\mu_g\mu_p + C_1)(2\sigma_{gp} + C_2)}{(\mu_g^2 + \mu_p^2 + C_1)(\sigma_g^2 + \sigma_p^2 + C_2)} \tag{8}$$

B. EXPERIMENTAL SETUP

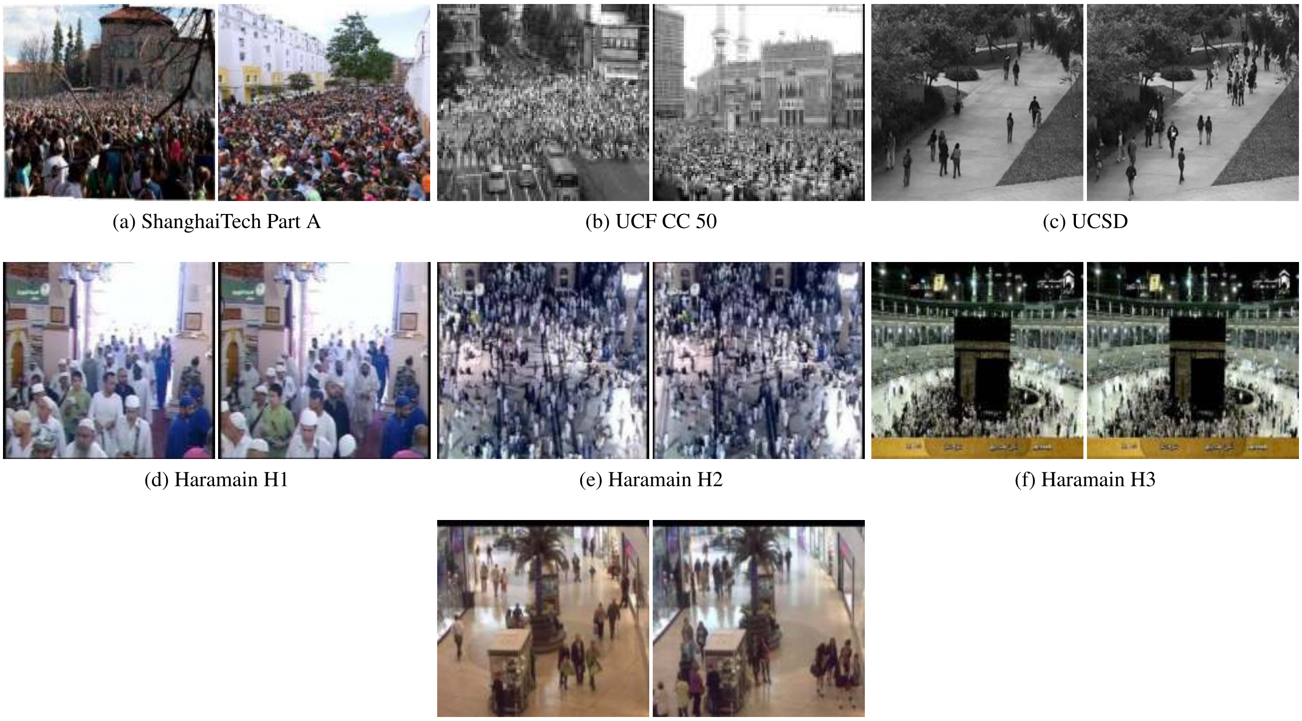
The U-ASD was tested on two image crowd counting benchmarks and three video crowd counting benchmarks (i.e., ShanghaiTech Part A, ShanghaiTech Part B, and UCF CC 50) and (i.e., UCSD, Mall, Haramain), respectively. Fig. 7 depicts some of their typical scenes. Table 2 lists the basic statistics of each dataset, and shows the total number of people in each dataset. As shown Table 2, the datasets have varying crowd densities, and ShanghaiTech Part A, ShanghaiTech Part B, and UCF CC 50 are highly imbalanced.

The training and evaluation were carried out in Python using PyTorch on a Tesla V100 GPU. Following [42], the original training images and frames, which have different resolutions as described in Table 2, are firstly resized to a resolution of 576 × 768, and the ground truths are formed at the same resolution.

1) Ground-truth generation: Since the CNN-based methods utilized for crowd counting require processing continuous data, and the available ground-truth information is discrete [16] thus, a conversion process is required to generate the density map and attention map information utilizing the discrete key points that represent the head annotations as shown in Fig. 8.

   - **Density map generation**: To obtain a density map (D_i) for each image in a dataset utilizing the available ground-truth information (labeled people heads), [25] is followed. The presence of a head at pixel p_i is reflected as a delta function $\delta(p - p_i)$. This allows the following interpretation of an image with N-labeled heads:

$$H(p) = \sum_{i=1}^{N}\delta(p - p_i) \tag{9}$$


(a) ShanghaiTech Part A

(b) UCF CC 50

(c) UCSD

(d) Haramain H1

(e) Haramain H2

(f) Haramain H3

(g) Mall

FIGURE 7. Examples of frames from the following datasets: ShanghaiTech Part A, UCF CC 50, UCSD, Haramain H1, Haramain H2, Haramain H3, and Mall, respectively.

This function can be convolved with a Gaussian kernel G to transform it into a continuous density function. Thus, the density can be formulated as:

$$F(p) = H(p) \times G_{\sigma}(p) \tag{10}$$

However, if the crowd is supposed to be uniformly distributed over each head, the average distance between the head and its nearest k neighbors estimates a rational approximation of the geometric distortion (resulting from the perspective effect). Consequently, the spread parameter σ for each individual in the picture should be determined based on the size of their head. A kernel with a window size of $\mu = 15$ and a spread parameter of $\sigma = 4$ are used in the experiments described in this paper.

- **Attention map generation**: The attention map (A_i) is generated following the methods in [30], [38] by first generating the density map with a larger spread parameter $\sigma = 6$. Then, a threshold to the corresponding density map is applied. The attention map can be obtained using the following formulated Equation:

$$A_i = \begin{cases} 0 & D_i < T \\ 1 & D_i \geqslant T \end{cases} \tag{11}$$

In our conducted experiments, the threshold is set to ($T = 0.001$). This threshold value was the best experimentally. Different threshold settings will change the performance, as shown in Table 3.

TABLE 3. Different threshold values to generate an attention map.

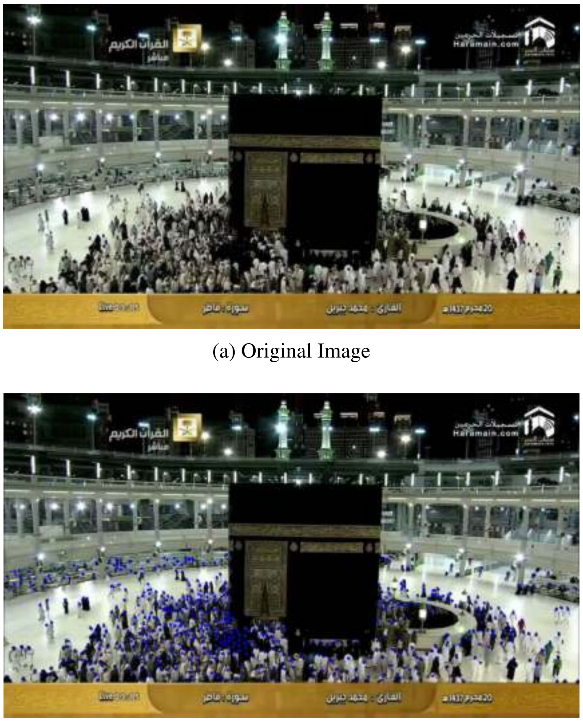
T	MAE	MSE
0.001	**7.5**	**12.4**
0.01	17.4	22.7
0.1	14.7	21.3

Fig. 9 shows a comparison between the density and attention maps. For visualization of the density map and the attention map, Fig. 9 (b) and (c) were created by Matplotlib *imshow*() function utilizing the jet cmap. Thus, the range values of the density map as well as 0(s) and 1(s) binary values of the attention map are mapping to the associated RGB value for the ‘‘jet’’ color scale.

2) Data augmentation: The images are indiscriminately cropped to 512 × 680. In like manner, density maps and binary maps are correspondingly resized, and the count labels are regenerated. Moreover, arbitrary horizontally flipping is utilized in the training step.
3) Implementation details: The MSE loss and BCE loss are used to train the whole U-ASD network, and the Adam optimizer is employed for optimization. The U-ASD model aims for optimizing the combined loss function as follows:

$$L_{Total} = \lambda_1 L_{bce} + \lambda_2 L_{mse} \tag{12}$$

where L_{mse} and L_{bce} are the loss functions for U and ASD, respectively, as mentioned in Section III C. λ_1 and λ_2 are parameters that are used as a balance between the loss values. The optimum values



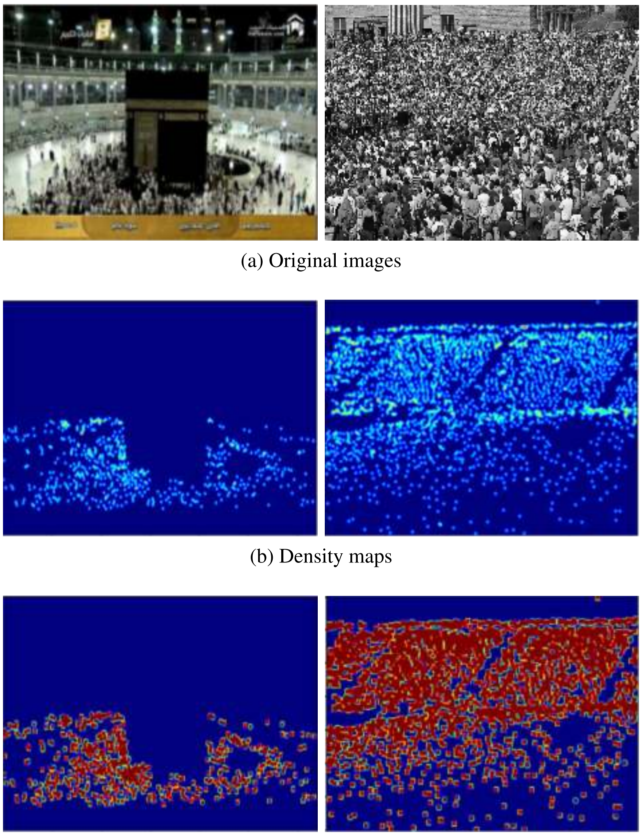
(a) Original Image

(b) Annotated Image

FIGURE 8. An example of head annotations in blue color from the Haramain H3 dataset.

for λ_1 and λ_2, which have been chosen empirically, are 20 and 1000, respectively. Fig. 10 represents the qualitative results of the U-ASD method on different test scenes. The sub-figures are, respectively, for each scene the: original image, ground-truth density map, and estimated density map. As shown in the sample results from UCSD, Mall, and Haramain H1, Fig. 10 (d, e, and f), the U-ASD model counts well not only under highly dense crowds but also counts well in sparse scenarios. As crowd density rises, people will appear to partially occlude one another, limiting the capacity of classic detection methods and prompting the development of density estimate models. Such situations can be noticed in Fig. 10 (a), (b), (c), (f), (h). Interestingly, the proposed model can locate these occluded people and count the crowds very well by producing high-quality density maps and thus providing accurate counting accuracy. The scenarios in the Mall dataset have strong perspective distortions, which results in significant variations in the scale and appearance of individual objects. Also, the occlusion resulted from some potted plants raises the complexity. As you can see in Fig. 10 (e), the produced density map from the U-ASD model on the Mall dataset locates the individual correctly and provides well counting.

4) Evaluation details: In the evaluation stage, a patch-based assessment as in [21] and [30] is used. The test images are cropped into patches and generated 9-overlapping units for each image. Then, a sliding window is run over the test image during the prediction process. Predictions are determined for each window before being aggregated to get the total count in the image.



(a) Original images

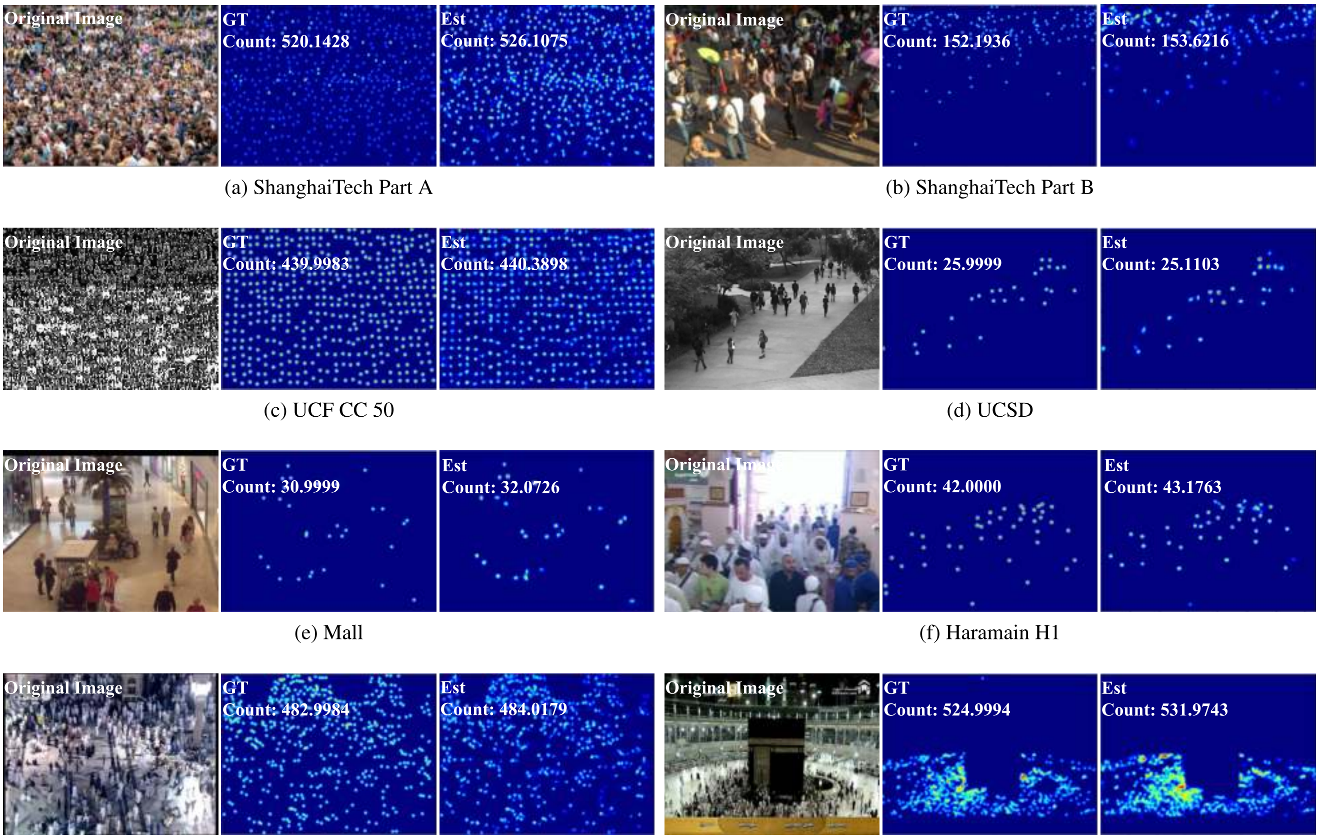
(b) Density maps

(c) Attention maps

FIGURE 9. The comparison between the density and attention maps. (a) Original images, Dataset: Haramain H3, and UCF CC 50, respectively, (b) Density map, (c) Attention map.

C. SHANGHAITECH DATASET EVALUATION

ShanghaiTech dataset [25] is a wide range and very challenging dataset, and it comprises two main parts. Part A contains 300 and 182 training and testing images, respectively. The images of this part have different resolutions and are gathered from the Internet. On the other hand, Part B comprises 400 and 316 training and testing images, respectively. The images of Part B have the same resolution of 768 × 1024 and have been collected from a metropolitan security camera. Table 4 shows the results of Part A and Part B of the ShanghaiTech dataset with other relevant mainstream methods (Zhang *et al.* [19], FCN [27], Flounder-Net [49], MCNN [25], Huang *et al.* [50], Cascaded-MTL [26], Switching-CNN [24], DecideNet [6], SaCNN [51], Wang *et al.* [40], DAN [52], ACSCP [31], CP-CNN [5], PCC Net [42], D-ConvNet [53], IG-CNN [54], L2R [55], HADF-Crowd [56], AU-CNN [57], CSRNet [20], AAFM [58], AWRFN [59], Zhang *et al.* [60], DENet [43] and U-ASD Net [ours]). Compared with other methods, U-ASD accomplishes the best MAE of 64.6 and the third-best MSE of 106.1 on Part A. In addition, our method outperforms



(a) ShanghaiTech Part A

(b) ShanghaiTech Part B

(c) UCF CC 50

(d) UCSD

(e) Mall

(f) Haramain H1

(g) Haramain H2

(h) Haramain H3

FIGURE 10. Qualitative results of U-ASD Net using different datasets with the ground truth density maps. Gt: Ground Truth Image, Est: Estimated Image.

all the state-of-the-art methods in Part B and accomplishes amazing results MAE of 7.5 and MSE of 12.4.

TABLE 4. Estimated errors on ShanghaiTech dataset with state-of-the-art methods.

Method	Part A		Part B	
	MAE ↓	**MSE ↓**	**MAE ↓**	MSE ↓
Zhang *et al.* [19]	181.8	277.7	32.0	49.8
FCN [27]	126.5	173.5	23.8	33.1
Flounder-Net [49]	113.0	166.1	22.4	39.2
MCNN [25]	110.2	173.2	26.4	41.3
Huang *et al.* [50]	-	-	20.2	35.6
Cascaded-MTL [26]	101.3	152.4	20.0	31.1
Switching-CNN [24]	90.4	135.0	21.6	33.4
DecideNet [6]	-	-	20.8	29.4
SaCNN [51]	86.8	139.2	16.2	25.8
Wang *et al.* [40]	88.5	147.6	17.6	26.8
DAN [52]	81.8	134.7	13.2	20.1
ACSCP [31]	75.7	102.7	17.2	27.4
CP-CNN [5]	73.6	106.4	20.1	30.1
PCC Net [42]	73.5	124.0	11.0	19.0
D-ConvNet [53]	73.5	112.3	18.7	26.0
IG-CNN [54]	72.5	118.2	13.6	21.1
L2R [55]	72.0	106.6	14.4	23.8
HADF-Crowd [56]	71.1	111.6	9.7	15.7
AU-CNN [57]	70.4	117.5	8.6	13.0
CSRNet [20]	68.2	115.0	10.6	16.0
AAFM [58]	67.1	104.2	10.6	15.8
AWRFN [59]	66.7	109.1	11.5	19.5
Zhang *et al.* [60]	-	-	8.3	12.9
DENet [43]	65.5	**101.2**	9.6	15.4
U-ASD Net [ours]	**64.6**	106.1	**7.5**	**12.4**

D. UCF CC 50 DATASET EVALUATION

UCF CC 50 dataset is created by [61], and it covers various views with different perspective distortion. UCF CC 50 is constituted of only fifty images but has large head annotations of 63,074, and the images differ in the number of individuals with a range from 96 to 4,633 with an average number of 1,279. Since this dataset has only fifty images, the state-of-the-art approaches utilize the traditional 5-fold cross-validation procedure to assess their methods [19], [20], [42], [61]. Thus, the 5-fold cross-validation is also applied to assess the proposed U-ASD method. Fig. 11 illustrates the estimated errors by applying 5-Fold Cross-Validation. As shown in Fig. 11, the average MAE and MSE are 232.3 and 217.8, respectively. Table 5 shows that the U-ASD method presents the third-best result in terms of the MAE metric and the best result in terms of the MSE metric with reducing the estimation MSE error by about 50.5 compared with the ASANet method.

E. UCSD DATASET EVALUATION

UCSD dataset [64] has several video frames for the same scene snapped by surveillance cameras with a resolution of 238 × 158. It includes 2,000 frames with a total of 49,885 annotated people. In this dataset, the number of individuals in the frames is sparse, with varying ranges of 11 − 46. To use the UCSD dataset in analyzing the

TABLE 5. Estimated errors using state-of-the-art methods on UCF CC 50 dataset.

Method	MAE ↓	MSE ↓
Zhang *et al.* [19]	467.0	498.5
Hydra-2s [62]	333.7	425.2
Hydra-3s [62]	465.7	371.8
MCNN [25]	377.6	509.1
Walach *et al.* [22]	364.4	341.4
FCN [27]	338.6	424.5
Cascaded-MTL [26]	322.8	397.9
Switching-CNN [24]	318.1	439.2
SaCNN [51]	314.9	424.8
DAN [52]	309.6	402.6
CP-CNN [5]	295.8	320.9
IG-CNN [54]	291.4	349.4
ACSCP [31]	291.0	404.6
D-ConvNet [53]	288.4	404.7
L2R [55]	279.6	388.9
CSRNet [20]	266.1	397.5
SAN [21]	258.4	334.9
AWRFN [59]	257.3	337.2
TEDnet [28]	249.4	354.5
AAFM [58]	247.1	329.4
PCC Net [42]	240.0	315.5
DENet [43]	241.9	345.4
W-Net [30]	201.9	309.2
ASANet [63]	**185.5**	268.3
U-ASD Net [ours]	232.3	**217.8**

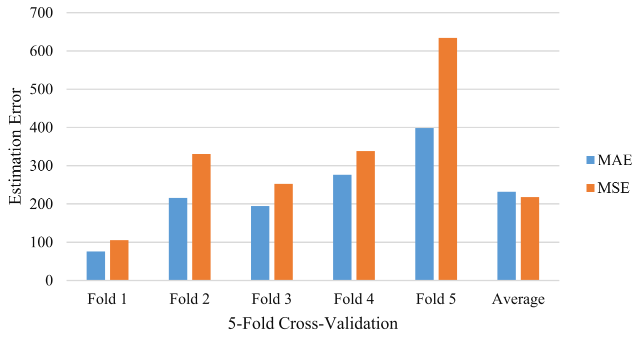


FIGURE 11. Estimated errors based on UCF CC 50 dataset by applying 5-fold cross-validation.

performance of the U-ASD method, the original settings in [64] are followed. The image sequences 601-1400 are used as the training set, and the remaining 1200 image sequences as the testing set. The results of the UCSD dataset are recorded in Table 6. The results of U-ASD are comparable with the state-of-the-art methods. The U-ASD has obtained the best MSE with 2.1.

TABLE 6. Estimated errors on UCSD dataset with state-of-the-art methods.

Method	MAE ↓	MSE ↓
Gaussian process regression [64]	2.2	8.0
Ridge Regression [65]	2.3	7.8
Cumulative Attribute Regression [66]	2.1	6.9
Hydra CNN [62]	1.7	-
Count forest [67]	**1.6**	4.4
ConvLSTM-nt [68]	1.7	3.5
Zhang *et al.* [19]	**1.6**	3.3
Switching-CNN [24]	**1.6**	**2.1**
U-ASD Net [ours]	1.7	**2.1**

TABLE 7. Estimated errors on Mall dataset with state-of-the-art methods.

Method	MAE ↓	MSE ↓
R-FCN [69]	6.0	5.5
Faster R-CNN [70]	5.9	6.6
Gaussian process regression [64]	3.7	20.1
Ridge regression [65]	3.6	19.0
Cumulative Attribute Regression [66]	3.4	17.7
MoCNN [23]	2.8	13.4
Count forest [67]	2.5	10
Weighted VLAD [71]	2.4	9.1
ACM-CNN [41]	2.3	3.1
MCNN+SEG+LR [72]	2.2	2.8
Bi-ConvLSTM [68]	2.1	7.6
Exemplary-Density [73]	**1.8**	2.7
U-ASD Net [ours]	**1.8**	**2.2**

TABLE 8. Estimated errors on the Haramain dataset.

Dataset	Without Cross-Validation		5-Fold Cross-Validation	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓
Haramain H1	2.1	2.4	**1.5**	**2.3**
Haramain H2	37.1	37.6	**7.8**	**8.6**
Haramain H3	4.8	5.6	10.6	14.1
Average	14.7	15.2	**6.6**	**8.3**

F. MALL DATASET EVALUATION

Mall dataset [65] is recorded inside a shopping centre by a public surveillance camera. This dataset poses some challenges, such as glass surface reflections and lighting conditions. The first 800 video sequences are used for training, while the other remaining 1,200 frames are used for testing, as described in [65]. A comparison against R-FCN [69], Faster R-CNN [70], Gaussian process regression [64], Ridge regression [65], Cumulative Attribute Regression [66], MoCNN [23], Count forest [67], Weighted VLAD [71], ACM-CNN [41], MCNN+SEG+LR [72], Bi-ConvLSTM [68], and Exemplary-Density [73], is given in Table 7. The U-ASD Net achieves the best performance.

G. HARAMAIN DATASET EVALUATION

The Haramain dataset includes various scenes at the holy haram in Mecca and Al-Madinah. People from all over the globe gather at the holy haram places for the sake of worship. Therefore, maintaining people's comfort while praying is considered a major management goal. About more than three million people visited the holy haram in Madinah each year. It covers an area of over 98,000 m^2 and has 42 multi-door entrances [74]. Consequently, maintaining a fine flow at all areas and entrances is a challenging task. Estimating the number of people in the crowd scenes helps to smooth the distribution of up to 167,000 people throughout the holy haram at a time.

To help addressing the crowd management in holy places, the Haramain dataset with its manual annotations is introduced, consisting of three parts for three different scenes. The first and second parts, called H1 and H2, respectively, include 70 and 60 image sequences from two scenes at Madinah mosque. The third part, called H3, comprises 60 image sequences from al-sahn area at al-haram al-sharif

TABLE 9. The detailed information of the U-ASD net and the main state-of-the-art methods on ShanghaiTech Part A dataset, U-ASD net* uses the nearest upsampling in the U-net part.

Methods	**MAE ↓**	**MSE ↓**	**PSNR ↑**	**SSIM ↑**	**Parameters**	**Runtime (ms)**	**Device**	**Pre-train**
Cascaded-MTL [26]	126.5	173.5	-	-	**0.12M**	**3**	TITAN-X	
Switching-CNN [24]	90.4	135	21.91	0.67	15.1M	153	-	✓
CP-CNN [5]	73.6	106.4	21.72	0.72	62.9M	5113	-	✓
PCC Net [42]	73.5	124	22.78	0.74	0.55M	89	1080Ti	
U-ASD Net* [ours]	69.7	106.8	41.09	**0.96**	31.4M	62	Tesla V100	✓
U-ASD Net [ours]	**64.6**	**106.1**	**41.41**	**0.96**	31.4M	94	Tesla V100	✓

mosque in Mecca, Saudi Arabia, during the pilgrimage season. The resolutions for each part and other details are shown in Table 2. Since the annotation process requires a lot of time, the length of video clips for this dataset has been limited, and 5-fold cross-validation is applied. Fig. 12 shows the estimated errors by applying 5-Fold Cross-Validation. Table 8 shows the results of the proposed U-ASD on the Haramain dataset. As clearly seen, applying the 5-fold cross-validation improves performance metrics by 8.1 and 6.9 on average for MAE and MSE metrics.

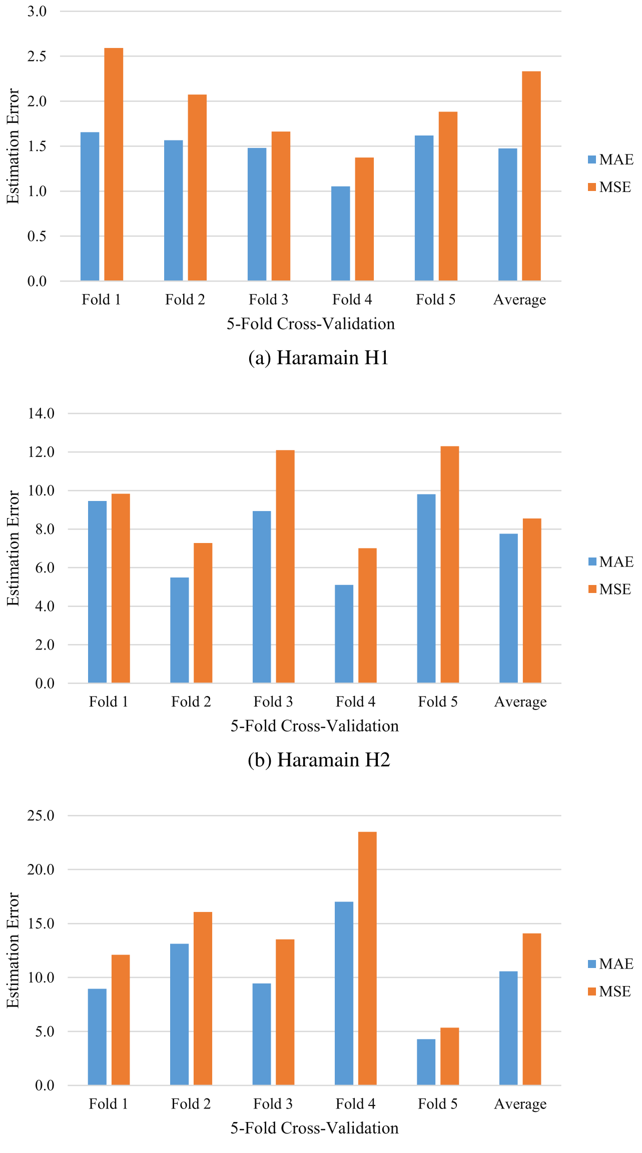
V. DISCUSSION AND ANALYSIS

The specifics of the proposed U-ASD model were contrasted with the state-of-the-art methods (Cascaded-MTL [26], Switching-CNN [24], CP-CNN [5], and PCC Net [42]) to demonstrate the superiority of our method. The four main metrics for evaluating density estimation efficiency are mentioned and calculated in Table 9 on the ShanghaiTech Part A: MAE, MSE, PSNR, and SSIM. As can be observed, U ASD Net is the best. The integration of U-Net with ASD-Net is responsible for this performance since it allows the whole model to implicitly identify all crowd scenarios and respond to diverse crowd images in a highly scenario-specific manner. U-ASD Net* is the model, which uses the nearest upsampling in the U-Net part. As clear in Table 9, the U-ASD provides better performance in terms of the counting accuracy (i.e., MAE and MSE), which comes at the expense of the runtime. In the experiments, the computational complexity, in terms of the number of parameters and training runtime, and the quality of the estimated density maps are also measured. Further details are in the next subsections.

A. COMPUTATIONAL COMPLEXITY

To reduce the complexity of the U-ASD Net, the VGG16-bn network (except the fully connected layers) is used for the encoder part of the U-Net and as a backbone for the ASD branches. In addition, for simplicity and to avoid adding complexity to the U-ASD Net, the original ASD Net is used without extra layers, except adding a nearest upsample layer at the output of the net to fuse it with the output from the U-Net.

Table 9 includes information on the computation complexity in terms of the number of parameters and execution runtime. Even though Cascaded-MTL [26] presents the lightest model with only 0.12M parameters and 3ms runtime among other models [5], [24], [42], it has the worst estimation performance. During the evaluation phase, U-ASD takes 94ms to process a 512 × 680 frame from ShanghaiTech Part A dataset on one Tesla V100 GPU. Since humans, in general, do not move so fast as well as each frame does not require to be analyzed, this runtime speed is adequate for several realistic applications [75]. Moreover, comparing the U-ASD with Pretrain models on ImageNet, the U-ASD provides a faster execution time than Switching-CNN and CP-CNN models.

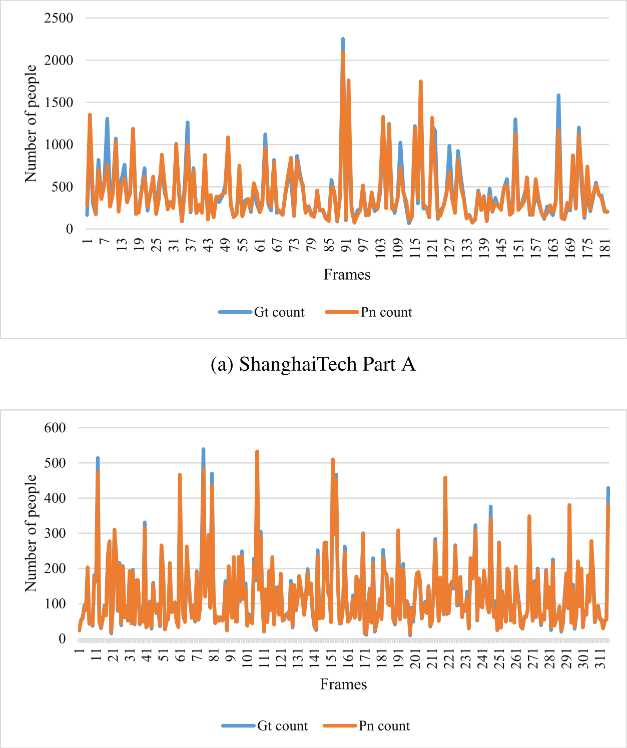


(a) Haramain H1

(b) Haramain H2

(c) Haramain H3

FIGURE 12. Estimated errors based on Haramain dataset by applying 5-fold cross-validation.



(a) ShanghaiTech Part A

(b) ShanghaiTech Part B

FIGURE 13. **The actual ground truth Gt and the predicted number Pn of crowd for ShanghaiTech dataset.**

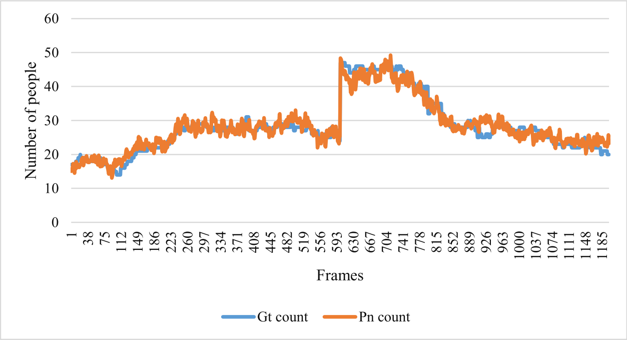


FIGURE 14. **The actual ground truth Gt and the predicted number Pn of crowd for UCSD dataset.**

Thus, taking into account the performance metrics (MAE, MSE, PSNR, and SSIM), and the number of parameters, the proposed U-ASD Net is very competitive.

B. QUALITY OF THE PREDICTED DENSITY MAP

To test the quality of the estimated density maps produced by U-ASD Net, the PSNR and SSIM were computed on ShanghaiTech Part A and Part B, UCF CC 50 and UCSD datasets for the MCNN [25], CP-CNN [5], CSRNet [20], ADCrowdNet [76], PCC Net [42], and U-ASD methods. Table 10 shows the PSNR and SSIM comparison. Clearly, the U-ASD offers the best structural integrity.

The estimations over time of each frame in ShanghaiTech Part A, ShanghaiTech Part B, UCSD, and Mall datasets concerning their ground truth are illustrated in Figs. 13, 14, and 15.

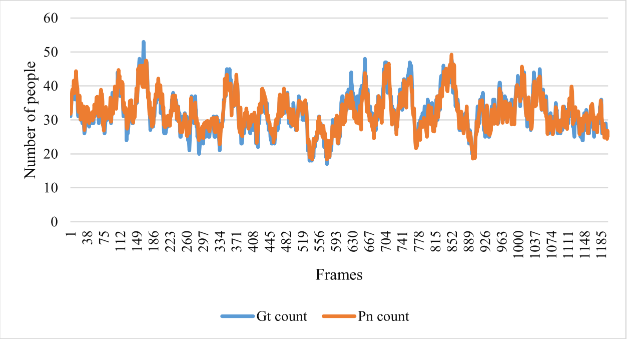
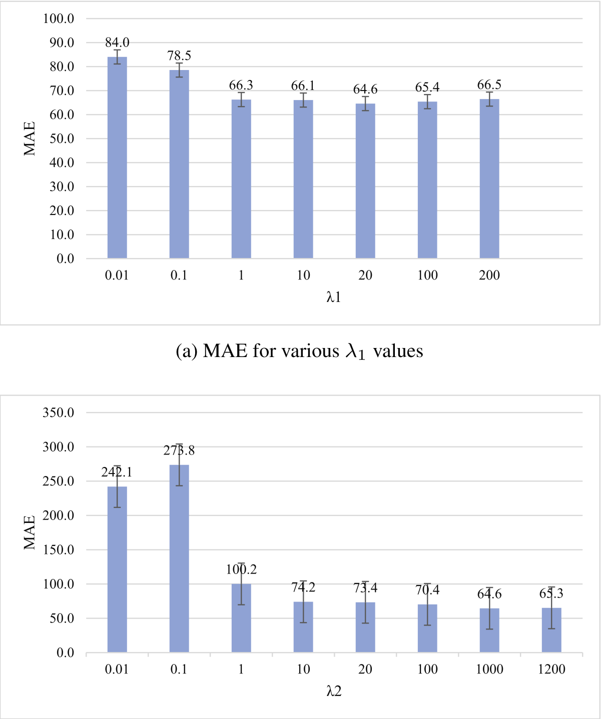


FIGURE 15. **The actual ground truth Gt and the predicted number Pn of crowd for Mall dataset.**



(a) MAE for various λ_1 values

(b) MAE for various λ_2 values

FIGURE 16. **MAE comparisons on Part A of the ShanghaiTech dataset for various λ_1 and λ_2 values.**

Interestingly, the prediction counts are almost identical to the ground truth counts.

C. PARAMETER λ_1 AND λ_2 STUDY

Comparative experiments on Part A of the ShanghaiTech dataset were conducted in order to determine the best values of λ_1 and λ_2 in Equation 12. Fig. 16 (a) illustrates that as the value of λ_1 increases, the MAE error value decreases, and the lowest error is acquired at $\lambda_1 = 20$. The error then increases since the weight of the L_{mse} loss becomes too significant in comparison to the L_{bce} loss. As a result, in our experiments, 20 is identified to λ_1. Similarly, as shown in Fig. 16 (b), the lowest MAE is acquired when λ_2 is specified by 1000.

D. THE PERFORMANCE OF U-ASD NET COMPONENTS

In the conducted experiments, it is noted that training the U-Net without the ASD Net in ShanghaiTech Part B achieved

TABLE 10. PSNR and SSIM comparison to demonstrate the quality of the predicted density map.

Dataset	MCNN [25]		CP-CNN [5]		CSRNet [20]		ADCrowdNet [76]		PCC Net [42]		U-ASD [ours]	
	PSNR	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
ShanghaiTech Part A	-	0.52	21.72	0.72	23.79	0.76	24.48	0.88	22.78	0.74	**41.41**	**0.96**
ShanghaiTech Part B	-	-	-	-	27.02	0.89	29.35	0.97	-	-	**49.47**	**0.99**
UCF CC 50			-	-	18.76	0.52	20.08	0.81	-	-	**37.17**	**0.91**
UCSD	-	-	-	-	20.02	0.86	26.39	0.93	-	-	**52.78**	**1**
Mall	-	-	-	-	-	-	-	-	-	-	**54.06**	**1**

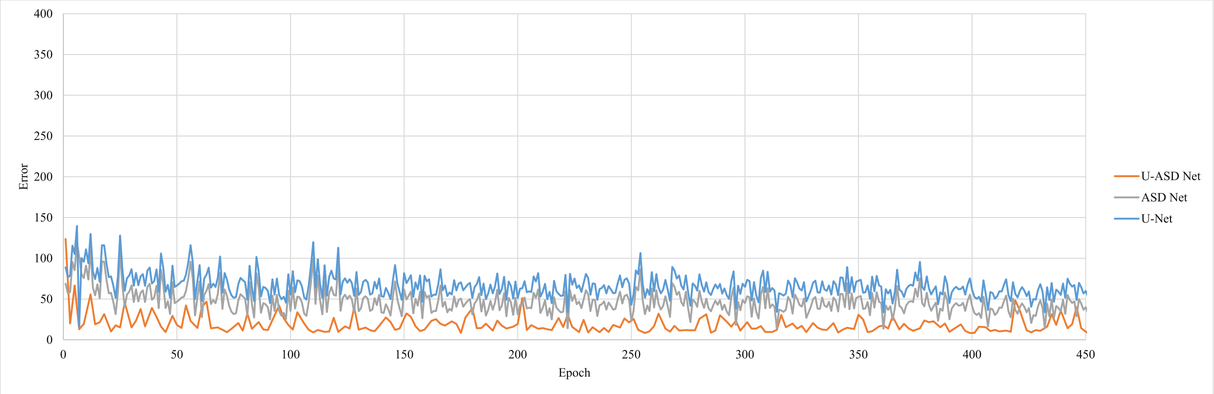


FIGURE 17. The curves of testing loss for U-Net, ASD Net, and U-ASD Net. After epoch 7, the U-Net's performance degrades, and the loss increases. Thus, the training of U-Net is stopped at this early stopping. U-ASD has the smoothest convergence curve and lowest error.

TABLE 11. Comparison of different U-ASD Net components using ShanghaiTech Part B dataset.

Network	MAE	MSE	PSNR	SSIM
ASD Net	13.6	24.7	24.91	0.54
U-Net	16.4	25.0	47.98	**0.99**
U-ASD Net	**7.5**	**12.4**	**49.47**	**0.99**

the best MAE value at epoch number 7, as shown in Fig. 17. After epoch number 7, the U-Net drastically degrade the counting performance and the loss goes up. Thus, the training of U-Net is stopped at this early stopping. This was the main reason the ASD Net was introduced as a binary classifier. The ASD Net, when independently trained, counts better than U-Net, whereas the quality of the estimated density map is lower than the U-Net. Both networks are combined using a combined loss function as described in Equation 12 (i.e., BCE loss and MSE loss), and the whole U-ASD Net is trained in an end-to-end fashion. As shown in Table 11, integrating U and ASD networks helps in increasing the counting accuracy and improve the quality of the produced density maps.

VI. CONCLUSION

This paper proposes an end-to-end trainable hybrid modified network architecture, named U-ASD Net, by integrating two novel architectures designed for image segmentation and crowd counting. The proposed U-ASD model has the ability to predict precise and high-quality density maps at half resolution compared to the input. The PSNR and SSIM metrics have proven the superiority of the proposed model in generating high-quality density maps. Moreover, the proposed model contributes in alleviating the drawbacks present in the state-of-the-art methods by addressing both sparse and dense scenes for crowd counting efficiently.

In the modified U-Net, the up-sampling algorithm is changed from nearest to max-unpooling for upsampling using the memorized indices used in U-Net. This accomplishes high counting accuracy. The proposed model achieves the lowest count error in terms of the MAE in ShanghaiTech Part A, Part B, and Mall datasets with 64.6, 7.5, and 1.8, respectively. Moreover, it achieves the lowest count error in terms of the MSE in ShanghaiTech Part B, UCF CC 50, UCSD, and Mall datasets with 12.4, 217.8, 2.1, 2.2, respectively. In addition, the proposed model accomplishes the best quality density maps on all the utilized datasets.

To assist in addressing crowd management and control in the holy places at Mecca and Al-Madinah, a new dataset, named Haramain dataset, is introduced, which consists of three parts for three different scenes. The proposed U-ASD model is applied in this dataset, and all the MAE, MSE, PSNR, and SSIM metrics have shown promising results.

Extensive experiments on four benchmark datasets and comparisons with recent state-of-the-art methods presented the substantial improvements accomplished by the proposed model.

ACKNOWLEDGEMENT

The authors extend their appreciation to the Deputyship of Research and Innovation, Ministry of Education in Saudi Arabia, for funding this research work through project number 758. The authors also would like to thank the Research Management Center of Universiti Teknologi Malaysia for managing this fund under vot. no. 4C396.

REFERENCES

[1] A. Hafeezallah and S. Abu-Bakar, "Crowd counting using statistical features based on curvelet frame change detection," *Multimedia Tools Appl.*, vol. 76, no. 14, pp. 15777–15799, Jul. 2017.

[2] A. A. H. Allah, S. A. A. Bakar, and W. A. Orfali, ‘‘Curvelet transform subdifference image for crowd estimation,’’ in *Proc. IEEE Int. Conf. Control Syst., Comput. Eng. (ICCSCE)*, Nov. 2014, pp. 502–506.

[3] A. Al-Dhamari, R. Sudirman, and N. H. Mahmood, ‘‘Abnormal behavior detection using sparse representations through sequential generalization of k-means,’’ *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 29, no. 1, pp. 152–168, Jan. 2021.

[4] D. Kang, Z. Ma, and A. B. Chan, ‘‘Beyond counting: Comparisons of density maps for crowd analysis tasks-counting, detection, and tracking,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1408–1422, May 2019.

[5] V. A. Sindagi and V. M. Patel, ‘‘Generating high-quality crowd density maps using contextual pyramid CNNs,’’ in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1861–1870.

[6] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, ‘‘DecideNet: Counting varying density crowds through attention guided detection and density estimation,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5197–5206.

[7] A. Al-Dhamari, R. Sudirman, and N. H. Mahmood, ‘‘Transfer deep learning along with binary support vector machine for abnormal behavior detection,’’ *IEEE Access*, vol. 8, pp. 61085–61095, 2020.

[8] A. Al-Dhamari, R. Sudirman, N. H. Mahmood, N. H. Khamis, and A. Yahya, ‘‘Online video-based abnormal detection using highly motion techniques and statistical measures,’’ *Telkomnika*, vol. 17, no. 4, pp. 2039–2047, 2019.

[9] Z. Wang, W. Li, Y. Shen, and B. Cai, ‘‘4-D SLAM: An efficient dynamic Bayes network-based approach for dynamic scene understanding,’’ *IEEE Access*, vol. 8, pp. 219996–220014, 2020.

[10] O. Ronneberger, P. Fischer, and T. Brox, ‘‘U-Net: Convolutional networks for biomedical image segmentation,’’ in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[11] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, ‘‘Stacked deconvolutional network for semantic segmentation,’’ *IEEE Trans. Image Process.*, early access, Jan. 25, 2019, doi: 10.1109/TIP.2019.2895460.

[12] G. Lin, A. Milan, C. Shen, and I. Reid, ‘‘RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[13] E. Wang, M. Zhang, X. Cheng, Y. Yang, W. Liu, H. Yu, L. Wang, and J. Zhang, ‘‘Deep learning-enabled sparse industrial crowdsensing and prediction,’’ *IEEE Trans. Ind. Informat.*, vol. 17, no. 9, pp. 6170–6181, Sep. 2021.

[14] H. Yin, Z. Yu, L. Wang, J. Wang, L. Han, and B. Guo, ‘‘ISI-ATasker: Task allocation for instant-sensing-instant-actuation mobile crowd sensing,’’ *IEEE Internet Things J.*, early access, Jul. 6, 2021, doi: 10.1109/JIOT.2021.3095160.

[15] V. Ranjan, H. Le, and M. Hoai, ‘‘Iterative crowd counting,’’ in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 270–285.

[16] X. Wu, Y. Zheng, H. Ye, W. Hu, T. Ma, J. Yang, and L. He, ‘‘Counting crowds with varying densities via adaptive scenario discovery framework,’’ *Neurocomputing*, vol. 397, pp. 127–138, Jul. 2020.

[17] K. Simonyan and A. Zisserman, ‘‘Very deep convolutional networks for large-scale image recognition,’’ 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[18] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, ‘‘Fast crowd density estimation with convolutional neural networks,’’ *Eng. Appl. Artif. Intell.*, vol. 43, pp. 81–88, Aug. 2015.

[19] C. Zhang, H. Li, X. Wang, and X. Yang, ‘‘Cross-scene crowd counting via deep convolutional neural networks,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 833–841.

[20] 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.

[21] 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.

[22] E. Walach and L. Wolf, ‘‘Learning to count with CNN boosting,’’ in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 660–676.

[23] S. Kumagai, K. Hotta, and T. Kurita, ‘‘Mixture of counting CNNs: Adaptive integration of CNNs specialized to specific appearance for crowd counting,’’ 2017, *arXiv:1703.09393*. [Online]. Available: http://arxiv.org/abs/1703.09393

[24] D. B. Sam, S. Surya, and R. V. Babu, ‘‘Switching convolutional neural network for crowd counting,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4031–4039.

[25] 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.

[26] V. A. Sindagi and V. M. Patel, ‘‘CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting,’’ in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.

[27] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, ‘‘Fully convolutional crowd counting on highly congested scenes,’’ 2016, *arXiv:1612.00220*. [Online]. Available: http://arxiv.org/abs/1612.00220

[28] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao, ‘‘Crowd counting and density estimation by trellis encoder-decoder networks,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6133–6142.

[29] V. Nekrasov, C. Shen, and I. Reid, ‘‘Light-weight RefineNet for real-time semantic segmentation,’’ 2018, *arXiv:1810.03272*. [Online]. Available: http://arxiv.org/abs/1810.03272

[30] V. K. Valloli and K. Mehta, ‘‘W-Net: Reinforced U-Net for density map estimation,’’ 2019, *arXiv:1903.11249*. [Online]. Available: http://arxiv.org/abs/1903.11249

[31] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, ‘‘Crowd counting via adversarial cross-scale consistency pursuit,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5245–5254.

[32] V.-S. Huynh, V.-H. Tran, and C.-C. Huang, ‘‘Iuml: Inception U-Net based multi-task learning for density level classification and crowd density estimation,’’ in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 3019–3024.

[33] N. Ilyas, A. Shahzad, and K. Kim, ‘‘Convolutional-neural network-based image crowd counting: Review, categorization, analysis, and performance evaluation,’’ *Sensors*, vol. 20, no. 1, p. 43, Dec. 2019.

[34] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, ‘‘U2-Net: Going deeper with nested U-structure for salient object detection,’’ *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.

[35] M. Zhao, J. Zhang, C. Zhang, and W. Zhang, ‘‘Leveraging heterogeneous auxiliary tasks to assist crowd counting,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12736–12745.

[36] R. Imtiaz, T. M. Khan, S. S. Naqvi, M. Arsalan, and S. J. Nawaz, ‘‘Screening of glaucoma disease from retinal vessel images using semantic segmentation,’’ *Comput. Electr. Eng.*, vol. 91, May 2021, Art. no. 107036.

[37] F. Wang, J. Sang, Z. Wu, Q. Liu, and N. Sang, ‘‘Hybrid attention network based on progressive embedding scale-context for crowd counting,’’ 2021, *arXiv:2106.02324*. [Online]. Available: http://arxiv.org/abs/2106.02324

[38] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, and T. Yao, ‘‘Dual path multi-scale fusion networks with attention for crowd counting,’’ 2019, *arXiv:1902.01115*. [Online]. Available: http://arxiv.org/abs/1902.01115

[39] Z. Yan, R. Zhang, H. Zhang, Q. Zhang, and W. Zuo, ‘‘Crowd counting via perspective-guided fractional-dilation convolution,’’ *IEEE Trans. Multimedia*, early access, Jun. 30, 2021, doi: 10.1109/TMM.2021.3086709.

[40] L. Wang, W. Shao, Y. Lu, H. Ye, J. Pu, and Y. Zheng, ‘‘Crowd counting with density adaption networks,’’ 2018, *arXiv:1806.10040*. [Online]. Available: http://arxiv.org/abs/1806.10040

[41] Z. Zou, Y. Cheng, X. Qu, S. Ji, X. Guo, and P. Zhou, ‘‘Attend to count: Crowd counting with adaptive capacity multi-scale CNNs,’’ *Neurocomputing*, vol. 367, pp. 75–83, Nov. 2019.

[42] 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.

[43] 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.

[44] M. Mathieu, C. Couprie, and Y. LeCun, ‘‘Deep multi-scale video prediction beyond mean square error,’’ 2015, *arXiv:1511.05440*. [Online]. Available: http://arxiv.org/abs/1511.05440

[45] A. K. Al-Dhamari and K. A. Darabkh, ‘‘Block-based steganographic algorithm using modulus function and pixel-value differencing,’’ *J. Softw. Eng. Appl.*, vol. 10, no. 1, pp. 56–77, 2017.

[46] H. Hiary, K. Eddin, M. S., and A. Al-Dhamari, ‘‘A hybrid steganography system based on LSB matching and replacement,’’ *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 9, pp. 374–380, 2016.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[48] K. A. Darabkh, A. K. Ai-Dhamari, and I. F. Jafar, "A new steganographic algorithm based on multi directional PVD and modified LSB," *J. Inf. Technol. Control*, vol. 46, no. 1, pp. 16–36, 2017.

[49] J. Chen, S. Xiu, X. Chen, H. Guo, and X. Xie, "Flounder-Net: An efficient CNN for crowd counting by aerial photography," *Neurocomputing*, vol. 420, pp. 82–89, Jan. 2021.

[50] S. Huang, X. Li, Z. Zhang, F. Wu, S. Gao, R. Ji, and J. Han, "Body structure aware deep crowd counting," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1049–1059, Mar. 2018.

[51] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1113–1121.

[52] H. Li, X. He, H. Wu, S. A. Kasmani, R. Wang, X. Luo, and L. Lin, "Structured inhomogeneous density map learning for crowd counting," 2018, *arXiv:1801.06642*. [Online]. Available: http://arxiv.org/abs/1801.06642

[53] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, and G. Zheng, "Crowd counting with deep negative correlation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5382–5390.

[54] D. B. Sam, N. N. Sajjan, R. V. Babu, and M. Srinivasan, "Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3618–3626.

[55] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7661–7669.

[56] N. Ilyas, B. Lee, and K. Kim, "HADF-crowd: A hierarchical attention-based dense feature extraction network for single-image crowd counting," *Sensors*, vol. 21, no. 10, p. 3483, May 2021.

[57] D. Wu, Z. Fan, and M. Cui, "Average up-sample network for crowd counting," *Appl. Intell.*, pp. 1–13, May 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10489-021-02470-8

[58] Z. Duan, H. Chen, and J. Deng, "AAFM: Adaptive attention fusion mechanism for crowd counting," *IEEE Access*, vol. 8, pp. 138297–138306, 2020.

[59] S. Peng, L. Wang, B. Yin, Y. Li, Y. Xia, and X. Hao, "Adaptive weighted crowd receptive field network for crowd counting," *Pattern Anal. Appl.*, vol. 24, no. 2, pp. 805–817, May 2021.

[60] S. Zhang, H. Li, and W. Kong, "A cross-modal fusion based approach with scale-aware deep representation for RGB-D crowd counting and density estimation," *Expert Syst. Appl.*, vol. 180, Oct. 2021, Art. no. 115071.

[61] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.

[62] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 615–629.

[63] X. Chen, H. Yan, T. Li, J. Xu, and F. Zhu, "Adversarial scale-adaptive neural network for crowd counting," *Neurocomputing*, vol. 450, pp. 14–24, Aug. 2021.

[64] A. B. Chan, Z.-S. John Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[65] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, 2012, p. 3.

[66] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2467–2474.

[67] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "COUNT forest: CO-voting uncertain number of targets using random forest for crowd density estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3253–3261.

[68] F. Xiong, X. Shi, and D.-Y. Yeung, "Spatiotemporal modeling for crowd counting in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5151–5159.

[69] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," 2016, *arXiv:1605.06409*. [Online]. Available: http://arxiv.org/abs/1605.06409

[70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*. [Online]. Available: http://arxiv.org/abs/1506.01497

[71] 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. 2018.

[72] J. He, X. Wu, J. Yang, and W. Hu, "CPSPNet: Crowd counting via semantic segmentation framework," in *Proc. IEEE 32nd Int. Conf. Tools With Artif. Intell. (ICTAI)*, Nov. 2020, pp. 1104–1110.

[73] Y. Wang and Y. Zou, "Fast visual object counting via example-based density estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3653–3657.

[74] A. A. H. Allah, S. A. Abu-Bakar, and W. A. Orfali, "Sub-difference image of curvelet transform for crowd estimation: A case study at the Holy Haram in Madinah," *Res. J. Appl. Sci., Eng. Technol.*, vol. 11, no. 7, pp. 740–745, Nov. 2015.

[75] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1774–1783.

[76] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, "ADCrowdNet: An attention-injective deformable convolutional network for crowd understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3225–3234.

ADEL HAFEEZALLAH received the B.Sc. and M.Sc. degrees in electrical engineering from King Abdulaziz University (KAU), Saudi Arabia, and the Ph.D. degree in electrical engineering from Universiti Teknologi Malaysia (UTM), Malaysia. Currently, he is an Assistant Professor with the College of Engineering, Taibah University, and a Researcher for one of the Ministry of Education's international collaboration initiatives for crowd management. His research interests include signal processing, computer vision, and machine learning.

AHLAM Al-DHAMARI received the B.Sc. degree in computer engineering from Hodeidah University, Yemen, the M.Sc. degree in computer engineering and networks from the University of Jordan, Jordan, and the Ph.D. degree in electrical engineering from Universiti Teknologi Malaysia (UTM), Malaysia. Currently, she is a Researcher with Universiti Teknologi Malaysia under the Postdoctoral Fellowship scheme for the project "Smart Crowd Surveillance and Management System for Pilgrimages." Her research interests include image and video processing, computer vision, machine learning, deep learning, computer architectures, and crowd analysis and management.

SYED ABD RAHMAN ABU-BAKAR (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Clarkson University, Potsdam, New York, USA, the M.S.E.E. degree from Georgia Tech, and the Ph.D. degree from the University of Bradford, U.K. He has been with the Faculty of Engineering, School of Electrical Engineering, Universiti Teknologi Malaysia, since 1992, where he is currently a Full Professor with the Electronics and Computer Engineering Division. In 2004, he formed the Computer Vision, Video and Image Processing Research Laboratory and has become the Head since then. He has published more than 150 scientific articles both at national and international levels. His research interests include computer vision and image processing with applications in video-based security and surveillance, medical image processing, and biometrics. In 2019, he received the Meritorious Regional Chapter Service Award from the IEEE Signal Processing Society. He was the Chair of the IEEE Signal Processing Society Malaysia Chapter, from 2014 to 2018.