A baseline pipeline for pedestrians analysis in public environments

Leonardo Zini
Pietro Martinello
Giovanni Casari

01

Our goal

With the advent of digitalization, numerous security cameras have begun to appear in urban centers, particularly in busy areas. This research aims to provide an initial tool for improving the productivity of current visual systems available for surveillance in such populated areas. In fact, the use of artificial intelligence models could lead to substantial time savings to the operator in following the movements of the individuals being viewed, and in the eventual investigation regarding their identity.

Our proposal

Tracking by detection

Using SoTA architecture yolo v7 as detector, we have trained a neural network that performs tracking of pedestrians. Results have been compared also with a model based on HOG descriptors.

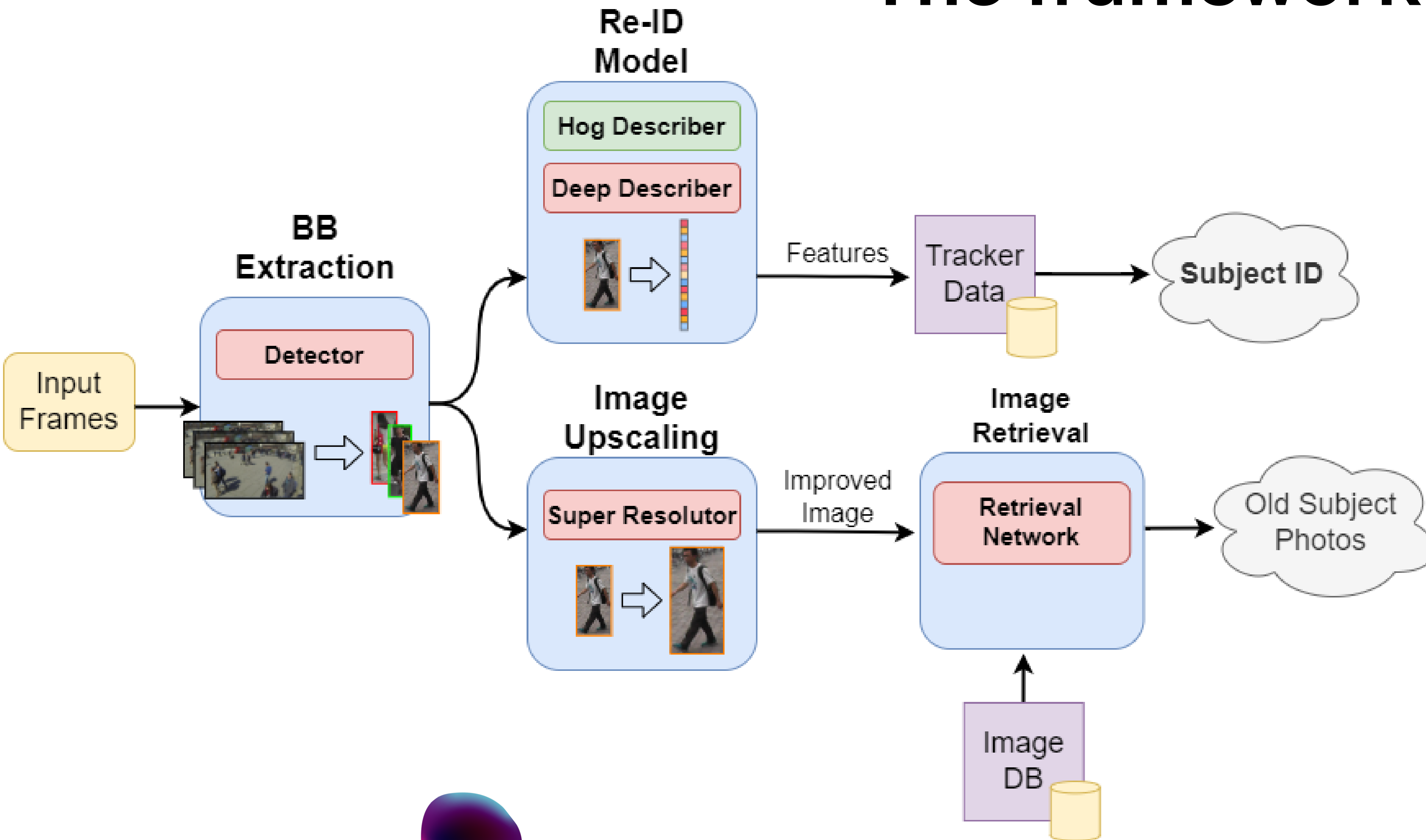
Super Resolution

After tracking, we want to enhance a small crop of pedestrians in order to increase visual quality. We have done it using a existing GAN architectures and classic image processing operators.

Image Retrieval

System is also able to retrieve the same individuals also from pictures taken in different contexts, with different clothes.
Long ReID scenarios

The framework



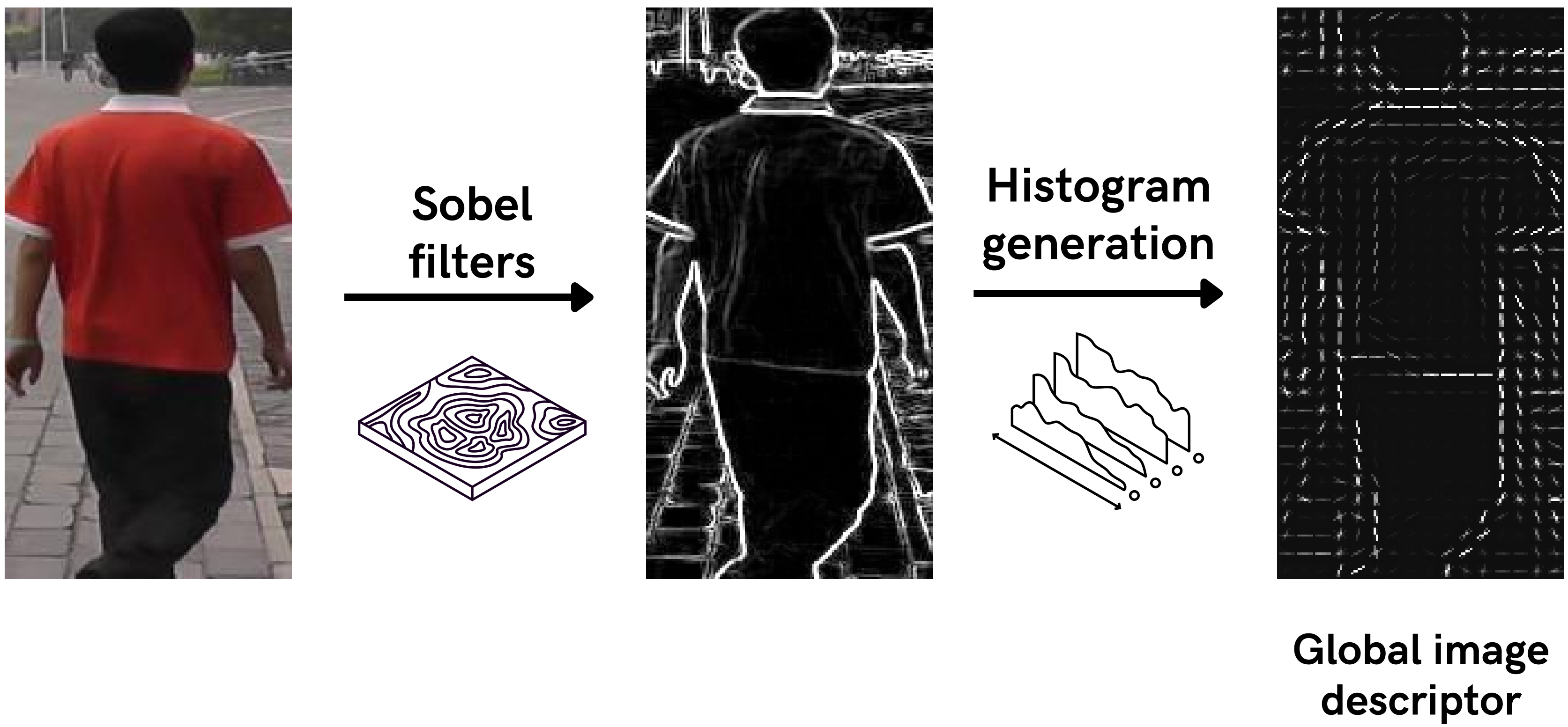
Detection

We opted for using a pre-trained network to perform the first step of out pipeline, the pedestrian detection.
After a research regarding balances between performance and computational inference costs of current state-of-the-art methods, we chose to adopt YOLOv7-tiny model as detector, given its low inference costs, which are designed for real-time tracking, but still a very high throughput.

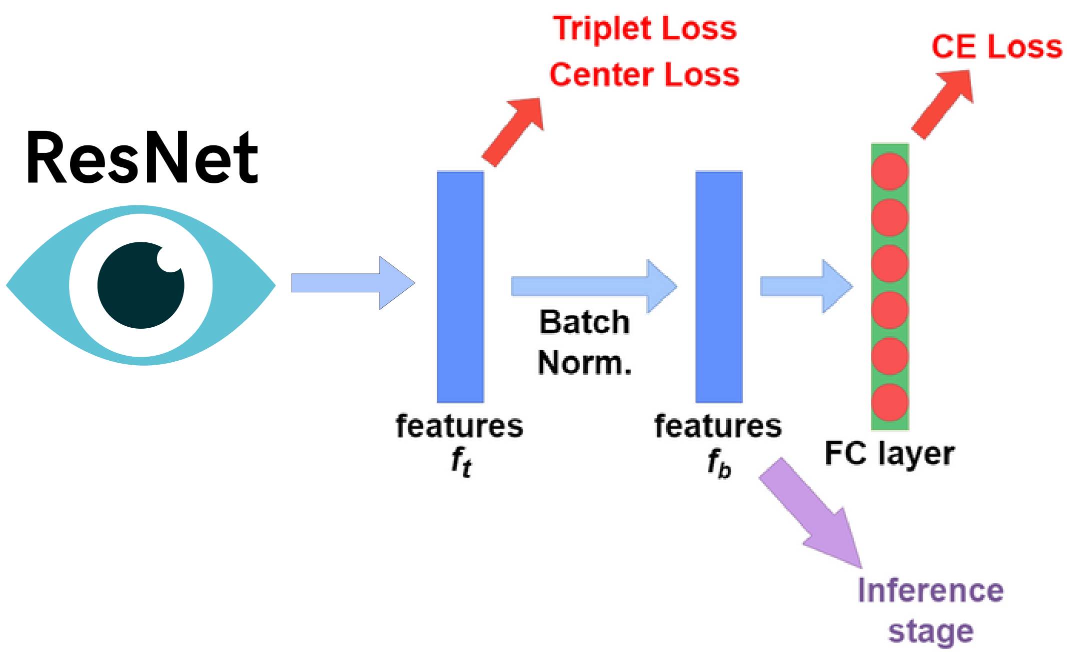
03

Tracking of pedestrians

HOG-based ReID



Deep ReID



$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \alpha \mathcal{L}_{\mathrm{Tr}} + \beta \mathcal{L}_{\mathrm{C}}$$

cross-entropy loss

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{N} q_i log\left(\frac{exp(s_i)}{\sum_{j=1}^{N} exp(s_j)}\right) \begin{cases} q_i = 0, & y \neq i \\ q_i = 1, & y = i \end{cases}$$

triplet loss

$$\mathcal{L}_{\mathrm{Tr}} = max(0,\ d_p - d_n + \gamma)$$

center loss

$$\mathcal{L}_{\mathrm{C}} = \sum_{j=1}^{N} \|f_{t_j} - c_{y_j}\|_2^2$$

Datasets:

MOTSynth

130k imgs, 1000 ids

Market1501

13k imgs, 750 ids

Training results

	Name:	A	B	C	D	E	F	G	H	I
	Model:	ResNet18				ResNet50				
	Training:	50e Mot	100e Mot	50e Mar	100e Mar	50e Mot	100e Mot	50e Mar	100e Mar	50e Mot + 50e Mar
MOTSynth	rank-1	92.3	92.8	82.2	83.3	92.6	**93.2**	79.0	76.7	83.6
	mAP	66.8	68.0	45.8	48.0	68.8	**69.2**	42.6	40.8	48.9
	mINP	20.6	21.8	9.0	9.2	22.6	**23.0**	8.0	7.5	9.9
Market 1501	rank-1	94.1	94.7	96.9	**97.6**	92.8	91.9	94.7	95.0	94.2
	mAP	19.5	20.2	44.9	**50.7**	19.5	18.3	37.6	36.7	40.3
	mINP	0.6	0.7	5.5	**8.2**	0.6	0.5	4.0	3.6	4.1
MARS	rank-1	82.9	83.2	89.8	**91.4**	82.3	81.9	86.9	86.5	86.9
	mAP	28.4	29.2	49.7	**54.5**	27.9	26.8	43.4	42.2	43.7
	mINP	1.4	1.4	6.8	**9.0**	1.4	1.3	4.8	4.2	4.2

Table 2: Results (%) on models trained differently

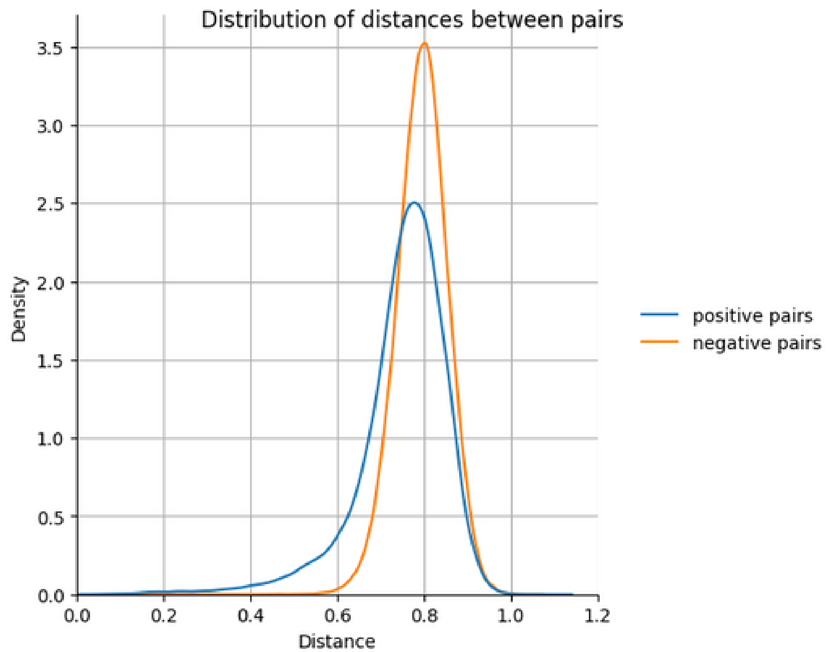
HOG vs Deep on MARS dataset

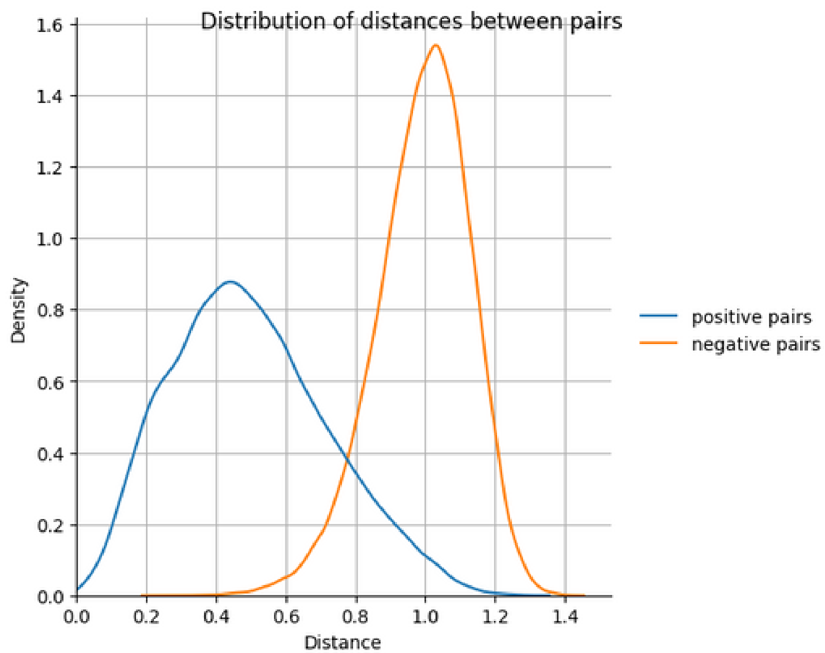
HOG

Rank-1	mAP	mINP
45.0	7.1	0.2

Our deep model

Rank-1	mAP	mINP
91.4	54.5	9.0

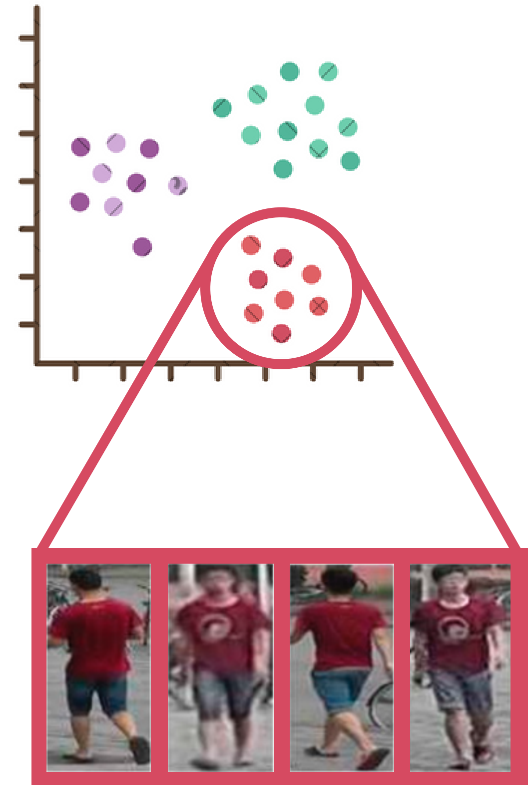




PeopleDB

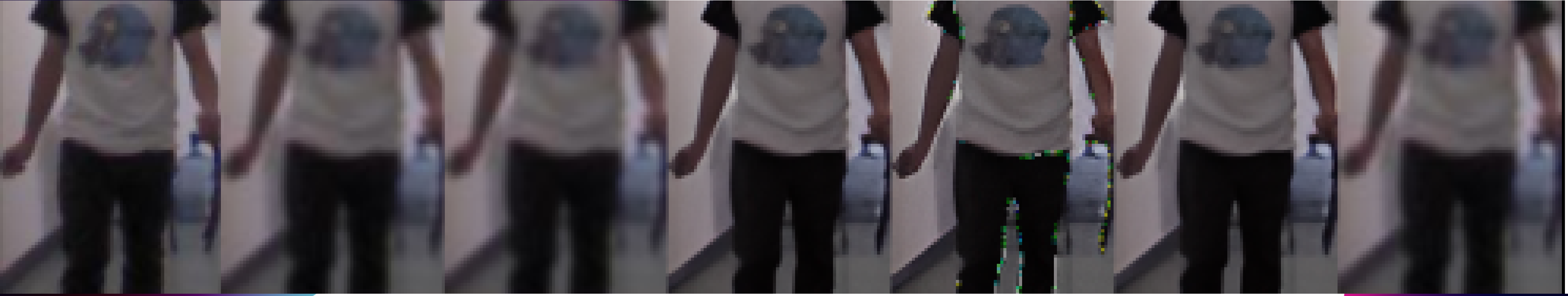
To allow correct re-identification, the feature vectors representing the different people will be stored inside the PeopleDB.

To be able to label correctly a new feature vector, the distance between it and the midpoint of each person stored will be calculated.
If the distance exceeds a certain threshold, a new identity will be created, otherwise that person model will be updated.



04

Super Resolution

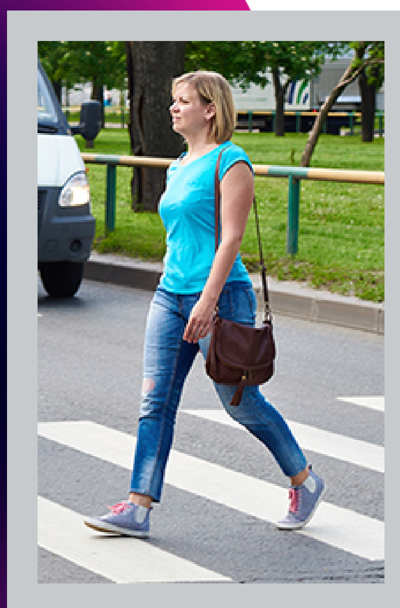


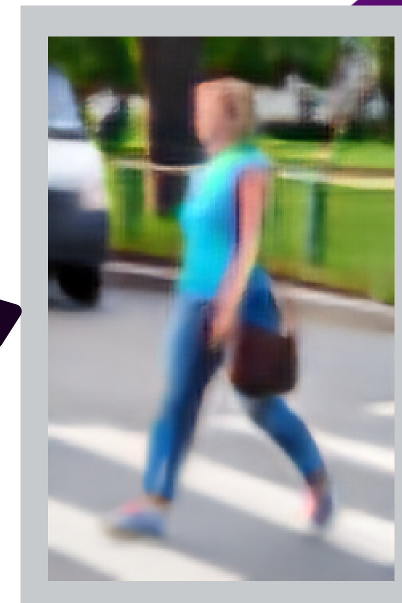
Super Resolution

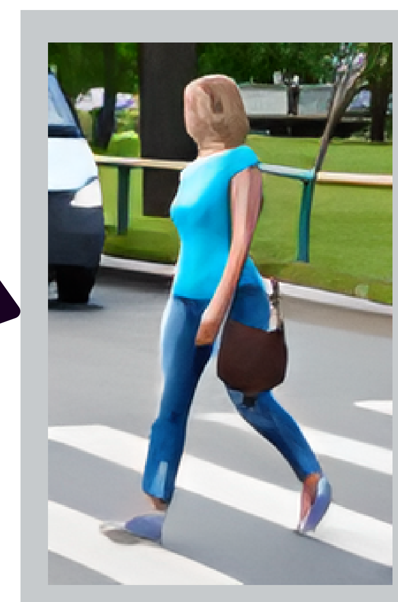
After understanding the problem of image degradation, we compare different pre-trained model, in order to chose the best for our purpose. We compare the model in reconstructing the image using the PSNR measure.

Since the model not only has to reconstuct the image in order to be more human-affable. We perform some image processing, substantially sharpening and gaussian, and compared the performace also based on the retrieval.

Methods comparisons







Here, we checked which of the two chosen model performs better, both from a metric and perceptual point of view.

Method	PSNR
PDM-SR	17,1
Real-ESRGAN	**20,7**
Bicubic	11,2
PyrUp	11,7

05

Image Retrieval

Our ultimate goal was to create a system capable of retrieving images from previous recordings of a specified person.

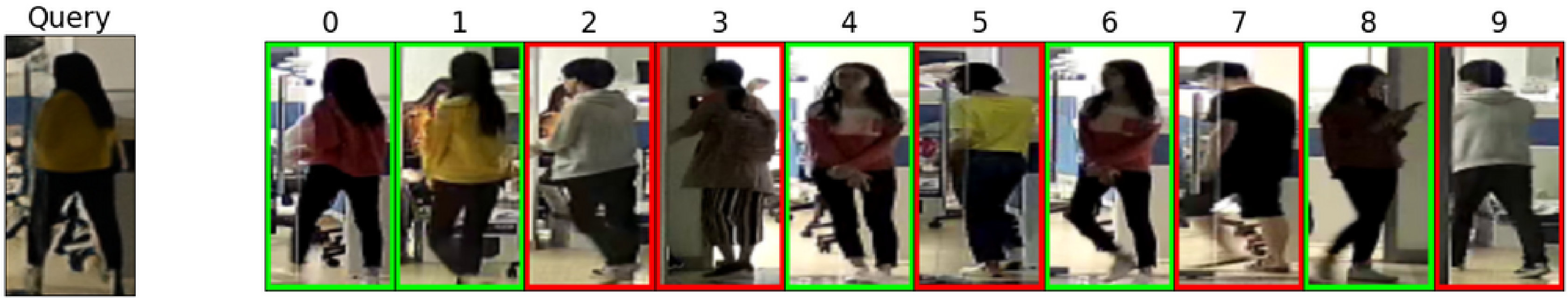
We then have to deal with the long-term people re-identification problem, that poses more challenges then the short-term counterpart, as we must expect a person to change his clothes .

Image Retrieval

We initially tried to use our ReID model in order to archive this task, the results weren't bad, but surely improvable.

We then used the ReIDCaps network, one of the currently sota architectures for this task, and the results improved .

The similarity distance has also be tested. We compared the result using both the Euclidean and the Cosine one, and the last one gave better results.



In conclusion..

		Downsampled			Bicubic			Superres		
		mAP	rank1	rank5	mAP	rank1	rank5	mAP	rank1	rank5
Our ReID Model	Euc	10.9	59.0	77.3	10.9	58.0	78.0	10.6	55.8	77.1
	Cos	**12.4**	**61.9**	**80.1**	**11.9**	**59.0**	**78.5**	**11.0**	54.9	76.5
ReIDCaps	Euc	8.8	54.2	71.8	8.8	50.9	71.4	10.1	62.7	83.8
	Cos	8.8	54.2	71.8	8.8	50.9	71.4	10.1	**62.7**	**83.8**

Thank you for your attention.