Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks

Mélodie Boillet[1,2], Christopher Kermorvant[1,2] and Thierry Paquet[2]

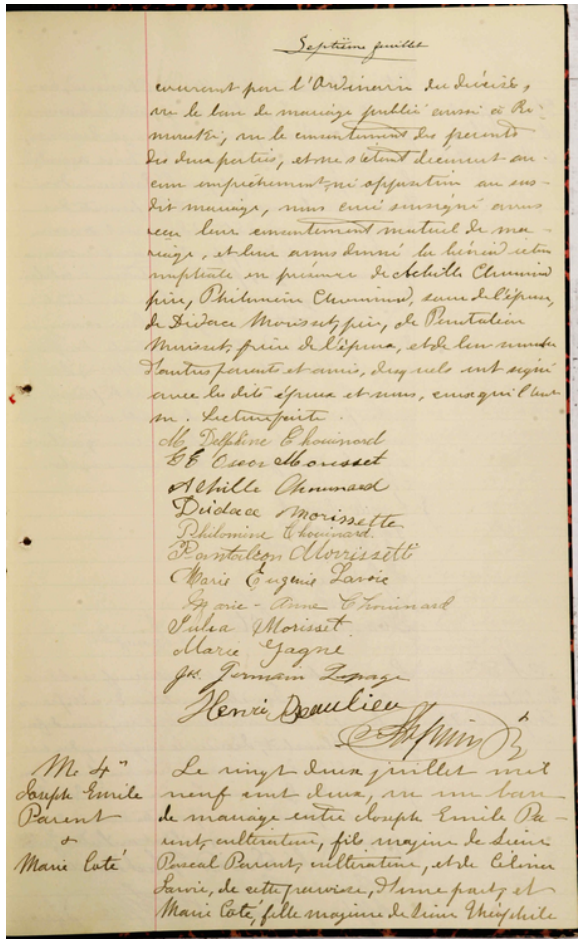
[1]Teklia SAS, Paris, France
[2]LITIS, Rouen-Normandy University, France

SIFED - 6th July 2020

Table of contents

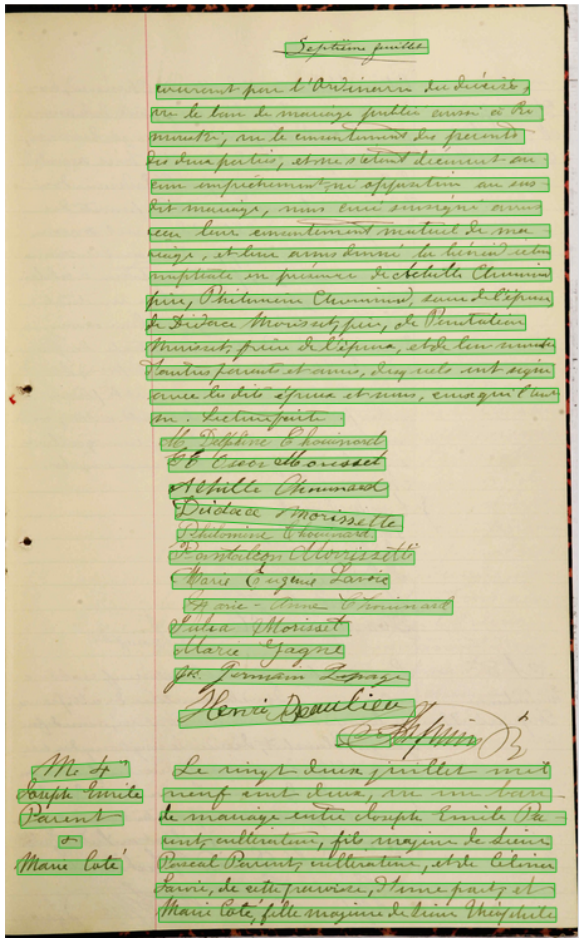
Text line segmentation

- **Goal:** detect the text lines of an image;
- Application: apply a text recognition system on the detected text lines.



Text line segmentation

- **Goal:** detect the text lines of an image;
- Application: apply a text recognition system on the detected text lines.



State-of-the-art

- dhSegment [Oliveira2018]:
  - CNN + Resnet-50 pre-trained on ImageNet;
  - Multi-task: text line detection, ornament detection...;
  - Good results on various datasets: DIVA, cBAD...
- Yang et al. [Yang2017]:
  - Multimodal FCN + use of the text content;
  - Good results on modern documents: DSSE, SectLabel...
- Moysset et al. [Moysset2015]:
  - Recurrent network;
  - Text line segmentation of paragraphs;
  - 2 labels: line and interline.

Problems of dhSegment

- Needs a lot of annotated data;
- Good results but can still be improved;
- Too long to analyse a whole corpus:
  ∼66 days for 2M images (Balsac dataset).

Is pre-training on natural scene images the most suitable for working on document images ?

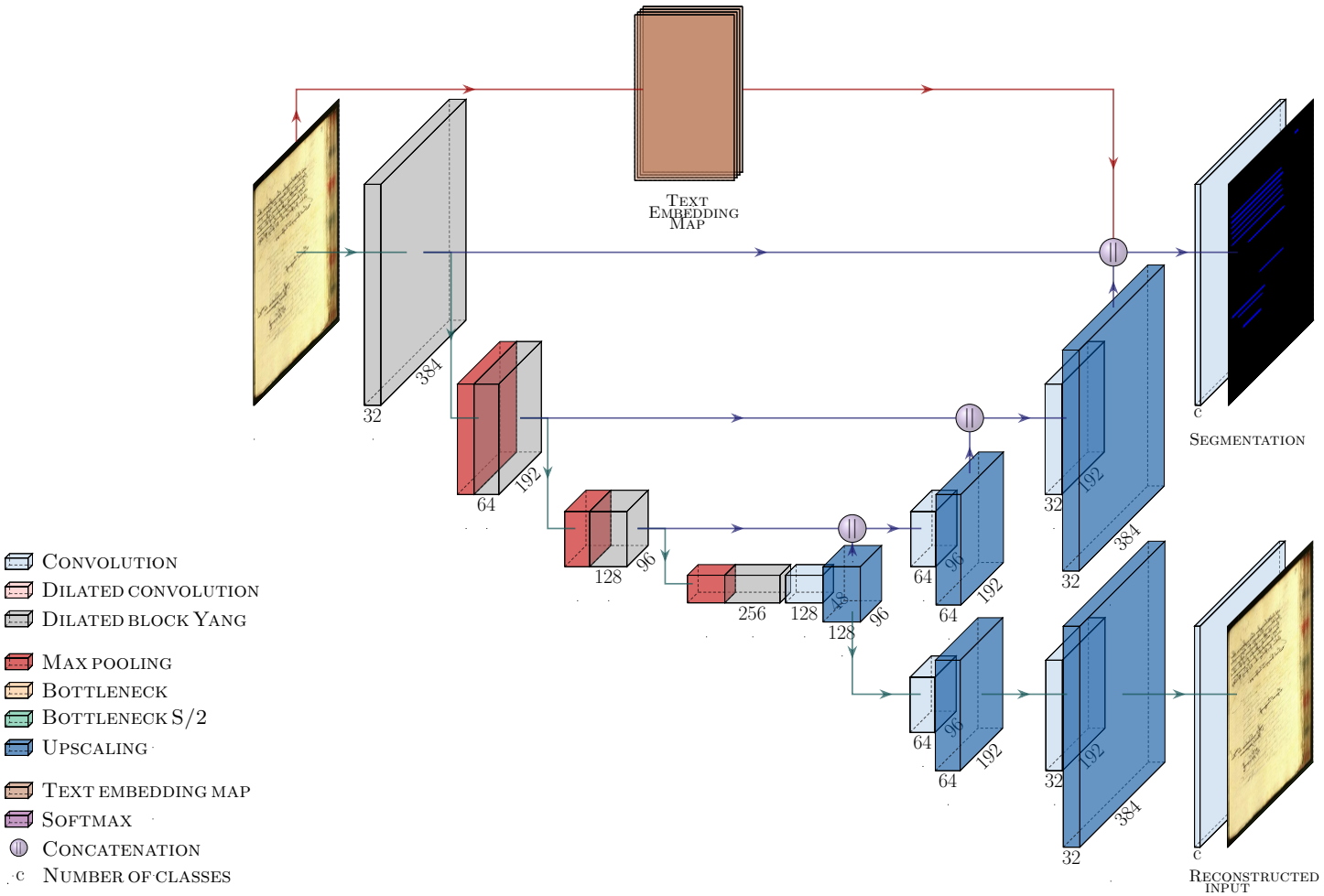
Main goal

Analyse the impact of a pre-training step on the line segmentation task.

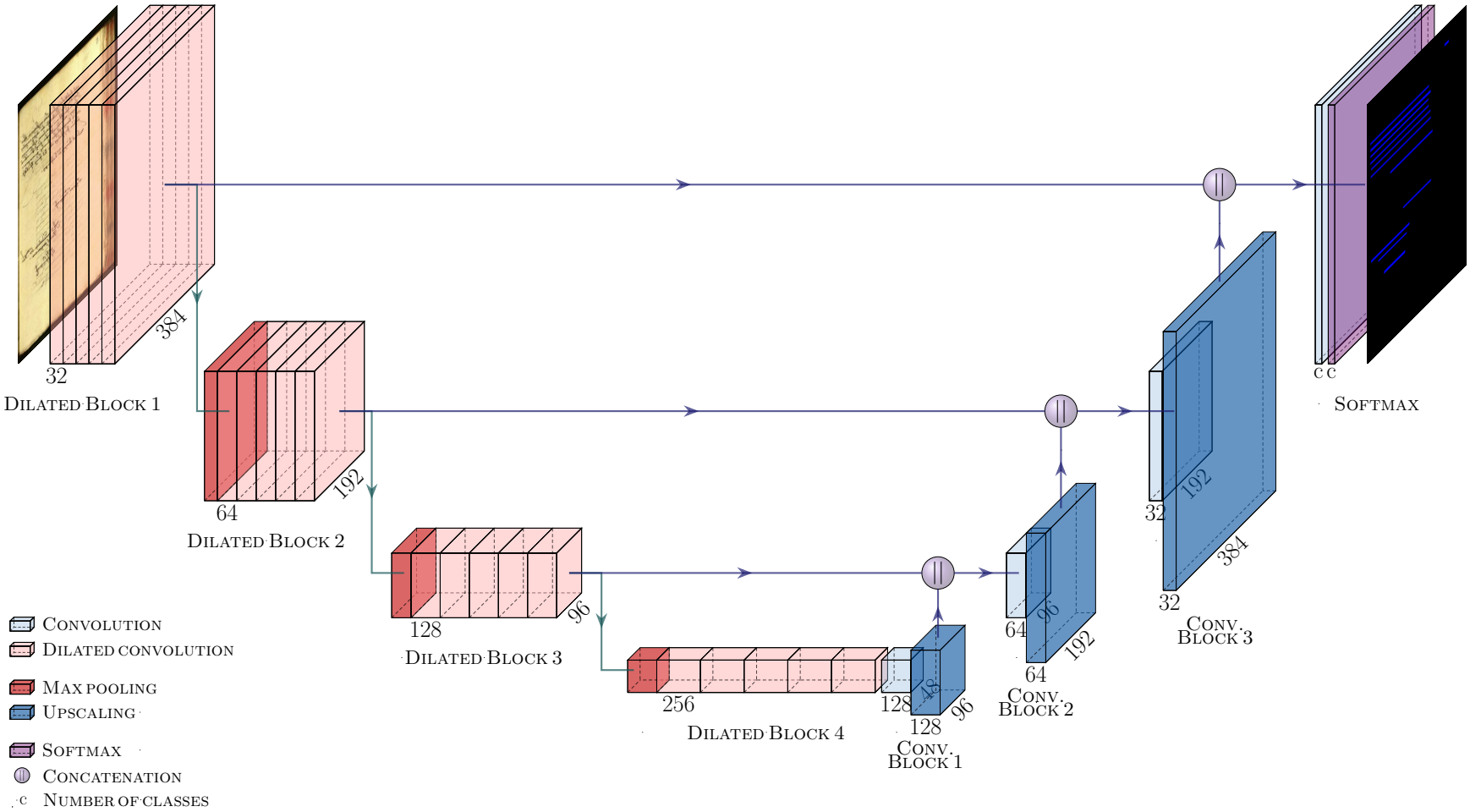
We want to develop a model:

- Containing no pre-trained part learnt on natural scene images;
- Having less parameters than dhSegment and a reduced prediction time;
- Yielding higher accuracy than dhSegment.

Yang's architecture [Yang2017]



Architecture of our model

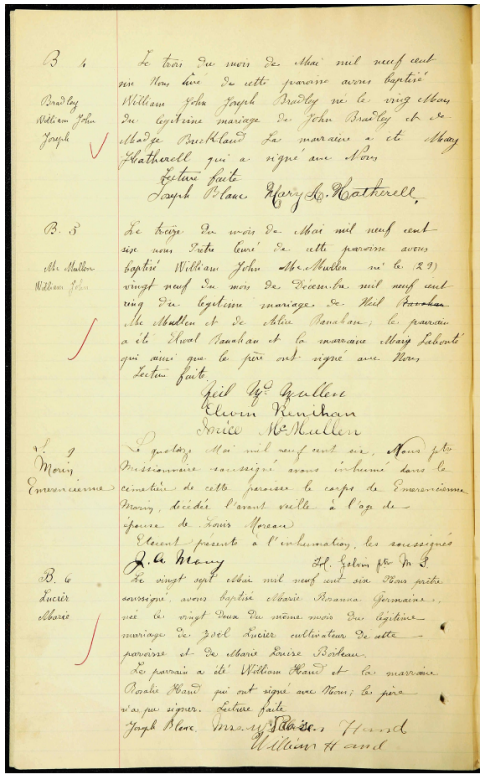


Implementation details

- Input image size: *384*x*384 px*;
- Batch normalization + dropout layers after all convolutions;
- Use of concatenations to help the detection of small objects;
- Post-processing: thresholding + removal of small connected components.

Datasets

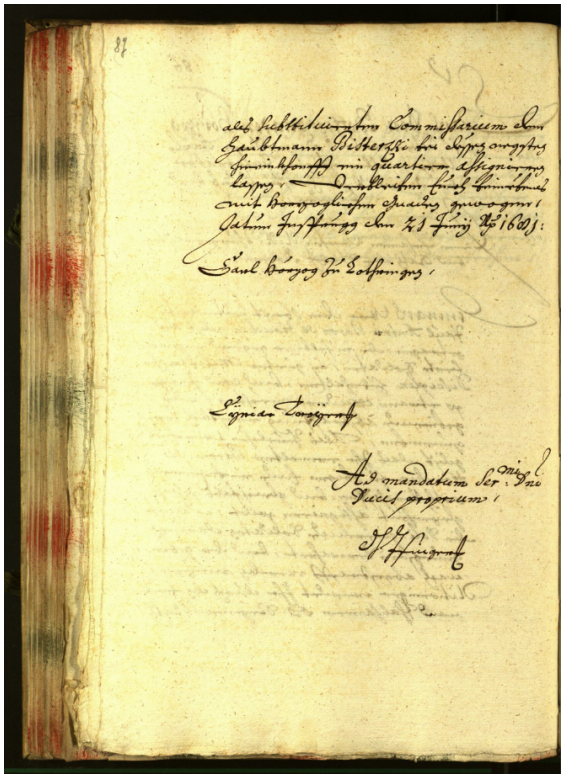
Balsac:
913 annotated images



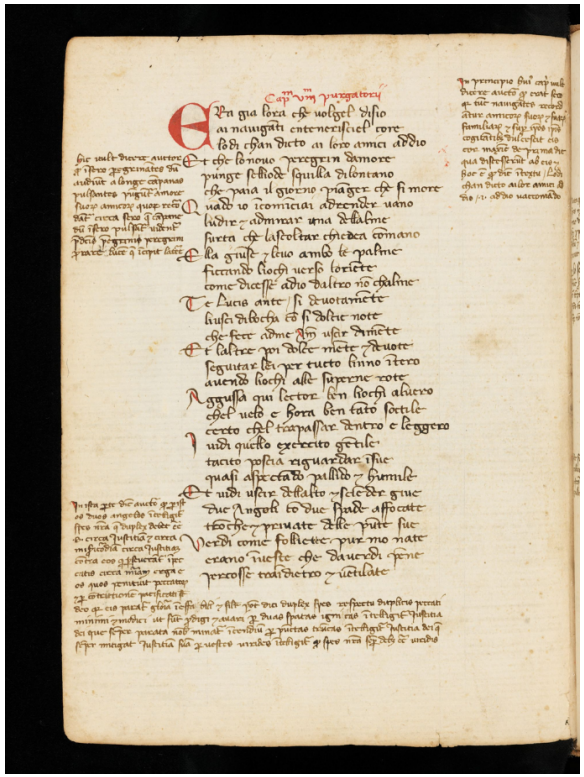
Horae [Boillet2019]:
557 annotated images

Datasets II

READ-BAD [Grüning2017]:
2036 annotated images



DIVA-HisDB [Simistira2016]:
120 annotated images



Comparison with dhSegment

Dataset	Model	IoU	Pr	Rec	F1	Time[1]
Balsac	dhSegment	73.78	92.07	78.76	84.81	66.3
	Our	**83.79**	**94.80**	**87.86**	**91.11**	**9.2**
Horae	dhSegment	**65.22**	71.70	**89.29**	82.32	18.8
	Our	63.95	**78.38**	80.45	**84.93**	**2.3**
READ-Simple	dhSegment	**64.55**	**85.04**	71.85	**77.25**	8.4[2]
	Our	64.03	81.76	**75.60**	76.66	**1.0**[2]
READ-Complex	dhSegment	52.91	79.28	59.16	69.27	10.6[2]
	Our	**54.40**	**83.62**	**61.97**	**73.16**	**1.3**[2]
DIVA-HisDB	dhSegment	74.24	**92.41**	79.10	85.19	NA
	Our	**75.71**	92.14	**80.88**	**86.09**	NA

	dhSegment	Our
Number of parameters	32.8M(9.36M)	**4.1M**

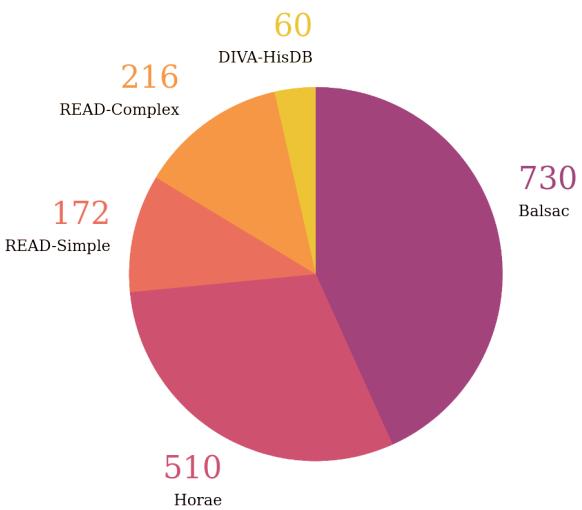
[1] Prediction time (GPU GeForce RTX 2070 8G) in days to analyse the whole corpus.

[2] Estimation based on the manuscripts sizes without *BHIC* and *Unibas*.

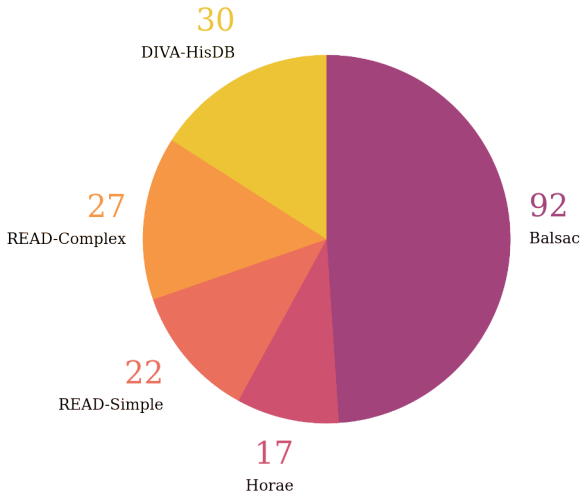
Multiple document dataset

Does pre-training on document images improve the performances ?

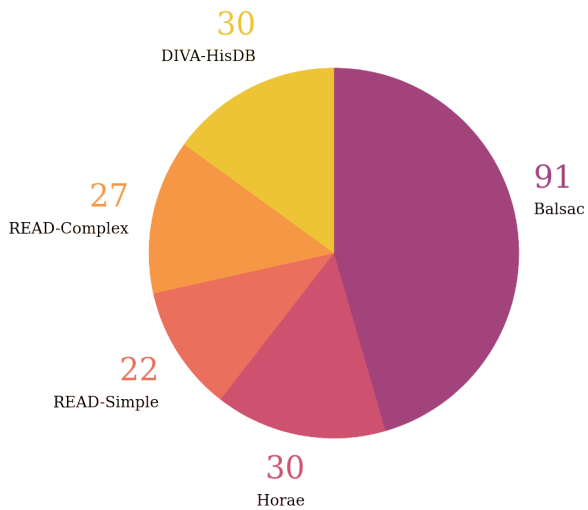
Training



Validation



Test



Comparison with dhSegment

Data	Model	IoU	Pr	Rec	F1
Balsac	dhSegment	73.78	92.07	78.76	84.81
	dhSegment PT	74.02	91.89	79.09	84.95
	Our	83.79	**94.80**	87.86	91.11
	Our PT	**84.87**	94.25	**89.49**	**91.75**
Horae	dhSegment	65.22	71.70	**89.29**	82.32
	dhSegment PT	60.69	**80.94**	73.65	81.99
	Our	63.95	78.38	80.45	84.93
	Our PT	**68.81**	80.31	84.80	**88.62**
READ-Simple	dhSegment	64.55	85.04	71.85	77.25
	dhSegment PT	65.07	**88.34**	71.56	**80.72**
	Our	64.03	81.76	75.60	76.66
	Our PT	**68.14**	83.19	**78.05**	79.45
READ-Complex	dhSegment	52.91	79.28	59.16	69.27
	dhSegment PT	53.34	**85.51**	57.80	68.47
	U-FCN	54.40	83.62	61.97	73.16
	U-FCN PT	**60.28**	81.03	**68.17**	**78.30**
DIVA-HisDB	dhSegment	74.24	**92.41**	79.10	85.19
	dhSegment PT	73.00	91.56	78.28	84.32
	U-FCN	**75.71**	92.14	80.88	**86.09**
	U-FCN PT	74.72	89.43	**82.20**	85.44

Conclusion

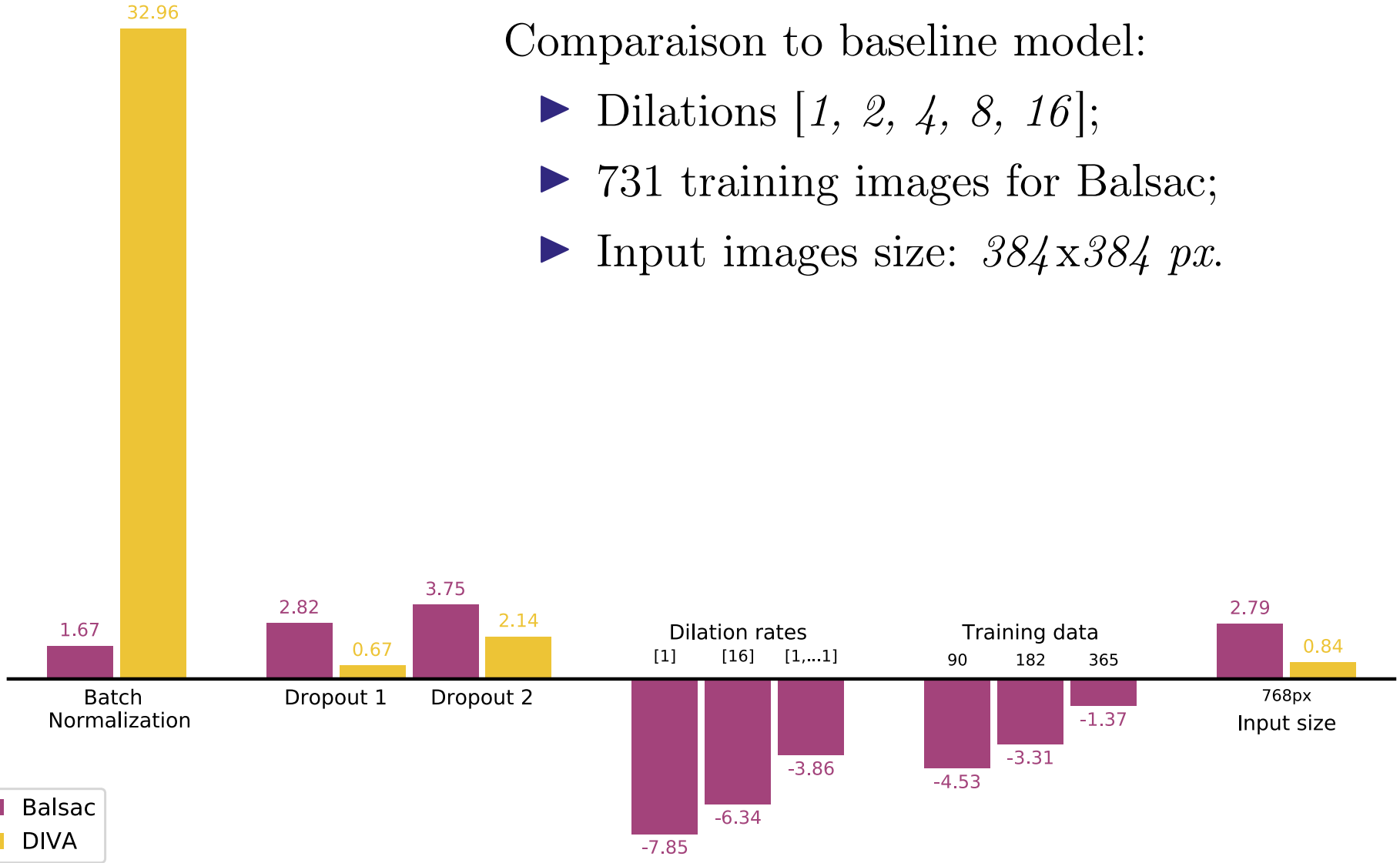
Does pre-training on document images improve the performances ?
YES

Intersection-over-Union :

- ✓ +5 percentage points on Horae and READ-Complex;
- ✓ +4 percentage points on READ-Simple;
- ≈ Similar performances on Balsac;
- ✗ −1 percentage point on DIVA-HisDB.

Our results are overall better than dhSegment's (except for the precision metric).

Ablation study



Comparaison to baseline model:

- Dilations [*1, 2, 4, 8, 16*];
- 731 training images for Balsac;
- Input images size: *384*x*384 px.*

Ablation study II

- Adding Batch Normalization and Dropout layers after all convolutions improves the performances;
- Using dilation rates [*1, 2, 4, 8, 16*] allows to have a bigger receptive field;
- Adding more training images and using bigger input images also improves the results.

BUT still better results (77.42 %) than SOTA dhSegment (73.78 %) with only 90 images.

Conclusion

We designed a model:

- Lighter than dhSegment;
- Giving most of the time better results;
- Having a reduced prediction time: up to 8 times faster.

Future work:

- Test our architecture on datasets with more than 2 classes;
- Build an historical model trained on various historical documents.

Bibliography

[Oliveira2018] Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. "dhSegment: A generic deep-learning approach for document segmentation". In: *Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on*. IEEE. 2018, pp. 7–12

[Yang2017] Xiao Yang et al. "Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Network". In: vol. abs/1706.02337. 2017. arXiv: 1706.02337. URL: http://arxiv.org/abs/1706.02337

[Moysset2015] B. Moysset et al. "Paragraph text segmentation into lines with Recurrent Neural Networks". In: *2015 13th ICDAR*. Aug. 2015, pp. 456–460. DOI: 10.1109/ICDAR.2015.7333803

[Boillet2019] Mélodie Boillet et al. "HORAE: An Annotated Dataset of Books of Hours". In: *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*. HIP '19. Sydney, NSW, Australia: Association for Computing Machinery, 2019, 7–12. ISBN: 9781450376686. DOI: 10.1145/3352631.3352633. URL: https://doi.org/10.1145/3352631.3352633

[Grüning2017] Tobias Grüning et al. "READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents". In: *CoRR* abs/1705.03311 (2017). arXiv: 1705.03311. URL: http://arxiv.org/abs/1705.03311

[Simistira2016] F. Simistira et al. "DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts". In: *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Oct. 2016, pp. 471–476. DOI: 10.1109/ICFHR.2016.0093

Thresholding

Post-processing after dhSegment:

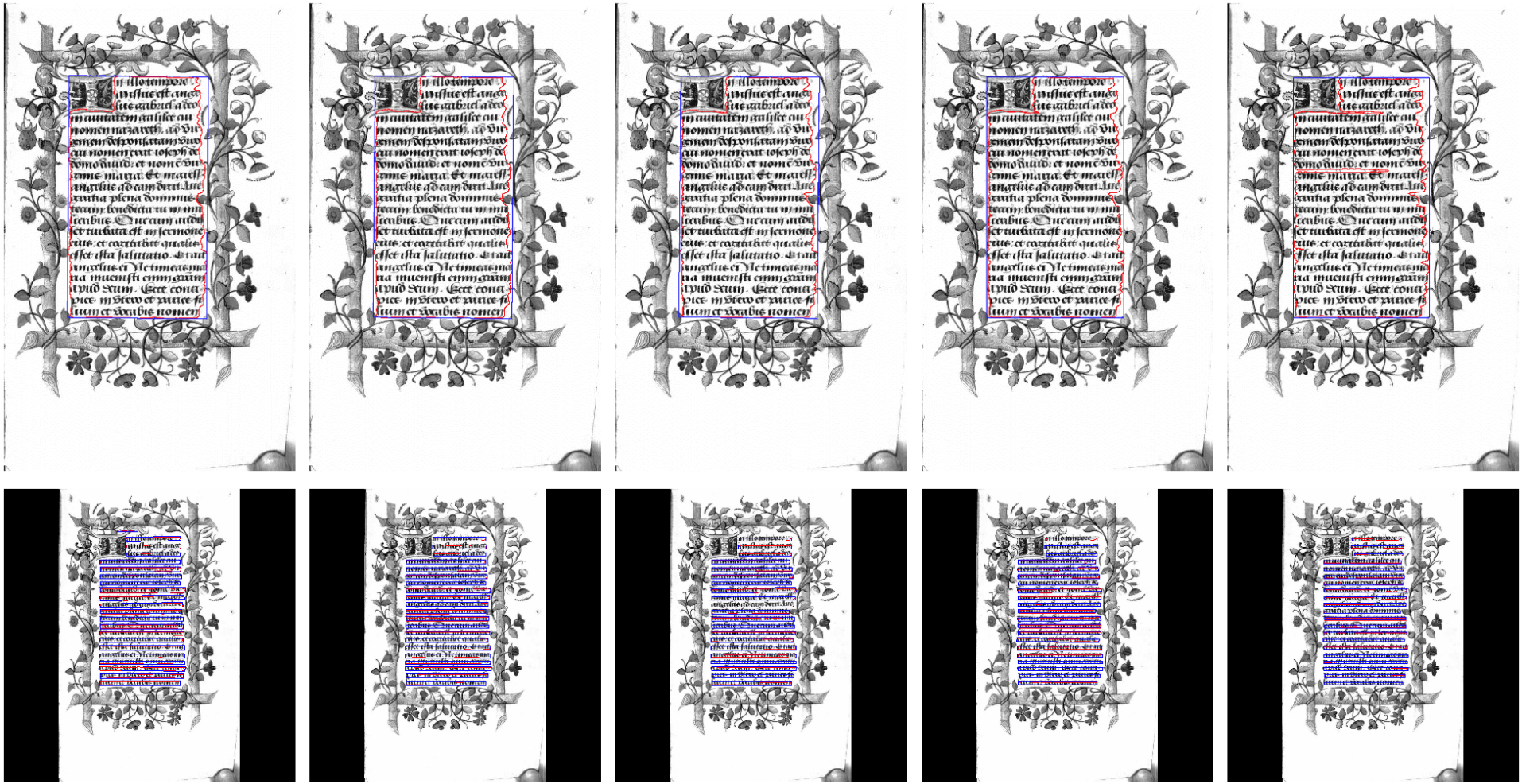
- Probabilities thresholding to keep the highest;
- Removal of the small connected components.

To be comparable, we used the same post-processing after our model.

Does this thresholding have a real impact on our results ? If so, how to optimize this threshold ?

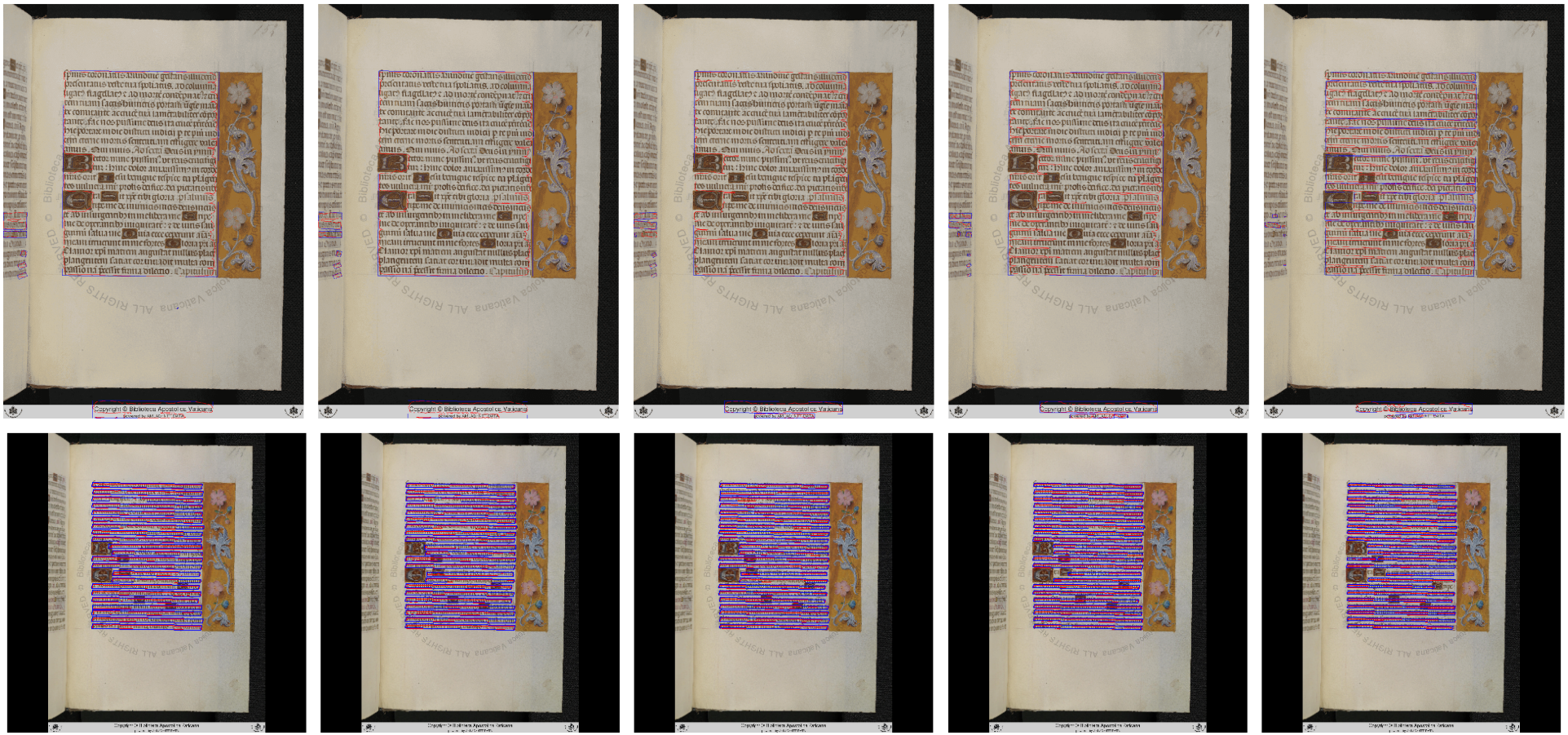
Thresholding impact

0.5 0.6 0.7 0.8 0.9



Thresholding impact

0.5 0.6 0.7 0.8 0.9



Thresholding impact II

- Choice of the threshold essential for dhSegment;
- Thresholding step obsolete for U-FCN:
  - $\rightarrow$ no need to choose a threshold;
  - $\rightarrow$ assign the class having the highest probability.