Loria
Laboratoire lorrain de recherche en informatique et ses applications

Semi-supervised learning through adversary networks for baseline detection

Romain Karpinski
Abdel Belaïd

Neural prediction method

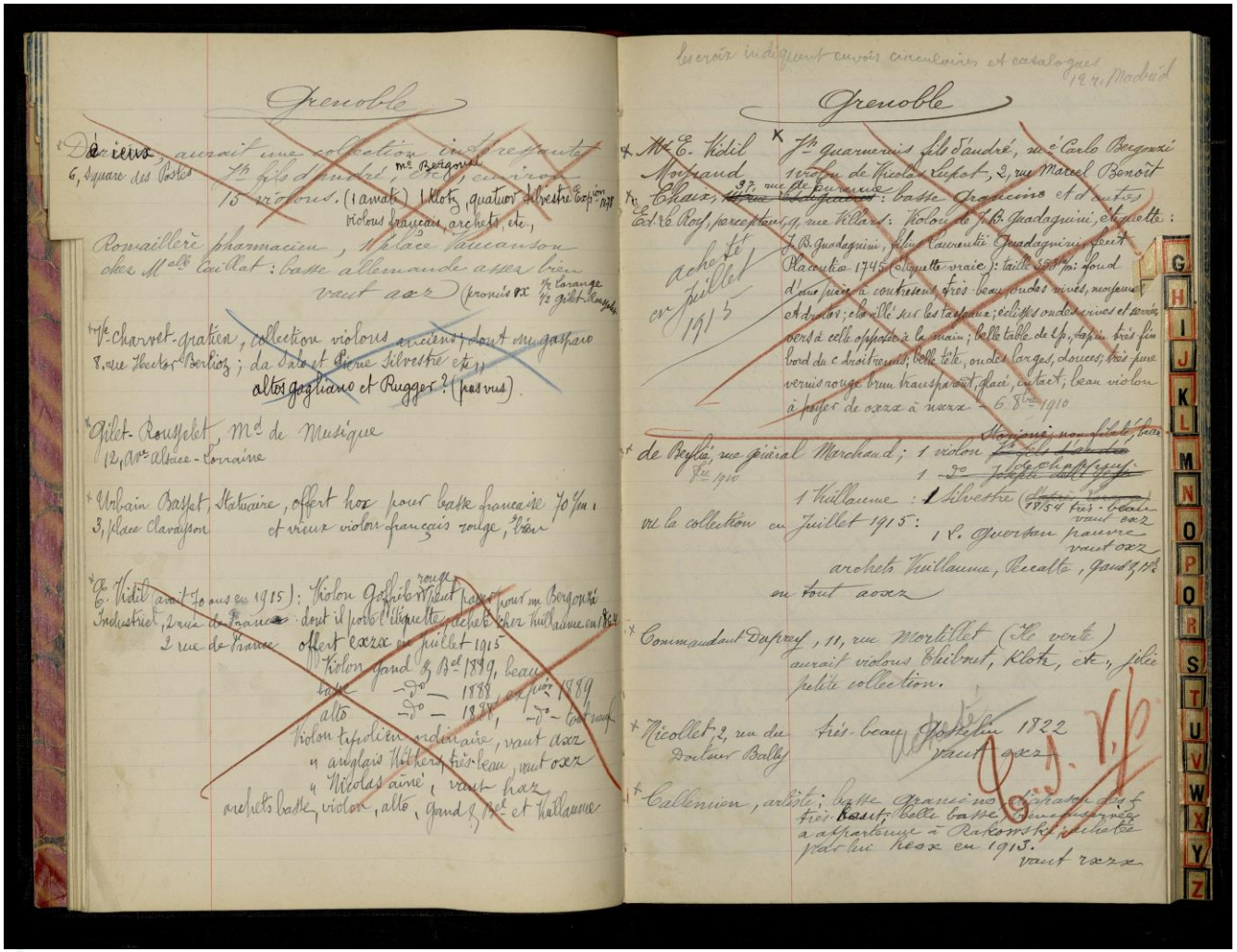
The context

- Request from the Museum of Music at Paris (MMP)
- Extract textual content of violin sales records
- Names of famous violins and luthiers
  - Example: "Amati", “Garneri”, "Stradivari“, etc.
- Financial value of instruments during transactions

Gand, Bernardel, Caressa et Français

A century and a half of history: 1816 to 1944

11 000 images, 2G



with transactions and well structured writing style

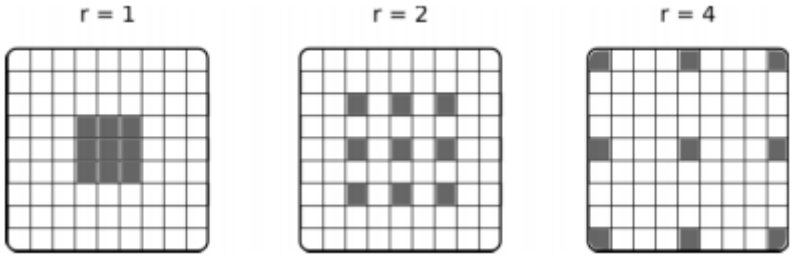
Line extraction

- Traditional method
  - Supervised approach → learning from examples
- Examples
  - Chosen to get a good representation
  - Labeled by hand
  - ➔ Very expensive
- Our aim
  - Improve results when there is few labeled data
  - Semi-supervised approach to take advantage of unlabeled images

Loria

The state of the art
Semantic labeling

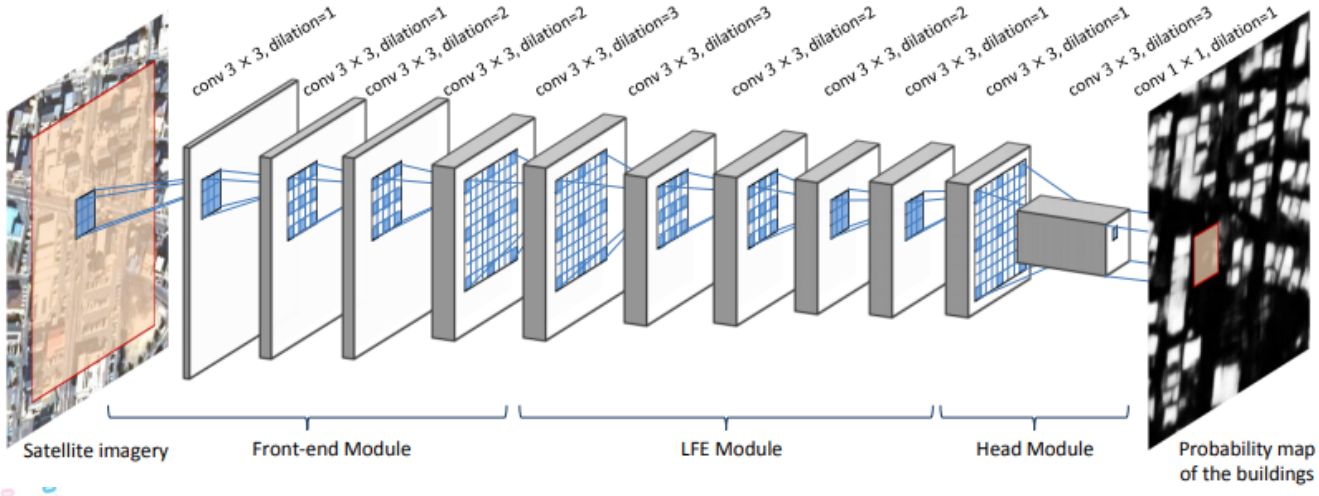
- Renton et al.*: VGG16 ➔ X-Heights
  - VGG16 modified ➔ fully convolutional
  - Max pooling ➔ dilated convolutions

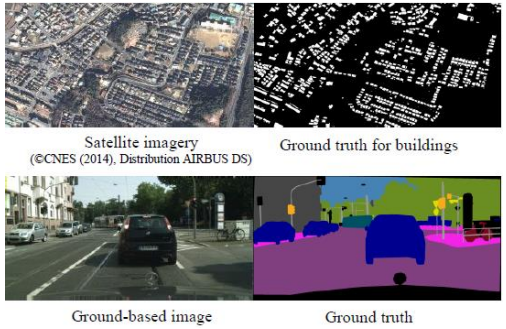


** Hamaguchi et al. "Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery", 2018, https://arxiv.org/pdf/1709.00179

The state of the art
Semantic labeling

- Hamaguchi et al. *: Aggregate local features with decreasing dilation factor





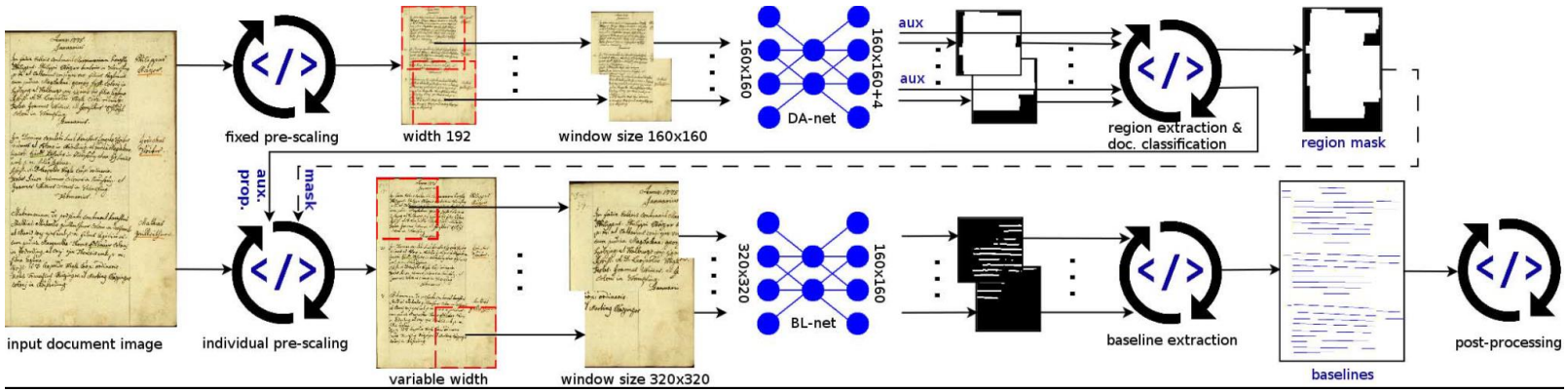
Remote sensing datasets (cars, sidewalks, buildings…)

Local feature extraction (LFE) module attached on top of dilated front-end module

* Hamaguchi et al. "Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery", 2018, https://arxiv.org/pdf/1709.00179

The state of the art

- Fink et al.*: CNN-based sliding window approach

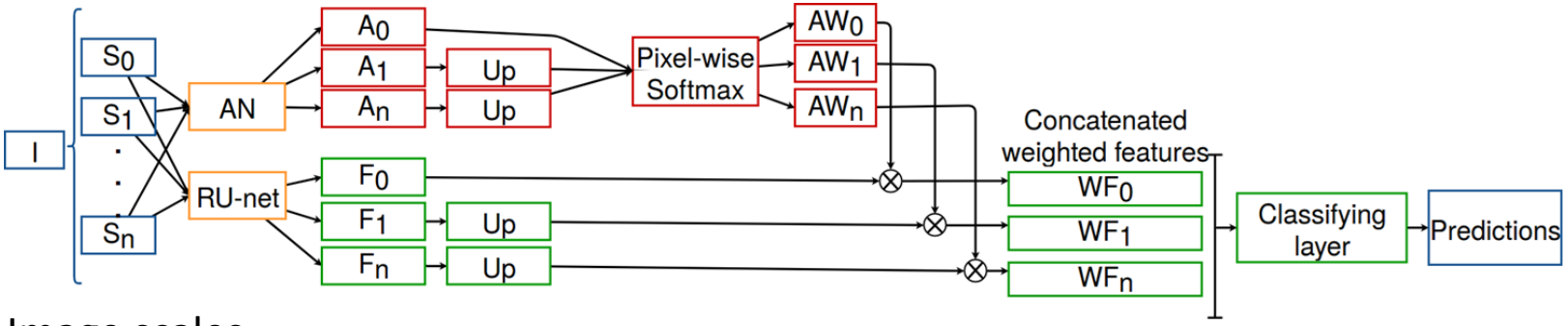


  - Preprocessing
    - U-net on patches for simple layout analysis
  - Post-processing
    - Pruning some bad candidate lines and joining segments to form good baselines

*Fink, Michael, et al. "Baseline Detection in Historical Documents Using Convolutional U-Nets." *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018.

The state of the art

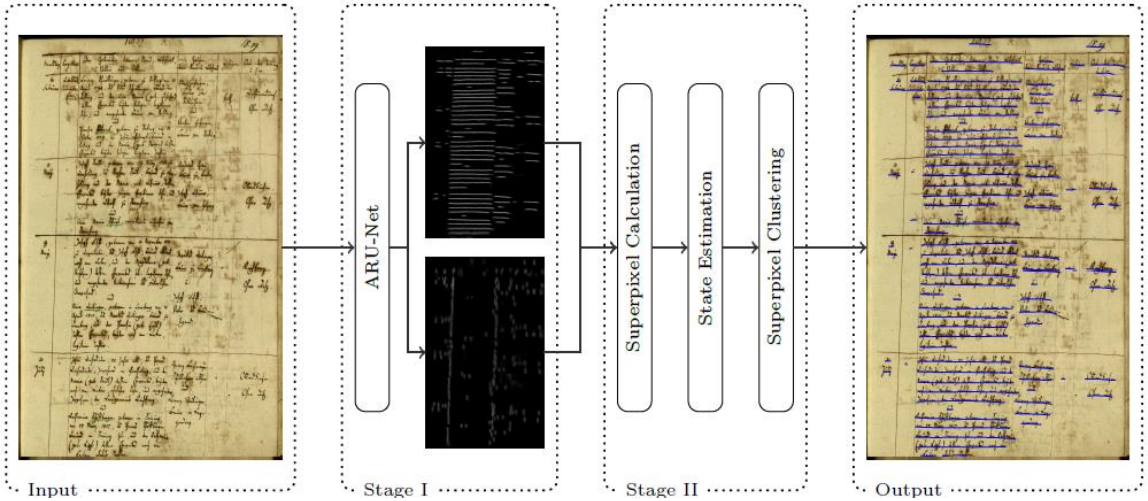
- Grüning et al.*: ARU-net: universal method for pixel labeling
  - The attention mechanism is added by successive passes on an image at different scales
  - Feature maps weighted by those of attention



*T. Grüning et al. "A Two-Stage Method for Text Line Detection in Historical Documents." *arXiv preprint arXiv:1802.03345* (2018)

The state of the art

- Grüning et al.*: (continuation)
  - Pre-processing
    - Pixel labeling: 3 classes: baseline, line separator, other
    - Network predictions ➔ bottom-up clustering to build baselines
  - Post-processing
    - Prediction is binarized before selecting points of interest
    - Use curvature and distance conditions to group points
  - Very effective: heavy & task specific



*T. Grüning et al. "A Two-Stage Method for Text Line Detection in Historical Documents." *arXiv preprint arXiv:1802.03345* (2018)

The state of the art

- Performance comparison

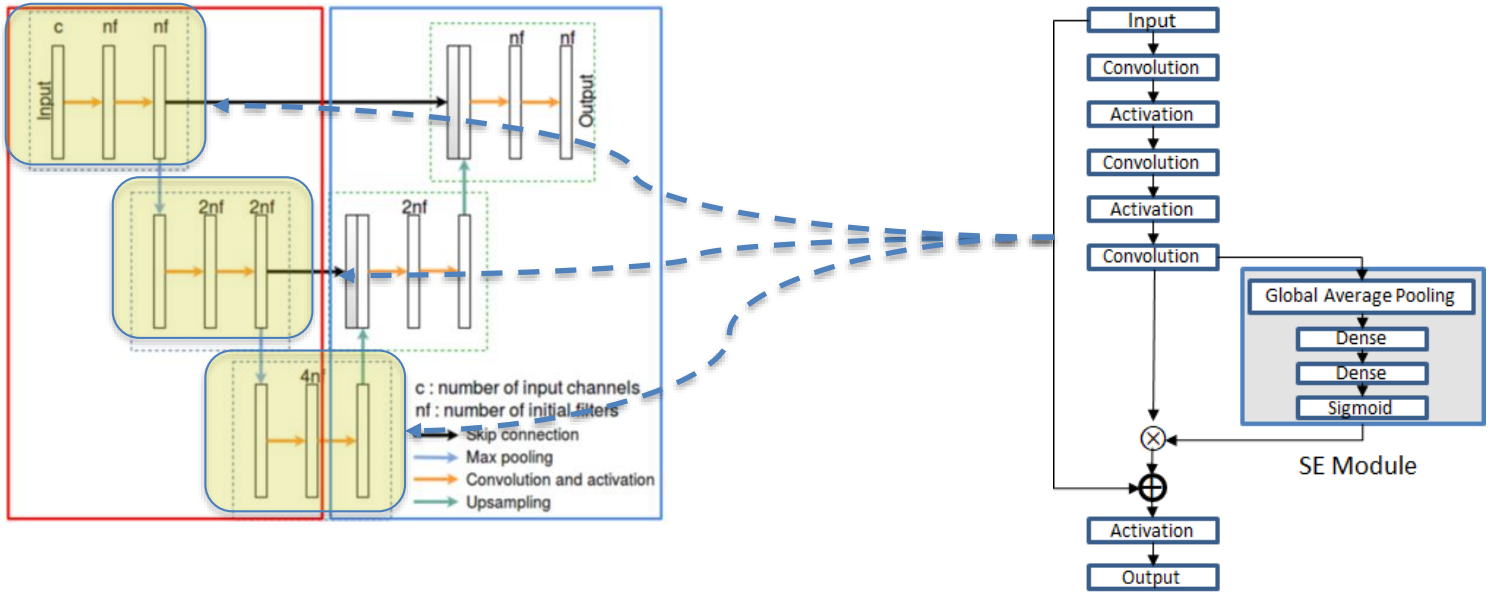
Method	Precision	Recall	FMeasure
Renton	78,0	83,6	90,7
Fink	97,3	97,0	97,1
Grüning	97,7	98,0	97,8
Grüning	96,15	96,36	96,26

* Result obtained from his code, with simple processing

Loria

Proposed approach

- Given the success of U-Net:
  - Added different modules
- First module: Squeeze & excitation (Hu et al. 2017)
  - A mechanism that allows the network to perform feature recalibration



1) Input Fmaps reduced to one (Global average pooling), 2) 2 dense layers to model the relation between feature maps, 3) a sigmoid to weight the original input FMaps

Hu et al. Squeeze-and-excitation networks. arXiv preprint arXiv :1709.01507, 7, 2017. 7, 36, 81

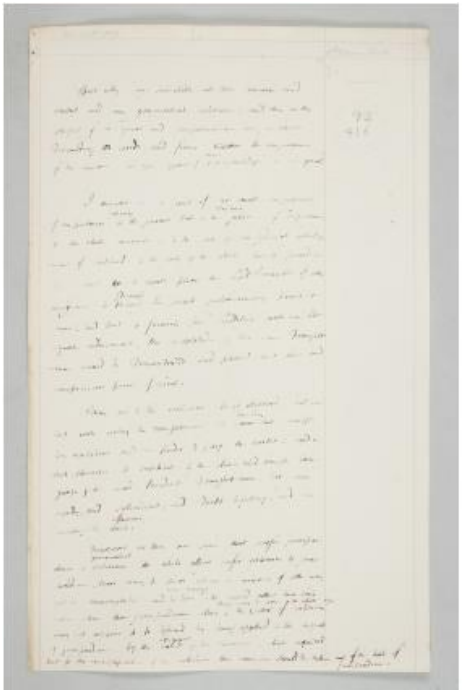
Proposed approach

- Second module: Post-processing
  - Simple clustering method to extract baselines, compared to Gruning
  - Different steps:
    1. Baseline prediction binarized: filter the low probabilities
    2. Mean Shift algorithm*: to group the points close to each other
    3. DBScan algorithm**: to group formed groups into larger ones
    4. Points of final groups plotted before extracting components
    5. Orientation of each of the related components is calculated
    6. Baselines calculated by performing polynomial regression from related components

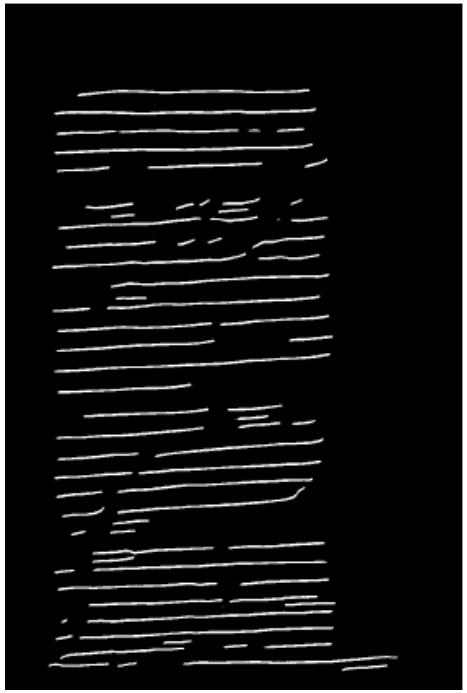
* Dorin Comaniciu and Peter Meer. Mean shift : A robust approach toward feature space analysis. IEEE TPAMI, 24(5) :603–619, 2002.

** Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, volume 96, pp. 226–231, 1996.

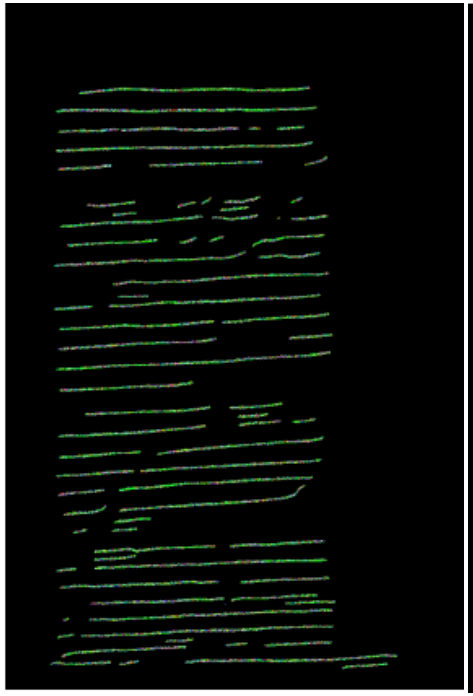
Experiments on cBAD



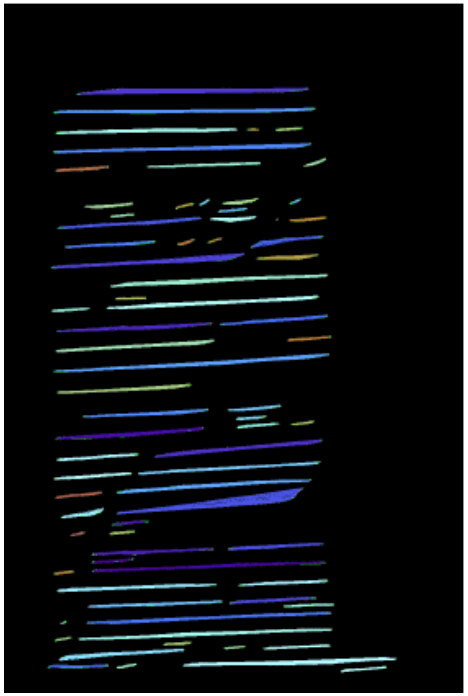
Original image



Prediction binarization

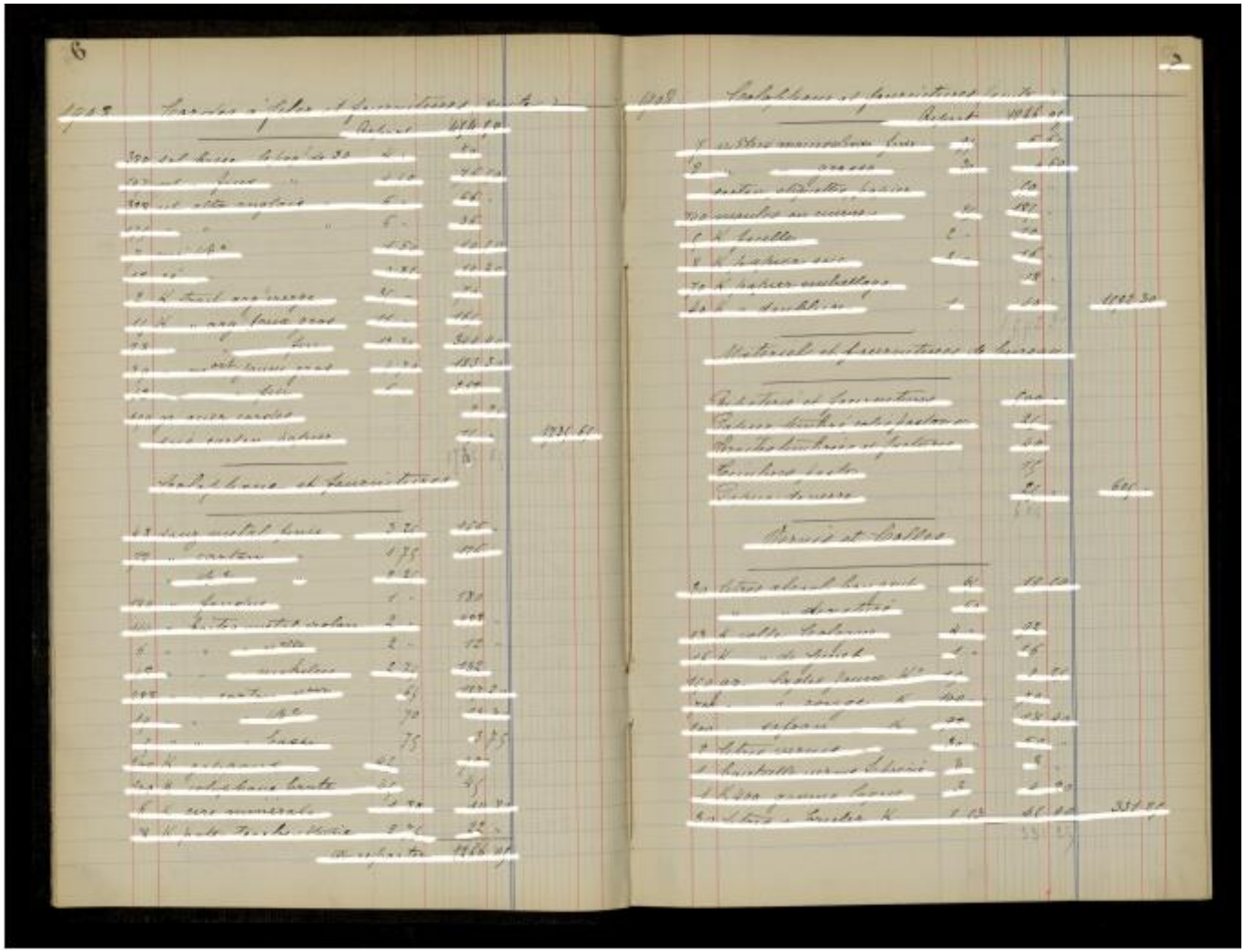


Mean shift clustering



DBScan

Experiments on MMP



Few missing lines, no graphic lines

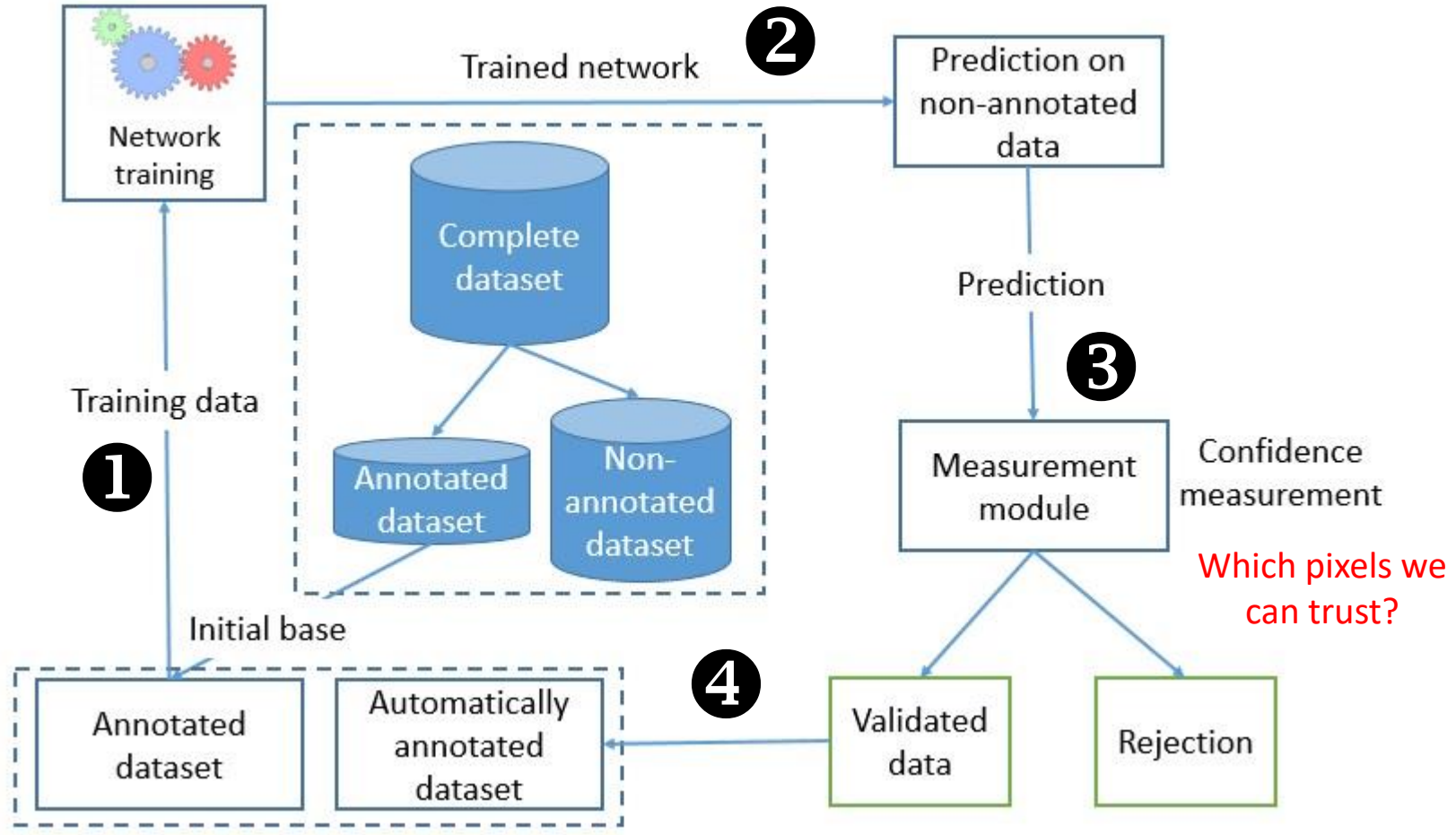
Experiments: varying the parameters

Different improvements have been reached

Exp.	Action	Precision	Recall	F-Measure	Cross-E-V
A	RU-net – SE	88,92	88,99	88,96	0,0416
B	RU-net + SE	90,16	89,61	89,89	0,0421
C	Crossed summed entropy replaced by average crossed entropy	88,42	90,52	89,46	0,0379
D	Attention distributed according to the characteristics extracted	89,91	90,37	90,14	0,0395
E	Impact of image resizing during learning	90,21	95,69	92,87	0,0337
F	Impact of image rotations	89,69	95,16	92,35	0,0357
G	Decreasing the learning rate	92,17	94,94	93,54	0,0351
H	Polynomial decreasing vs. exponential decreasing	89,24	95,23	92,14	0,0336
I	Curriculum learning	**92,47**	94,53	93,49	0,0243
J	Data augmentation	91,89	95,62	**93,62**	**0,1904**

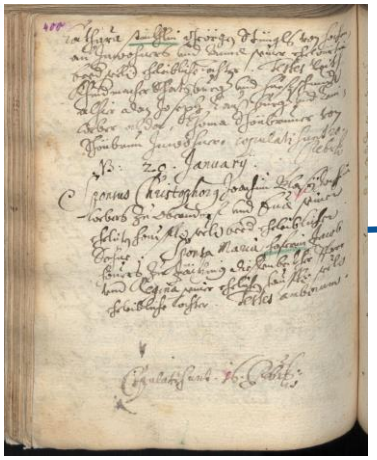
Semi-supervised learning

- First idea



Second idea: Generative adversarial network:
Goodfellow et al.*

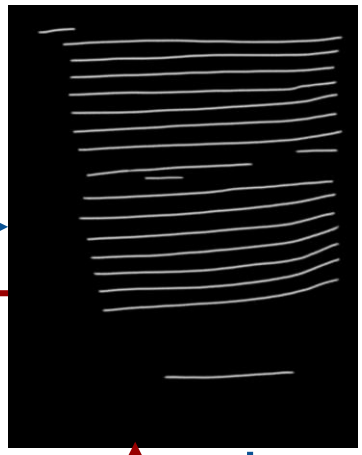
Raw data: non annotated



Must deceive the discriminator

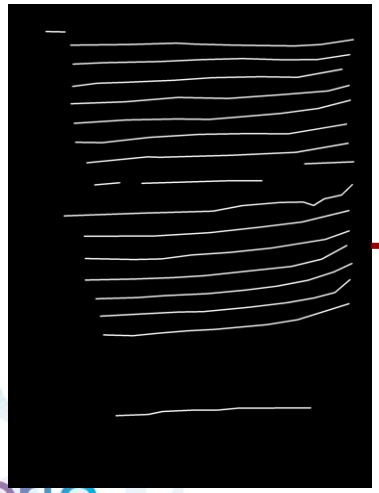
Generator

Baselines



Uses it to update its parameters

Annotated data



Decision transmitted to the generator

Learns the distribution of predictions

Discriminator

Gives the probability that the entry comes from the truth

Learns to make distinction between truth and prediction

* I. Goodfellowet al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

A min-max problem defined by V(G,D)

$$min_G\ max_D\ V(G,D) = E_{y\sim p_{data}(y)}[\log(D(y)] + E_{z\sim p_z(z)}[1 - \log(D(G(z))]$$

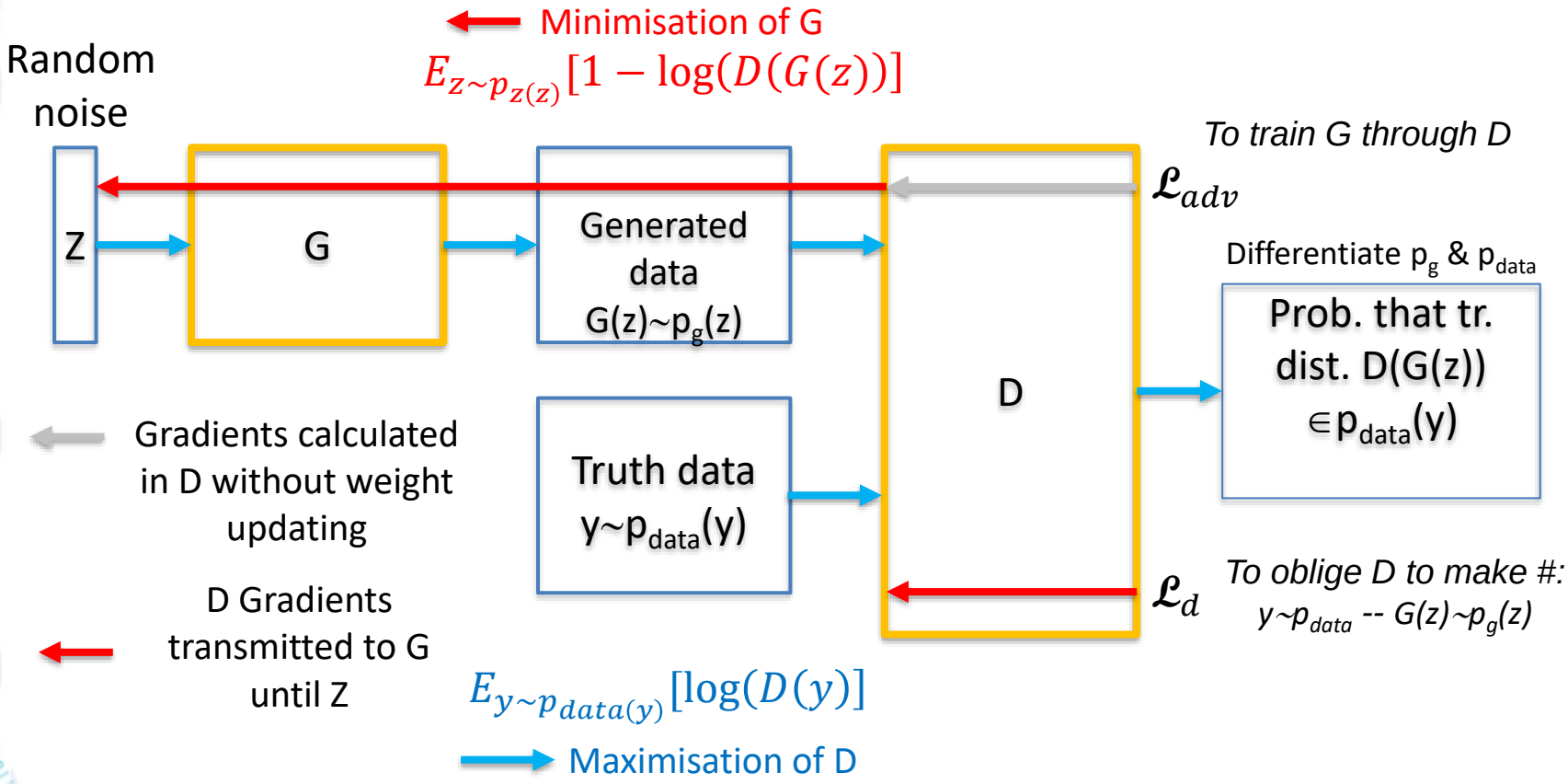
D must maximize the probability that an element of the truth belongs to the truth

G must minimize the error by deceiving the discriminator

- ? to find a balance between the two networks
- Generator not able to give predictions
  - ➔ the discriminator takes over
- Discriminator not able to make a good distinction
  - ➔ the generator will learn erroneous differences

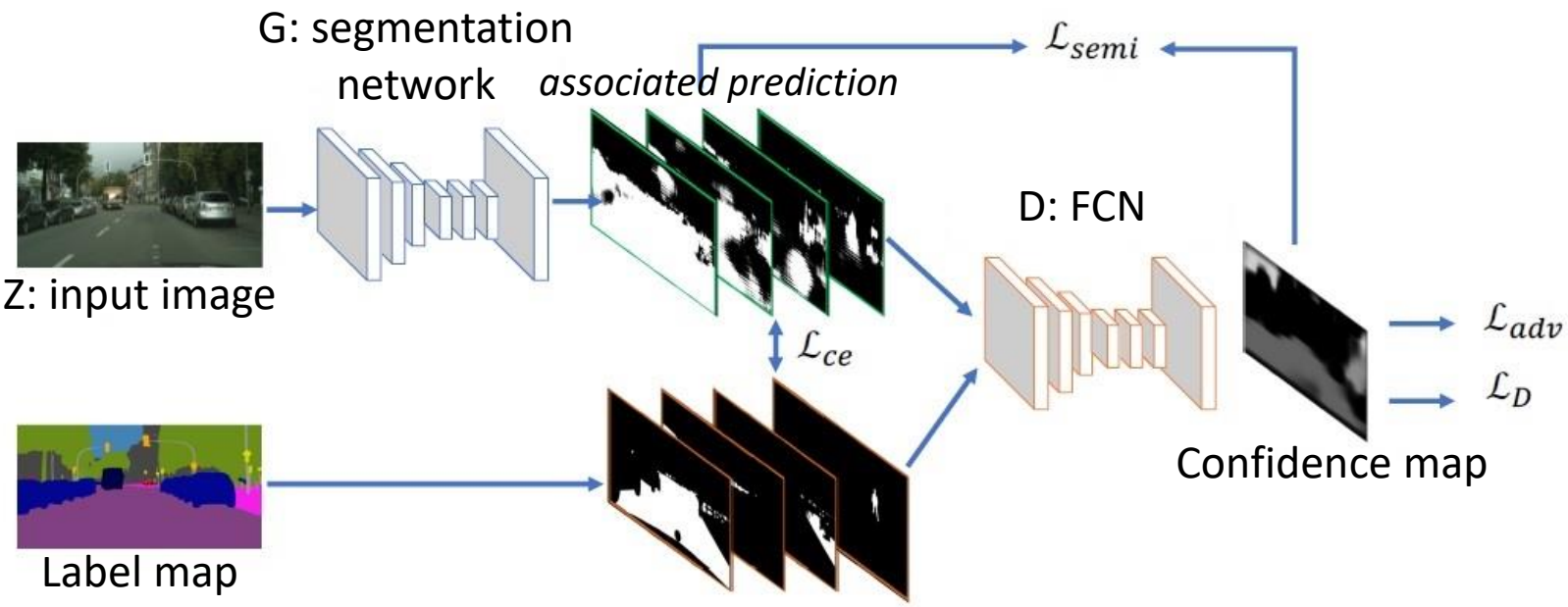
Loria

Schematic example



G absorbs # between its p_g distribution and p_{data} captured by D

Inspired by Hung et al*



- G: Segmentation network: any network designed for semantic labeling
- D: same role: make the # between predicted (baseline) & truth image
  - Trained using the loss $\mathcal{L}_{\mathcal{D}}$, not a unique value but a Proba for each pixel ➔ FCN
- They optimize the segmentation network using three loss functions
  - $\mathcal{L}_{ce}$ on the segmentation ground truth,
  - $\mathcal{L}_{adv}$ to fool the discriminator,
  - $\mathcal{L}_{semi}$ based on the confidence map

Loria

W-C. Hung, Y-H. Tsai, Y-T. Liou, Y-Y Lin, and M-H. Yang. Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv :1802.07934, 2018. 53, 54, 55, 57, 60, 73

In the configuration of Hung et al*

- Loss function of G:
  - $\mathcal{L}_g = \mathcal{L}_{ce} + \lambda_{adv}\, \mathcal{L}_{adv}$
    - $\mathcal{L}_{ce}$ = H(y,G(z)): cross entropy of G
    - λ_{adv}: weight given to the adversarial training
    - $\mathcal{L}_{adv}$ = H(y,D(G(z))): cross entropy of the adversarial training
- In addition, they consider:
  - D(G(z)): confidence measure in detecting reliable areas:
    - Reliable (correct) areas: have the same distribution as p_{data}

➔ a semi-supervised learning on non-annotated images

W-C. Hung, Y-H. Tsai, Y-T. Liou, Y-Y Lin, and M-H. Yang. Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv :1802.07934, 2018. 53, 54, 55, 57, 60, 73

System of Hung et al.

- In semi supervised learning:
  1. A non-annotated image z = S_n is labeled by the generator G(z)
  2. D: determines areas of the image that correspond to D(G(z))
     - Output of D thresholded to obtain a confidence mask :
       - I(D(G(z)) > T_d, with I(.) the indicator function
     - Prediction of G: I(G(z)) > T_g= $\hat{Y}$ to obtain the class of each pixel

$$\mathcal{L}_{semi} = \sum_{i=1}^{m} \sum_{i=1}^{|C|} \text{I(D(G(z)) > T)}.\ \hat{Y} \times \log(\text{G(z)})$$

m: number of areas

|C|: number of classes

- $\mathcal{L}_g = \mathcal{L}_{ce} + \lambda_{adv}\, \mathcal{L}_{adv} + \lambda_{semi}\, \mathcal{L}_{semi}$

Our System

- Same loss functions as in Hung et al.
  - The cross-entropy replaced by the mean cross-entropy
    - ($\mathcal{L}_{ce}$ is different)
  - We use the average entropy
    - the images are not of the same size, we give them the same size
    - If we average, we consider that the images are of equal importance
- The architectures of G and D are different

Loria

Experiments

- How the system can be driven by this architecture ?
- Start: Impact of the number of elements, in conventional config.
  - Dataset : cBAD-A: 755 images, 522 for training
  - Training data: 100% 50% and 25%
    - Even with half, the results are close to L0 exp. with full data

➔ the network has a good ability to generalize with few data

Experiment	Quantity (%)	P (%)	R(%)	F(%)
L0	100	91.91	95.6	93.62
L1	50	91.26	94.85	93.02
L2	25	85.6	93.53	89.4

Experiments: Adversarial & semi-supervised training

- 4 exp. to analyze the impact of the adversarial training by varying…

E	Q	Dx	λ_{adv}	λ_{semi}	P	R	F	Comment
L0	100	-	-	-	91.91	95.6	93.62	
L4	25	No	0.01	0	87.2	94.29	90.6	Add Gaussian noise to baselines, to make them more resembling to predictions
L5	25	Yes	0.01	0	87.74	94.82	91.4	Test if concatenation of the source image with its prediction (Dx=yes) helps D to determine correct regions
L6	25	Yes	0.001	0	84.26	95.47	89.51	Use 0.001 to see if decreasing of Lambda decreases perf.
L7	50	Yes	0.01	0.1	91.75	95.5	93.59	Noticed that similar perf. to using 50% data

Conclusion

- Lot of experiments around U-Net
  - Training by adversarial networks improves the quality of predictions
- Consequent gain of performance with few data
  - First objective reached
- Our work lays the foundations
  - for the development of a semi-supervised learning method (use of non-labeled images)
- The results obtained on the documents of the Philharmonie Museum in Paris are promising

  > 90%, using just 200 labeled images on 11 000

Loria