Characterizing Scattered Occlusions for Effective Dense-Mode Crowd Counting

Khalid J Almalki[1,2] Baek-Young Choi[1] Yu Chen[3] Sejun Song[1]
[1]School of Computing and Engineering, University of Missouri-Kansas City, MO, USA
[2]College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia
[3]Dept. of Electrical and Computer Engineering, Binghamton University, Binghamton, NY, USA
[1]{kjaf3f, choiby, sjsong}@umsystem.edu,[2]k.almalki@seu.edu.sa, [3]ychen@binghamton.edu

Abstract

We propose a novel deep learning approach for effective dense crowd counting by characterizing scattered occlusions, named CSONet. CSONet recognizes the implications of event-induced, scene-embedded, and multitudinous obstacles such as umbrellas and picket signs to achieve an accurate crowd analysis result. CSONet is the first deep learning model for characterizing scattered occlusions of effective dense-mode crowd counting to the best of our knowledge. We have collected and annotated two new scattered occlusion object datasets, which contain crowd images occluded with umbrellas (cso-umbrellas dataset) and picket signs (cso-pickets dataset). We have designed and implemented a new crowd overfit reduction network by adding both spatial pyramid pooling and dilated convolution layers over modified VGG16 for capturing high-level features of extended receptive fields. CSONet was trained on the two new scattered occlusion datasets and the ShanghaiTech A and B datasets. We also have built an algorithm that merges scattered object maps and density heatmaps of visible humans to generate a more accurate crowd density heatmap output. Through extensive evaluations, we demonstrate that the accuracy of CSONet with scattered occlusion images outperforms over the state-of-art existing crowd counting approaches by 30% to 100% in both mean absolute error and mean square error.

1. Introduction

Crowd counting is becoming an increasingly important issue of computer vision, as it has many applications in the context of smart cities especially pertaining to public safety. The lack of proper crowd safety control and management often leads to human casualties and infectious disease (i.e., COVID-19) spreading at densely crowded political, entertaining, and religious events. Hence, automated crowd interpretation using AI techniques [5, 20, 42] is becoming an increasingly critical task for many practical crowd safety applications [13, 25, 30, 34]. Although many CNN-based methods have been proposed to improve the performance on complex crowd images to deal with variations in scale, perspective, and image resolution [21, 1, 22, 29, 32, 37, 38], they still have significant limitations in the face of occlusions that partially impede sight of individuals in a crowd scene. Crowd images are often scattered with occlusions that make it difficult to identify all human heads in the scene. As illustrated in Figure 1, the types of fixed environmental obstacles such as buildings, big trees, and walls are constrained to specific parts of a image, thus can be easily excluded from the crowd counting area. However, the interpretations of event-induced, scene-embedded, and multitudinous obstacles, namely *Scattered Occlusions (SO)*, such as umbrellas and picket signs are challenging, as they can obscure the sight of one or more individuals entirely or partially depending on crowd size and density as well as occlusion types [33]. Despite its commonness in many mass gathering scenes such as sport events, political rallies or protests, existing approaches fail to do accurate human counting in the presence of SO in crowd images.

In this paper, we propose a novel deep learning approach for effective dense-mode crowd counting by characterizing scattered occlusions (CSONet). CSONet effectively recognizes event-induced, scattered, and multitudinous occlusions and applies the effect to improve crowd counting accuracy and crowd density mapping quality. Specifically, CSONet tackles the dense-mode crowd scenarios such as people under umbrellas and behind pickets, which can hide people according to the event and recurring patterns in various ways. CSONet is an efficiently trained model using a simple convolutional structure comprised of three components. First, the Scattered Occlusion Datasets (SOD) component generates two new crowd counting datasets that contain diffused umbrella (cso-umbrellas dataset) and picket (cso-pickets dataset) occlusion objects in the crowd images. SOD also trains the model and outputs umbrella and picket heatmaps. Second, a network for Crowd Overfit Reduction (COR) is added on the well-trained VGG16-based CSRNet

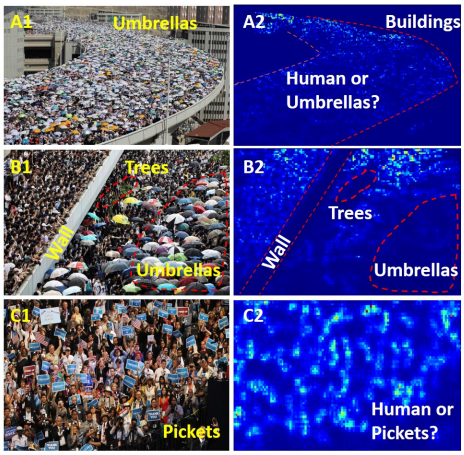

Figure 1: Crowd Map with Occlusion Objects (Overpass, Buildings, Walls, Fences, Trees, Umbrellas, and Pickets).

architecture [19] to reduce the Mean Absolute Error (MAE) and Mean Squared Error (MSE). We use the first ten layers of VGG16 to extract features from the crowd images. The extracted VGG16 features are grouped in a Spatial Pyramid Pooling (SPP) layer using average pooling in two different receptive fields (6×6 and 12×12) to soften overfitting. The Dilated Convolution Layers (DCL) outputs a predicted count and density map, which improves the crowd density prediction. Finally, a Scattered Occlusion Mapper (SOM) is implemented to combine the SO object heatmap with the human crowd heatmap to generate an accurate crowd density map and the crowd count. Using multiple datasets (cso-umbrellas dataset, cso-pickets dataset, and ShanghaiTech datasets A (dense-mode) and B (sparse-mode)), we demonstrate that CSONet's accuracy outperforms existing techniques such as SPN [7], ASNet [15] and CSRNet [19]. Our main objective is to achieve higher accuracy with the SO. CSONet reaches 100% better MAE and MSE for cso-umbrellas(MAE-U and MSE-U) and 30% better MAE and MSE for cso-pickets (MAE-P and MSE-P) than CSRNet. CSONet also achieves 64% better MAE and 80% better MSE than SPN for umbrella dataset and 46% better MAE and MSE than ASNet for picket dataset. To the best of our knowledge, this is the first work that adaptively estimates the number of people occluded by objects scattered throughout a crowd scene to accurately quantify the total counts of people in a crowd image. The main contributions of this work include:

- We have designed and developed a CSONet architecture, which is the first deep learning model for characterizing scattered occlusions of effective dense-mode crowd counting to the best of our knowledge.
- We have investigated the impact and challenges of SO in CNN crowd counting methods by collecting and annotating two new SO datasets, containing crowd images occluded with umbrellas (cso-umbrellas dataset) and picket signs (cso-pickets dataset).
- We have implemented COR by adding SPPL and DCL over modified VGG16 layers, which deploys a deeper CNN for capturing high-level features of extended receptive fields. COR was trained on the two new SO object datasets and the ShanghaiTech A and B datasets.
- We have built an algorithm that merges scattered object heatmaps and visible human heatmaps to generate a more accurate crowd density output.

The rest of the paper is organized as follows. Section 2 introduces related work for crowd counting in both traditional and CNN based approaches. Section 3 presents the proposed CSONet architecture, including the training details. In Section 4, we demonstrate the experiments by evaluation metrics and discuss the results. Finally, Section 5 concludes the paper.

2. Related Works

There have been significant studies and remarkable improvements made in crowd counting and density estimation. Traditional non-machine learning methods can be broadly classified into three categories, namely, detection-based, regression-based, and density estimation-based approaches [24]. Despite various advancements, those approaches have shortcomings of complexities and limited accuracy. In recent years, researchers mostly have adopted machine learning techniques to overcome those weaknesses. In this section, we briefly highlight noticeable prior studies.

2.1. Traditional Methods

A number of early methods have attempted to tackle the challenges of crowd counting and density estimation via implementing detection-based approaches. Generally, these methods use a detector or classifier to recognize a human's whole or body part to estimate the crowd count. Dollar et al. [9] applied a sliding window detector to extract the features from the input image and determine the human count. Most of the methods focused on extracting features, such as histograms of oriented gradients HOG [8], and Haar wavelets [35] from the crowd images to learn the density and the count. However, the counting results of the whole body methods perform poorly in highly crowded images. Although a part-based detector is proposed to detect the density of people in a crowd [10, 18], these methods still face difficulties in locating people, especially when the crowd in a scene is highly occluded or densely populated, as it happens often in various events. Regression-based approaches have been proposed to tackle the limitations of the detection-based method, concentrating on the difficulty of

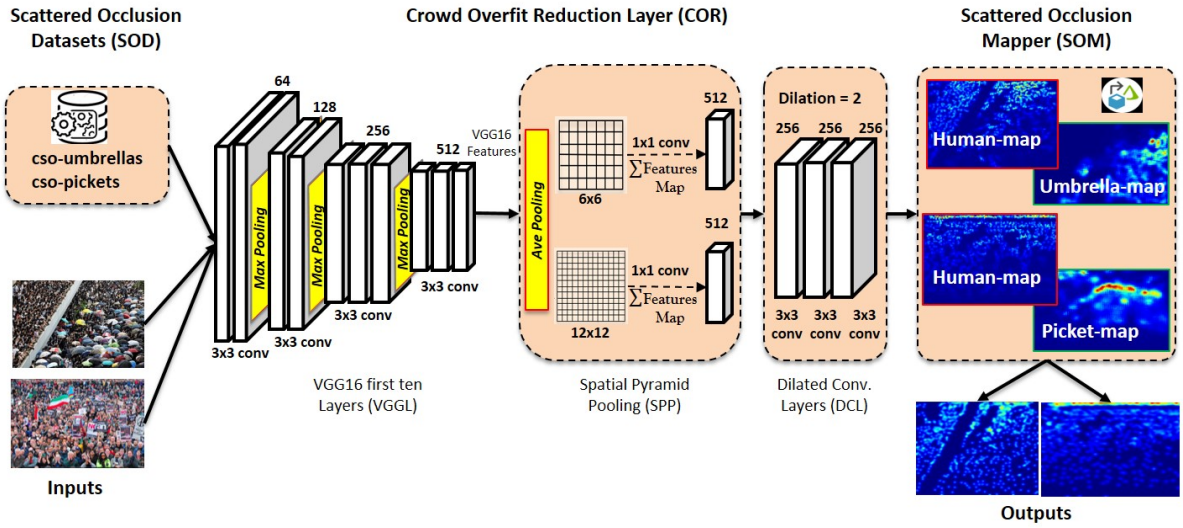

Figure 2: CSONet Architecture.

detecting the count in a highly dense crowd scene. Regression methods aim to learn the mapping between extracted features from the image and the count number of objects [6, 4]. Regression methods typically have two main components: low-level feature extraction and regression modeling [28]. Density estimation-based methods are another approach for crowd counting and density estimation. Researchers have successfully addressed the issues of occlusion and clutter by using regression-based methods. Nevertheless, some of the existing techniques overlooked spatial information, which affected the result of counting. In contrast, Lempitsky et al. [17] proposed a supervised learning framework to estimate the count of objects in images. They used a linear mapping technique that focuses on the density through learning the relationship between the local image features and object density maps. However, Pham et al. [27] observed the limitation of the linear mapping and proposed a random forest framework to learn a non-linear mapping between local image features and density maps.

2.2. CNN-based Approaches

Several studies have proposed Convolutional Neural Networks (CNN) based approaches for crowd counting and density estimation. Those methods have obtained a significant improvement in crowd counting and density estimation addressing various kinds of challenges, such as perspective, image resolutions, occlusions, and non-uniform environment. Here, we briefly summarize various of the recent methods for crowd counting and density estimation in terms of their CNN architectures. A multi-column CNN architecture called MCNN was proposed in [40] to estimate the crowd count and density map in an arbitrary crowd image. The CrowdNet method proposed in [2] is considered one of the early CNN based architectures inspired by VGG16 [31]. The CrowdNet combined convolutional networks and a shallow network to learn robust scale features and to generate the density maps. Cao et al. [3] present an encoder-decoder network called a scale aggregation network (SANet). The encoder layer extracts the multi-scale features, and the decoder will generate high-resolution density maps. Also, training loss is introduced by combining euclidean loss and local pattern consistency loss, which contributed to improving the final count. Li et al. [19] introduce the congested scene recognition network (CSRNet), which is one of the state-of-the-art in terms of performance among the ones inspired by VGG16. It consists of two essential components: CNN as the front-end layers and dilated convolution layers as the back-end. Zhang et al. [39] proposed a method that generates a probability map and presents the high expectations indicated in locations where heads are possible to be present. CANet [23] proposed a deep network architecture that performs multi-level feature comparison between the support and the query images and iterative refinements of the results. Chen et al. [7] proposed a scale pyramid network (SPN), which consists of a single column structure to extract multiple-scale features by dilated convolutions with various rates. ASNet [15] is also considered as a state-of-the-art method. It contains a density attention network that generates attention masks, and then provides it to attention scaling network in order to generate scaling factors outputting attention-based density.

Despite improvements achieved by such recent approaches, the accuracy of crowd counting can significantly diminished in the presence of SOs in a crowd scene. Our method address the very issue of SOs and achieves a high accuracy of crowd counting in the presence of such SOs.

3. Proposed Architecture

The proposed design aims to characterize scattered occlusions to improve the accuracy of crowd counting as well as the quality of crowd density mapping. In this section, we introduce the CSONet architecture that consists of a network for Scattered Occlusion Datasets (SOD), Crowd Overfit Reduction (COR), and Scattered Occlusion Mapper (SOM), as depicted in Figure 2. SOD creates two new scattered occlusion object datasets and trains on them. COR deploys a deeper CNN for capturing high-level features with larger receptive fields. SOM generates high-quality crowd density maps.

3.1. Scattered Occlusion Datasets (SOD)

In the Scattered Occlusion Datasets (SOD) component, we build two new datasets and perform CSONet training with these new datasets and two well-known public datasets.

3.1.1 Datasets and Experimental Settings

Our goal is to investigate the impact and challenges of Scattered Occlusion (SO) objects in the CNN crowd counting methods. However, there has been no crowd image dataset available focusing on SO objects such as umbrellas and pickets. Hence, we have created new SO object datasets and trained our network CSONet on them. The generated dataset consists of the cso-umbrellas dataset and the cso-pickets dataset. They were collected from two resources. First, both umbrella and picket crowd images were mainly downloaded from Google images by running web search scripts with various keywords, including "umbrellas" ("crowd with umbrellas" and "crowd in the rain") and "pickets" ("demonstration" and "protest"). Second, the cso-umbrellas dataset images are partially converted from the Hajj event videos, an annual Islamic pilgrimage to Mecca, Saudi Arabia, during the summer, where the crowd holds umbrellas.

The **cso-umbrellas dataset** contains 250 crowd images and a total of 27,697 umbrella annotations. Among them, 170 images were used for training, and 80 images were used for testing. The **cso-pickets dataset** consists of 200 images and 9,681 picket annotations. 130 images were used for training, and 70 images were used for testing. To conduct comparisons with the existing state-of-the-art crowd counting method, we also train and test on the ShanghaiTech A and B datasets. The **ShanghaiTech dataset** is a large-scale crowd counting dataset containing 1198 images with 330,165 head annotations. It consists of two parts: ShanghaiTech A and ShanghaiTech B. Part A includes 482 dense-crowd images that have been collected randomly from the Internet. 300 images were used for training, and the remaining 182 images were used for testing. Part B has 716 sparse-crowd images, which were taken on busy streets in Shanghai. 400 images were used for training, and 316 were used for testing. Table 1 demonstrates a summary of the statistics of the datasets.

Datasets	Images	Annotations	Avg. Count	Max. Count	Avg. Resolution
Shanghai A	482	241,677	501	3,139	589 x 868
Shanghai B	716	88,488	123	578	768 x 1024
cso-umbrellas	250	27,697	111	862	561 x 783
cso-pickets	200	9,681	48	386	728 x 969

Table 1: Summary of statistics of the datasets.

3.1.2 Ground-Truth Generation

We have annotated all the images to generate the density map "ground-truth". We have applied the geometry-adaptive Gaussian kernels [40] as defined below to generate the density map for each crowd image. The labeled objects' locations in the original image are converted to the ground-truth density map $F(x)$ as follows:

$$F(x) = \sum_{t=1}^{N} \delta(x - x_i) * G_{\sigma i}(x), \text{ with } \sigma_i = \beta \overline{\mathrm{d_i}} \qquad (1)$$

where N is the number of object annotations in the image, x_i is referring to each object in a given image, and $\overline{d}_i$ indicates the average distance of k-nearest neighbors. Also, the delta function $\delta(x - x_i)$ is convolved with a Gaussian kernel with the standard deviation parameter σ_i to generate density heatmaps.

3.1.3 Training details

We have trained the CSONet structure in an end-to-end manner. Adam optimizer [16] is used as an optimization method to train CSONet with a learning rate of 1e-5 and a momentum of 0.9. Performing multiple experiments starting from 1e-4 to 1e-9, we found that 1e-5 is the ideal learning rate. In addition, we used other recommended training hyper-parameters, including a batch of size 32 and an epoch number of 100.

3.2. Crowd Overfit Reduction (COR) Layer

In this subsection, we explain the network structure of the Crowd Overfit Reduction (COR) layer that consists of three components, including VGG16 Layers (VGGL) [31], Spatial Pyramid Pooling Layers (SPPL) [14], and Dilated Convolution Layers (DCL). We start with the VGG16 network, which was initially designed for large-scale natural image classification. VGG16 has thirteen convolutional layers and three fully connected layers. However,

we have modified the VGG16 network, which learns ten convolutional layers with max-pooling, does two SPPs with average-pooling and applies three dense models in DCL.

3.2.1 Modified VGG16 Network Layer

We apply the first ten convolutional layers and three max-pooling layers of VGG16 to extract the crowd features. The VGG16 is employed to ensure excellent learning performance in object classification and detection, which has been used by various practices such as CSRNet [19] and DADNet [12]. The input images commence with a fixed size by 224×224 pixel RGB image at the first convolutional layer. As illustrated in Figure 2, the images sequentially pass through a stack of 3×3 kernel convolutional layers with different filter depths (64, 128, 256, and 512, respectively) and three max-pooling layers of 2×2 pixel windows in-between to create VGG16 features.

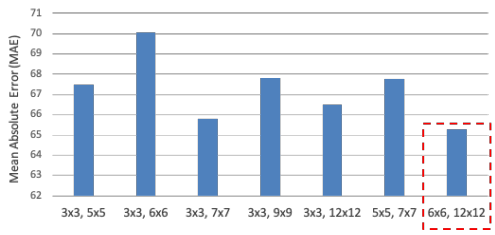

Figure 3: Various receptive fields performance.

3.2.2 Spatial Pyramid Pooling Layers (SPPL)

We have implemented Spatial Pyramid Pooling Layers (SPPL) to process the features extracted from the first ten VGG16 layers, which improves the semantic segmentation results [41] in the density map. We apply average pooling instead of max-pooling to reduce the overfitting and level the prediction results. In particular, the average pooling layer assigns the extracted VGG16 features into two different receptive fields (6×6 and 12×12), followed by a 1×1 convolutional layer. As presented in Figure 3, among the various sets of receptive field experimental results, the receptive fields of (6×6 and 12×12) achieve the lowest Mean Absolute Error (MAE) value.

3.2.3 Dilated Convolution Layers (DCL)

The Dilated Convolution Layers (DCL) is the last function of the COR structure. Generally, DCL is widely used in various computer vision processes to promote crowd density predictions and improve semantic segmentation results. Moreover, DCL maintains the exponential expansion of the receptive field without reducing the resolution [36]. We implement DCL with three convolution layers using the same depth filter of 256 and a kernel size of 3×3. We set the dilation rate to two to gain better performance. However, we transfer the feature maps to these smooth layers to produce the CSONet outputs, predicted crowd count, and density heatmap. The motivation for utilizing the DCL in COR was to promote the dense prediction in congested images.

3.3. Scattered Occlusion Mapper (SOM)

Scattered Occlusion Mapper (SOM) is the last component of CSONet architecture. It generates a high-quality crowd density heatmap and an accurate crowd count by merging Scattered Occlusion (SO) object data with human crowd data. Most of the existing crowd counting methods require a visible head to detect and count the number of people, which cannot delimit individuals under umbrellas or behind pickets. A simple one-to-one mapping won't work as the SO object's impact on visual saliency for an image depends on the size, density, mobility type, flow direction, and velocity. As illustrated in Figure 4, an umbrella's effect is different in the sparse and dense crowd scenarios. An earlier study proposed an illustration for crowd counting per unit [33]. We propose a procedure for estimating the number of people under umbrellas or behind pickets in a particular crowd event. Our analysis shows that each umbrella covers zero to three people, and each picket occludes zero to two people corresponding to the Occlusion Object to Human Ratio (OHR). Also, we assume that the number of SO objects cannot be more than the original human count. However, those effects converge into similar values in the high-density images. As shown in Figure 5, the average number of people under an SO object mainly depends on the OHR. Therefore, a formula is proposed to count the total number of people in an SO image:

$$T_{human} = D_{human} + (D_{so} * \alpha) \tag{2}$$

where T_{human} is the total predicted crowd count in an image, D_{human} is the detected human count. D_{so} indicates the number of predicted objects in an image (umbrellas D_u or pickets D_p). An α can be measured by using the ground truth values named MSOI (Measured SO Impact). Also, it is estimated as an SO Impact (SOI) value. According to Figure 5, an α value is chosen from the SOI value according to the OHR. For example, if OHR is 40 % (i.e., human count : SO object count = 100 : 40), α is 2.

4. Experiments

We test the proposed CSONet using multiple different datasets, including two new SO object datasets (cso-umbrellas and cso-pickets), along with two public crowd datasets, ShanghaiTech A [40] and ShanghaiTech B [40]. In this section, the evaluation metrics are introduced and then SO evaluations are conducted to analyze the efficacy

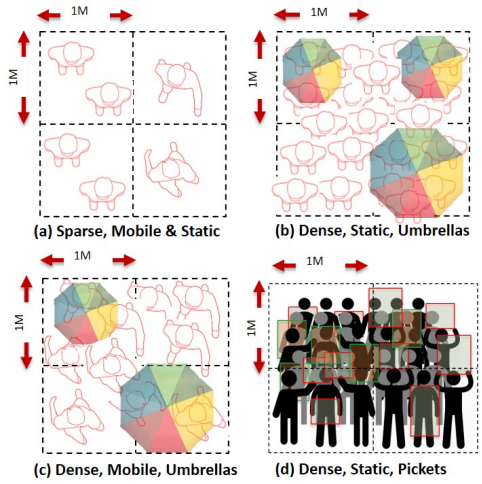

Figure 4: Crowd image annotations with different mode, type, and object.

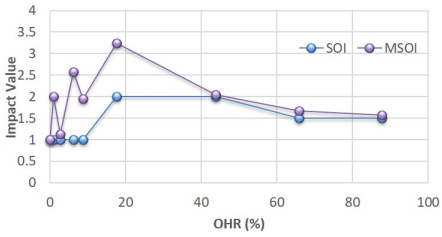

Figure 5: Scattered Occlusion (SO) impact values.

of the proposed model. We evaluate and compare the performance of CSONet to various crowd counting methods including SPN, ASNet, and CSRNet. The CSONet prototype was implemented using the Pytorch framework [26]. All of the experiments were conducted on an NVIDIA GeForce GTX 1080 Ti.

4.1. Evaluation Metrics

Both Mean Absolute Error (MAE) and Mean Squared Error (MSE) are adopted in our performance testing. These metrics are broadly used in crowd counting to evaluate the accuracy of the measurement performance.

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|Y_i - Y_i^G| \tag{3}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(Y_i - Y_i^G)^2} \tag{4}$$

where N is the number of test images, Y_i is the predicted number, and the Y^G is the ground-truth counts of the test image i.

We also compute the Structural Similarity Index (SSIM) [43], which is a metric used to measure the similarity between two images. The SSIM value ranges from 0 to 1, equaling 1 if the two images are identical.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

Following the preprocessing method given by [43], Eq. (5) measures the similarity between two images. Where x is the estimated density-map, and y is the ground-truth. C_1 and C_2 are small constants, used to avoid division by zero. In addition, μ_x and σ_x^2 are the local mean and variance estimations of x, and σ_{xy} is the local covariance estimation. μ_y and σ_y^2 are computed similarly.

For SO evaluation, we use ERT (error against real ground truth) in Eq. (6).

$$ERT = |V_i - OH_i^G| \tag{6}$$

where V_i refers to the detected human count, the OH_i^G presents the real ground-truth (RGT) of the test image i.

4.2. Scattered Occlusion Evaluations

The experiment is designed to evaluate SO object detection's performance and the accuracy of the estimated number of people occluded by the SO. We investigate how significantly umbrellas and pickets impact the accuracy of crowd counting and density estimation. For this purpose, we use an original crowd image with 114 people and continuously increase SO object annotations over the image from 0 to 87.7 % (Occlusion Object vs. Human Ratio (OHR)), as shown in Table 2 and Figure 6. CSRNet [19] is selected as a baseline to compare and evaluate the prediction accuracy of the proposed work. Table 2 shows the statistics of the experimental results with various scenarios and methods. The Real Ground Truth (RGT) value is the original number of people in the crowd image (i.e., 114 people) and the known number of SO objects (umbrellas (U) or Pickets (P)) placed on the image (i.e., from 0 to 100 umbrellas or pickets). We also use a Detected Ground Truth (DGT) of the number of SO objects (U/P) and the number of remaining visible humans heads (H), based on manual Matlab-based annotation analysis. DGT-U is the DGT after umbrella annotation, and DGT-P is the DGT after picket annotation. As the number of SO object annotation increases, the number of visible heads decreases due to occlusion. Also, the SO object count accuracy reduces due to many overlaps (i.e., only 72 umbrellas are detected after applying 100 umbrellas). We run both CSRNet and CSONet to find the number of humans and SO objects in a crowd image. As presented in Figure 6, after applying 75 SO object annotations (65.8% of OHR), there are almost no visible human heads. However, as CSONet applies the SO object impacts (SOI) for

its final crowd counting according to Eq. (2), its prediction results are as good as RGT.

OHR(%)	RGT	DGT-U		DGT-P		CSRNet		CSONet	
U/P	U/P	H	U	H	P	U	P	U	P
0	0	114	0	114	0	119	119	113	113
0.9	1	112	1	111	2	107	107	113	115
2.6	3	108	4	106	3	105	105	115	112
6.1	7	97	7	100	7	85	85	102	106
8.8	10	94	11	94	9	88	88	102	105
17.5	20	57	17	68	19	68	86	95	97
43.9	50	15	49	30	46	18	55	110	106
65.8	75	6	64	9	70	14	13	105	110
87.7	100	0	72	0	73	7	4	110	110

Table 2: Experimental results with SO objects.

Figure 6: Crowd images with SO object annotations.

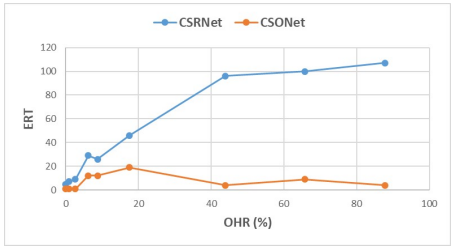

Figure 7: Error against RGT umbrella annotations.

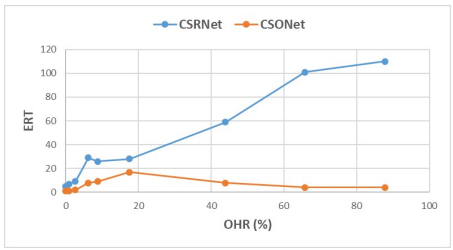

Figure 8: Error against RGT picket annotations.

Figures 7 and 8 compare the crowd counting performance of CSRNet and CSONet in the aspect of ERT in Eq. (6) for umbrella and picket annotations, respectively. The ERT of CSONet is much lower than the ERT of CSRNet. The ERTs of CSRNet significantly increase when OHR increases. However, the ERT of CSONet does not increase for all OHR. Therefore, CSONet's crowd counting performance is much more stable and accurate than CSRNet, indicating that merging the human and SO density heatmaps is critical for better crowd count accuracy.

4.3. Performance Comparison

The performance in MAE and MSE metrics with ShanghaiTech datasets (i.e., MAE-A means MAE with ShanghaiTech A) of the existing state-of-the-art crowd counting solutions, including CP-CNN [32], CSRNet [19], PCC Net [11], SPN [7], and ASNet [15] are compared in Table 3. It shows that the most recent ASNet achieves the least MAE-A and MSE-A. SPN is as good as ASNet, which is 27% better than other earlier approaches such as CSRNet. ASNet did not evaluate ShanghaiTech B dataset (sparse-mode), as they are interested in counting densely populated crowd with ShanghaiTech A (dense-mode) dataset.

Method	MAE-A	MSE-A	MAE-B	MSE-B
CP-CNN [32]	**73.6**	**106.4**	20.1	30.1
CSRNet [19]	**68.2**	**115.0**	10.6	16.0
PCC Net [11]	**73.5**	**124.0**	11.0	19.0
SPN [7]	**61.7**	**99.5**	9.4	14.4
ASNet [15]	**57.78**	**90.13**	-	-

Table 3: Performance comparisons of different methods on ShanghaiTech A (dense-mode) and B (sparse-mode).

Table 4 presents the crowd counting accuracy results with the new SO object datasets. We choose 80 umbrella and 70 picket images from cso-umbrellas and cso-pickets datasets, respectively, and tested them with SPN, ASNet, CSRNet, and CSONet to obtain the MAE, MSE, and SSIM values (i.e., MAE-U means MAE for the umbrella images). According to the density of cso-umbrellas and cso-pickets datasets in Table 1, the MAE and MSE with the ShanghaiTech dataset in Table 3 align with the results in Table 4. For example, MAE-P and MSE-P maintain lower values due to the cso-picket images are sparse. Also, MAE-U and MSE-U of CSRNet are slightly higher than SPN and ASNet. CSONet's performance in terms of accuracy is significantly better than the other methods. For example, CSONet achieves 100% better MAE and MSE for cso-umbrellas(MAE-U and MSE-U) and 30% better MAE and MSE for cso-pickets (MAE-P and MSE-P) than CSRNet. CSONet also achieves 64% better MAE and 80% better MSE than SPN for umbrella dataset and 46% better MAE and MSE than ASNet for picket dataset. The SSIM measures the similarity between the ground-truth and the estimated density-map images. Although SOs already impact

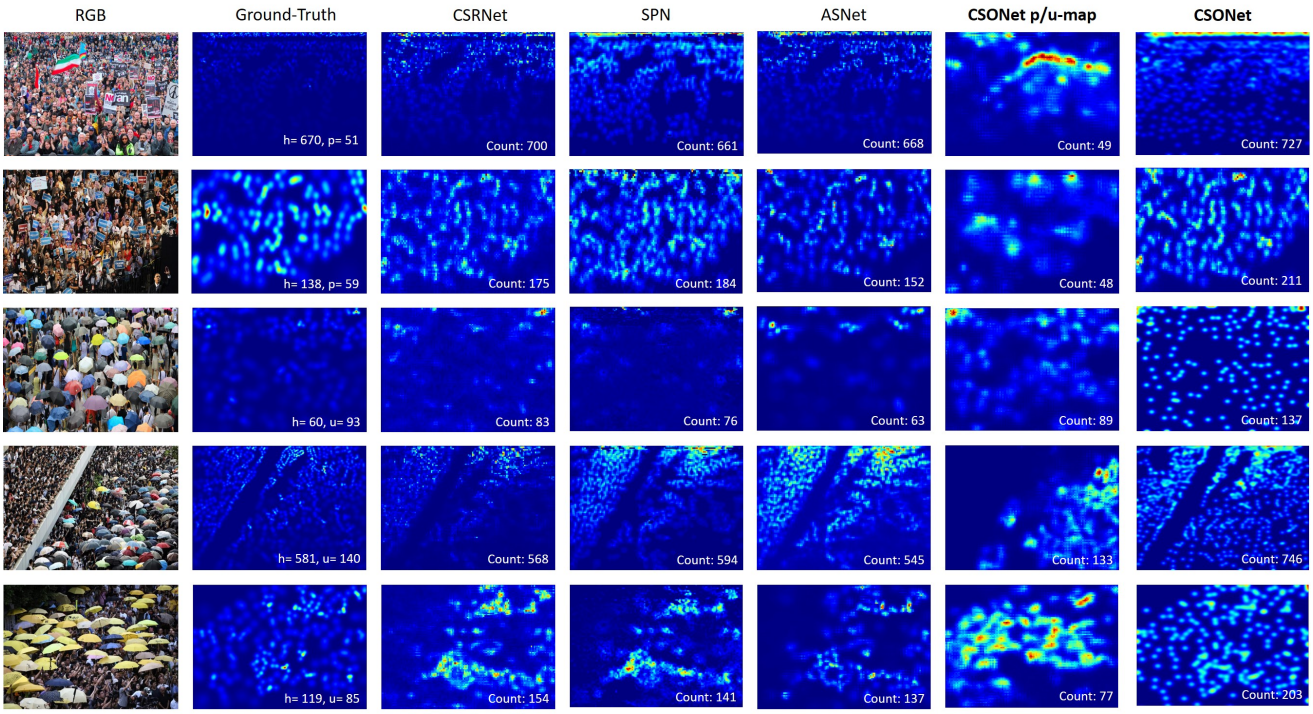

Figure 9: Five crowd image samples are randomly selected from the SO object datasets and evaluated with CSRNet, SPN, ASNet, and CSONet. We display the density maps and counts of each sample. For a given RGB image, from the left column, the detected ground-truth (DGT) density map shows the heads (H) and umbrellas (U)/ pickets(P) counts. The CSRNet, SPN, and ASNet prediction results (count and density map) are in columns 3, 4, and 5, respectively. The last two columns present CSONet results. The h-map is the human map and count, and p/u-maps are the detected SO objects (pickets/umbrellas). Finally, the CSONet map and count demonstrate the estimation of human count and density map, which applies the SOI (i.e., under umbrellas or behind pickets) in a particular crowd event.

Method	MAE-U	MAE-P	MSE-U	MSE-P	SSIM-U	SSIM-P	SSIM-A
CSRNet [19]	73.6	19.1	135.9	35.5	0.83	0.92	0.76
SPN [7]	59.9	24.2	120.1	42.3	0.85	0.90	-
ASNet [15]	71.5	21.3	133.7	40.5	0.81	0.88	-
CSONet	**36.5**	**14.6**	**66.5**	**28.4**	**0.87**	**0.94**	**0.91**

Table 4: Performance comparisons of CSRNet, SPN, ASNet, and CSONet with cso-umbrellas and cso-pickets datasets.

the DGT images, the CSONet still creates a higher structural similarity than the other methods. Figure 9 presents overall performance results of crowd density heatmaps and crowd counts (human, umbrella, and picket) with five SO image samples. According to the heatmaps of DGT, CSRNet, SPN, and ASNet, the area covered by SOs are shown by low density. However, the CSONet adjusts those areas by identifying p/u heatmaps and overlaying them to human heatmaps, which results in more accurate crowd counting.

5. Conclusions

We proposed an architecture for scattered occlusion characterization called CSONet for efficient crowd counting and high-quality density heatmap generation. We first generated, annotated, and trained two new scatter occlusion object datasets, the cso-umbrellas dataset, and the cso-pickets dataset. We then implemented CSONet using spatial pyramid pooling and dilated convolutional layers to expand the receptive field without losing resolution in the congested scenes. CSONet recognizes event-induced, scattered, and multitudinous occlusions and applies the effect to a human crowd map to generate an accurate crowd count and high-quality density-map. Through extensive evaluations, we demonstrated that the accuracy of CSONet outperforms over the state-of-art existing crowd counting approaches.

References

[1] Deepak Babu Sam, Neeraj N Sajjan, R Venkatesh Babu, and Mukundhan Srinivasan. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3618–3626, Salt Lake City, UT, USA, 2018. IEEE. 1

[2] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 640–644, Amsterdam, The Netherlands, 2016. Association for Computing Machinery. 3

[3] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 757–773, Cham, 2018. Springer. 3

[4] Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *2009 IEEE 12th international conference on computer vision*, pages 545–551, Kyoto, Japan, 2009. IEEE, IEEE. 3

[5] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, Portland, OR, USA, 2013. IEEE. 1

[6] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *BMVC*, volume 1, page 3, 2012. 3

[7] Xinya Chen, Yanrui Bin, Nong Sang, and Changxin Gao. Scale pyramid network for crowd counting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1941–1950. IEEE, 2019. 2, 3, 7, 8

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893, San Diego, CA, USA, USA, 2005. IEEE, IEEE. 2

[9] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011. 2

[10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009. 2

[11] Junyu Gao, Qi Wang, and Xuelong Li. Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 7

[12] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1823–1832, Nice, France, 2019. Association for Computing Machinery. 5

[13] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, Providence, RI, USA, 2012. IEEE, IEEE. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 4

[15] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4706–4715, 2020. 2, 3, 7, 8

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *in Proceedings of the IEEE International Conference on Learning Representations (ICLR)*, May 2015. 4

[17] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, pages 1324–1332, Red Hook, NY, 2010. 3

[18] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Tampa, FL, USA, 2008. IEEE, IEEE. 2

[19] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, Salt Lake City, UT, USA, 2018. IEEE. 2, 3, 5, 6, 7, 8

[20] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, UT, USA, 2018. IEEE. 1

[21] Lei Liu, Wenjing Jia, Jie Jiang, Saeed Amirgholipour, Yi Wang, Michelle Zeibots, and Xiangjian He. Denet: A universal network for counting crowd with varying densities and scales. *IEEE Transactions on Multimedia*, 2020. 1

[22] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1774–1783, Seoul, Korea (South), Korea (South), 2019. IEEE. 1

[23] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, Long Beach, CA, USA, 2019. IEEE. 3

[24] Chen Change Loy, Ke Chen, Shaogang Gong, and Tao Xiang. Crowd counting and profiling: Methodology and evaluation. In *Modeling, simulation and visual analysis of crowds*, pages 347–382. Springer, 2013. 2

[25] Yu-Jen Ma, Hong-Han Shuai, and Wen-Huang Cheng. Spatiotemporal dilated convolution with uncertain matching for

video-based crowd estimation. *IEEE Transactions on Multimedia*, 2021. 1

[26] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, 6, 2017. 6

[27] Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo Okada. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3253–3261, Santiago, Chile, 2015. IEEE. 3

[28] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Crowd counting using multiple local features. In *2009 Digital Image Computing: Techniques and Applications*, pages 81–88, Melbourne, VIC, Australia, 2009. IEEE, IEEE. 3

[29] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4031–4039, Honolulu, HI, USA, 2017. IEEE, IEEE. 1

[30] Jing Shao, Chen-Change Loy, Kai Kang, and Xiaogang Wang. Slicing convolutional neural network for crowd video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5620–5628, Las Vegas, NV, USA, 2016. IEEE. 1

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *in Proceedings of the IEEE International Conference on Learning Representations (ICLR)*, May 2015. 3, 4

[32] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1879–1888, Venice, Italy, 2017. IEEE. 1, 7

[33] G Keith Still. *Introduction to crowd science*. CRC Press, 2014. 1, 5

[34] HY Swathi, G Shivakumar, and HS Mohana. Crowd behavior analysis: a survey. In *2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT)*, pages 169–178, Bangalore, India, 2017. IEEE, IEEE. 1

[35] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 2

[36] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *in Proceedings of the IEEE International Conference on Learning Representations (ICLR)*, 2016. 5

[37] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841, Boston, MA, USA, 2015. IEEE. 1

[38] Lu Zhang, Miaojing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1113–1121, Lake Tahoe, NV, USA, 2018. IEEE, IEEE. 1

[39] Youmei Zhang, Chunluan Zhou, Faliang Chang, Alex C Kot, and Wei Zhang. Attention to head locations for crowd counting. In *International Conference on Image and Graphics*, pages 727–737, Cham, 2019. Springer, Springer. 3

[40] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, Las Vegas, NV, USA, 2016. IEEE. 3, 4, 5

[41] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, Honolulu, HI, USA, 2017. IEEE. 5

[42] Tao Zhao, Ram Nevatia, and Bo Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE transactions on pattern analysis and machine intelligence*, 30(7):1198–1211, 2008. 1

[43] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6