Multi-Level Bottom-Top and Top-Bottom Feature Fusion for Crowd Counting

Vishwanath A. Sindagi Vishal M. Patel
Department of Electrical and Computer Engineering,
Johns Hopkins University, 3400 N. Charles St, Baltimore, MD 21218, USA
{vishwanathsindagi,vpatel36}@jhu.edu

Abstract

Crowd counting presents enormous challenges in the form of large variation in scales within images and across the dataset. These issues are further exacerbated in highly congested scenes. Approaches based on straightforward fusion of multi-scale features from a deep network seem to be obvious solutions to this problem. However, these fusion approaches do not yield significant improvements in the case of crowd counting in congested scenes. This is usually due to their limited abilities in effectively combining the multi-scale features for problems like crowd counting. To overcome this, we focus on how to efficiently leverage information present in different layers of the network. Specifically, we present a network that involves: (i) a multi-level bottom-top and top-bottom fusion (MBTTBF) method to combine information from shallower to deeper layers and vice versa at multiple levels, (ii) scale complementary feature extraction blocks (SCFB) involving cross-scale residual functions to explicitly enable flow of complementary features from adjacent conv layers along the fusion paths. Furthermore, in order to increase the effectiveness of the multi-scale fusion, we employ a principled way of generating scale-aware ground-truth density maps for training. Experiments conducted on three datasets that contain highly congested scenes (ShanghaiTech, UCF_CROWD_50, and UCF-QNRF) demonstrate that the proposed method is able to outperform several recent methods in all the datasets.

1. Introduction

Computer vision-based crowd counting [8, 17, 26, 27, 36, 44, 48, 56, 68, 69, 74, 77] has witnessed tremendous progress in the recent years. Algorithms developed for crowd counting have found a variety of applications such as video and traffic surveillance [15, 21, 38, 59, 64, 71, 72], agriculture monitoring (plant counting) [35], cell counting [22], scene understanding, urban planning and environmental survey [11, 68].

Crowd counting from a single image, especially in congested scenes, is a difficult problem since it suffers from multiple issues like high variability in scales, occlusions, perspective changes, background clutter, etc. Recently, several convolutional neural network (CNN) based methods [3, 7, 34, 43, 48, 49, 51, 56, 69, 74] have attempted to address these issues with varying degree of successes. Among these issues, the problem of scale variation has particularly received considerable attention from the research community. Scale variation typically refers to large variations in scale of the objects being counted (in this case heads) (i) within image and (ii) across images in a dataset. Several other related tasks like object detection [6, 16, 23, 30, 37, 45] and visual saliency detection [10, 14, 41, 73] are also affected by such effects. However, these effects are more evident especially in crowd counting in congested scenes. Furthermore, since the annotation process for highly congested scenes is notoriously challenging, the datasets available for crowd counting typically provide only x, y location information about the heads in the images. Since the scale labels are unavailable, training the networks to be robust to scale variations is much more challenging. In this work, we focus on addressing the issue of scale variation and missing scale information from the annotations.

CNNs are known to be relatively less robust to the presence of such scale variations and hence, special techniques are required to mitigate their effects. Using features from different layers of a deep network is one approach that has been successful in addressing this issue for other problems like object detection. It is well known that feature maps from shallower layers encode low-level details and spatial information [6, 13, 29, 42, 67], which can be exploited to achieve better localization. However, such features are typically noisy and require further processing. Meanwhile, deeper layers encode high-level context and semantic information [6, 13, 29, 42] due to their larger receptive field sizes, and can aid in incorporating global context into the network. However, these features lack spatial resolution, resulting in poor localization. Motivated by these observations, we believe that high-level global semantic informa-

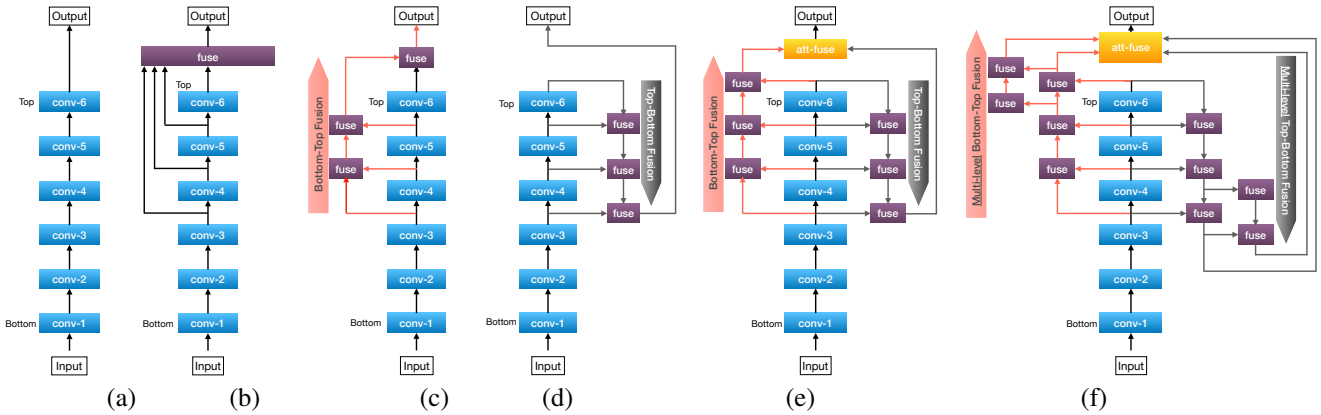


Figure 1. Illustration of different multi-scale fusion architectures: (a) No fusion, (b) Fusion through concat or add, (c) Bottom-top fusion, (d) Top-bottom fusion, (e) Bottom-top and top-bottom fusion, (f) Multi-level bottom-top and top-bottom fusion (proposed).

tion and spatial localization play an important role in generating effective features for crowd counting, and hence, it is important to fuse features from different layers in order to achieve lower count errors.

In order to perform an effective fusion of information from different layers of the network, we explore different fusion architectures as shown in Fig. 1*(a)-(d)*, and finally arrive at our proposed method (Fig. 1*(f)*). Fig. 1*(a)* is a typical deep network which processes the input image in a feed-forward fashion, with no explicit fusion of multi-scale features. The network in Fig. 1*(b)* extracts features from multiple layers and fuses them simultaneously using a standard approach like addition or concatenation. With this configuration, the network needs to learn the importances of features from different layers automatically, resulting in a sub-optimal fusion approach. As will be seen later in Section 5.2, this method does not produce significant improvements as compared to the base network.

To overcome this issue, one can choose to progressively incorporate detailed spatial information into the deeper layers by sequentially fusing the features from lower to higher layers (bottom-top) as shown in Fig. 1*(c)* [58]. This fusion approach explicitly incorporates spatial context from lower layers into the high-level features of the deeper layers. Alternatively, a top-bottom fusion (Fig. 1*(d)*) [47] may be used that involves suppressing noise in lower layers, by propagating high-level semantic context from deeper layers into them. These approaches achieve lower counting errors as compared to the earlier configurations. However, both of these methods follow uni-directional fusion which may not necessarily result in optimal performance. For instance, in the case of bottom-top fusion, noisy features also get propagated to the top layers in addition to spatial context. Similarly, in the case of top-bottom fusion, the features from the top layer may end up suppressing more than necessary details in the lower layers. Variants of these top-bottom approaches and bottom-top approaches have been proposed for other problems like semantic segmentation and object detection [12, 32, 40, 52].

Recently, a few methods [66, 76] have demonstrated superior performance on other tasks by using multi-directional fusion technique (Fig. 1*(e)*) as compared to uni-directional fusion. Motivated by the success of these methods on their respective tasks, we propose a multi-level bottom-top and top-bottom fusion (MBTTBF) technique as shown in Fig 1*(f)*. By doing this, more powerful features can be learned by enabling high-level context and spatial information to be exchanged between scales in a bidirectional manner. The bottom-top path ensures flow of spatial details into the top layer, while the top-bottom path propagates context information back into the lower layers. The feedback through both the paths ensures that minimal noise is propagated to the top layer in the bottom-top direction, and also that the context information does not over-suppress the details in the lower layers. Hence, we are able to effectively aggregate the advantages of different layers and suppress their disadvantages. Note that, as compared to existing multi-directional fusion approaches [66, 76], we propose a more powerful fusion technique that is multi-level and aided by scale-complementary feature extraction blocks (see Section 3.2). Additionally, the fusion process is guided by a a set of scale-aware ground-truth density maps (see Section 3.3), resulting in scale-aware features.

Furthermore, we propose a scale complementary feature extraction block (SCFB) which uses cross-scale residual blocks to extract features from adjacent scales in such a way that they are complementary to each other. Traditional fusion approaches such as feature addition or concatenation are not necessarily optimal because they simple merge the features and have limited abilities to extract relevant information from different layers. In contrast, the proposed scale complementary extraction enables the network to compute relevant features from each scale.

Lastly, we address the issue of missing scale-information

in crowd-datasets by approximating the same based on the crowd-density levels and superpixel segmentation principles. Zhang *et al*. [74] also estimate the scale information, however, they rely on heuristics based on the nearest number of heads. In contrast, we combine information from the annotations and super-pixel segmentation of the input image in a Markov Random Field (MRF) framework [25].

The proposed counting method is evaluated and compared against several recent methods on three recent datasets that contain highly congested scenes: ShanghaiTech [74], UCF_CROWD_50[17], and UCF-QNRF [19]. The proposed method outperforms all existing methods by a significant margin.

We summarize our contributions as follows:

- A multi-level bottom-top and top-bottom fusion scheme to effectively merge information from multiple layers in the network.
- A scale-complementary feature extraction block that is used to extract relevant features form adjacent layers of the network.
- A principled way of estimating scale-information for heads in crowd-counting datasets that involves effectively combining annotations and super-pixel segmentation in a MRF framework.

2. Related work

Compared to traditional approaches ([9, 17, 22, 24, 39, 46, 65]), recent methods have exploited Convolutional neural networks (CNNs) [2, 5, 38, 48, 48, 56, 60, 62, 69, 74] to obtain dramatic improvements in error rates. Typically, existing CNN-based methods have focused on design of different architectures to address the issue of scale variation in crowd counting. Switching-CNN, proposed by Babu *et al*. [48], learns multiple independent regressors based on the type of image patch and has an additional switch classifier to automatically choose the appropriate regressor for a particular input patch. More recently, Sindagi *et al*. [56] proposed Contextual Pyramid CNN (CP-CNN), where they demonstrated significant improvements by fusing local and global context through classification networks. For a more elaborate study and discussion on these methods, interested readers are referred to a recent survey [57] on CNN-based counting techniques.

While the these methods build techniques that are robust to scale variations, more recent methods have focused on other aspects such as progressively increasing the capacity of the network based on dataset [3], use of adversarial loss to reduce blurry effects in the predicted output maps [49, 56], learning generalizable features via deep negative correlation based learning [51], leveraging unlabeled data for counting by introducing a learning to rank framework [34], cascaded feature fusion [43] and scale-based feature aggregation [7], weakly-supervised learning for crowd counting [58]. Recently, Idrees *et al*. [19] created a new large-scale high-density crowd dataset with approximately 1.25 million head annotations and a new localization task for crowded images.

Most recently, several methods have focused on incorporating additional cues such as segmentation and semantic priors [61, 75], attention [31, 54, 58], perspective [50], context information respectively [33], multiple-views [70] and multi-scale features [20] into the network. Wang *et al*. [63] introduced a new synthetic dataset and proposed a SSIM based CycleGAN [78] to adapt the synthetic datasets to real world dataset.

3. Proposed method

In this section, we discuss details of the proposed multi-level feature fusion scheme along with the scale complementary feature extraction blocks. This is followed by a discussion on the estimation of head sizes using the MRF framework.

3.1. Multi-level bottom-top and top-bottom Fusion (MBTTBF)

The proposed method for crowd counting is based on the recently popular density map estimation approach [22, 39, 65], where the network takes image as an input, processes it and produces a density map. This density map indicates the per-pixel count of people in the image. The network weights are learned by optimizing the L_2 error between the predicted density map and the ground truth density map. As discussed earlier, crowd counting datasets provide x, y locations and these are used to create the ground-truth density maps for training by imposing 2D Gaussians at these locations:

$$D_i(x) = \sum_{x_g \in S} \mathcal{N}(x - x_g, \sigma), \tag{1}$$

where σ is the Gaussian kernel's scale and S is the list of all locations of people. Integrating the density map over its width and height produces the total count of people in the input image.

Fig 2 illustrates the overview of the proposed network. We use VGG16 [53] as the backbone network. Conv1 - conv5 in Fig. 2 are the first five convolutional layers of the VGG16 network. The last layer conv6 is defined as $\{M_2 - C_{512,128,1} - R\}$[1]). As it can be observed from this figure, the network consists of primarily three branches: (i) main branch (VGG16 backbone), (ii) multi-level bottom-top fusion branch, and (iii) multi-level top-bottom fusion

[1] M_s denotes max-pooling with stride s, $C_{N_i,N_o,k}$ is convolutional layer (where N_i = number of input channels, N_o = number of output channels, $k \times k$ = size of filter), R is activation function (ReLU).

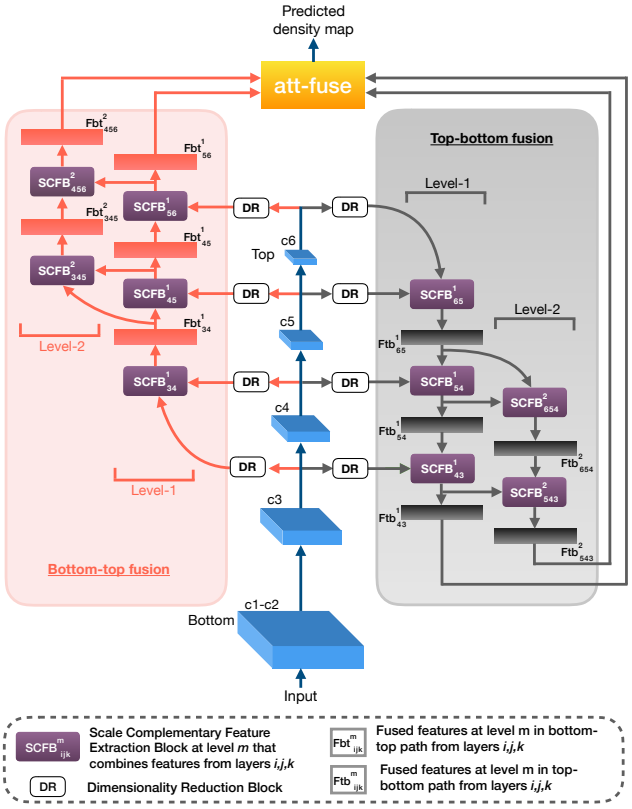


Figure 2. Overview of the proposed multi-level top-bottom and bottom-top fusion method for crowd counting.

branch. The input image is passed through the main branch and multi-scale features from conv3-conv6 layers are extracted. These multi-scale features are then forwarded through dimensionality reduction (DR) blocks that consists of 1×1 conv layers to reduce the channel dimensions to 32.

The feature maps extracted from the lower conv layers of the main branch contain detailed spatial information which are important for accurate localization, whereas the feature maps from higher layers contain global context and high-level information. The information contained in these different layers are fused with each other in two separate fusion branches: multi-level bottom-top branch and multi-level top-bottom branch.

Multi-level bottom-top fusion: The bottom-top branch hierarchically propagates spatial information from the bottom layers to the top layers. This branch has two levels of fusion. In the first level, features from the main branch are progressively forwarded through a series of scale complementary feature extraction blocks ($SCFB^1_{34}$-$SCFB^1_{45}$-$SCFB^1_{56}$). First, $SCFB^1_{34}$ combines the feature maps from conv3 and conv4 to produce enriched feature maps Fbt^1_{34}. These features are then combined with conv5 features of the main branch through $SCFB^1_{45}$ to produce Fbt^1_{45}. Finally, these feature maps are combined with conv6 feature maps through $SCFB^1_{56}$ to produce Fbt^1_{56}.

Further, we add another level of bottom-top fusion path which progressively combines features from the first level through another series of scale complementary feature extraction blocks ($SCFB^2_{345}$-$SCFB^2_{456}$). Specifically, Fbt^1_{34} and Fbt^1_{45} are combined through $SCFB^2_{345}$ to produce Fbt^2_{345}. Finally, Fbt^2_{345} is combined with Fbt^1_{56} through $SCFB^2_{456}$ to produce Fbt^2_{456}. The two levels of fusion together form a hierarchy of fusion paths.

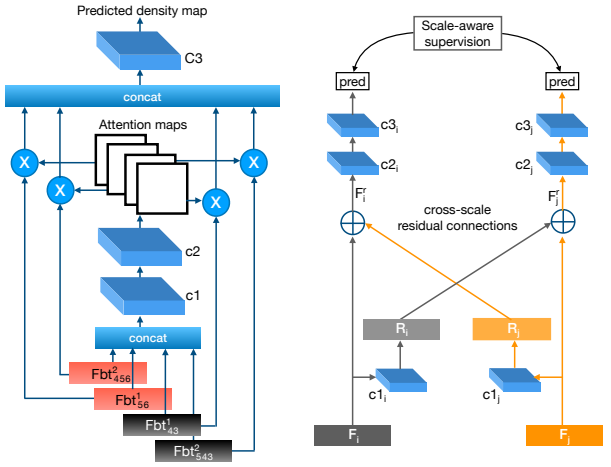
Multi-level top-bottom fusion: The bottom-top branch while propagating spatial information to the top layers, inadvertently passes noise information as well. To overcome this, we add a top-bottom fusion path that hierarchically propagates high-level context information into the lower layers. Similar to the bottom-top path, the top-bottom path also consists of two levels of fusion. In the first level, features from the main branch are progressively forwarded through a series of scale complementary feature extraction blocks ($SCFB^1_{65}$-$SCFB^1_{54}$-$SCFB^1_{43}$). First, $SCFB^1_{65}$ combines the feature maps from conv6 and conv5 to produce enriched feature maps Ftb^1_{65}. These features are then combined with conv4 features of the main branch through $SCFB^1_{54}$ to produce Ftb^1_{54}. Finally, these feature maps are combined with conv3 feature maps through $SCFB^1_{43}$ to produce Ftb^1_{43}.

The second level of bottom-top fusion path progressively combines features from the first level through another series of scale complementary feature extraction blocks ($SCFB^2_{654}$-$SCFB^2_{543}$). Specifically, Ftb^1_{65} and Ftb^1_{54} are combined through $SCFB^2_{654}$ to produce Ftb^2_{654}. Finally, Ftb^2_{654} is combined with Ftb^1_{43} through $SCFB^2_{543}$ to produce Fbt^2_{543}. Again, the two levels of fusion together form a hierarchy of fusion paths in the top-bottom module.

Self attention-based fusion: The features produced by the bottom-top fusion (Fbt^1_{56} and Fbt^2_{456}), although refined, may contain some unnecessary background clutter. Similarly, the features (Ftb^1_{43} and Ftb^2_{543}) produced by the top-bottom fusion may over suppress the detail information in the lower layers. In order to further suppress the background noise in the bottom-top path and avoid over-suppression of detail information due to the top-bottom path, we introduce a self-attention based fusion module at the end that combines feature maps from the two fusion paths. Given the set of feature maps (Fbt^1_{56} , Fbt^2_{456}, Ftb^1_{43} and Ftb^2_{543}) from the fusion branches, the attention module concatenates them and forwards them through a set of conv layers ($\{C_{128,16,3} - R - \{C_{16,4,1}\}$[1]) and a sigmoid layer to produces an attention maps with four channels, with each channel specifying the importance of the corresponding feature map from the fusion

branch. The attention maps are calculated as follows: $A = sigmoid(cat(F^1_{56}, F^2_{456}, F^1_{43}, F^2_{543}))$.

These attention maps are then multiplied element-wise to produce the final feature map: $F_f = A^1 \odot F^1_{56} + A^2 \odot F^2_{456} + A^3 \odot F^1_{43} + A^4 \odot F^2_{543}$, where $\odot$ denotes element-wise multiplication. This self-attention module effectively combines the advantages of the two paths, resulting in more powerful and enriched features. Fig. 3*(a)* shows the self-attention block used to combine different feature maps. The final features F_f are then forwarded through 1×1 conv layer to produce the density map Y_{pred}.



(a) (b)

Figure 3. (a)Attention fuse module. (b) Scale complementary feature extraction block (SCFB).

3.2. Scale complementary feature extraction block (SCFB)

In this section, we describe the scale complementary feature extraction block that is used to combine features from adjacent layers in the network. Existing methods such as feature addition or concatenation are limited in their abilities to learn complementary features. This is because features of adjacent layers are correlated, and this results in some ambiguity in the fused features. To address this issue, we introduce scale complementary feature extraction block as shown in Fig. 3*(b)*. This block enables extraction of complementary features from each of the scales being fused. The initial conv layers $c1_i, c1_j, c2_i, c2_j$ in Fig. 3*(b)* are defined as $\{C_{32,32,3} - R\}$[1], where as the final conv layers $c3_i, c3_j$ are defined as $\{C_{32,1,1} - R\}$[1].

The SCFB consists of cross-scale residual connections (R_i and R_j) which are followed by a set of conv layers. The individual branches in the SCFB are supervised by scale-aware supervision (which is now possible due to the scale estimation framework discussed in Section 3.3). More specifically, in order to combine feature maps F_i, F_j from layers i, j, first the corresponding cross-scale residual features F^r_i, F^r_j are estimated and added to the original feature maps F_i, F_j to produce $\hat{F}_i, \hat{F}_j$, *i.e.*, $\hat{F}_i = F_i + F^r_j$ and $\hat{F}_j = F_j + F^r_i$. These features are then forwarded through a set of conv layers, before being supervised by the scale-aware ground-truth density maps Y^s_i, Y^s_j. By adding these intermediate supervisions and introducing the cross-scale residual connections, we are able to compute complementary features from the two scales in the form of residuals. This reduces the ambiguity as compared to the existing fusion methods. For example, if a feature map F_i from a particular layer/scale i is sufficient enough to obtain perfect prediction, then the residual F^r_j is simply driven towards zero. Hence, involving residual functions reduces the ambiguity as compared to the existing fusion techniques.

Figure 4. Scale aware ground truth density maps imposed on the input image. The overall density map is divided into four maps based on the size/scale of the heads. The first image (leftmost) has density corresponding to the smallest set of heads, whereas the last image (rightmost) has densities corresponding to the largest set of heads.

In order to supervise the SCFBs, we create scale-aware ground-truth density maps based on the scales/sizes estimated as described in Section 3.3. Annotations in a particular image are divided into four categories based on the corresponding head sizes, and these four categories are used to create four separate ground-truth density maps ($Y^s_3, Y^s_4, Y^s_5 and Y^s_6$) for a particular image. Fig. 4 shows the four scale-aware ground-truth density maps for two sample images. It can be observed that the first ground-truth (left) has labels corresponding to the smallest heads, where as the last ground-truth (right) has labels corresponding to the largest heads. These maps ($Y^s_3, Y^s_4, Y^s_5 and Y^s_6$) are used to provide intermediate supervision to feature maps coming from conv layers 3,4,5 and 6 coming from the main branch in SCFBs.

3.3. Head size estimation using MRF framework

As discussed earlier, the ground truth density maps for training the CNNs are created by imposing 2D Gaussians at the head locations (Eq. (1)) provided in the dataset. The scale/variance of these Gaussians needs to be decided based on the heads size. Existing methods either assume constant

variance [56] or estimate the variance based on the number of nearest heads [74]. Assuming constant variance results in ambiguity in the density maps and hence, prohibits the network to learn scale relevant features. Fig. 5*(a)* shows the scales for annotations assuming constant variance. On the other hand, estimating the variance based on nearest neighbours leads to better results in regions of high density. However, in regions of low density, the estimates are incorrect leading to ambiguity in such regions (as shown in Fig. 5*(b)*).

To overcome these issues, we propose a principled way of estimating the scale or variance by considering the input images which were not exploited earlier. We leverage color cues from the input image and combine them with the annotation data to better estimate the scale. Specifically, we first over-segment the input image using a super-pixel algorithm (SLIC [1]) and then combine with watershed segmentation [4] resulting from the distance transform of the head locations in an MRF framework. The size of the segments resulting from this procedure are then used to estimate the scale of the corresponding head lying in that segment. Fig. 5*(c)* shows the scales/variances estimated using the proposed method. It can be observed that this method performs better in both sparse and dense regions.

(a) (b) (c)

Figure 5. Scale estimation comparison. Scale estimated using (a) Constant scale (b) Nearest neighbours (c) Our method.

4. Details of implmentation and training

The network weights are optimized in and end-to-end fashion. We use Adam optimizer with a learning rate of 0.00005 and a momentum of 0.9. We add random noise and perform random flipping of images for data augmentation. We use mean absolute error (MAE) and mean squared error (MSE) for evaluating the network performance. These metrics are defined as: $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|$ and $MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|^2}$ respectively, where N is the total number of test images, y_i is the ground-truth/target count of people in the image and y_i' is the predicted count of people in to the i^{th} image. Supervision is provided to the network at the final level as well as at intermediate levels in the SCFBs using Euclidean loss. At the final level, the network is supervised by the overall density map (consisting of annotations corresponding to all the heads), whereas the paths in the SCFBs are supervised by the corresponding scale-aware ground-truths.

5. Experiments and results

In this section, we first analyze the different components involved in the proposed network through an ablation study. This is followed by a detailed evaluation of the proposed method and comparison with several recent state-of-the-art methods.

5.1. Datasets

We use three different congested crowd scene datasets (ShanghaiTech [74], UCF_CROWD_50[17] and UCF-QNRF [19]) for evaluating the proposed method. The ShanghaiTech [74] dataset contains 1198 annotated images with a total of 330,165 people. This dataset consists of two parts: Part A with 482 images and Part B with 716 images. Both parts are further divided into training and test datasets with training set of Part A containing 300 images and that of Part B containing 400 images. The UCF_CC_50 is an extremely challenging dataset introduced by Idrees *et al*. [17]. The dataset contains 50 annotated images of different resolutions and aspect ratios crawled from the internet. The UCF-QNRF [19] dataset, introduced recently by Idrees *et al*., is a large-scale crowd dataset containing 1,535 images with 1.25 million annotations. The images are of high resolution and are collected under a diverse backgrounds such as buildings, vegetation, sky and roads. The training and test sets in this dataset consist of 1201 and 334 images, respectively.

5.2. Ablation Study

We perform a detailed ablation study to understand the effectiveness of various fusion approaches described earlier. The ShanghaiTech Part A and UCF-QNRF datasets contain different conditions such as high variability in scale, occluded objects and large crowds, *etc*. Hence, we used these datasets for conducting the ablations. The following configurations were trained and evaluated:

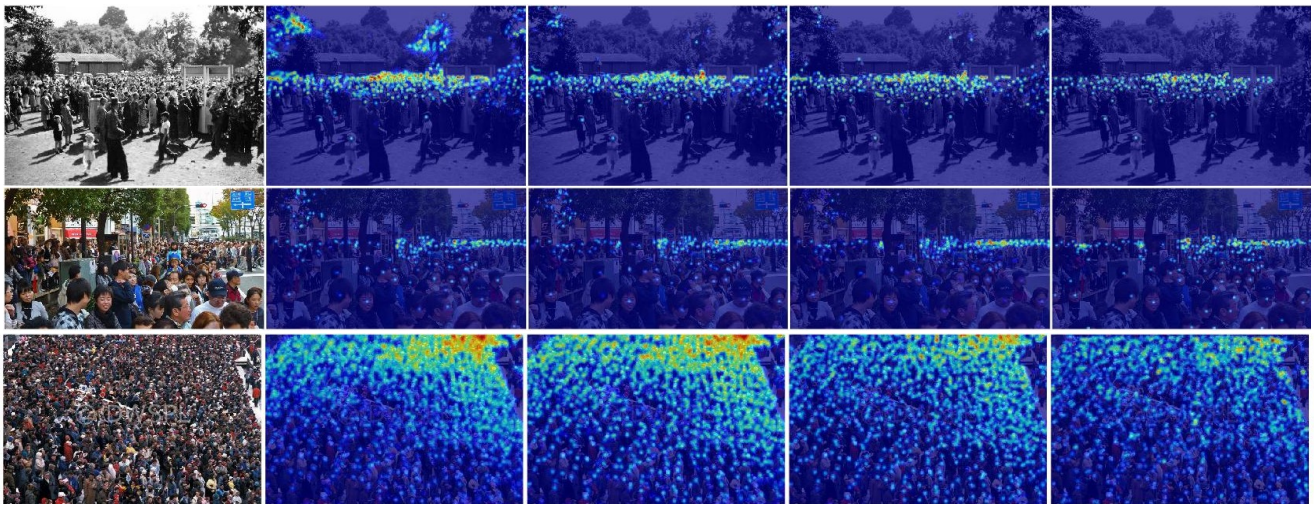
(i) *Baseline*: VGG16 network with $conv6$ at the end (Fig. 1(a)),

(ii) *Baseline + fuse-a*: Baseline network with multi-scale feature fusion using feature addition (Fig. 1(b)),

(iii) *Baseline + fuse-c*: Baseline network with multi-scale feature fusion using feature concatenation (Fig. 1(b)),

(iv) *Baseline + BT + fuse-c*: Baseline network with bottom-top multi-scale feature fusion using feature concatenation (Fig. 1(c)),

(v) *Baseline + TB + fuse-c*: Baseline network with



(a) (b) (c) (d) (e)

Figure 6. Ablation study results: (a) Input, (b) Simple feature concatenation (experiment-ii), (c) Bottom-top and top-bottom fusion (experiment - vi), (d) MBTTF (experiment - viii), (e) Ground-truth density map.

Table 1. Ablation study results.

Dataset	Shanghaitech-A[74]		UCF-QNRF[19]	
Method	MAE	MSE	MAE	MSE
Baseline (Fig. 1a)	78.3	126.6	150.2	220.1
Baseline + fuse-a (Fig. 1b)	73.6	118.4	140.3	210.8
Baseline + fuse-c (Fig. 1b)	73.4	115.6	135.2	200.2
Baseline + BT + fuse-c (Fig. 1c)	68.1	122.2	114.1	185.2
Baseline + TB + fuse-c (Fig. 1d)	70.2	118.5	120.1	188.1
Baseline + BTTB + fuse-c (Fig. 1e)	66.9	112.2	115.4	174.5
Baseline + MBTTB + fuse-c (Fig. 1f)	63.2	108.5	105.5	169.5
Baseline + MBTTB + SCFB-NS (Fig. 2)	62.5	105.1	102.1	168.1
Baseline + MBTTB + SCFB (Fig. 2)	60.2	94.1	97.5	165.2

top-bottom multi-scale feature fusion using feature concatenation (Fig. 1(d)),
(vi) *Baseline + BTTB + fuse-c*: Baseline network with bottom-top and top-bottom multi-scale feature fusion using feature concatenation (Fig. 1(e)),
(vii) *Baseline + MBTTB + fuse-c*: Baseline network with multi-level bottom-top and top-bottom multi-scale feature fusion using feature concatenation (Fig. 1(f)),
(viii) *Baseline + MBTTB + SCFB-NS*: Baseline network with multi-level bottom-top and top-bottom multi-scale feature fusion using SCFB, without using scale-aware supervision (Fig. 2)
(ix) *Baseline + MBTTB + SCFB*: Baseline network with multi-level bottom-top and top-bottom multi-scale feature fusion using SCFB (Fig. 2)

The quantitative results of the ablation study are shown in Table 1. As it can be observed, simple fusion scheme of addition/concatenation (experiments (i) and (ii)) of multi-scale features at the end, does not yield significant improvements as compared to the baseline network. This is due to the reason that in case of feature fusion at the end, the supervision directly affects the initial conv layers in the main branch, which may not be necessarily optimal.

However, when the features are fused in either bottom-top/top-bottom fashion, the results improve considerably, when compared to the baseline. Since this kind of fusion sequentially propagates the information in a particular direction, the initial conv layers do not get affected directly. The bottom-top and top-bottom (experiment (vi)) further improves the performance. The multi-level bottom-top and top-bottom configuration, in which an additional level of bottom-top and top-bottom fusion path is added (experiment-vii), reduces the count error further, signifying the importance of the multi-level fusion paths.

Next, we replace the fusion blocks in experiment-vii with the SCFB blocks, which amounts to the proposed method as shown in Fig. 2 (experiment viii). However, the SCFB blocks are not supervised by the scale-aware ground-truths. The use of these blocks enables the network to propagate relevant and complementary features along the fusion paths, thus leading to improved performance. Finally, we provide scale-aware ground-truth as supervision signal to the SCFB blocks (experiment - ix), which results in further improvements as compared to without scale-aware supervision.

Fig. 6 shows qualitative results for different fusion configurations. Due to space constraints and also to explain better, we show the results of experiments (iii) *Baseline + fuse-c*, (vi) *Baseline + BTTB + fuse-c*, (ix) *Baseline + MBTTB + SCFB* only. It can be observed from Fig. 6*(b)*, that simple concatenation of feature maps results in lot of background noise and loss of details in the final predicted density map, indicating that such an approach is not effective. The bottom-top and top-bottom approach, shown in Fig. 6*(c)* results in the refined density maps, however, they still contain some amount of noise and loss of details. Lastly, the results of experiment (ix) as shown in Fig. 6*(d)*

which have more details where necessary with much lesser background clutter as compared to earlier configurations.

5.3. Comparison with recent methods

In this section, we present the results of the proposed method and compare them with several recent approaches on the three different datasets described in Section 5.1.

Comparison of results the ShanghaiTech and UCF_CROWD_50 datasets are presented in Table 2 and 3 respectively. The proposed method achieves the best results among all the existing methods on the ShanghaiTech Part A dataset and the UCF_CROWD_50 dataset. On the ShanghaiTech B dataset and UCF_CROWD_50dataset, our method achieves a close 2^{nd} position, only behind CAN [33].

Table 2. Comparison of results on ShanghaiTech [74].

	Part A		Part B	
Method	MAE	MSE	MAE	MSE
Switching-CNN [48] (CVPR-17)	90.4	135.0	21.6	33.4
TDF-CNN [47] (AAAI-18)	97.5	145.1	20.7	32.8
CP-CNN [56] (ICCV-17)	73.6	106.4	20.1	30.1
IG-CNN [3] (CVPR-18)	72.5	118.2	13.6	21.1
Liu *et al*. [34] (CVPR-18)	73.6	112.0	13.7	21.4
CSRNet [28] (CVPR-18)	68.2	115.0	10.6	16.0
SA-Net [7] (ECCV-18)	67.0	104.5	8.4	13.6
ic-CNN [43] (ECCV-18)	69.8	117.3	10.7	16.0
ADCrowdNet [31] (CVPR-19)	63.2	98.9	8.2	15.7
RReg [61] (CVPR-19)	63.1	96.2	8.7	13.5
CAN [33] (CVPR-19)	**61.3**	100.0	**7.8**	**12.2**
Jian *et al*. [20] (CVPR-19)	64.2	109.1	8.2	**12.8**
HA-CCN [58] (TIP-19)	62.9	**94.9**	8.1	13.4
MBTTBF-SCFB (proposed)	**60.2**	**94.1**	**8.0**	15.5

Results on the recently released large-scale UCF-QNRF [19] dataset are shown in Table 4. We compare our results with several recent approaches. The proposed achieves the best results as compared to other recent methods on this complex dataset, thus demonstrating the significance of the proposed multi-level fusion method.

Qualitative results for sample images from the ShanghaiTech dataset are presented in Fig. 7.

6. Conclusion

We presented a multi-level bottom-top and top-bottom fusion scheme for overcoming the issues of scale variation that adversely affects crowd counting in congested scenes. The proposed method first extracts a set of scale-complementary features from adjacent layers before propagating them hierarchically in bottom-top and top-bottom fashion. This results in a more effective fusion of features from multiple layers of the backbone network. The effectiveness of the proposed fusion scheme is further enhanced by using ground-truth density maps that are created in a principled way by combining information from the image and location annotations in the dataset. In comparison to existing fusion schemes and state-of-the-art counting methods, the proposed approach is able to achieve significant improvements when evaluated on three popular crowd counting datasets.

Table 3. Comparison of results on UCF_CROWD_50[18].

	UCF_CROWD_50	
Method	MAE	MSE
Switching-CNN [48] (CVPR-17)	318.1	439.2
TDF-CNN [47] (AAAI-18)	354.7	491.4
CP-CNN [56] (ICCV-17)	295.8	320.9
IG-CNN [3] (CVPR-18)	291.4	349.4
D-ConvNet [51] (CVPR-18)	288.4	404.7
Liu *et al*. [34] (CVPR-18)	289.6	408.0
CSRNet [28] (CVPR-18)	266.1	397.5
ic-CNN [43] (ECCV-18)	260.9	365.5
SA-Net-patch [7] (ECCV-18)	258.5	334.9
ADCrowdNet [31] (CVPR-19)	266.4	358.0
CAN [33] (CVPR-19)	**212.2**	**243.7**
Jian *et al*. [20] (CVPR-19)	249.9	354.5
HA-CCN [58] (TIP-19)	256.2	348.4
MBTTBF-SCFB (ours)	**233.1**	**300.9**

Table 4. Comparison of results on the UCF-QNRF datastet [19].

Method	MAE	MSE
CMTL [55] (AVSS-17)	252.0	514.0
MCNN [74] (CVPR-16)	277.0	426.0
Switching-CNN [48] (CVPR-17)	228.0	445.0
Idrees *et al*. [19] (ECCV-18)	132.0	191.0
Jian *et al*. [20] (CVPR-19)	113.0	188.0
CAN [33] (CVPR-19)	107.0	183.0
HA-CCN [58] (TIP-19)	118.1	180.4
MBTTBF-SCFB (ours)	**97.5**	**165.2**

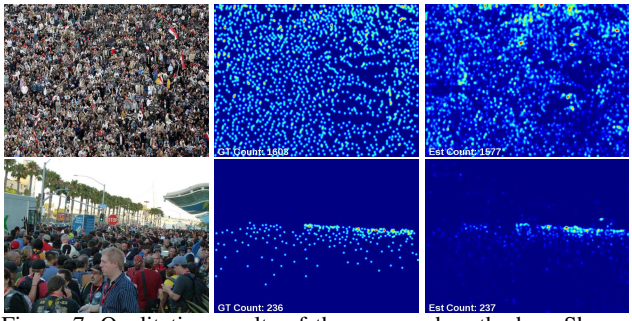


Figure 7. Qualitative results of the proposed method on ShanghaiTech [74] *First column:* Input. *Second column:* Ground truth *Third column:* Predicted density map.

Acknowledgment

This work was supported by the NSF grant 1922840.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, Sabine Süsstrunk, et al. Slic superpixels. *Ecole Polytechnique Fédéral de Lausssanne (EPFL), Tech. Rep*, 149300:155–162, 2010. 6

[2] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *European Conference on Computer Vision*, pages 483–498. Springer, 2016. 3

[3] 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. 1, 3, 8

[4] Serge Beucher et al. The watershed transformation applied to image segmentation. *SCANNING MICROSCOPY-SUPPLEMENT-*, pages 299–299, 1992. 6

[5] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 640–644. ACM, 2016. 3

[6] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision*, pages 354–370. Springer, 2016. 1

[7] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *European Conference on Computer Vision*, pages 757–773. Springer, 2018. 1, 3, 8

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

[9] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *European Conference on Computer Vision*, 2012. 3

[10] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018. 1

[11] Geoffrey French, Mark Fisher, Michal Mackiewicz, and Coby Needle. Convolutional neural networks for counting fish in fisheries surveillance video. In *British Machine Vision Conference Workshop*. BMVA Press, 2015. 1

[12] Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision*, pages 519–534. Springer, 2016. 2

[13] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015. 1

[14] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 1

[15] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[16] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 951–959, 2017. 1

[17] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013. 1, 3, 6

[18] Haroon Idrees, Khurram Soomro, and Mubarak Shah. Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):1986–1998, 2015. 8

[19] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *European Conference on Computer Vision*, pages 544–559. Springer, 2018. 3, 6, 7, 8

[20] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder network. *arXiv preprint arXiv:1903.00853*, 2019. 3, 8

[21] Di Kang, Zheng Ma, and Antoni B Chan. Beyond counting: Comparisons of density maps for crowd analysis tasks-counting, detection, and tracking. *arXiv preprint arXiv:1705.10118*, 2017. 1

[22] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, pages 1324–1332, 2010. 1, 3

[23] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE transactions on Multimedia*, 20(4):985–996, 2018. 1

[24] 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 *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008. 3

[25] Stan Z Li. Markov random field models in computer vision. In *European conference on computer vision*, pages 361–370. Springer, 1994. 3

[26] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(3):367–386, 2015. 1

[27] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2014. 1

[28] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 1091–1100, 2018. 8

[29] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 1

[30] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. 1

[31] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. *arXiv preprint arXiv:1811.11968*, 2018. 3, 8

[32] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 2

[33] 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, 2019. 3, 8

[34] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3, 8

[35] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. Tasselnet: Counting maize tassels in the wild via local counts regression network. *Plant Methods*, 13(1):79, 2017. 1

[36] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, volume 249, page 250, 2010. 1

[37] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4875–4884, 2017. 1

[38] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision*, pages 615–629. Springer, 2016. 1, 3

[39] 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, 2015. 3

[40] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European Conference on Computer Vision*, pages 75–91. Springer, 2016. 2

[41] Vasili Ramanishka, Abir Das, Jianming Zhang, and Kate Saenko. Top-down visual saliency guided by captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7206–7215, 2017. 1

[42] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2017. 1

[43] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *European Conference on Computer Vision*, pages 278–293. Springer, 2018. 1, 3, 8

[44] Mikel Rodriguez, Ivan Laptev, Josef Sivic, and Jean-Yves Audibert. Density-aware person detection and tracking in crowds. In *2011 International Conference on Computer Vision*, pages 2423–2430. IEEE, 2011. 1

[45] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *European conference on computer vision*, pages 186–201. Springer, 2016. 1

[46] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications, 2009. DICTA'09.*, pages 81–88. IEEE, 2009. 3

[47] Deepak Babu Sam and R Venkatesh Babu. Top-down feedback for crowd counting convolutional neural network. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 8

[48] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 3, 8

[49] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3

[50] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7279–7288, 2019. 3

[51] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3, 8

[52] Abhinav Shrivastava, Rahul Sukthankar, Jitendra Malik, and Abhinav Gupta. Beyond skip connections: Top-down modulation for object detection. *arXiv preprint arXiv:1612.06851*, 2016. 2

[53] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 3

[54] Vishwanath Sindagi and Vishal Patel. Inverse attention guided deep crowd counting network. *arXiv preprint*, 2019. 3

[55] Vishwanath A. Sindagi and Vishal M. Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 IEEE International Conference on*. IEEE, 2017. 8

[56] Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 3, 6, 8

[57] 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*, 2017. 3

[58] Vishwanath A Sindagi and Vishal M Patel. Ha-ccn: Hi-

erarchical attention-based crowd counting network. *arXiv preprint arXiv:1907.10255*, 2019. 2, 3, 8

[59] Evgeny Toropov, Liangyan Gui, Shanghang Zhang, Satwik Kottur, and José MF Moura. Traffic flow from a low frame rate city camera. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 3802–3806. IEEE, 2015. 1

[60] Elad Walach and Lior Wolf. Learning to count with cnn boosting. In *European Conference on Computer Vision*, pages 660–676. Springer, 2016. 3

[61] Jia Wan, Wenhan Luo, Baoyuan Wu, Antoni B Chan, and Wei Liu. Residual regression with semantic prior for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4036–4045, 2019. 3, 8

[62] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1299–1302. ACM, 2015. 3

[63] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. *arXiv preprint arXiv:1903.03303*, 2019. 3

[64] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *IEEE International Conference on Computer Vision*. IEEE, 2017. 1

[65] Bolei Xu and Guoping Qiu. Crowd density estimation based on rich features and random projection forest. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016. 3

[66] Fan Yang, Xin Li, Hong Cheng, Yuxiao Guo, Leiting Chen, and Jianping Li. Multi-scale bidirectional fcn for object skeleton extraction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[67] Rajeev Yasarla and Vishal M. Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[68] Beibei Zhan, Dorothy N Monekosso, Paolo Remagnino, Sergio A Velastin, and Li-Qun Xu. Crowd analysis: a survey. *Machine Vision and Applications*, 19(5-6):345–357, 2008. 1

[69] 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, 2015. 1, 3

[70] Qi Zhang and Antoni B Chan. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8297–8306, 2019. 3

[71] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. Understanding traffic density from large-scale web camera data. In *IEEE Computer Vision and Pattern Recognition*. IEEE, 2017. 1

[72] Shanghang Zhang, Guanhang Wu, João P. Costeira, and José M. F. Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *IEEE International Conference on Computer Vision*. IEEE, 2017. 1

[73] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 714–722, 2018. 1

[74] 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, 2016. 1, 3, 6, 7, 8

[75] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12736–12745, 2019. 3

[76] Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3080–3088, 2018. 2

[77] Feng Zhu, Xiaogang Wang, and Nenghai Yu. Crowd tracking with dynamic evolution of group structures. In *European Conference on Computer Vision*, pages 139–154. Springer, 2014. 1

[78] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3