Crowd Counting: A view around state-of-the-art methods

Anh Nguyen Tuan, Hoang Ha Van

University of Information Technology - VNU-HCM

November 27, 2023

Table of Contents

Table of Contents

Crowd Counting

- Estimating the number of people or objects in a given scene or image.

Crowd Counting

- Estimating the number of people or objects in a given scene or image.
  - Input: An image containing crowd.
  - Output: The estimated number of people in the image

Crowd Counting

- Estimating the number of people or objects in a given scene or image.
  - Input: An image containing crowd.
  - Output: The estimated number of people in the image

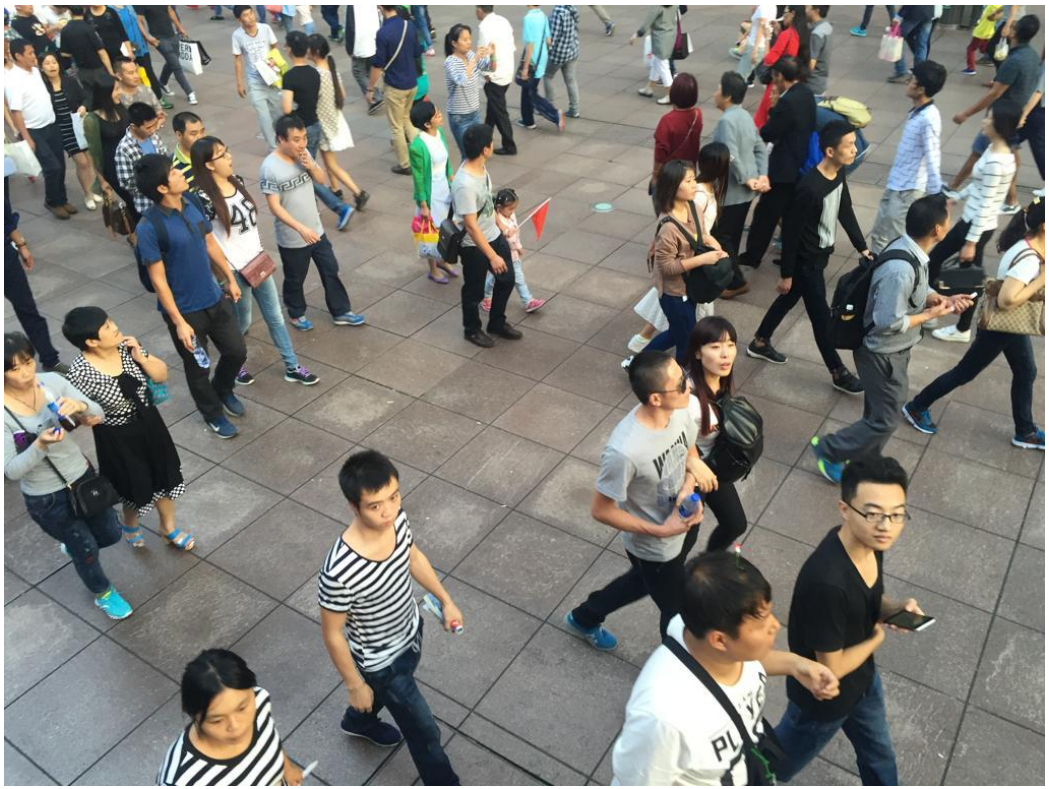

Figure 1: Number of people estimated: 43

Crowd Counting

- Estimating the number of people or objects in a given scene or image.
  - Input: An image containing crowd.
  - Output: The estimated number of people in the image
- The problem encounters a common challenge in computer vision.

Density map based counting

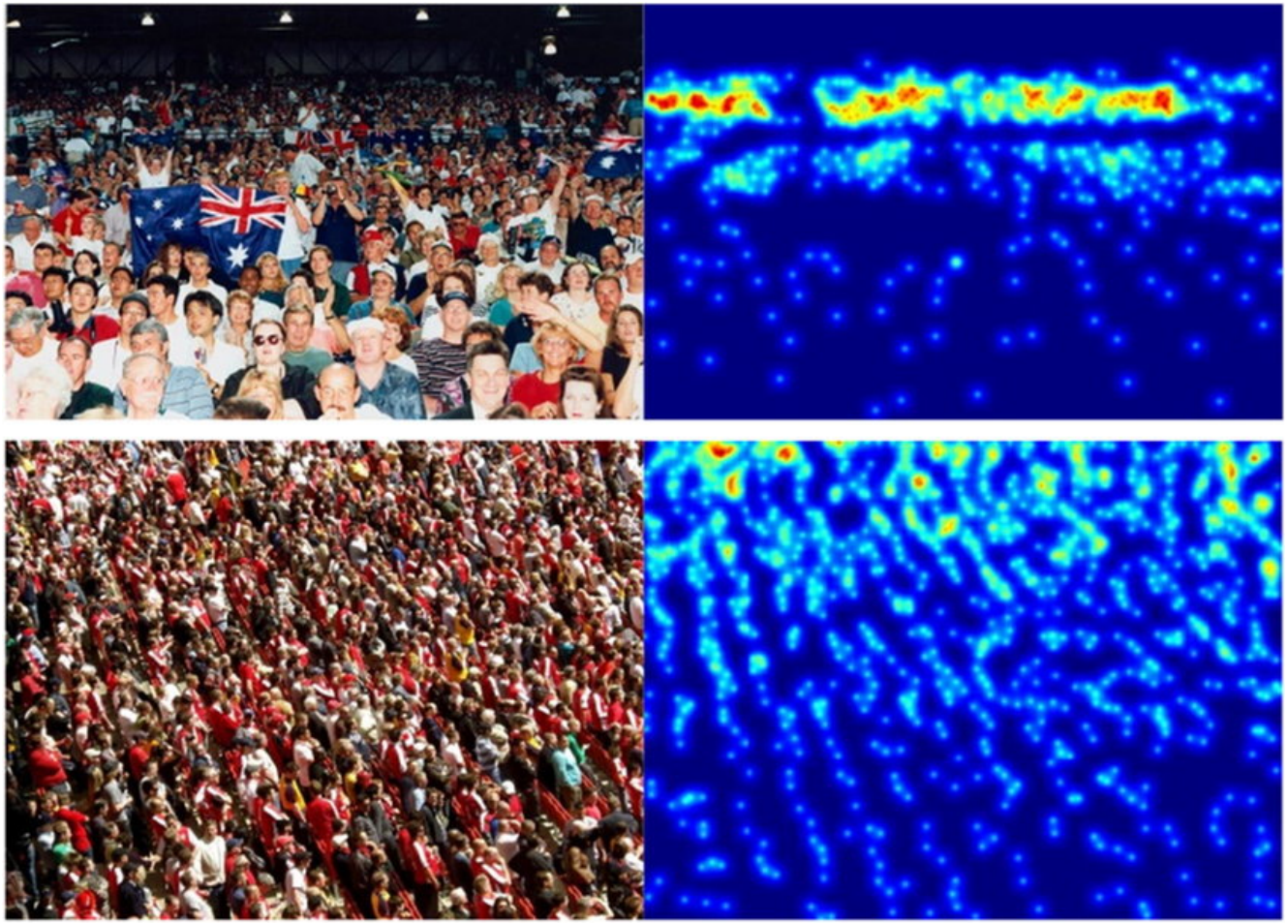

Figure 2: Density map

Density map based counting

A head at pixel $\mathbf{x}_i \mapsto \delta(\mathbf{x} - \mathbf{x}_i)$. Then an image with N heads labeled:

$$F(\mathbf{x}) = \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i).G_{\sigma_i}(\mathbf{x}), \quad \text{with } \sigma_i = \beta \overline{d^i}$$

Note: G is Gaussian kernel, $\overline{d^i}$ is average distances from $\mathbf{x}_i$ to k nearest neighbors.
We call F be ground truth density map.
Let $f(X_i; \Theta)$ is density map generated by model corresponding to image X_i, where Θ is a set of learnable parameters.
The loss function is defined as follows:

$$L(\Theta) = \frac{1}{2M} \sum_{i=1}^{M} ||f(X_i; \Theta) - F_i||_2^2$$

After that, train and train..

Table of Contents

Multi-Column Convolutional Neural Network (MCNN)

- Paper: Single-Image Crowd Counting via Multi-Column Convolutional Neural Network (CVPR 2016)
- Architecture:

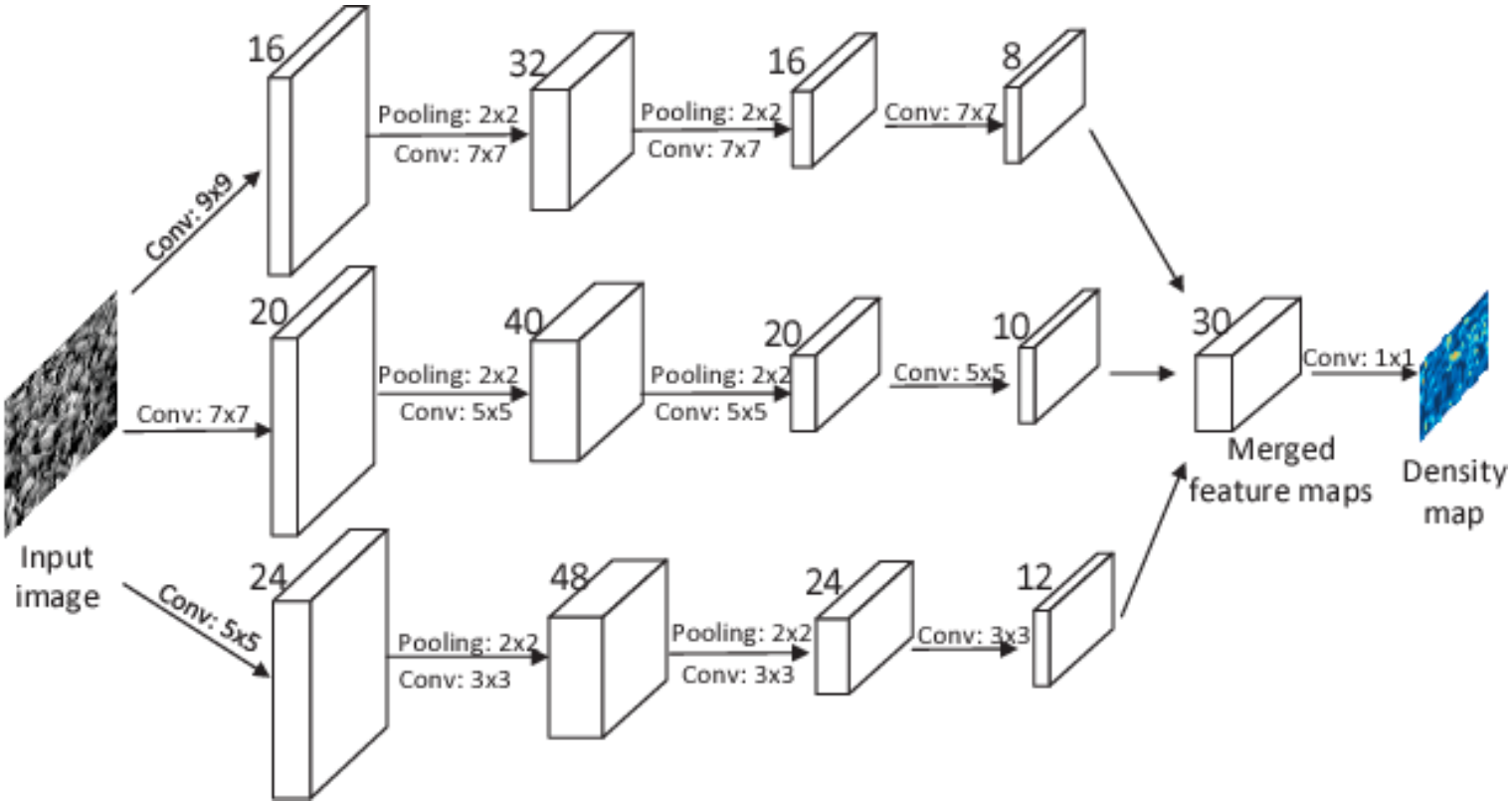

Figure 3: MCNN's architecture

Generalized Loss Function

- Paper: A Generalized Loss Function for Crowd Counting and Localization (CVPR 2021)

Generalized Loss Function

Figure 4: Optimal Transport

Generalized Loss Function

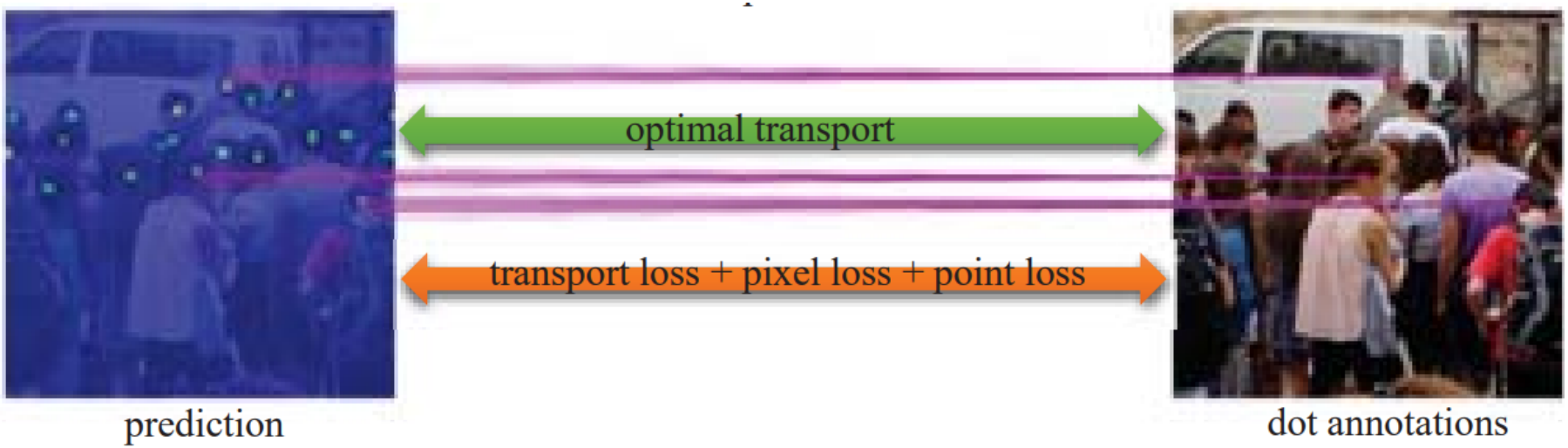

Figure 5: Loss's style

Generalized Loss Function

Loss function is based on the entropic-regularized unbalanced optimal transport cost:

$$\mathcal{L}_{\mathbf{C}}^{\tau}(A, B) = \min_{\mathbf{P} \in \mathbb{R}_{+}^{n \times m}} \{\langle \mathbf{C}, \mathbf{P} \rangle - \epsilon H(\mathbf{P}) + \tau D_1(\mathbf{P}1_m | \mathbf{a}) + \tau D_2(\mathbf{P}^\top 1_n | \mathbf{b})\}$$

With $\mathbf{C} \in \mathbb{R}_{+}^{n \times m}$ is the transport cost matrix, $\mathbf{P}$ is the transport matrix,
$H(\mathbf{P}) = \sum_{ij} P_{ij} \log(P_{ij})$ is the entropic regularization term,
$D_1(\mathbf{P}1_m | \mathbf{a})$ is the pixel-wise loss, $D_2(\mathbf{P}^\top 1_n | \mathbf{b})$ is the point-wise loss.

Residual Network (ResNet-50)

- Paper: Rethinking Spatial Invariance of Convolutional Networks for Object Counting (CVPR 2022)
- Architecture:

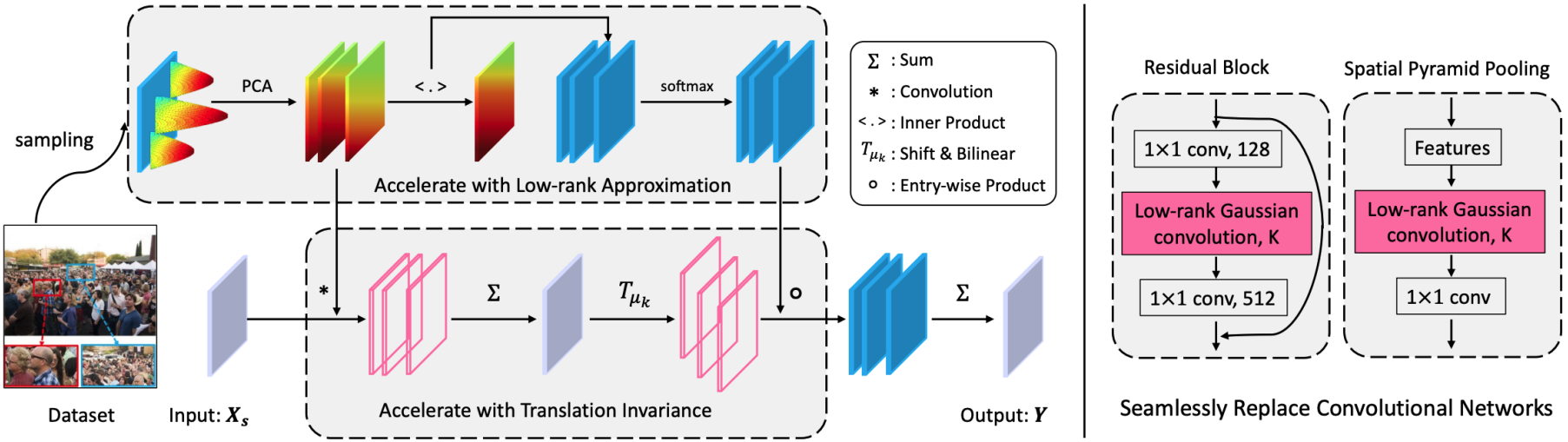

Figure 6: Low-rank Gaussian convolutional layer

Table of Contents

ShanghaiTech

ShanghaiTech comprises 1198 annotated images, with a total of 330,165 people with annotated head centers:

- Part A: 482 randomly crawled images from the Internet.
- Part B: 716 images captured from busy metropolitan streets in Shanghai.

Each image has an attached ".mat" file containing the centers of the annotated heads of the people appearing in this image.

	Part A	Part B
Train	300	400
Test	182	316

Table 1: Train and test sets of ShanghaiTech dataset.

ShanghaiTech

ShanghaiTech comprises 1198 annotated images, with a total of 330,165 people with annotated head centers:

- Part A: 482 randomly crawled images from the Internet.
- Part B: 716 images captured from busy metropolitan streets in Shanghai.

Each image has an attached ".mat" file containing the centers of the annotated heads of the people appearing in this image.

	Part A	Part B
Train	300	400
Test	182	316

Table 1: Train and test sets of ShanghaiTech dataset.

ShanghaiTech

ShanghaiTech comprises 1198 annotated images, with a total of 330,165 people with annotated head centers:

- Part A: 482 randomly crawled images from the Internet.
- Part B: 716 images captured from busy metropolitan streets in Shanghai.

Each image has an attached ".mat" file containing the centers of the annotated heads of the people appearing in this image.

	Part A	Part B
Train	300	400
Test	182	316

Table 1: Train and test sets of ShanghaiTech dataset.

ShanghaiTech

ShanghaiTech comprises 1198 annotated images, with a total of 330,165 people with annotated head centers:

- Part A: 482 randomly crawled images from the Internet.
- Part B: 716 images captured from busy metropolitan streets in Shanghai.

Each image has an attached ".mat" file containing the centers of the annotated heads of the people appearing in this image.

	Part A	Part B
Train	300	400
Test	182	316

Table 1: Train and test sets of ShanghaiTech dataset.

ShanghaiTech

ShanghaiTech comprises 1198 annotated images, with a total of 330,165 people with annotated head centers:

- Part A: 482 randomly crawled images from the Internet.
- Part B: 716 images captured from busy metropolitan streets in Shanghai.

Each image has an attached ".mat" file containing the centers of the annotated heads of the people appearing in this image.

	Part A	Part B
Train	300	400
Test	182	316

Table 1: Train and test sets of ShanghaiTech dataset.

ShanghaiTech

	Resolution	Num	Max	Min	Total
Part A	different	482	3139	33	241,677
Part B	768×1024	316	578	9	88,488

Table 2: The statistics of ShanghaiTech dataset.

Table of Contents

Results

	Part A		Part B	
	MAE	MSE	MAE	MSE
MCNN	114.96	172.66	33.00	48.08
VGG-19 (L2)	70.81	114.77	9.83	14.41
VGG-19 (GLoss)	64.79	108.49	8.32	14.18
ResNet-50	107.72	188.39	7.90	12.51

Table 3: Results on ShanghaiTech dataset.

Demo

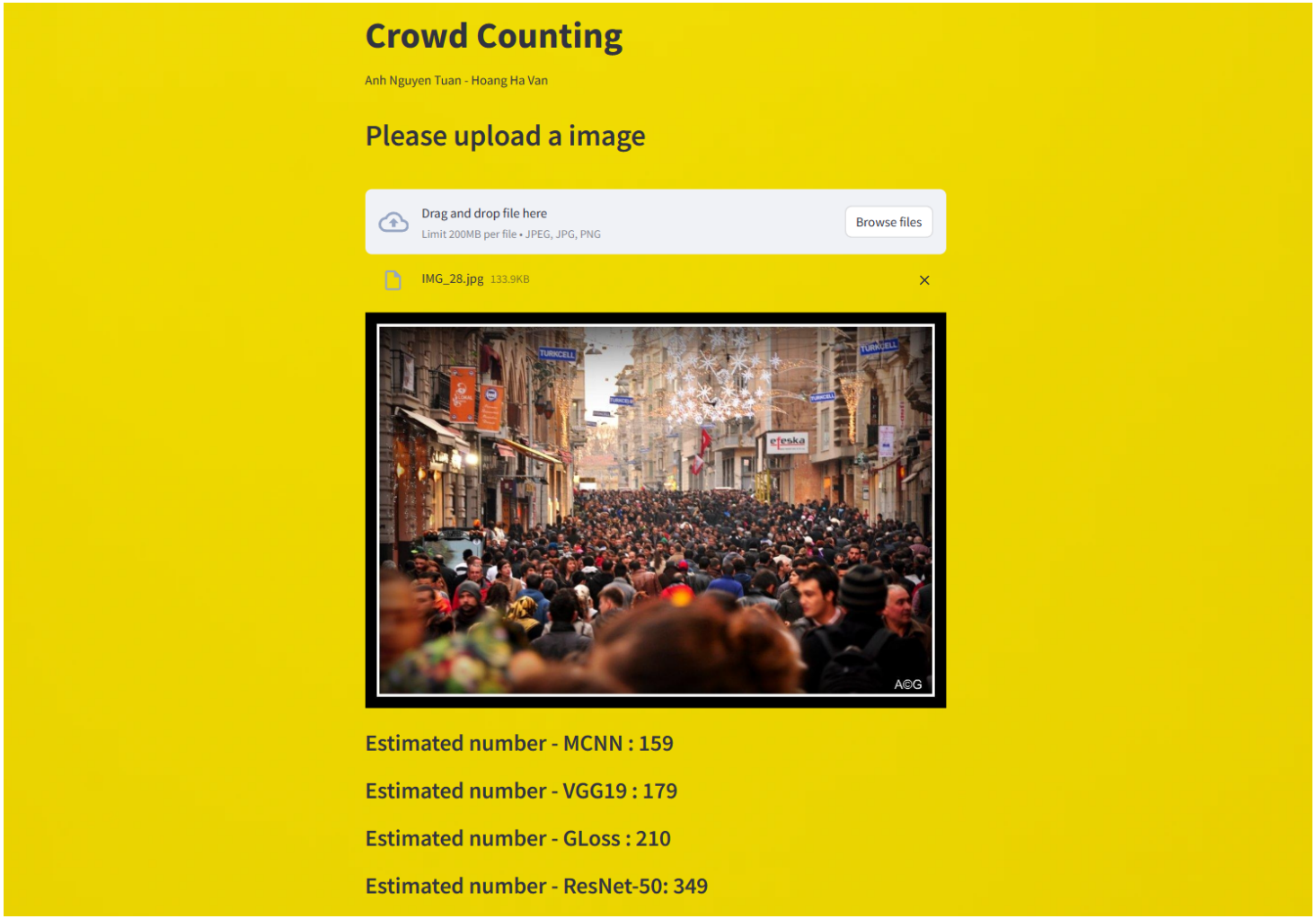

Figure 7: Crowd Counting web app demo

Table of Contents

Conclusion

- GLoss significantly improves the estimation performance.
- Resnet-50 model trained on part A had results that seem to be not good.

Conclusion

- GLoss significantly improves the estimation performance.
- Resnet-50 model trained on part A had results that seem to be not good.

Thank you!