# Evaluating Crowd Density Estimators Via Their Uncertainty Bounds

Jennifer Vandoni, Emanuel Aldea, Sylvie Le Hégarat-Mascle

## To cite this version:

Jennifer Vandoni, Emanuel Aldea, Sylvie Le Hégarat-Mascle. Evaluating Crowd Density Estimators Via Their Uncertainty Bounds. 2019 IEEE International Conference on Image Processing (ICIP), Sep 2019, Taipei, Taiwan. pp.4579-4583, 10.1109/ICIP.2019.8803522 . hal-02338016

**HAL Id: hal-02338016**

**https://hal.science/hal-02338016**

Submitted on 29 Oct 2019

# EVALUATING CROWD DENSITY ESTIMATORS VIA THEIR UNCERTAINTY BOUNDS

*Jennifer Vandoni, Emanuel Aldea and Sylvie Le Hégarat-Mascle*

SATIE - CNRS UMR 8029, Paris-Sud University, Paris-Saclay University, France

## ABSTRACT

In this work, we use the Belief Function Theory which extends the probabilistic framework in order to provide uncertainty bounds to different categories of crowd density estimators. Our method allows us to compare the multi-scale performance of the estimators, and also to characterize their reliability for crowd monitoring applications requiring varying degrees of prudence.

***Index Terms***— density estimation, crowd counting, multi-scale evaluation, uncertainty bounds

## 1. INTRODUCTION

Understanding crowd systems and predicting their evolution is of paramount importance when considering the world population growth and urbanization rates. One of the reasons for which the urban infrastructure sector has not fully taken advantage of vast available video data is the difficulty to extract accurately microscopic and macroscopic observations in high-density conditions. Although it does not require accurate target localization, density estimation inside crowds is still a challenging problem, due to phenomena such as strong occlusion and visual homogeneity. However, deep learning advancements significantly improved the state-of-the-art performance (see [1] for a comprehensive survey). Recent methods are mostly based on the estimation of a density map whose integral over a region provides the number of people within it, in such a way to incorporate spatial information directly into the learning process (e.g. MCNN [2], Cascaded-MTL [3], Hydra CNN [4], CSRNet [5]). In the vast majority of the proposed works, the estimator evaluation is performed at image scale, with error metrics such as MAE or MSE [1]. However, from the point of view of modelling the crowd as a dynamical system, accurate *local* densities are required in order to characterize wave-like propagation phenomena. The error metrics above are related to large scale statistics, which do not apply to small scales due to compensation between overestimating and underestimating the density in different areas. An additional limitation of current density estimators is the absence of an uncertainty range provided along with the scalar density. Ranges on the pedestrian count are greatly needed, as the trade-off between safety concerns and optimal use of infrastructure capacity promotes different levels of congestion in different contexts. In statistical learning theory, the uncertainty of classification or regression processes has been studied based on how models are affected by the skewed distribution and noise of training data, by inaccurate training data labeling or by the regularization policy. SVM output uncertainty has been studied [6] due to algorithm's convenient generalization ability coupled to the simple underlying principle. In [7] a new strategy was introduced for addressing the epistemic uncertainty estimation in Deep Neural Networks by approximating the probabilistic output distribution using dropout during inference.

In this line of research, we propose a generic approach for evaluating the uncertainty of the output of a crowd density estimator. As a second contribution, we apply the proposed evaluation on a multiscale domain derived from the image lattice, which allows us to characterize the estimator performance locally as well. We show that we are able to compare different learning algorithms across the scale space and to provide density estimations with bounded uncertainties.

## 2. EVIDENTIAL CNN-ENSEMBLE

**FE+LFE network.** Following recent advancements on density estimation [5], we propose a fully convolutional network which makes use of dilated convolutions instead of pooling layers, in order to preserve the output resolution in presence of small targets. However, as highlighted in [8], aggressively increasing dilation factors through the network layers is detrimental in aggregating local features. By taking inspiration from this latter work we propose a network which is composed of two parts, i.e. a front end (FE) module with increasing dilation factors to consider larger context around small objects, and a local feature extractor (LFE) module with decreasing dilation factors to enforce the spatial consistency of the output by gathering spatial information. Moreover, unlike [8], we add batch normalization before ReLU activation functions for faster convergence. The structure of the proposed FE+LFE network is detailed in Table 1. The number of filters per layer is kept small to avoid overfitting since we intend to be able to train the network with relatively small datasets. Note that we employ a ReLU activation function also after the last layer. This has the effect of a zero-threshold; nevertheless, it has beneficial effects on backpropagation with

This work was funded by French ANR grant ANR-15-CE39-0005.

|  | Layers - part 1 |  | Layers - part 2 |
|---|---|---|---|
| FE | Conv $3 \times 3$, $F = 16$, $D = 1$<br>Conv $3 \times 3$, $F = 32$, $D = 1$<br>Conv $3 \times 3$, $F = 32$, $D = 2$<br>Conv $3 \times 3$, $F = 64$, $D = 2$ | LFE | Conv $3 \times 3$, $F = 64$, $D = 2$<br>Conv $3 \times 3$, $F = 64$, $D = 2$<br>Conv $3 \times 3$, $F = 64$, $D = 1$<br>Conv $3 \times 3$, $F = 64$, $D = 1$ |
|  | Conv $3 \times 3$, $F = 64$, $D = 3$ |  | Conv $1 \times 1$, $F = 1$, $D = 1$ |

**Table 1**: Architecture of the FE+LFE network. $F$ is the number of filters and $D$ is the dilation factor of dilated convolutions. Each convolutional layer is followed by batch normalization (except for the last one) and ReLU activation function.

respect to a simple post-processing thresholding. The local density estimation is therefore enhanced, since the network loses its tendency to add noise to compensate between low and high values.

Finally, a L2 loss function is used between the estimated density map and the ground-truth derived by placing a Gaussian on each head center as in [2]. Since we know the geometry of the scene, we apply perspective correction as in [9] instead of geometry-adaptive kernels.

**Building a CNN-ensemble.** Recently ensemble techniques have been successfully exploited by the deep learning community since they allows for more robust predictions as well as for a measure of predictive uncertainty, i.e. the confidence of the network with respect to its prediction (which in our case represents the likelihood of head presence). In the context of Bayesian Neural Networks (BNNs), the authors of [7] developed a new theoretical framework called *MC-dropout* casting dropout [10] as approximate Bayesian inference in Gaussian processes. This method overcomes the major limitations of BNNs that generally require prohibitive computational costs [11, 12, 13]. In [7] instead, after training the network, Monte Carlo (MC) methods are used at inference time to draw samples from a Bernoulli distribution across the network weights, by performing *T* stochastic forward passes through the network with dropout. The ensemble is thus composed by *T* different realizations given by dropping out different units of the network at each forward pass. Another ensemble approach (although non-Bayesian) has been recently proposed in [14], where a *deep ensemble* is derived by training the same network on the same data but with different random weight initializations. Compared to MC-dropout, this method has nonetheless the immediate drawback of requiring multiple training of the network.

In this work, we derive a CNN-ensemble relying on MC-dropout. We train the network once and then we sample the posterior distribution over the weights using dropout at inference time, obtaining *T* different realization maps $\hat{\mathscr{M}}_1, \ldots, \hat{\mathscr{M}}_T$, outputs of different dropout-perturbed versions of the original network. Classically, the mean map $\mathscr{M}_\mu$, given by the mean value evaluated independently for each pixel, would be interpreted as the final prediction map, while the standard deviation map $\mathscr{M}_\sigma$ would be interpreted as an estimate of the predictive uncertainty. However, we propose to work in the Belief Function (BF) framework [15, 16], that we consider more suited to model the specific imprecision of each different realization obtained with dropout, allowing us to derive the uncertainty bounds for density estimation.

**Modeling imprecision with BFT.** To handle both the uncertainty provided by the classification and the related imprecision that may exist due to the specific classifier and/or data used in the training process, Belief Function Theory (BFT), also called evidential theory, is designed to handle a larger hypothesis set than the probabilistic one. Denoting by $\Theta$ the discernment frame, i.e. the set of mutually exclusive hypotheses of cardinality $|\Theta|$, belief functions are defined on the powerset $2^\Theta$. In our setting, denoting by $H$ and $\overline{H}$ the two mutually exclusive (*singleton*) hypotheses *"Head"* and *"Not Head"*, the discernment frame is $\Theta = \{H, \overline{H}\}$ while $2^\Theta = \{\emptyset, H, \overline{H}, \{H, \overline{H}\}\}$. Classically, the *mass* function $m$ is the *Basic Belief Assignment* (BBA) that satisfies $\forall A \in 2^\Theta$, $m(A) \in [0, 1]$, $\sum_{A \in 2^\Theta} m(A) = 1$. The hypotheses associated to non-null mass functions are called *focal elements*. BBAs that have only singleton hypotheses as focal elements are called Bayesian BBAs.

Now, we want to consider the imprecision that possibly arises performing inference on unknown images with a model learned by a neural network, by modelling the pixel-wise classification outputs as BBAs. We therefore exploit the the CNN-ensemble composed by the *T* realizations obtained with MC-dropout. Firstly, we derive Bayesian BBA maps $\mathcal{M}_1^{\mathcal{B}}, \ldots, \mathcal{M}_T^{\mathcal{B}}$, where a BBA is associated to each pixel $\mathbf{x}$ of every realization, so that we obtain $T$ maps of BBAs $\{m_{\mathbf{x},t}^{\mathcal{B}}\}_{\mathbf{x} \in \mathcal{P}}$, where $\mathcal{P}$ is the pixel domain and $t \in \{1, \ldots, T\}$. These Bayesian BBA maps are 4-layer images where each layer corresponds to the mass value of any hypothesis in $\{\emptyset, H, \overline{H}, \Theta\}$ respectively. For example, $\mathcal{M}_t^{\mathcal{B}}(A)$ corresponds to the layer image associated to hypothesis $A$ for the realization (source) $t$. In this preliminary Bayesian BBA allocation, layer images corresponding to non-singleton hypotheses are null by definition, whereas for each source $t$, with $t = 1, \ldots, T$: $\mathcal{M}_t^{\mathcal{B}}(H) = \hat{\mathscr{M}}_t$, and $\mathcal{M}_t^{\mathcal{B}}(\overline{H}) = 1 - \hat{\mathscr{M}}_t$.

In order to account for the reliability of the pixel-wise prediction given by every source, we perform a pixel-wise tailored discounting, namely a *generalization* of each BBA on the basis of its reliability [15]. To evaluate this latter, noting that the median has been shown to be a more robust estimator than the average in presence of outliers, for each source $t$ we compute a discounting coefficient map $\Gamma_t \ : \ \{\gamma_{\mathbf{x},t}\}_{\mathbf{x} \in \mathcal{P}}$ such that a different coefficient $\gamma_{\mathbf{x},t}$ is associated to every pixel of each source,

$$\Gamma_t = \alpha \left( 1 - \left( \left| \hat{\mathscr{M}}_t - \text{median} \left( \left\{ \hat{\mathscr{M}} \right\}_1^T \right) \right| \right) \right). \qquad (1)$$

In this way, we discount more pixels whose value is more distant to the median value among the $T$ realizations, since they are supposed to be less representative (even possibly outliers). The $\alpha$ parameter is a scaling factor which allows us to control the amount of discounting. Applying the proposed discounting, we derive the following BBAs map for every source $t$:

$\forall A \in \{H, \overline{H}\}$,

$$\left\{ \begin{array}{lcl} \mathcal{M}_t(\emptyset) & = & \{0\}_{\mathbf{x}\in\mathcal{P}}\,, \\ \mathcal{M}_t(A) & = & \Gamma_t \star \mathcal{M}_t^{\mathcal{B}}(A), \\ \mathcal{M}_t(\Theta) & = & \{1\}_{\mathbf{x}\in\mathcal{P}} - \mathcal{M}_t(H) - \mathcal{M}_t(\overline{H}), \end{array} \right. \tag{2}$$

where $M_1 \star M_2$ represents the Hadamard product between matrices $M_1$ and $M_2$.

To combine the $T$ different maps to obtain a single output map $\mathcal{M}$ with BBAs associated to each pixel $\mathbf{x}$, i.e. $\{m_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{P}}$, we use the conjunctive combination rule [16]. In our case where $|\Theta| = 2$, the analytic result may be easily derived: $\forall A \in \{H, \overline{H}\}$,

$$\left\{ \begin{array}{lcl} m_{\mathbf{x}}(A) & = & \sum\limits_{\substack{(B_1,\ldots,B_T)\in\{A,\Theta\}^T, \\ \exists t\in[1,T] s.t. B_t = A}} \prod_{t=1}^{T} m_{\mathbf{x},t}(B_t)\,, \\ m_{\mathbf{x}}(\Theta) & = & \prod_{t=1}^{T} m_{\mathbf{x},t}(\Theta)\,, \\ m_{\mathbf{x}}(\emptyset) & = & 1 - m_{\mathbf{x}}(H) - m_{\mathbf{x}}\left(\overline{H}\right) - m_{\mathbf{x}}(\Theta)\,. \end{array} \right. \tag{3}$$

The result is thus a four-layer map $\mathcal{M}$ of BBAs $m_{\mathbf{x}}$, that can be used to derive evidential measures of uncertainty about the network prediction. To this extent, we can obtain the ignorance map as $\mathcal{M}(\Theta)$, that represents the remaining ignorance which has been decreased by the combination but not completely solved, indicating a lack of sufficient information during training to perform a reliable prediction. Likewise, $\mathcal{M}(\emptyset)$ is often interpreted as a conflict map [17], and presents higher values for pixels whose prediction completely disagrees through the various realizations.

Finally, in every pixel $\mathbf{x}$ the decision is taken from $m_{\mathbf{x}}$. Pignistic probability [16] may be used to give a probabilistic interpretation to the BBAs. Since in our setting $|\Theta| = 2$, $\forall A \in \{H, \overline{H}\}$, $BetP_{\mathbf{x}}(A) = \frac{1}{1-m_{\mathbf{x}}(\emptyset)}\left(m_{\mathbf{x}}(A) + \frac{m_{\mathbf{x}}(\Theta)}{2}\right)$. This allows us to assign a $BetP_{\mathbf{x}}(H)$ value to the resulting BBA associated to each pixel $\mathbf{x}$ that will be differently normalized on the basis of its conflict value, $m_{\mathbf{x}}(\emptyset)$.

Then, other functions are in a one-to-one relationship with $m_{\mathbf{x}}$, and can be used either for decision or for some computations, namely the *Plausibility* (*Pl*) and the *Belief* (*Bel*) functions. In this particular setting where $|\Theta| = 2$ applying a normalization to the BBAs (so that $m_{\mathbf{x}}(\emptyset) = 0$), they are defined as: $Bel_{\mathbf{x}}(A) = \frac{1}{1-m_{\mathbf{x}}(\emptyset)}\left(m_{\mathbf{x}}(A)\right)$, and $Pl_{\mathbf{x}}(A) = \frac{1}{1-m_{\mathbf{x}}(\emptyset)}\left(m_{\mathbf{x}}(A) + m_{\mathbf{x}}(\Theta)\right)$. These functions may also be interpreted as upper and lower probabilities respectively [15] and they check the duality property: $\forall A \in 2^{\Theta}, Pl_{\mathbf{x}}(A) = 1 - Bel_{\mathbf{x}}(\overline{A})$ (where $\overline{A}$ represents the complement of $A$ with respect to $\Theta$).

## 3. DENSITY UNCERTAINTY FOR BOUNDING PEDESTRIAN COUNTS

In this work we propose a multiscale evaluation strategy which computes for each considered scale $\mathcal{S}$ indicators based on all squared subdomains $S \in \mathcal{S}_i$. These indicators use the derived upper and lower density bounds $\underline{s}(S)$, $\overline{s}(S)$: $\underline{s}(S) = w\sum_{\mathbf{x}\in S} Bel_{\mathbf{x}}(H)$ and $\overline{s}(S) = w\sum_{\mathbf{x}\in S} Pl_{\mathbf{x}}(H)$. The factor $w$ relating the numerical output to the actual pedestrian count is 1 for networks trained on actual density maps, but in the general case it may be determined as in [18] on a validation set consisting of $BetP(H)$ maps. We then calculate for $\mathcal{S}_i$ the *prediction error probability* (PEP) as:

$$PEP_i = \Big|\{S \in \mathcal{S}_i | g(S) \in [\underline{s}(S), \overline{s}(S)]\}\Big| / |\mathcal{S}_i|, \tag{4}$$

and the *relative imprecision* (RI) interval as:

$$RI_i = \Big(\sum_{S\in\mathcal{S}_i} (\overline{s}(S) - \underline{s}(S))/g(S)\Big) / |\mathcal{S}_i|, \tag{5}$$

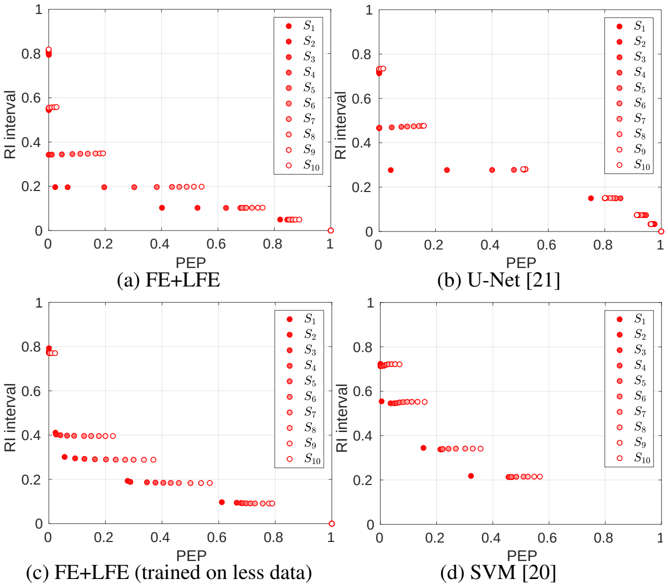
where $g(S)$ is the ground-truth count over $S$. In our work, we take $\mathcal{S}_1$ as the set of the largest possible squares which fit the image space, and then we use a scale factor $\delta$ to reduce the square side for subsequent scales.

The RI criterion highlights the size of the imprecision interval around the estimated count, while the PEP criterion indicates the error rate of the prediction, namely whether the ground-truth count for the considered region is outside the estimated interval. Thus, a two-axis plot presenting the evolution of RI vs. PEP across multiple scales and for different estimators allows one to compare them and to select an operating point with an explicit uncertainty tied to a desired error rate. In order to compute the values required by Eqs. (4) and (5), the process may be accelerated significantly by using the Integral Histogram [19] trick, given that the most intensive task is to compute sums over rectangular supports defined in the bounded image space.

## 4. EXPERIMENTAL RESULTS

We validated our proposed approach on high-density crowd images acquired at Makkah during Hajj [20]. Besides evaluating the proposed FE+LFE network, we compared it to U-Net [21], originally introduced for medical image segmentation and very effective even on relatively small training datasets as in our case (35 crowd images). The two networks are trained by using Adam stochastic optimizer with a learning rate of $7\times10^{-3}$ (FE+LFE) and of $10^{-2}$ (U-Net). Additionally, we perform data augmentation and early stopping in order to limit overfitting. A CNN-ensemble of size $T = 10$ is then obtained by applying dropout at inference time in the central layers as in [22] with probability $p_{\text{drop}} = 0.5$.

Figures 1a and 1b show the results when applying the proposed uncertainty bound evaluation for the FE+LFE and U-Net networks respectively. Ideally, an estimator should predict with a high confidence (low PEP) that the estimated count is within a small RI interval. One may increase the size of the RI interval by decreasing the $\alpha$ parameter in Eq. (1), in order to obtain better prediction accuracy (at the expense of



(a) FE+LFE

(b) U-Net [21]

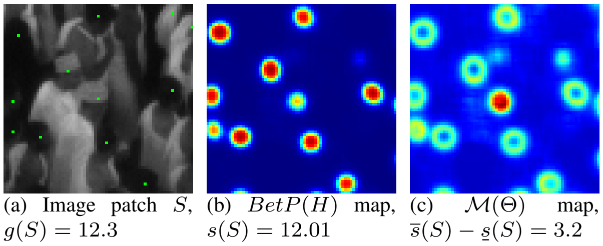
(c) FE+LFE (trained on less data)

(d) SVM [20]

**Fig. 1**: Density estimator evaluation with the proposed RI vs. PEP plot at multiple scales and with different discounting amounts. Each horizontal cluster corresponds to a different discounting factor.

a larger RI). We tested different discounting factors, corresponding to horizontally aligned clusters of dots. For each cluster, each dot depicts the performance obtained at a different scale, with a scale factor $\delta = 1.1$, $\mathcal{S}_1$ being the largest scale. Both networks perform better at larger scales, due to error compensation. The proposed FE+LFE network outperforms U-Net, showing the importance of preserving spatial information without pooling operations in presence of small targets, while increasing at the same time the contextual information with dilations.

To stress the independence of the proposed evaluation approach with respect to the classifier used, Fig. 1d shows the results of the density estimation obtained with SVM using active learning (AL) as in [20], where an SVM-ensemble is built iteratively by training SVMs with different descriptors on selected informative samples. The imprecision derives both from possible errors in the calibration procedure to obtain probability estimates out of SVM scores, and from the score heterogeneity in the image space. Moreover, Fig. 1c shows the results obtained training the proposed FE+LFE network with a smaller amount of data (i.e. the pool of unlabeled samples $\mathcal{U}$ available for AL in [20]). This allows us to perform two different types of analysis. Firstly, we can perform a fairer comparison between the two classifiers. To this extent, we notice that FE+LFE, even when trained on less data, outperforms the SVM-based approach, especially at larger scales. Nonetheless, the two methods exhibit almost identical performance when considering the smaller scales. Secondly, it is interesting to evaluate the same network trained with different amounts of data. According to Figs. 1a and 1c, we see that a larger training set is beneficial for density estimation especially at larger scales. However, considering smaller scales, the performance gap is consistently reduced, indicating thus a implicit limit in the network capacity (increasing the number of layers and/or filters per layer could help, paying attention to overfitting).



(a) Image patch $S$, $g(S) = 12.3$

(b) $BetP(H)$ map, $s(S) = 12.01$

(c) $\mathcal{M}(\Theta)$ map, $\overline{s}(S) - \underline{s}(S) = 3.2$

**Fig. 2**: Results of the density estimation map (Fig. 2b) along with the estimated uncertainty bounds (Fig. 2c). The input data and the ground truth annotations are shown in Fig. 2a.

Figure 2a shows an image patch with corresponding ground-truth count (obtained after Gaussian smoothing). Figure 2b shows the resulting $BetP(H)$ map which represents the scalar density estimation map, while Fig. 2c shows the imprecision map $\mathcal{M}(\Theta)$ (in our case for pixel $\mathbf{x}$ the imprecision value $Pl_{\mathbf{x}}(H) - Bel_{\mathbf{x}}(H)$ is equal to $m_{\mathbf{x}}(\Theta)$). The values in $\mathcal{M}(\Theta)$ may be interpreted as the predictive uncertainty, and provide a bound for the density estimation itself. For the given region $S$ indeed, by integrating over the $BetP(H)$ map we obtain the estimated number of people within it. Similarly, integrating over the $\mathcal{M}(\Theta)$ map we obtain the imprecision interval $\overline{s}(S) - \underline{s}(S)$. Then, the corresponding RI interval is given by $(\overline{s}(S) - \underline{s}(S))/g(S) = 0.26$, so that we can conclude that in $S$ there are $12.01 \pm 13\%$ heads, i.e. $s(S) \in [10.4, 13.6]$. Moreover, from Fig. 2c we can notice that, in addition to head edges, ignorance is particularly high on heads with lower gradient on the borders and strong clutter, reflecting in a smaller confidence about the prediction. Finally as expected, we underline the desirable effect of ignorance being higher in circularly-shaped areas (e.g. shoulders, or round dark blobs) which are similar to heads, even if they have a low corresponding score.

## 5. CONCLUSION

We proposed a strategy for associating an uncertainty interval to crowd density estimation using BFT. A new evaluation method taking into account the output uncertainty at multiple scales was proposed as well. The results show that our contributions are effective in characterizing the multi-scale performance of different density estimators. Our work opens a promising avenue for crowd safety applications which account for estimation uncertainty during planning and monitoring. Future work will be devoted to applying our evaluation to other widely used density estimation networks such as MCNN or CSRNet across more datasets.

## 6. REFERENCES

[1] 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*, vol. 107, pp. 3–16, 2018.

[2] 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*, 2016, pp. 589–597.

[3] 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 14th IEEE International Conference on*. IEEE, 2017, pp. 1–6.

[4] Daniel Onoro-Rubio and Roberto J López-Sastre, "Towards perspective-free object counting with deep learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 615–629.

[5] 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*, 2018, pp. 1091–1100.

[6] Philippe Xu, Franck Davoine, Hongbin Zha, and Thierry Denoeux, "Evidential calibration of binary svm classifiers," *International Journal of Approximate Reasoning*, vol. 72, pp. 55–70, 2016.

[7] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[8] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1442–1450.

[9] Jennifer Vandoni, Emanuel Aldea, and Sylvie Le Hégarat-Mascle, "Active learning for high-density crowd count regression," in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–6.

[10] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[11] Radford M Neal, *Bayesian learning for neural networks*, vol. 118, Springer Science & Business Media, 2012.

[12] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[13] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra, "Weight uncertainty in neural networks," *arXiv preprint arXiv:1505.05424*, 2015.

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, 2017, pp. 6402–6413.

[15] Glenn Shafer, *A mathematical theory of evidence*, vol. 1, Princeton university press Princeton, 1976.

[16] Philippe Smets and Robert Kennes, "The transferable belief model," *Artificial intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[17] Marie Lachaize, Sylvie Le Hégarat-Mascle, Emanuel Aldea, Aude Maitrot, and Roger Reynaud, "Evidential split-and-merge: Application to object-based image analysis," *International Journal of Approximate Reasoning*, vol. 103, pp. 303–319, 2018.

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

[19] Fatih Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 829–836.

[20] Jennifer Vandoni, Emanuel Aldea, and Sylvie Le Hégarat-Mascle, "Evidential query-by-committee active learning for pedestrian detection in high-density crowds," *International Journal of Approximate Reasoning*, vol. 104, pp. 166–184, 2019.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.