Investigating the Effect of using Synthetic and Semi-synthetic Images for Historical Document Font Classification

Konstantina Nikolaidou[1], Richa Upadhyay[1], Mathias Seuret[2], Marcus Liwicki[1]
[1] *Machine Learning Group,* Luleå University of Technology, Sweden, firstname.lastname@ltu.se
[2] *Pattern Recognition Lab*, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany, firstname.lastname@fau.de

Highlights

Use font classification dataset of early printed books [21]

Pre-process it to get image crops

Create baseline dataset of 10K crops

Investigate SOTA pre-trained CNN architectures on baseline

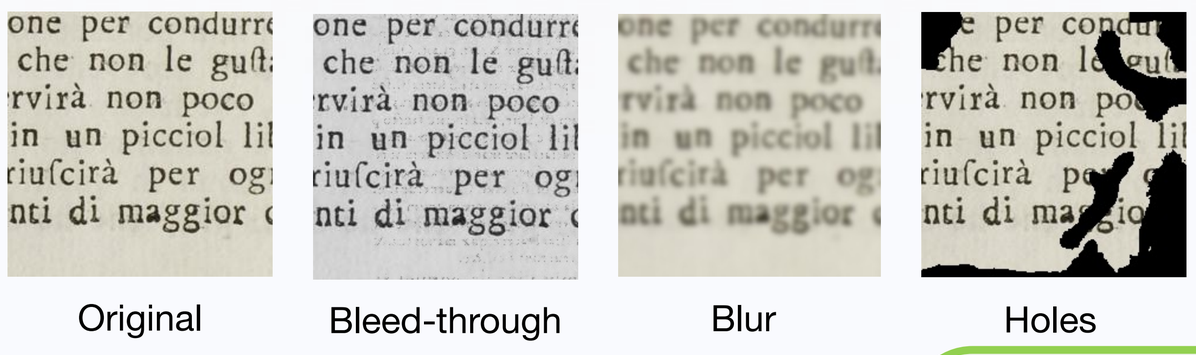
and on baseline with additional 60K:

- semi-synthetic images from DocCreator [10]
- synthetic images from OpenGAN [4]
- real data samples

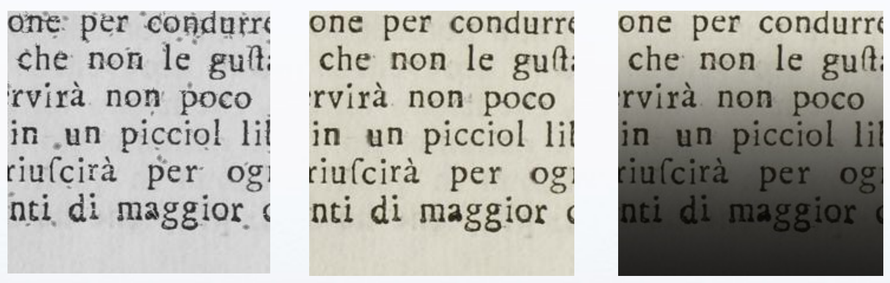
Dataset	ResNet50	DenseNet	EfficientNet
Baseline	95.78 ± 0.55	96.91 ± 0.41	96.79 ± 0.10
+DocCreator	**97.93 ± 0.21**	**98.30 ± 0.69**	**97.85 ± 0.24**
+GAN	95.41 ± 0.33	96.81 ± 0.43	96.40 ± 0.42
+Real	96.57 ± 0.23	96.89 ± 0.54	97.16 ± 0.12

Classification **accuracy** on the competition test set.

Semi-synthetic data

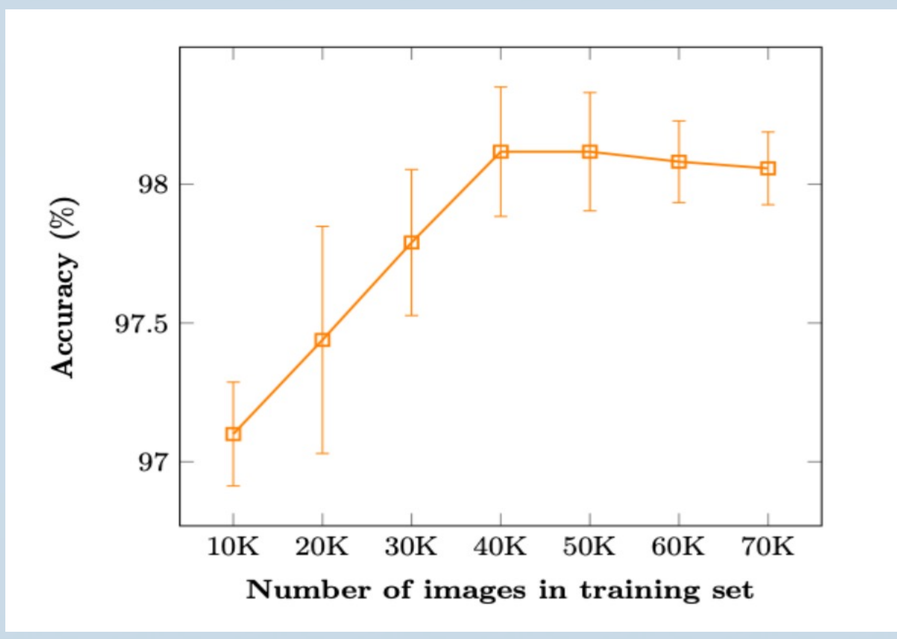


Original | Bleed-through | Blur | Holes



Character degradation | Phantom character | Shadow

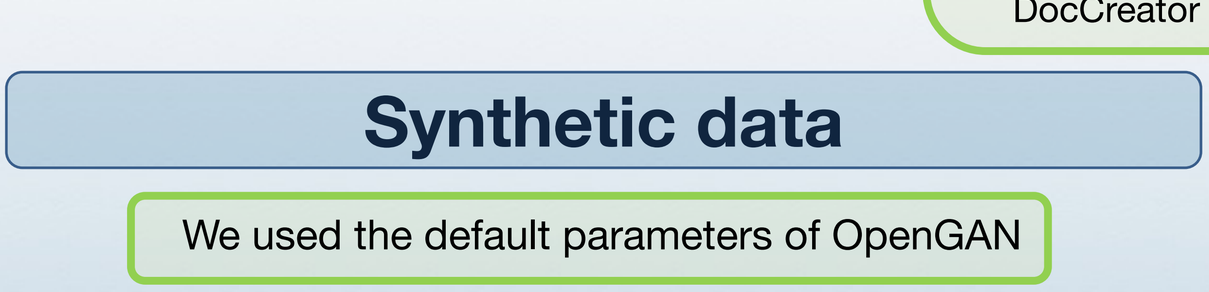
We created 1 instance per degradation (6 total) using random parameter values in DocCreator



DenseNet accuracy starting with the 10K baseline samples and with gradual addition of DocCreator samples.

Synthetic data

We used the default parameters of OpenGAN



Key Takeaway

Domain knowledge provided by **DocCreator** degradations **surpasses** fully synthetic data created by a **GAN** and even contributes to the performance more than **real data**.

References

[21] Seuret et al. (2019). Dataset of Pages from Early Printed Books with Multiple Font Groups. *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*.
[10] Journet, et al. (2017). DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images. *J. Imaging, 3*, 62.
[4] Ditria et al. (2020). OpenGAN: Open Set Generative Adversarial Networks. *ACCV*.

https://github.com/koninik/Font_Classification