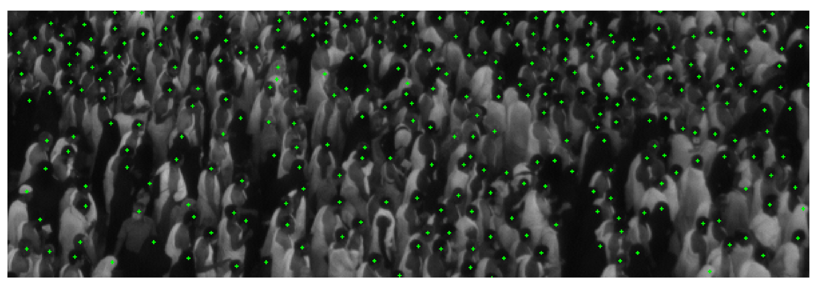
Evaluating Crowd Density Estimators via Their Uncertainty Bounds

Jennifer Vandoni, Emanuel Aldea, Sylvie Le Hégarat-Mascle
SATIE - CNRS UMR 8029, Paris-Sud University, Paris-Saclay University
{jennifer.vandoni, emanuel.aldea, sylvie.le-hegarat}@u-psud.fr

Context and objectives

Context:

- Crowd density estimation is a challenging problem due to phenomena such as strong occlusion and visual homogeneity
- Recent deep methods are mostly based on the estimation of a density map whose integral over a region provides the number of people within it
- The estimator evaluation is performed at image scale: compensation between overestimating and underestimating the density in different areas
- Absence of an uncertainty range provided along with the scalar density



Our objective:

- We use the Belief Function Theory 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
  - Characterize their reliability for crowd monitoring applications requiring varying degrees of prudence

Evidential CNN-ensemble

FE+LFE network:

- Fully convolutional encoder-decoder structure
- Front End (FE) module with increasing dilation factors to consider larger context around small objects
- Local Feature Extractor (LFE) module with decreasing dilation factors to enforce the spatial consistency of the output [Ham+18]
- BatchNorm + ReLU activation functions
- ReLU after the last layer: zero-threshold effect with beneficial effects on backpropagation

	Layers - part 1		Layers - part 2
FE	Conv 3×3, $F=16$, $D=1$	LFE	Conv 3×3, $F=64$, $D=2$
	Conv 3×3, $F=32$, $D=1$		Conv 3×3, $F=64$, $D=2$
	Conv 3×3, $F=32$, $D=2$		Conv 3×3, $F=64$, $D=1$
	Conv 3×3, $F=64$, $D=2$		Conv 3×3, $F=64$, $D=1$
	Conv 3×3, $F=64$, $D=3$		Conv 1×1, $F=1$, $D=1$

Building a CNN-ensemble:

- We derive a CNN-ensemble relying on MC-dropout [GG16], obtaining T different realization maps $\hat{\mathscr{M}}_1,\dots,\hat{\mathscr{M}}_T$
- Traditional methods interpret the mean map $\mathscr{M}_\mu$ as the final prediction map and the standard deviation map $\mathscr{M}_\sigma$ as an estimate of the predictive uncertainty
- We instead rely on Belief Function Theory

Belief Function Theory (BFT):

- BFT extends probabilistic approaches by modeling *imprecision* in addition to *uncertainty*
- Larger hypotheses set: $2^\Theta=\{\emptyset,H,\bar{H},\{H,\bar{H}\}\}$, H=*"Head"* and $\bar{H}$=*"Not Head"*
- Basic Belief Assignment (BBA): function m s.t. $\sum_{A\in 2^\Theta} m(A)=1, \forall A\in 2^\Theta,\ m(A)\in[0,1]$

Modeling imprecision with BFT:

- We exploit the T realizations obtained through MC-dropout
- We associate a BBA map to every realization t, i.e. a 4-layer images where each layer corresponds to the mass value of any hypothesis in $\{\emptyset,H,\bar{H},\Theta\}$
- Bayesian BBA map associated to each realization t, with $t=1,\dots,T$: $\mathcal{M}_t^{\mathcal{B}}(H)=\hat{\mathscr{M}}_t$, and $\mathcal{M}_t^{\mathcal{B}}(\bar{H})=1-\hat{\mathscr{M}}_t$
- Pixel-wise tailored discounting of each BBA on the basis of its reliability:
  - $\forall t$, 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:

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

  - We derive the discounted BBA maps for every source t applying Γ_t
- Conjunctive combination rule to combine the discounted BBA maps into a single output BBA map
  - $\mathcal{M}(\Theta)$: ignorance map (lack of sufficient information during training to perform a reliable prediction)
  - $\mathcal{M}(\emptyset)$: conflict map (higher values for pixels whose prediction completely disagrees through the various realizations)
- Decision through belief functions: $\forall A\in\{H,\bar{H}\}$,
  - Final probabilistic decision: $BetP_{\mathbf{x}}(A)=\frac{1}{1-m_{\mathbf{x}}(\emptyset)}\left(m_{\mathbf{x}}(A)+\frac{m_{\mathbf{x}}(\Theta)}{2}\right)$
  - Belief (lower bound): $Bel_{\mathbf{x}}(A)=\frac{1}{1-m_{\mathbf{x}}(\emptyset)}\left(m_{\mathbf{x}}(A)\right)$
  - Plausibility (upper bound): $Pl_{\mathbf{x}}(A)=\frac{1}{1-m_{\mathbf{x}}(\emptyset)}\left(m_{\mathbf{x}}(A)+m_{\mathbf{x}}(\Theta)\right)$

Density uncertainty for bounding pedestrian count

Multiscale evaluation strategy:

For each considered scale $\mathcal{S}$ we compute indicators based on all squared subdomains $S\in\mathcal{S}_i$, by using the derived upper and lower density bounds $\underline{s}(S),\bar{s}(S)$:

$$\underline{s}(S)=w\sum_{\mathbf{x}\in S}Bel_{\mathbf{x}}(H)\quad\text{and}\quad\bar{s}(S)=w\sum_{\mathbf{x}\in S}Pl_{\mathbf{x}}(H)$$

- *Prediction Error Probability* (PEP):

$$PEP_i=\left|\{S\in\mathcal{S}_i|g(S)\in[\underline{s}(S),\bar{s}(S)]\}\right|/|\mathcal{S}_i|$$

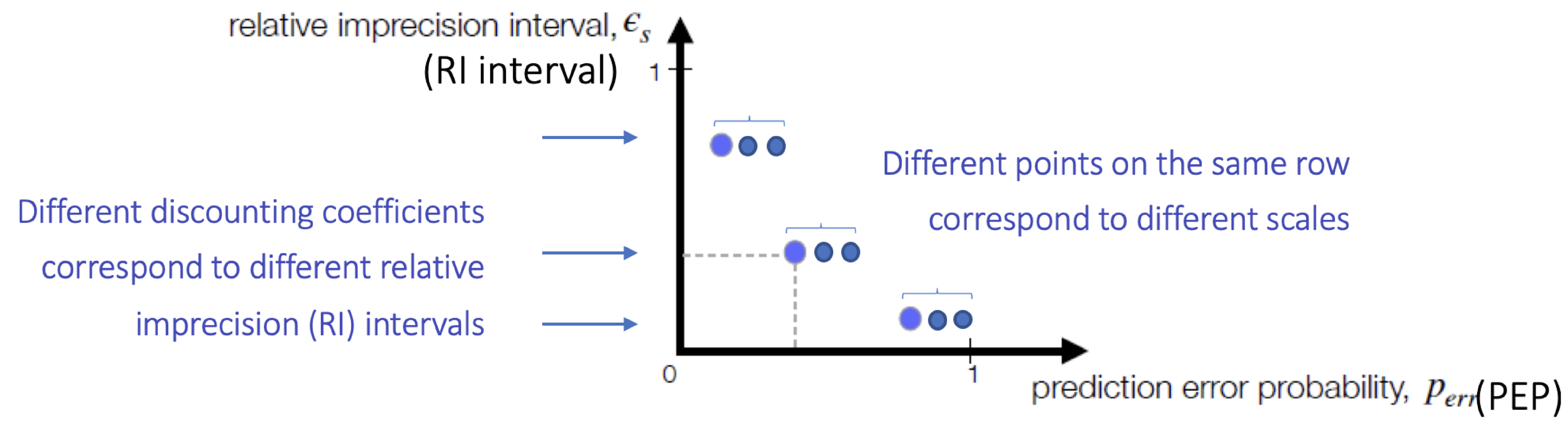
- *Relative Imprecision* (RI) interval:

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

where $g(S)$ is the ground-truth count over S

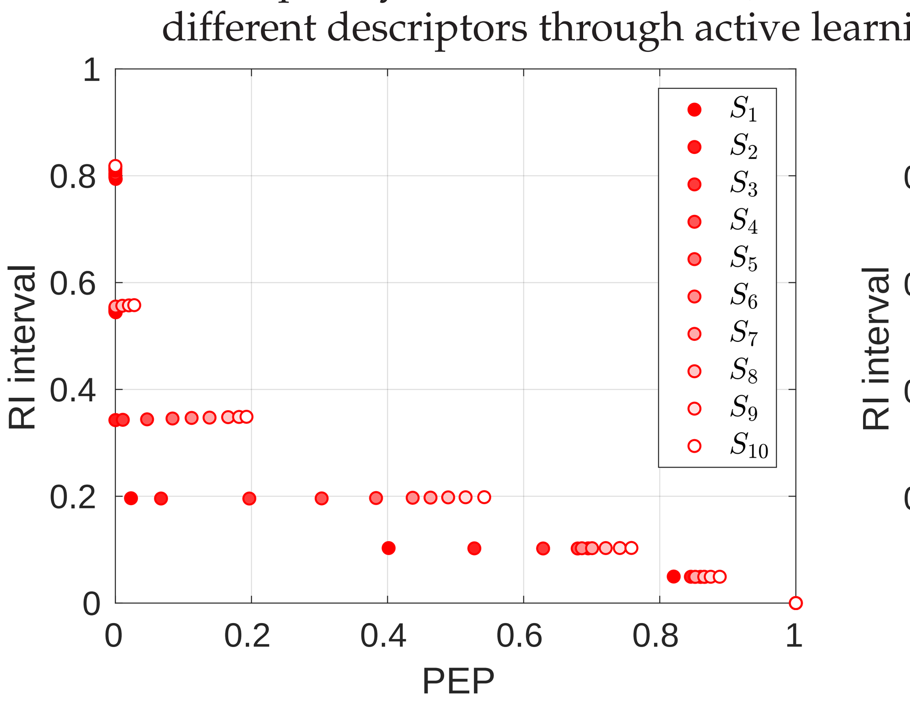
Experiments And Results

Proposed evaluation method for density estimators:

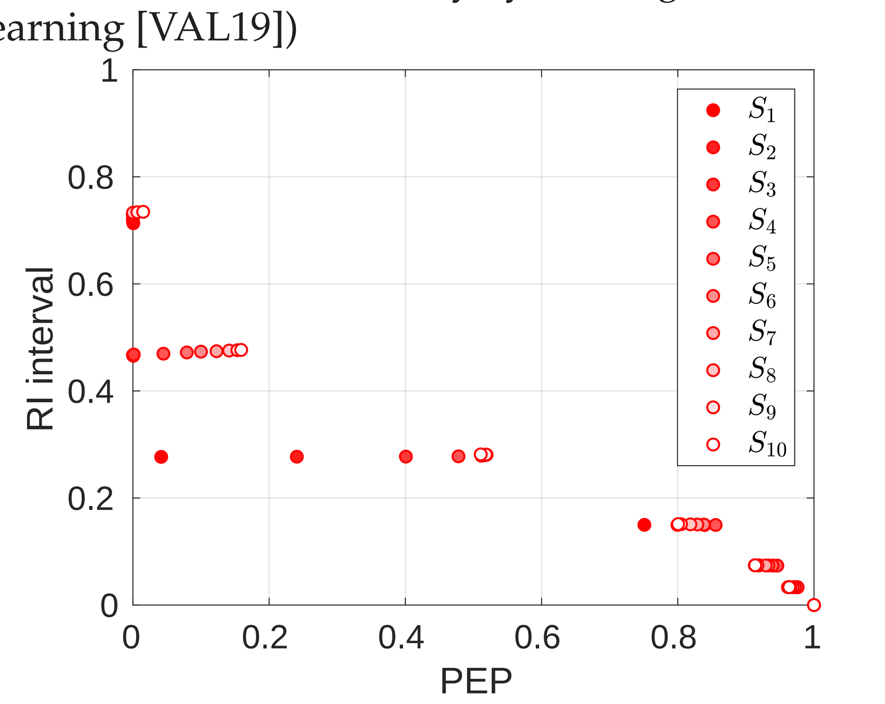


Comparison of different density estimators:

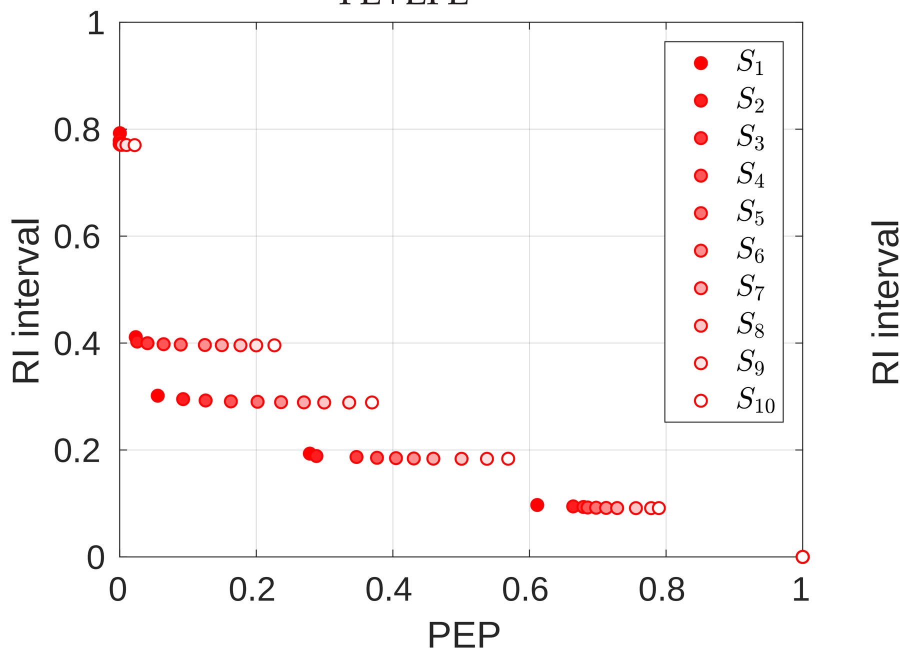
- CNN-ensemble derived using MC-dropout with $T=10$
- Comparison of the proposed FE+LFE network with respect to:
  - A different network (U-Net)
  - The same network trained on less data
  - A completely different classifier (SVM-ensemble built iteratively by training SVMs with different descriptors through active learning [VAL19])



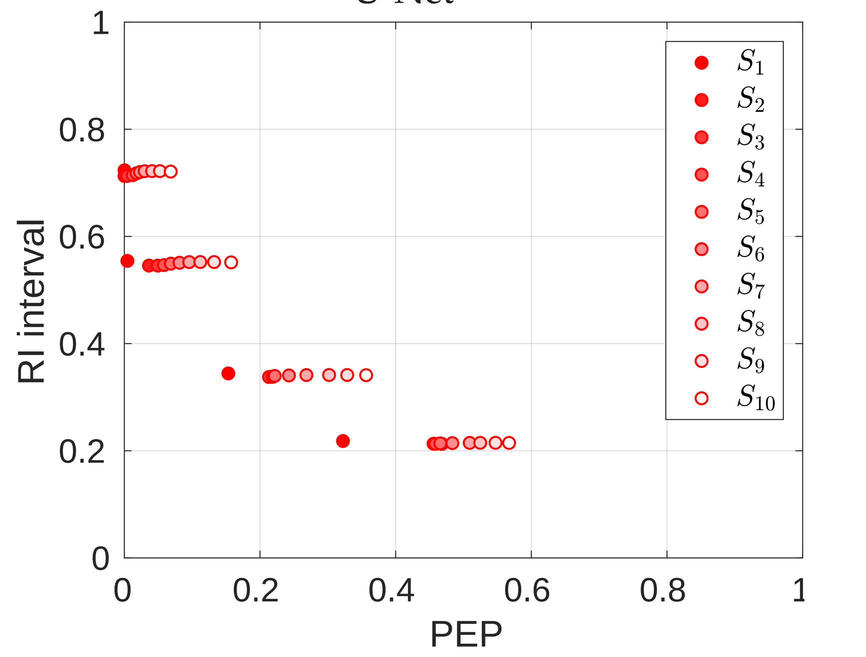
FE+LFE



U-Net



FE+LFE (trained on less data)



SVM [VAL19]

Visual example:

For given input data and ground truth annotations, results of the density estimation map along with the estimated uncertainty bounds:

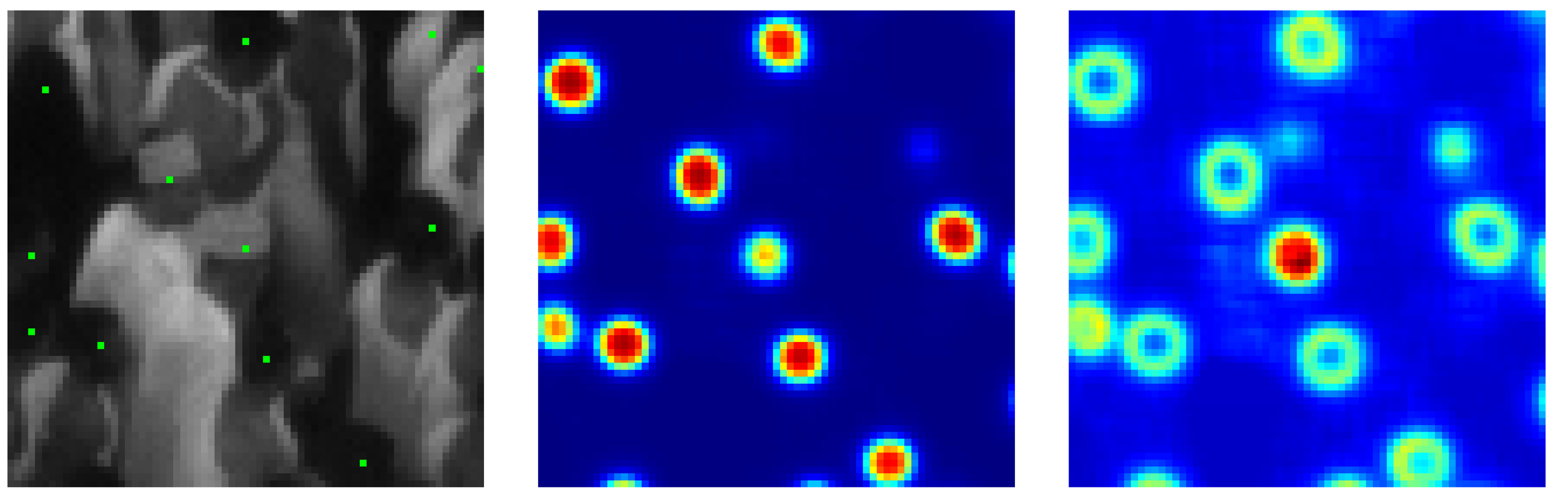


Image patch S, $g(S)=12.3$ | $BetP(H)$ map, $s(S)=12.01$ | $\mathcal{M}(\Theta)$ map, $\bar{s}(S)-\underline{s}(S)=3.2$

- RI interval: $(\bar{s}(S)-\underline{s}(S))/g(S)=0.26$, $\rightarrow$ in S there are $12.01\pm13\%$ heads, i.e. $s(S)\in[10.4,13.6]$
- Ignorance is particularly high on:
  - Head edges
  - Heads with lower gradient on the borders and strong clutter
  - Circularly-shaped areas (shoulders or round dark blobs) which are similar to heads

Conclusion

- We proposed a strategy for associating an uncertainty interval to crowd density estimation using BFT
- We proposed a new evaluation method taking into account the output uncertainty at multiple scales
- Our work opens a promising avenue for crowd safety applications which account for estimation uncertainty during planning and monitoring

References

[GG16] 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.

[Ham+18] Ryuhei Hamaguchi et al. "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.

[VAL19] Jennifer Vandoni, Emanuel Aldea, and Sylvie Le Hégarat-Mascle. "Evidential query-by-committee active learning for pedestrian detection in high-density crowds". In: *International Journal of Approximate Reasoning* 104 (2019), pp. 166–184.