Residual Regression with Semantic Prior for Crowd Counting

Jia Wan[1,2*] Wenhan Luo[2] Baoyuan Wu[2] Antoni B. Chan[1] Wei Liu[2]
[1]Department of Computer Science, City University of Hong Kong [2]Tencent AI Lab
{jiawan1998,whluo.china,wubaoyuan1987}@gmail.com abchan@cityu.edu.hk wl2223@columbia.edu

Abstract

Crowd counting is a challenging task due to factors such as large variations in crowdedness and severe occlusions. Although recent deep learning based counting algorithms have achieved a great progress, the correlation knowledge among samples and the semantic prior have not yet been fully exploited. In this paper, a residual regression framework is proposed for crowd counting harnessing the correlation information among samples. By incorporating such information into our network, we discover that more intrinsic characteristics can be learned by the network which thus generalizes better to unseen scenarios. Besides, we show how to effectively leverage the semantic prior to improve the performance of crowd counting. We also observe that the adversarial loss can be used to improve the quality of predicted density maps, thereby leading to an improvement in crowd counting. Experiments on public datasets demonstrate the effectiveness and generalization ability of the proposed method.

1. Introduction

Crowd counting plays a very important role in intelligent monitoring systems aiming at automatically detecting the crowd congestion. Technically, a solution to crowd counting takes input like an image or a video clip, and outputs a predicted number indicating the crowdedness in the input. This is very challenging because of issues like large variations of density, scale, perspective, and severe occlusion.

Traditional algorithms count crowd numbers by detection of people, which is extremely challenging under highly congested scenes due to severe occlusions (see Fig. 1). To avoid difficulties in explicitly detecting people, regression based approaches are proposed to directly estimate the crowd number by density related features. However, the performances of these algorithms are limited due to variations of density and scale. Recently, crowd counting algorithms have achieved great advances, especially those based on deep learning by predicting a density map and aggregating to a final count. Some of them attempt to deal with scale variations via different network structures with different receptive field sizes. Some of the methods utilize contextual information to improve the performance.

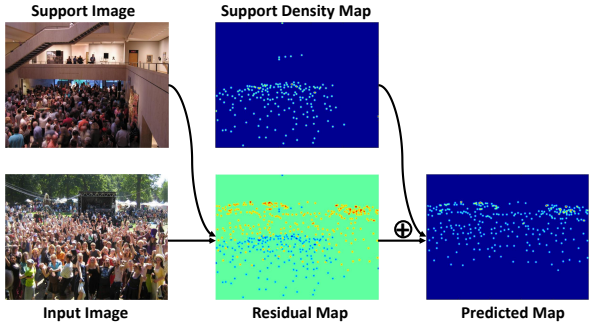

Figure 1. Residual regression attempts to predict the residual map (the difference of density maps) between the input image and the support image. By predicting various residual maps based on diverse support images, the generalization capability is boosted.

Unfortunately, most of the algorithms concentrate solely on the appearance of a single image but ignore the correlation information between image samples. Existing research works for other problems have shown that more intrinsic features can be learned by comparing samples to mine the correlation knowledge [29, 31]. To this end, we propose a residual regression framework, within which more effective features can be obtained by learning the difference between samples. The proposed algorithm can be considered as a data-driven learning method with prior knowledge [32, 6]. To be specific, we propose a novel algorithm to predict the density map by taking into account not only the appearance but also the residual maps (*i.e.*, the difference between density maps) between the input image and labeled images from a support set (shown in Fig. 1). As the support images are with different levels of crowdedness, comparing the concerned image with diverse support images will improve the generalization ability in the case of unseen scenarios. The final density map is estimated by fusing density maps that are predicted using both appearance and residual maps.

*This work was primarily done when Jia Wan was a research intern with Tencent AI Lab.

Furthermore, the semantic prior is effective for eliminating noisy areas, since the scenarios of crowd counting are usually of semantic structures (*e.g.*, there usually are sky, buildings, and trees). Observing this, the density of the pixels belonging to sky, wall, or trees should be close to zero. In this paper, we show how to utilize this kind of semantic prior effectively to improve crowd counting, even without fine-tuning the employed segmentation network to yield the semantic information. Specifically, we decrease the predicted density in the area without people by a factor. The weight of the area without people is also decreased while we calculate the loss. By doing so, the network is constrained to concentrate more on the area containing people, so the noisy false alarms can be eliminated to some extent.

Previous work has also shown that high-quality density maps are beneficial to the performance of crowd counting [27]. Inspired by this, we adopt an adversarial loss to improve the realism of the predicted density maps by discriminating them against real density maps. Also, we increase the resolution of the predicted density maps, and the quality-improved density maps are helpful for improving crowd counting performance .

In the experiments, the proposed approach outperforms several state-of-the-art algorithms and shows a satisfactory generalization capability when transferring to unseen scenes without fine-tuning.

Our contributions are three-fold: 1) We propose a novel approach of residual regression learning by comparing the concerned image with a set of support images to improve the generalization capability in the case of unseen scenarios. 2) We incorporate a semantic prior to eliminate the side effect of noisy false alarms in the predicted density maps. 3) An adversarial loss is adopted to enhance the quality of the predicted density maps, further improving the crowd counting performance.

2. Related Works

In general, most of the traditional algorithms are based on people detection and crowd number regression, while recent deep learning approaches typically count by estimating density map and then aggregating to final count numbers.

Most conventional detection algorithms detect either the whole body or parts for counting by detection [13] or tracking [18, 19]. [13] counts crowd by detecting human heads and shoulders. A shape learning process is proposed to detect and count individuals by [5]. Although detection has been advanced significantly with the development of deep learning, the detection of pedestrians under highly congested scenarios is still challenging. Apart from explicit detection, methods are developed attempting to count crowd by directly mapping a crowd image to a number. In these methods, hand-crafted features, like texture, gradient, foreground, and edge, are frequently used as low-level cues. Then, linear regression, random forest, or Gaussian process (GP) are utilized to predict the final crowd number. For instance, [3] develops a method counting the crowd number by holistic features and GP regression. A prior distribution is introduced in the proposed Bayesian Poisson regression to estimate the size of inhomogeneous crowds by [4]. However, [9] has shown that single features are insufficient to count crowd numbers in extremely crowd images due to large variations, clutters, and occlusions.

More recently, density map estimation of crowd images becomes more popular. [12] proposes to count local patches and then integrates them to the final count, which incorporates spatial information better for accurate counting. Motivated by this, most recent deep learning based approaches predict density maps and achieve a notable progress. To cope with density variations, a Multi-column Convolutional Neural Network (MCNN) composed of different CNNs with different kernel sizes is proposed by [35]. Detection and regression approaches are combined together to handle different kinds of scenes by [15]. Multiple CNNs with different receptive field sizes are proposed to deal with density variations, and a switch network is developed to choose the best one [23]. [34] proposes a scale-adaptive network which combines multi-scale features extracted from different layers to deal with scale and perspective changes. An Incrementally Growing CNN (IG-CNN) is developed to cope with large diversities in crowd images by [1]. [27] proposes to utilize global and local contexts to improve the performance. They discovered that high-quality density maps are useful for further decreasing the counting error. To improve the generalization ability, [25] trains a pool of decorrelated regressors. [17] learns from unlabeled data through prior knowledge for crowd counting. [33] proposes to transfer a well-trained model to a new target scene by a data-driven fine-tuning method. For more related works, the readers may refer to [28].

In contrast to previous approaches that predicted from a single image, our proposed approach mines the exemplar correlation knowledge among diverse samples, therefore generalizing better in unseen scenarios.

3. Our Approach

Typically, traditional methods predict density maps solely by the appearance of an input image, but ignore the relationship between samples. We argue that the correlation between samples is important. Therefore, we propose a residual regression algorithm which fuses appearance and correlation knowledge effectively. As shown in Fig. 2, density maps are predicted by simultaneously considering appearance and comparing the input image with a set of support images, and the final prediction is estimated by fusing these cues. During the estimation, a semantic prior is introduced to eliminate the false alarms in the area without peo-

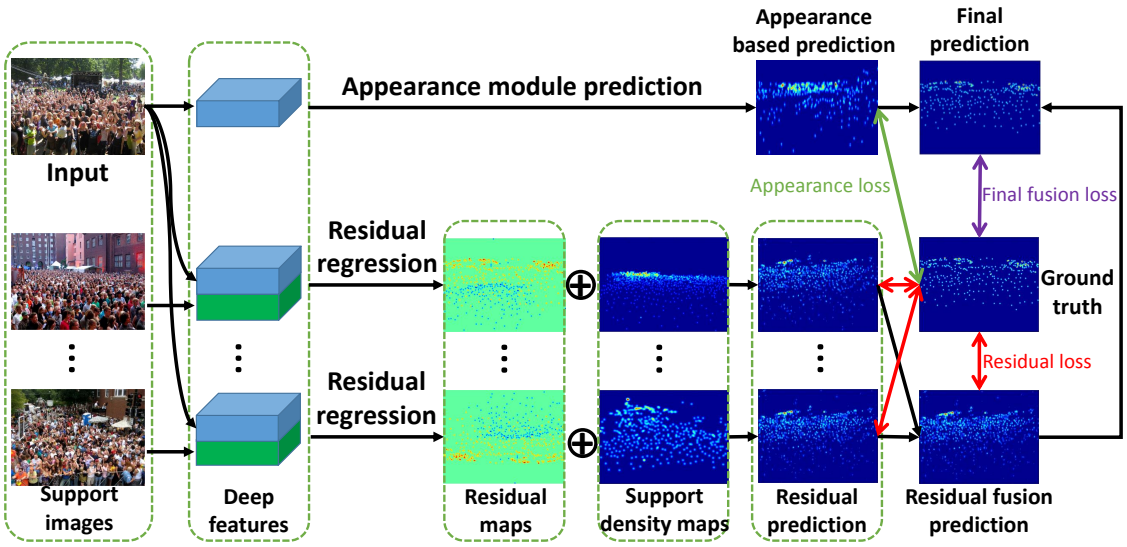

Figure 2. **The schematic of the proposed method**. Residual regression predicts a residual map (*i.e.*, the difference between density maps) between the input image and the support image. All the residual predictions are fused, and the final predicted map is calculated based on the fused residual prediction and the appearance-based prediction from the input image. Black arrows indicate data flow, while green, red, and purple two-way arrows respectively represent the appearance loss, the residual loss, and the final fusion loss (best viewed in color).

ple. Furthermore, we adopt an adversarial loss to improve the quality of density maps.

3.1. Counting by Residual Regression

In this section, we first present a traditional appearance-based counting model with a customized modification. Then, we propose a residual regression algorithm, in which the density map is predicted by comparing the input image with a labeled exemplar image. Finally, the appearance-based and residual-based predictions are fused to yield the final prediction.

3.1.1 Appearance-Based Prediction

In our practice, we adopt two typical networks, MCNN [35] and CSRNet [14], respectively, as the backbone networks, to predict the density map by sole appearance of the image. Given an image X_i as input, the network outputs a predicted density map. Formally, the appearance-based prediction is as follows,

$$\hat{Y}_i^a = F_a(X_i), \tag{1}$$

where X_i is the input image, $\hat{Y}_i^a$ denotes the appearance-based prediction, and $F_a(\cdot)$ is the mapping function approximated by the modified appearance-based network.

3.1.2 Residual-Based Prediction

Exemplar correlation is exploited via a novel regression approach by comparing between samples (see Fig. 2). First, a residual map is estimated from deep features extracted from the input image and an exemplar support image. Then the density map is calculated by adding the estimated residual map and the ground-truth density map of the support image.

Specifically, given a support set containing k labeled images $\{(X_1^s, Y_1^s), (X_2^s, Y_2^s), \cdots, (X_k^s, Y_k^s)\}$, features are firstly extracted by the appearance-based network. Then, the extracted deep features from the input image and a support image are concatenated and fed into a correlation learning network to predict a residual map between density maps of the input and support images. Based on the predicted k residual maps and the corresponding ground-truth density maps of the support images, k estimated density maps based on residual learning can be calculated and consequently fused.

Formally, given an input image X_i and a labeled support image (X_j^s, Y_j^s), the density map with regard to the support image with residual regression is

$$\hat{Y}_i^{r_j} = F_r\left(f_a(X_i), f_a\left(X_j^s\right)\right) + Y_j^s, \tag{2}$$

where $\hat{Y}_i^{r_j}$ is the density map predicted with the j-th support image X_j^s. $F_r(\cdot,\cdot)$ denotes the residual map prediction network, and $f_a(\cdot)$ denotes the feature extraction function defined by the appearance module (without linear function) described above.

Since we have k density maps predicted by comparing with different support images, we thus fuse them to generate the final residual-based density map $\hat{Y}_i^r$,

$$\hat{Y}_i^r = F_{rf}\left(\hat{Y}_i^{r_1}, \hat{Y}_i^{r_2}, \cdots, \hat{Y}_i^{r_k}\right), \tag{3}$$

where F_{rf} is a fusion network.

Table 1. The detailed configuration of the modified appearance module network F_a.

MCNN			CSRNet
C(16,9)-P	C(20,7)-P	C(24,5)-P	2×C(64,3)-P
C(32,7)-P	C(40,5)-P	C(48,3)-P	2×C(128,3)-P
C(16,7)	C(20,5)	C(24,3)	3×C(256,3)-P
C(16,7)	C(20,5)	C(24,3)	3×C(512,3,2)
T(8,4)	T(10,4)	T(12,4)	3×C(512,3,2)
T(4,4)	T(5,4)	T(6,4)	C(256,3,2) C(128,3,2) C(64,3,2)
C(1,1)			C(1,1)

Table 2. The detailed configuration of our proposed network. To encode multi-scale features, we use 3 branches similar to MCNN.

	Branch	Fusion
Residual map predictor F_r	C(16,3) C(8,5) C(4,7)	C(16,7)-C(8,5)-C(1,3)
Residual map fusion F_{rf}	C(8,1) C(4,3) C(2,5)	C(1,3)
Final map fusion F_{ff}	C(8,1) C(4,3) C(2,5)	C(1,3)

3.1.3 Density Maps Fusion

The density maps from the appearance and exemplar correlation are fused together to obtain the final density map. Instead of simply utilizing a 1×1 kernel, we propose a sophisticated network to embed spatial context effectively. Formally, the final prediction $\hat{Y}_i$ is estimated by fusing appearance and residual based predictions as follows,

$$\hat{Y}_i = F_{ff}\left(\hat{Y}_i^a, \hat{Y}_i^r\right), \tag{4}$$

where $F_{ff}(\cdot,\cdot)$ is the final fusion network.

3.2. Counting with Semantic Prior

In this section, we show how to effectively utilize a semantic prior to boost the performance of the proposed residual regression model for crowd counting. Intuitively, the pedestrian density of the semantic area without people, such as wall, trees, and sky, should be close to zero. To employ such a kind of semantic prior, we firstly generate the semantic map of the given image. We adopt a popular encoder-decoder model[1] pre-trained on the ADE20K dataset [36] to generate semantic maps. Note that, pixels in the generated semantic map are classified into two sets: the area without pedestrian and the area potentially containing pedestrian. Only if the confidence score is high enough, would the pixels be classified as the area without pedestrian. Otherwise, the pixels are classified as the area potentially containing pedestrian. The semantic map, which is the same size as the predicted density map, serves as an importance map for each pixel.

Our goal is to eliminate the side effect from false alarms in the area containing no people and simultaneously focus on the area potentially occupied by pedestrian. To maintain the attention in the area with pedestrian, we set the area pixel weight to 1. Ideally, we could naively set the area pixels without pedestrian to zero if the semantic map is generated accurately. However, as we directly utilize a segmentation network without fine-tuning on our dataset (we do not have segmentation labels for the crowd images), the semantic prediction may not be sufficiently accurate. To remedy this problem, we set the weight of pixels containing no pedestrian as a constant value $\sigma \in (0, 1]$, to decrease the density of the area containing no pedestrian by a factor instead of directly setting it to zero. This weighting scheme with discrimination is conducted via a semantic MSE metric which will be described later. By doing so, the side effect of false alarms like trees can be eliminated.

[1] https://github.com/CSAILVision/semantic-segmentation-pytorch

3.3. Network Architecture

The functions in Eqs. (1) - (4) are constructed as four components of a network: an appearance-based module (MCNN or CSRNet) $F_a(\cdot)$ learning deep features, a residual regressor $F_r(\cdot,\cdot)$ predicting the residual map, a residual map fusion module F_{rf} fusing density maps predicted from different support images, and a multi-cue fusion network $F_{ff}(\cdot,\cdot)$ conducting the fusion of density maps in terms of both appearance and residual regression into the final estimation.

The appearance-based modules are illustrated in Table 1. Here, C denotes a convolutional layer and the numbers in braces are the filter number, filter size, and dilation parameters (default=1). P represents a max-pooling layer which decreases resolution to half of the preceding layer. T is the fractionally-stridden convolutional layer which increases the resolution by a factor of two. PRelu [7] activation is appended to every convolutional layer except the last layer for activation. The remaining modules in the network are given in Table 2.

3.4. Loss Function

Given N training samples $\{X_i\}_{i=1}^N$, their corresponding semantic maps $\{M_i\}_{i=1}^N$, and the ground-truth density maps$\{Y_i\}_{i=1}^N$, the loss function of the proposed method is

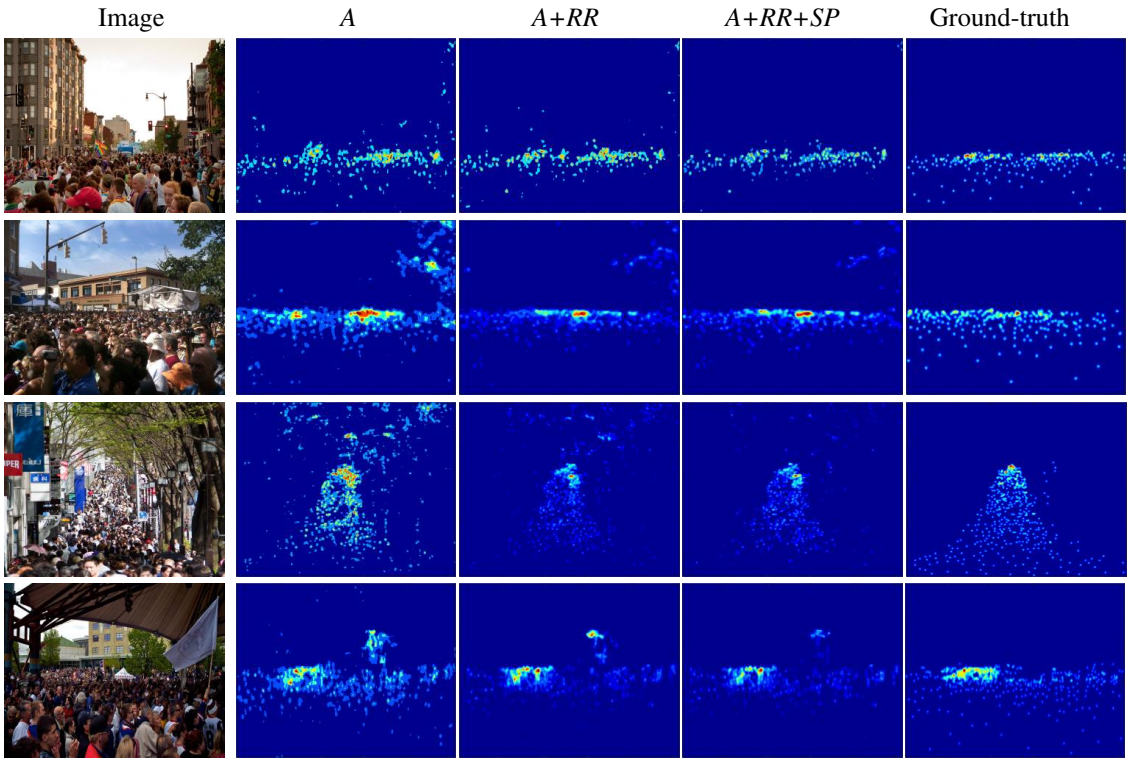

Figure 3. The qualitative comparison between variants of our approach: (*A*) *Appearance*, (*A+RR*) *Appearance + Residual Regression*, and (*A+RR+SP*) *Appearance + Residual Regression + Semantic Prior*.

defined as follows,

$$\mathcal{L} = \mathcal{L}_a + \alpha\mathcal{L}_r + \beta\mathcal{L}_{ff} + \eta\mathcal{L}_{ad}, \tag{5}$$

where $\mathcal{L}_a$, $\mathcal{L}_r$, $\mathcal{L}_{ff}$, and $\mathcal{L}_{ad}$ are the appearance loss, residual loss, final fusion loss, and adversarial loss.

3.4.1 Semantic MSE

As mentioned in Sec. 3.2, a semantic prior based MSE metric is proposed to decrease the influence due to potential false alarms. Specifically, the proposed Semantic MSE counts the area without pedestrian with lower weights. Given a prediction $\hat{Y}_i$, its semantic map M_i, and the corresponding ground-truth Y_i, the semantic MSE (S-MSE) is

$$f_{smse}(\hat{Y}_i, Y_i, M_i) = \|\hat{Y}_i \otimes M_i - Y_i \otimes M_i\|^2, \tag{6}$$

where the value of elements in M_i is either 1 (corresponding to pixels potentially containing person) or σ (corresponding to pixels without person), and $\otimes$ denotes the element-wise multiplication.

3.4.2 Appearance Loss

The appearance loss measures the discrepancy between the predicted density map $\hat{Y}_i^a$ and the ground-truth density map via S-MSE,

$$\mathcal{L}_a = \sum_{i=1}^{N} f_{smse}(\hat{Y}_i^a, Y_i, M_i)\,. \tag{7}$$

3.4.3 Residual Loss

Similar to the appearance loss, the semantic MSE is utilized to measure the difference between residual-based prediction and the ground-truth. Different from the appearance loss, the residual loss is rather complex since we obtain $k+1$ density maps as $\{\hat{Y}_i^{r_1}, \hat{Y}_i^{r_2}, \cdots, \hat{Y}_i^{r_k}, \hat{Y}_i^{r}\}$ based on k support images and the fusion module. The residual loss is defined as

$$\mathcal{L}_r = \sum_{i=1}^{N}\{[\sum_{j=1}^{k} f_{smse}(\hat{Y}_i^{r_j}, Y_i, M_i)] + f_{smse}(\hat{Y}_i^{r}, Y_i, M_i)\}. \tag{8}$$

3.4.4 Final Fusion Loss

The final fusion loss is to measure the distinction between the final prediction after fusion and the ground-truth,

$$\mathcal{L}_{ff} = \sum_{i=1}^{N} f_{smse}(\hat{Y}_i, Y_i, M_i). \tag{9}$$

Table 3. Experimental results on ShanghaiTech A. The arrows next to the metrics indicate the direction of better performance, *i.e.*, ↓ means that smaller values are better. The best results are shown in bold, and the second best results are indicated by underline. This also applies to the following tables.

Method	MAE ↓	MSE ↓
Cross-scene [33]	181.8	277.7
MCNN [35]	110.2	173.2
FCN [20]	126.5	173.5
Cascaded-MTL [26]	101.3	152.4
Switching-CNN [23]	90.4	135.0
CP-CNN [27]	73.6	106.4
ASACP [24]	75.7	102.7
Top-Down [22]	97.5	145.1
L2R [17]	73.6	112.0
CSRNet [14]	68.2	115.0
IG-CNN [21]	72.5	118.2
ic-CNN [21]	68.9	117.3
SANet (patch) [2]	<u>67.0</u>	104.5
SANet (image) [2]	88.1	134.3
SCNet [30]	71.9	117.9
Spatial-Aware [16]	69.3	<u>96.4</u>
Image Pyramid [10]	80.6	126.7
Ours (MCNN, A)	86.97	138.46
Ours (MCNN, A+RR)	79.72	119.9
Ours (MCNN, A+RR+SP)	75.9	118.1
Ours (MCNN, full)	72.6	114.3
Ours (CSRNet, A)	68.2	115.0
Ours (CSRNet, A+RR)	64.8	98.4
Ours (CSRNet, A+RR+SP)	64.2	98.0
Ours (CSRNet, full)	**63.1**	**96.2**

3.4.5 Adversarial Loss

To improve the quality of the predicted density map, an adversarial loss is exploited during training. Specifically, a shallow network is developed as the discriminator. The network configuration is C(64,3)-C(128,3)-P-C(256,3)-P-C(256,3)-C(256,3)-P-C(1,1)-Sigmoid.

Based on the developed discriminator, the adversarial loss is

$$\mathcal{L}_{ad} = \sum_{i=1}^{N} \left(log(Y_i) - log(\hat{Y}_i) \right), \tag{10}$$

where $\hat{Y}_i$ and Y_i are the final predicted and the ground-truth density maps.

4. Experiments

We firstly present the implementation details, then illustrate the dataset & metrics, and subsequently report the experimental results, including *ablation study*, *comparison with the state of the art*, and *cross-dataset evaluation*.

Table 4. Experimental results on ShanghaiTech B.

Method	MAE↓	MSE↓
Cross-scene [33]	32.0	49.8
MCNN [35]	26.4	41.3
FCN [20]	23.76	33.12
Cascaded-MTL [26]	20.0	31.1
Switching-CNN [23]	21.6	33.4
CP-CNN [27]	20.1	30.1
DecideNet [15]	20.75	29.42
ASACP [24]	17.2	27.4
Top-Down [22]	20.7	32.8
L2R [17]	14.4	23.8
CSRNet [14]	10.6	16.0
IG-CNN [1]	13.6	21.1
ic-CNN [21]	10.7	16.0
SANet [2]	**8.4**	<u>13.6</u>
SCNet [30]	9.3	14.4
Spatial-Aware [16]	11.1	18.2
Image Pyramid [10]	10.2	18.3
Ours (MCNN)	15.5	23.1
Ours (CSRNet)	<u>8.72</u>	**13.56**

4.1. Implementation Details

Density Map Synthesis. We synthesize density maps as ground truth following [35].

Generation of Support Set. The support set plays an important role in the proposed approach. Images in the set should exhibit as broad coverage, in terms of both crowdedness and spatial structure, as possible to boost the generalization ability. We first extract spatial crowd features from each training image. Specifically, we divide the image space into grid regions. The count value of each grid is the summation of pixel values in the covered region. Then these count values are concatenated as a vector as the crowd feature encoding the spatial information. Subsequently, the training images are grouped into several clusters by the *k*-means algorithm. For each cluster, we select the image which is closest to the cluster centroid as one support image in the set. The number of support images is set to 3, as empirically increasing the number leads to an increase of computational cost, albeit with a very limited performance boost.

Training. To stabilize the training of the proposed network, we firstly train appearance based module in the network and then the whole network is optimized with the pre-trained appearance based network. Note that, the learning rate is set to 0.0001 during the training of the appearance based network. When training the whole network, the learning rate for other components and the appearance module is 0.0001 and 0.00001 respectively. Empirically, σ is set to 0.5. α, β, and η are set to 1, 1, and 1×10^{-12}, respectively.

Table 5. Experimental results on WorldExpo. MAE is used for evaluation. Avg. is the average result over all testing scenes.

Method	S1	S2	S3	S4	S5	Avg.
Cross-scene [33]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [35]	3.4	20.6	12.9	12.0	8.1	11.6
SwitchingCNN [23]	4.4	15.7	10.0	11.0	5.9	9.4
CP-CNN [27]	2.9	14.7	10.5	10.4	5.8	8.86
CNN-pixel [11]	2.9	18.6	14.1	24.6	6.9	13.4
Body structure [8]	4.1	21.7	11.9	11.0	3.5	10.5
DecideNet [15]	**2.0**	13.1	8.9	17.4	4.8	9.2
ASACP [24]	2.8	14.1	9.6	**8.1**	2.9	**7.5**
Top-Down [22]	2.7	23.4	10.7	17.6	3.3	11.5
CSRNet [14]	2.9	11.5	8.6	16.6	3.4	8.6
IG-CNN [1]	2.6	16.1	10.2	20.2	7.6	11.3
ic-CNN [21]	17.0	12.3	9.2	**8.1**	4.7	10.3
SANet [2]	2.6	13.2	9.0	13.3	3.0	8.2
SpatialAware [16]	2.6	11.8	10.3	10.4	3.7	7.76
ImagePyramid [10]	2.5	16.5	12.2	20.5	2.9	10.9
Ours (MCNN)	2.2	**11.1**	11.3	15.8	2.8	8.7
Ours (CSRNet)	2.9	15.0	**7.2**	14.7	**2.6**	8.5

4.2. Datasets & Metrics

The evaluation is performed on three popular datasets: ShanghaiTech [35], Expo [33] and UCF_50 [9]. The ShanghaiTech dataset includes two parts A and B. ShanghaiTech A consists of 482 images with diverse resolutions and the crowd number varies from 33 to 3139. ShanghaiTech B contains 716 images (768×1024), and the crowd number varies from 9 to 518. WorldExpo is a real-world surveillance dataset containing 3980 labeled frames (576×720). UCF_50 is a very challenging dataset consisting of 50 extremely congested images (average of 1279 people, maximum of 4535) with different resolutions.

To provide quantitative evaluation, metrics of the Mean Absolute Error (MAE) and the Rooted Mean Squared Error (RMSE) are utilized:

$$MAE = \frac{1}{K}\sum_{i=1}^{K}|\hat{C}_i - C_i|, RMSE = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\|\hat{C}_i - C_i\|^2}, \tag{11}$$

where K is the size of the testing size. $\hat{C}_i$ and C_i are the predicted and ground-truth crowd counts, as computed from the corresponding density maps.

4.3. Ablation Study

An ablation study is performed on ShanghaiTech A to evaluate the effectiveness of each module. Based on MCNN and CSRNet, the following variants are compared:

1) *Appearance (A)*: a modified version of MCNN or CSRNet.

2) *Appearance + Residual Regression (A+RR)*: the combination of appearance and residual regression modules.

Table 6. Experimental results of cross-dataset evaluation.

Method	Shanghai Tech_B		Expo		UCF_50	
	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [35]	-	-	-	-	397.6	624.1
L2R [17]	-	-	-	-	337.6	434.3
Appearance	44.7	87.7	62.2	85.3	358.2	562.1
Ours(MCNN)	40.0	68.5	30.4	42.5	355.0	560.2

3) *Appearance + Residual Regression + Semantic Prior (A+RR+SP)*: the model utilizing appearance, residual regression and semantic prior.

4) *Ours (full)*: our proposed approach trained with adversarial loss.

The results of these variants are reported in Table 3. We observe that,

1) *A+RR* outperforms *A* with a significant advantage, indicating that after mining exemplar correlation knowledge, the network is more effective. Moreover, the performance of the sole *A* model is improved if compared with the original MCNN [35] – the learned deep features become more powerful after joint optimization with the pair-wise correlation information. Fig. 3 shows qualitative results of four exemplar images. The residual regression module makes the estimated density map closer to the ground truth.

2) The performance is improved comparing *A+RR+SP* with *A+RR*, suggesting that the embedded semantic prior is effective to eliminate false alarms. This is also confirmed by the comparison between the third and the fourth columns in Fig. 3. The embedded semantic prior eliminates false density in the area of sky (see the top row) and the tree (see the second and the third rows).

3) Our full method additionally reduces the MAE and MSE values, which demonstrates the effectiveness of the introduced adversarial loss.

4) With the replacement of MCNN by CSRNet, we achieve results which are similar to those of MCNN, showing that the proposed framework is effective even when based on a strong baseline.

4.4. Comparison with State-of-the-Arts

We conduct evaluation on ShanghaiTech Part A & B, WorldExpo and compare the results with the state-of-the-art algorithms shown in Table 3-5.

1) On ShanghaiTech A, Table 3 shows the proposed method achieves the best performance in terms of MAE and MSE which indicates the advantage of the proposed approach over others.

2) Similarly, on ShanghaiTech B, the proposed algorithm achieves better performance than most of the algorithms except SANet [2] as shown in Table 4. It is important to note that, the evaluation method in SANet is different from the standard one in the literature. The best performance of

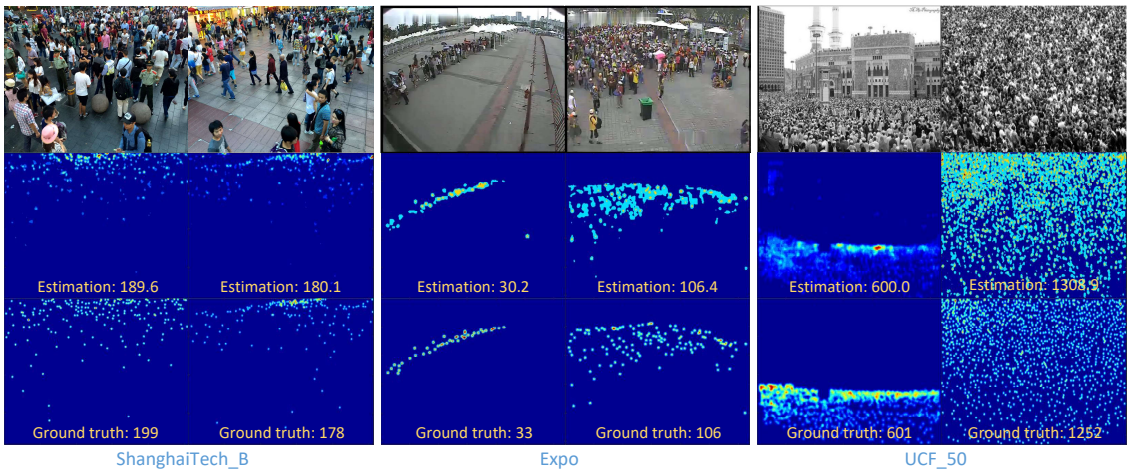

Figure 4. Exemplar images (top) of the employed datasets, the corresponding predictions (middle) by the proposed approach, and the ground-truth (bottom).

SANet is achieved by its patch-level evaluation. However, as Table 3 shows, when using the standard image-level evaluation, the performance of SANet degrades severely. Thus the performance of SANet using the standard evaluation protocol is expected to drop on the ShanghaiTech B dataset. Nevertheless, our method still achieves comparable performance as SANet (patch).

3) The dataset of WorldExpo is proposed to evaluate the generalization ability of algorithms in different scenarios. As shown in Table 5, the proposed method achieves the best performance on Scene 3 & 5 revealing that the proposed method can potentially transfer better to unseen scenarios.

In most cases, the proposed method achieves slightly superior performance compared to the start-of-the-art methods. We further analyze the qualitative results shown in Fig. 4. After correlation knowledge and semantic prior are included, density map estimation becomes more accurate. However, there also exists failure case. The fire hydrant in the example (the third column) is incorrectly classified as crowds since the appearance of fire hydrant is very similar to a single person. Therefore, mining hard negative examples can be further investigated in the future.

4.5. Cross-Dataset Evaluation

When applying crowd counting methods in real-world applications, the generalization ability is very important to ensure satisfactory performance in case of unseen scenes. To further evaluate the generalization ability of the proposed method, a cross-dataset experiment is conducted.

In this experiment, the source domain is ShanghaiTech A and the other datasets served as the target domains. The model is trained with ShanghaiTech A and tested on other datasets without fine-tuning. We report only the performance of full method with MCNN in Table 6 as UCF_50 contains only gray-scale images while CSRNet is trained with RGB images. Note that, for MCNN [35] and L2R [17] we cannot report results for ShanghaiTech B and Expo, as their models/codes are not public and we cannot find other source to quote from. The proposed method shows better performance than sole Appearance model based on MCNN [35], revealing that the generalization ability is improved after the appearance and residual regression models are jointly optimized. Compared with MCNN , the proposed approach achieves better performance on UCF_50 dataset. However, L2R is slightly superior to ours since additional data are used during training.

5. Conclusion

In this paper, a novel residual regression approach was proposed to incorporate correlation knowledge. This approach can learn more effective features, and shows a better generalization capability after joint optimization of appearance and correlation. In addition, the semantic prior was leveraged to compute the loss function, which was demonstrated to be effective to eliminate false alarms in crowd images. To further improve the quality of the predicted density maps, the adversarial loss was employed to regularize the predicted density maps. In the future exploration, crowd images could be synthesized to help the network transfer to unseen scenarios.

Acknowledgements

This work was partially supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. [T32-101/15-R] and CityU 11212518), and by a Strategic Research Grant from City University of Hong Kong (Project No. 7004887). We are grateful to the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

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, 2018.

[2] 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 734–750, 2018.

[3] A. B. Chan, Zhang-Sheng John Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[4] A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *International Conference on Computer Vision*, pages 545–551, 2009.

[5] W. Ge and R. T. Collins. Marked point processes for crowd counting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2913–2920, 2009.

[6] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (TOG)*, 26(3):4, 2007.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[8] Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, Shenghua Gao, Rongrong Ji, and Junwei Han. Body structure aware deep crowd counting. *IEEE Trans. Image Processing*, 27(3):1049–1059, 2018.

[9] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013.

[10] Di Kang and Antoni B. Chan. Crowd counting by adaptively fusing predictions from an image pyramid. In *British Machine Vision Conference*, page 89, 2018.

[11] Di Kang, Zheng Ma, and Antoni B Chan. Beyond counting: comparisons of density maps for crowd analysis tasks-counting, detection, and tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[12] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in neural information processing systems*, pages 1324–1332, 2010.

[13] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *International Conference on Pattern Recognition*, pages 1–4, 2008.

[14] 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, 2018.

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

[16] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 849–855, 2018.

[17] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] Wenhan Luo, Peng Sun, Fangwei Zhong, Wei Liu, Tong Zhang, and Yizhou Wang. End-to-end active object tracking and its real-world deployment via reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[19] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Xiaowei Zhao Tae-Kyun Kim. Multiple object tracking: A literature review. *CoRR*, abs/1409.7618, 2017.

[20] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E. O'Connor. Fully convolutional crowd counting on highly congested scenes. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 27–33, 2017.

[21] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *European Conference on Computer Vision*, pages 278–293, 2018.

[22] Deepak Babu Sam and R. Venkatesh Babu. Top-down feedback for crowd counting convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7323–7330, 2018.

[23] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4031–4039, 2017.

[24] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5245–5254, 2018.

[25] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5382–5390, 2018.

[26] Vishwanath A Sindagi and Vishal M Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–6, 2017.

[27] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *IEEE International Conference on Computer Vision*, pages 1879–1888, 2017.

[28] Vishwanath A Sindagi and Vishal M Patel. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3–16, 2018.

[29] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[30] Ze Wang, Zehao Xiao, Kai Xie, Qiang Qiu, Xiantong Zhen, and Xianbin Cao. In defense of single-column networks for crowd counting. In *British Machine Vision Conference*, page 78, 2018.

[31] Flood Sung Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[32] Jenny Yuen and Antonio Torralba. A data-driven approach for event prediction. In *European Conference on Computer Vision*, pages 707–720. Springer, 2010.

[33] Cong Zhang, Hongsheng Li, X. Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 833–841, 2015.

[34] L. Zhang, M. Shi, and Q. Chen. Crowd counting via scale-adaptive convolutional neural network. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1113–1121, 2018.

[35] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 589–597, 2016.

[36] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5122–5130, 2017.