A system for crowd counting in highly congested scenes geared towards Smart transportation systems

Cecilia V. Macias Hernandez, Lina Maria Aguilar Lobo and Gilberto Ochoa Ruiz

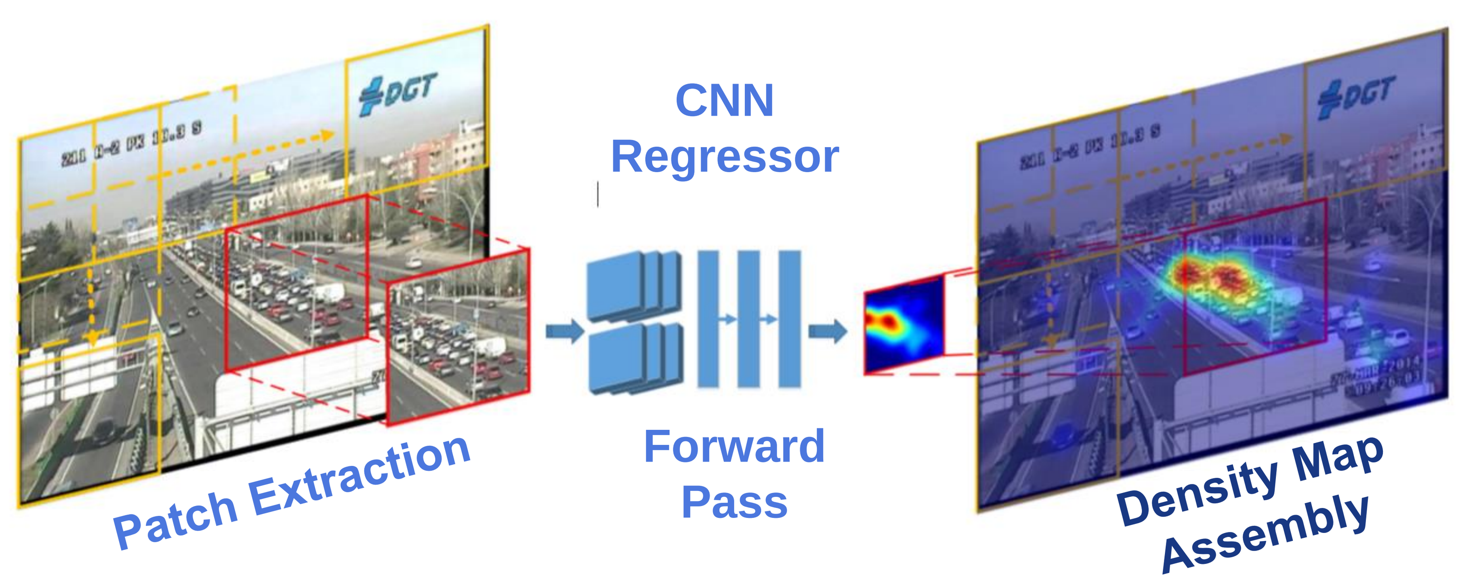
1. Context and Motivation

Crowd counting is a very important topic in computer vision and has many application areas **(i.e. surveillance, city planning, traffic monitoring)**

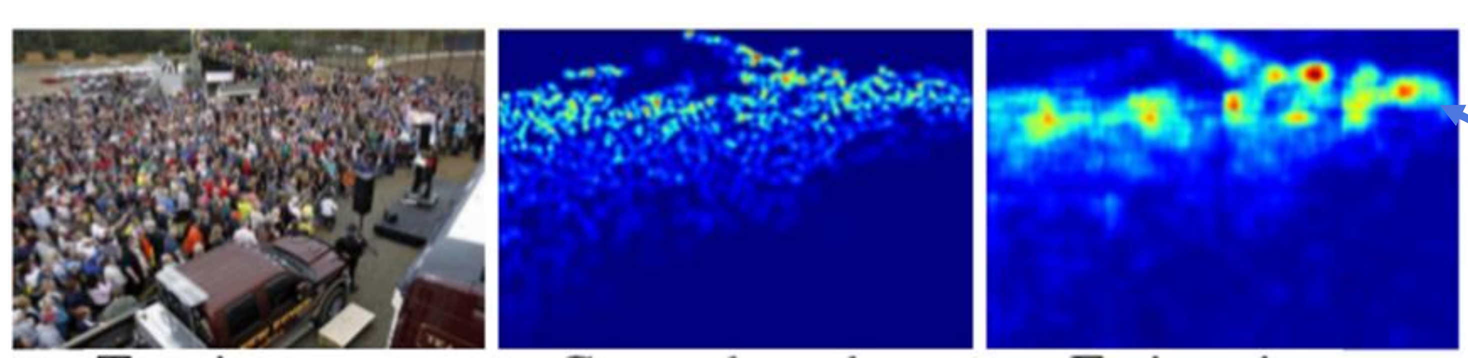
But is very challenging due to aspects such as

1. Effective **feature extraction**
2. **Annotation** is time consuming
3. Extreme **perspective distortions**
4. Variability of **camera viewpoints**
5. Variability in **illumination**
6. Varying **crowd distributions**
7. **Severe occlusions** among people

Most of the SOTA methods in **crowd counting** learn the a non-linear mapping from image features to a **density map,** from which the count can be estimated

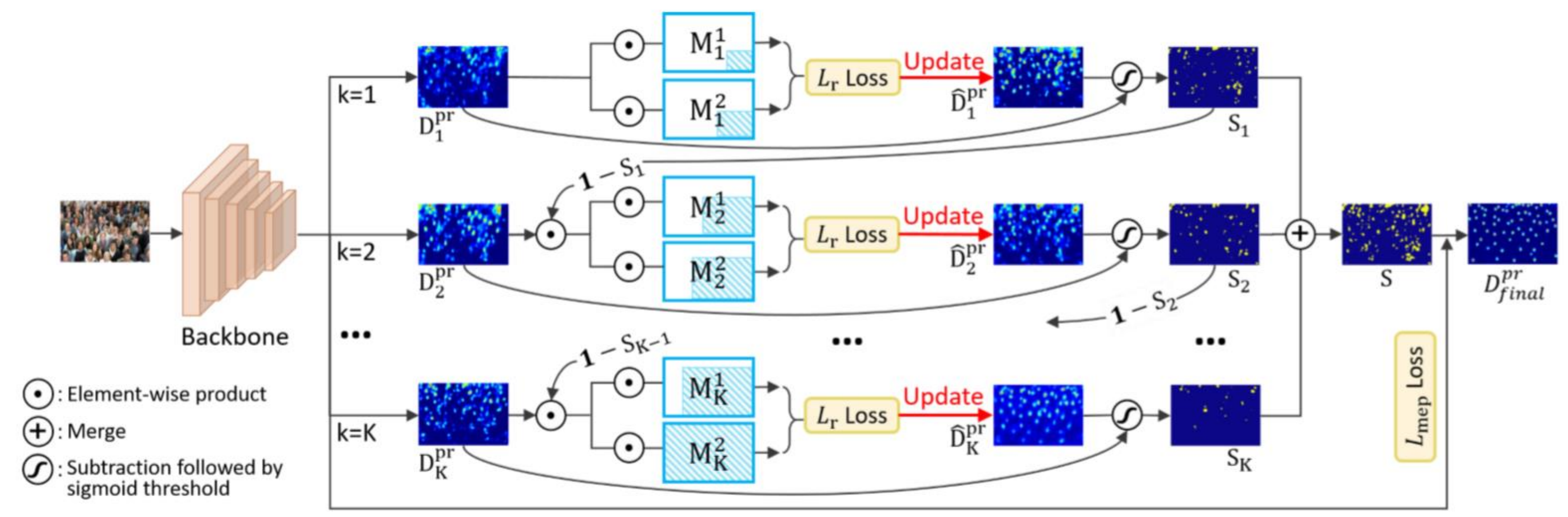


This mapping is learned by **CNN-based regressors** which are trained to produce the density map (DM). **Problem**: the quality of the DM is usually poor and it affect the **crowd count estimation**



2. Proposed Method

We propose a novel multi-branch architecture to iteratively refine the DMs from a **backbone regression network**



We make use of **saliency detection** concepts ranking methods to refine the quality of the DM

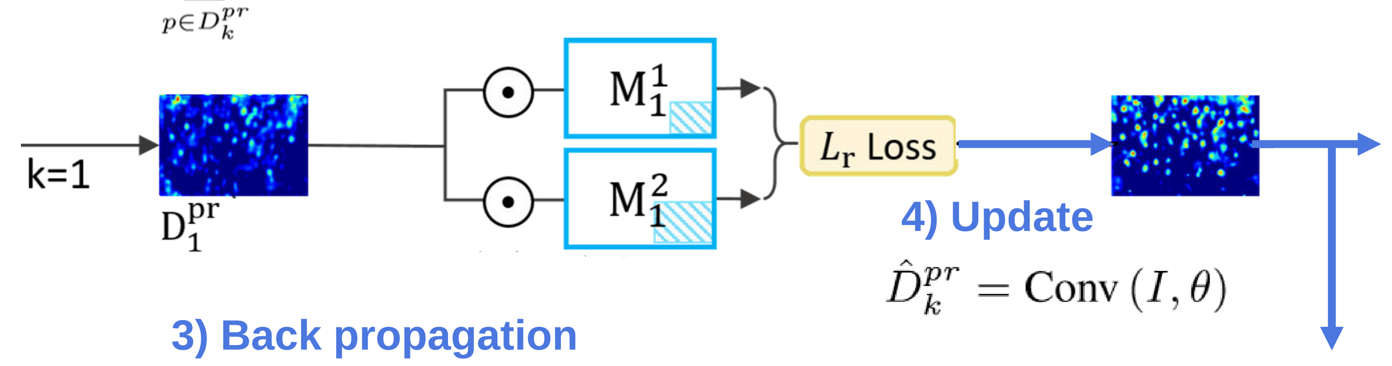
1) Mask generation

$$C\left(M_k^1\right)=\sum_{p\in D_k^{pr}}\left(D_k^{pr}\odot M_k^1\right)$$

$$C\left(M_k^2\right)=\sum_{p\in D_k^{pr}}\left(D_k^{pr}\odot M_k^2\right)$$

2) Pairwise Ranking Hinge Loss

$$L_r\left(C(M_k^1),C(M_k^2)\right)=\max\left(0,\ C(M_k^1)-C(M_k^2)+\xi\right)$$



4) Update

$$\hat{D}_k^{pr}=\mathrm{Conv}\left(I,\theta\right)$$

3) Back propagation

$$\nabla_\theta L_r=\begin{cases}0, & \text{if } C\left(M_k^1\right)-C\left(M_k^2\right)+\xi\leqslant 0,\\ \nabla_\theta C\left(M_k^1\right)-\nabla_\theta C\left(M_k^2\right), & \text{otherwise.}\end{cases}$$

5) Subregion Generation

$$S_k=\frac{1}{1+\exp\left(-\delta\left(\nabla D_k^{pr}-\boldsymbol{\Sigma}\right)\right)}$$

where $\nabla D_k^{pr}=|\hat{D}_k^{pr}-D_k^{pr}|$

6)Salient Region Detection

$$D_{k+1}^{pr}=D_{k+1}^{pr}\odot\left(\mathbf{1}-S_k\right)$$

7) Final Density Map Merging

$$S=\sum_{k=1}^{K}\{S_k\}$$

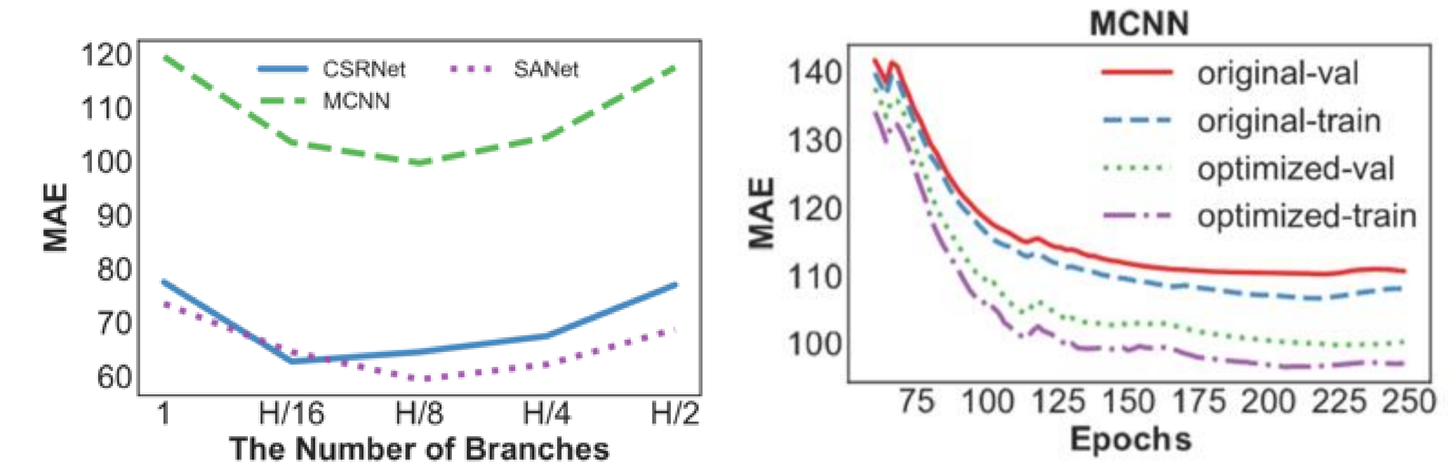
Overall training Goal $L_{global}=\sum_{k=1}^{K}L_r+L_{mep}+L_{vanilla}$

MEP loss function $L_{mep}\left(D^{pr},D^{gt}\right)=\frac{1}{N}\sum_{i=1}^{N}\left|\sum_{p\in S}D^{pr,i}(p)-\sum_{p\in S}D^{gt,i}(p)\right|$

We introduce two new **loss functions** to avoid blurring **artifacts** in the DM, which produces errors in the crowd counting, the system is **trainable end-to-end**

3 Results and Discussion

We performed several **ablation studies** to assess the best hyperparameters or our model (i.e. number of branches, window size) and combinations of the **loss functions**



We attained lower errors than other methods using the standard metrics for **regression models**

Table 1. Count errors for different methods ShanghaiTech Part A.

Method	Reference	MAE	MSE
MCNN	(Zhang et al., 2016)	181.8	277.7
SW-CNN	(Sam et al., 2017)	90.4	135.0
C-MTL	(Sindagi, 2017)	101.3	152.4
Ours		67.0	104.5

Future work: integrating these models in smart cameras within transportation systems