Paired Image to Image Translation for Strikethrough Removal from Handwritten Words

Raphaela Heil, Ekta Vats, Anders Hast

✉ raphaela.heil@it.uu.se

15th IAPR International Workshop on Document Analysis Systems 2022

2022-05-24, La Rochelle, France

UPPSALA
UNIVERSITET

Strikethrough Removal

removing strikethrough strokes, returning the original, clean words

Primary use case:

research questions for which struck-through (deleted) words are of interest, e.g. **genetic criticism**

→ pre-processing for humans and machines

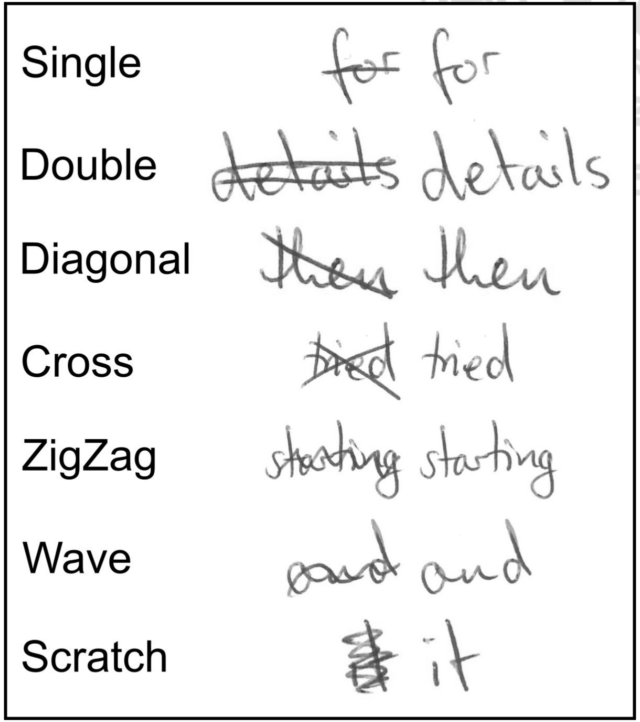


Fig.: Examples of different types of struck-through words and respective clean ground truth.

General Approach

- paired image to image translation
  - from source domain (struck words) to target domain (clean words)
- segmentation-based, i.e. focus on struck-through words

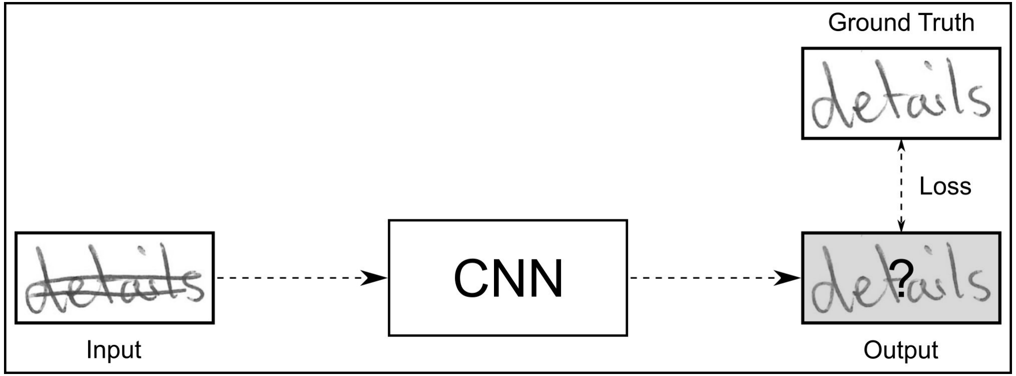


Fig.: Schematic view of paired image to image translation for strikethrough removal.

Examined Models

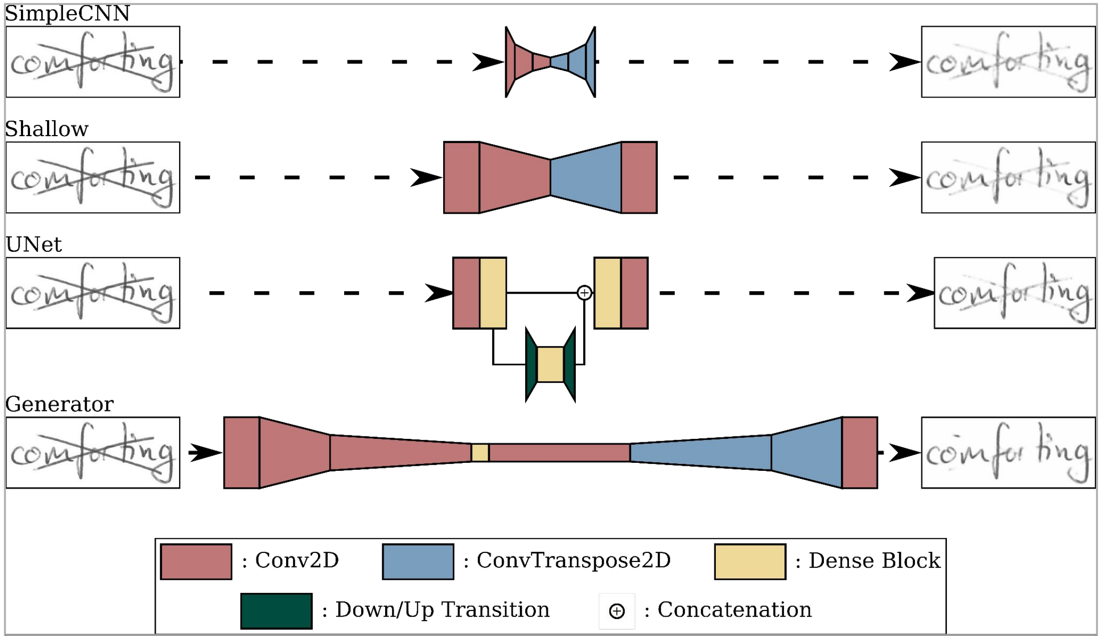


Fig.: Schematic overview over the four architectures, examined in a paired setting.

+ Attribute-Guided CycleGAN from [1]

→ unpaired approach

[1] Heil, R., Vats, E., Hast, A. (2021). Strikethrough Removal from Handwritten Words Using CycleGANs. ICDAR 2021

Examined Models

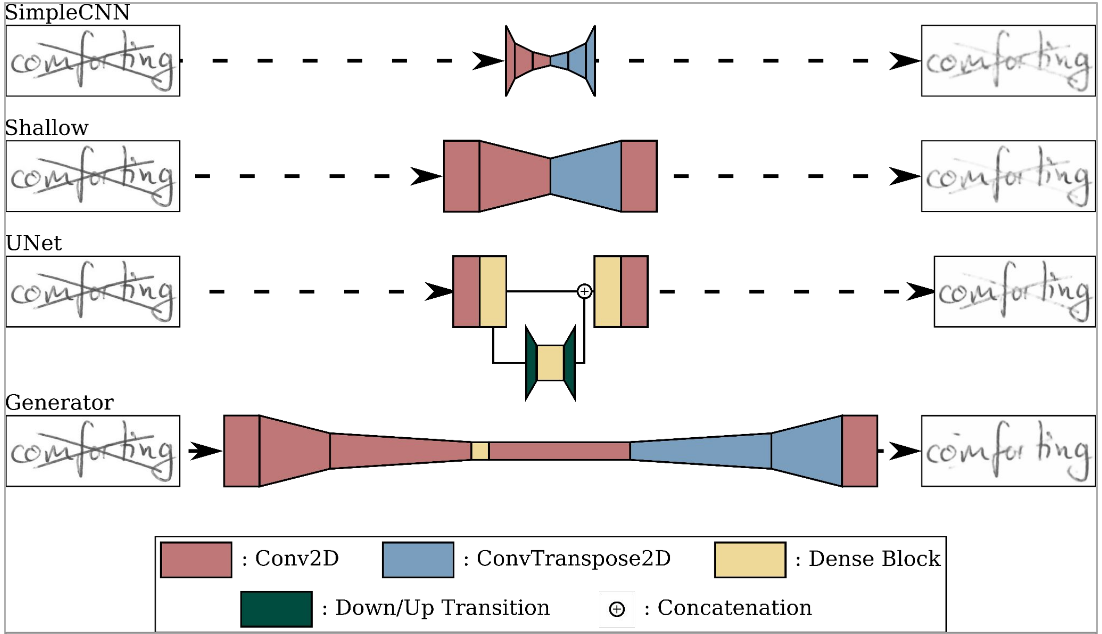


Fig.: Schematic overview over the four architectures, examined in a paired setting.

Model name	Parameter count
SimpleCNN	28 065
Shallow	154 241
UNet	181 585
Generator	1 345 217
Attribute-guided CycleGAN [1]	8 217 604

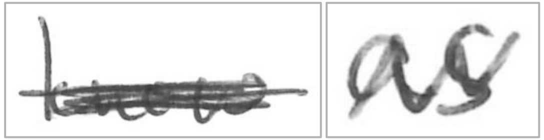
[1] Heil, R., Vats, E., Hast, A. (2021). Strikethrough Removal from Handwritten Words Using CycleGANs. ICDAR 2021

Datasets

- focus on synthetic data for training
  - ➢ genuine paired strikethrough data not easily obtained from conventional manuscripts
- evaluation on genuine data

- segmentation-based
- greyscale
- background removed

Datasets

Dataset	**IAM$_{synth}$**	**Dracula$_{real}$**	**Dracula$_{synth}$**
General Info	subset of IAM + synthetic strikethrough	handwritten copy of a portion of Bram Stoker's Dracula write → scan → strike through → scan → align	clean words from Dracula$_{real}$ train + synthetic strikethrough 5 sets of random strikethrough stroke augmentations per word
writers	multi (no set overlap)	single	single
# train	3066	126	5 x 126 and 630
# validation	273	126	N/A
# test	819	378	N/A
samples			

Experiments

1. Train on IAM_{synth} — test on IAM_{synth}
   ➢ base comparison with [1]

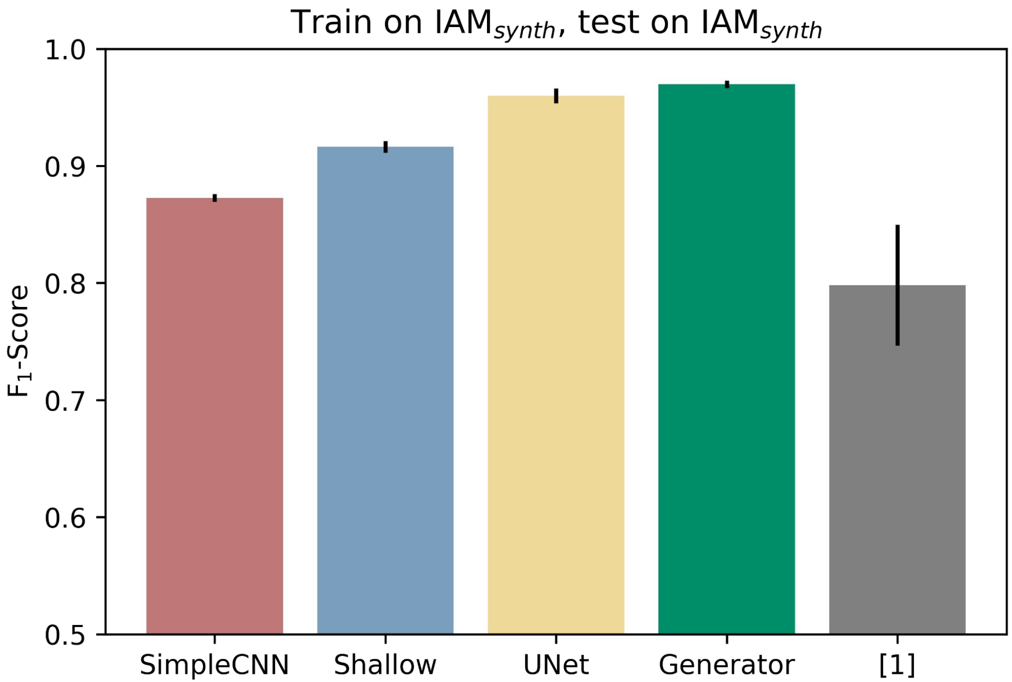
2. Train on IAM_{synth} — test on $Dracula_{real}$
   ➢ how well do the synthetic multi-writer models perform on (unseen) genuine single-writer data?

3. Train on individual sets of $Dracula_{synth}$ — test on $Dracula_{real}$
   ➢ what performance can be achieved with few (126) writer-specific synthetic samples?

4. Train on aggregated sets of $Dracula_{synth}$ — test on $Dracula_{real}$
   ➢ can results from 3. be improved by combining the sets?
   (→ same unique words images but larger variation in strikethrough strokes)

UPPSALA UNIVERSITET

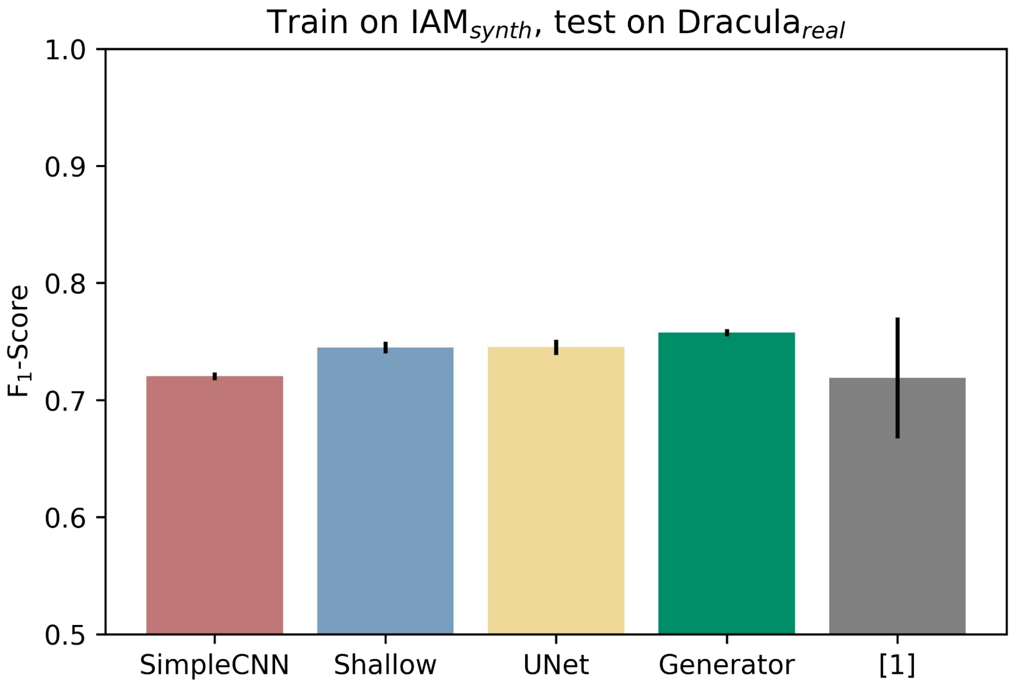
1. Train on IAM_{synth} — Test on IAM_{synth}



base comparison with [1]:

→ paired approaches outperform attribute-guided CycleGAN

2. Train on IAM_{synth} — Test on $Dracula_{real}$



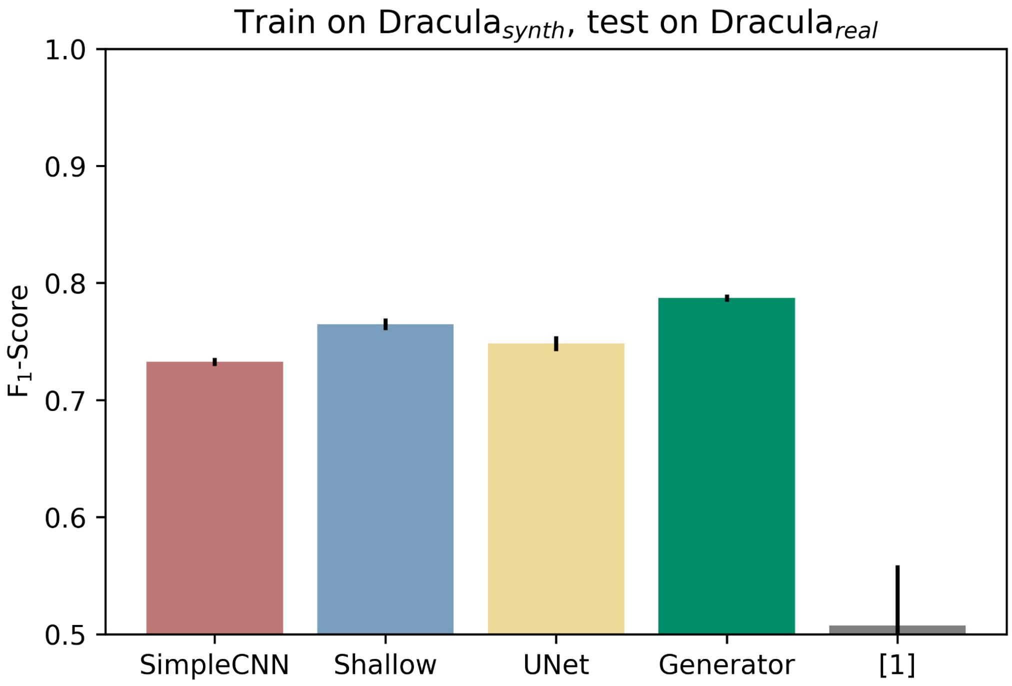
How well do the synthetic multi-writer models perform on (unseen) genuine single-writer data?

→ clear drop in performance

→ overall very similar performance across models

UPPSALA UNIVERSITET

3. Train on sets of Dracula$_{synth}$ — Test on Dracula$_{real}$

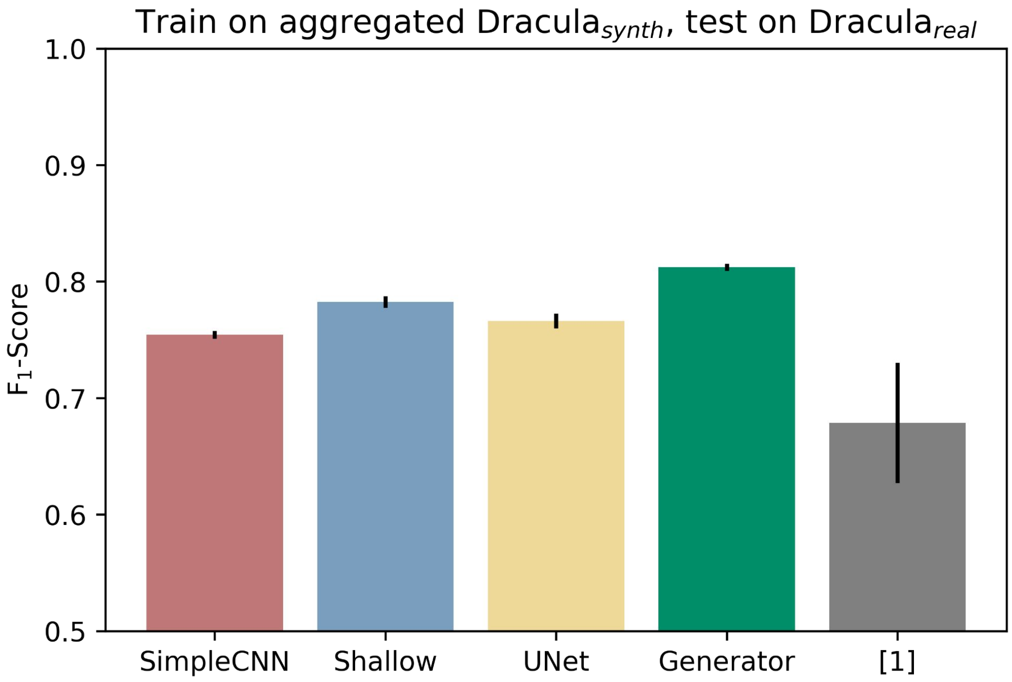


What performance can be achieved with few (126) writer-specific synthetic samples?

→ paired approaches outperform their IAM$_{synth}$ counterparts
→ drastic performance drop for attribute-guided CycleGAN

UPPSALA UNIVERSITET

4. Train on aggregated Dracula$_{synth}$ — Test on Dracula$_{real}$

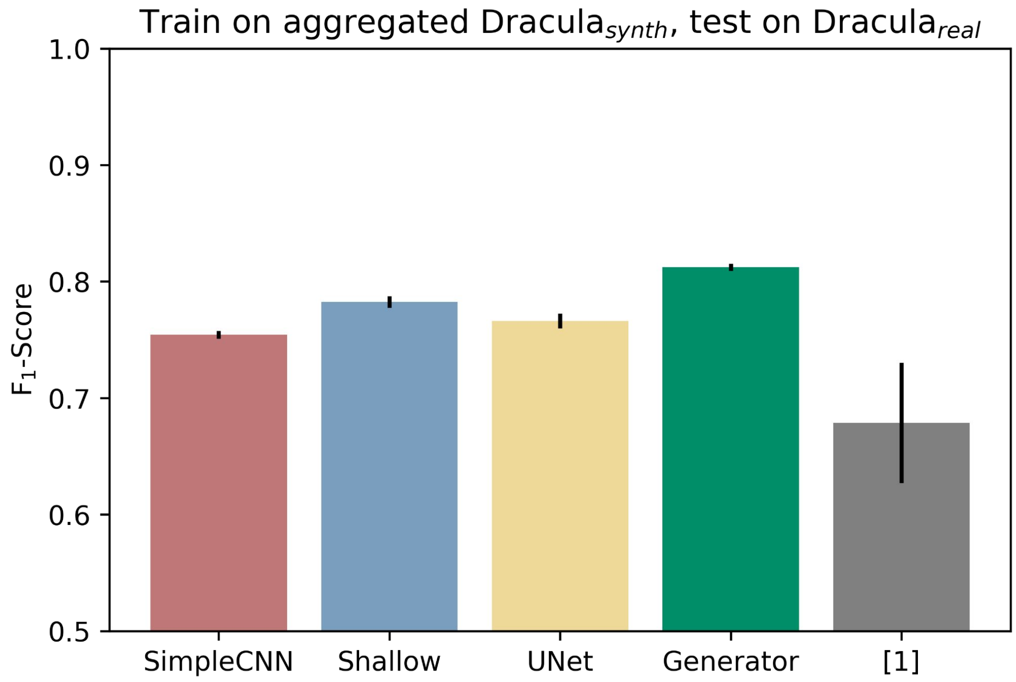


Can results from 3. be improved by combining the sets?

→ modest but consistent improvement for paired approaches
→ substantial improvement for attribute-guided CycleGAN

UPPSALA UNIVERSITET

4. Train on aggregated Dracula$_{synth}$ — Test on Dracula$_{real}$



Can results from 3. be improved by combining the sets?

Future work: impact of increasing the number of unique word images in train set

Qualitative Results (trained on aggregated Dracula$_{synth}$)

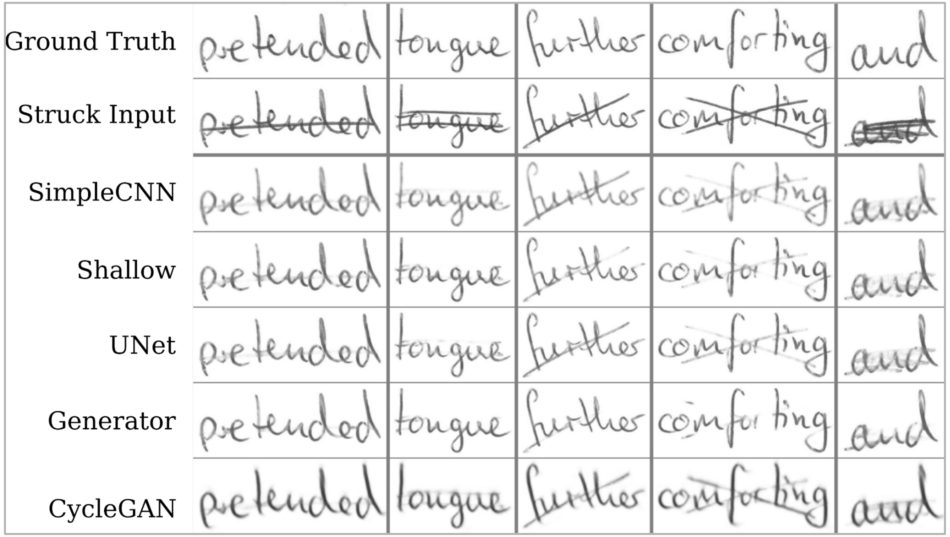


Fig.: Cherry-picked examples. Mean greyscale images, averaged over output from 30 separate models each.

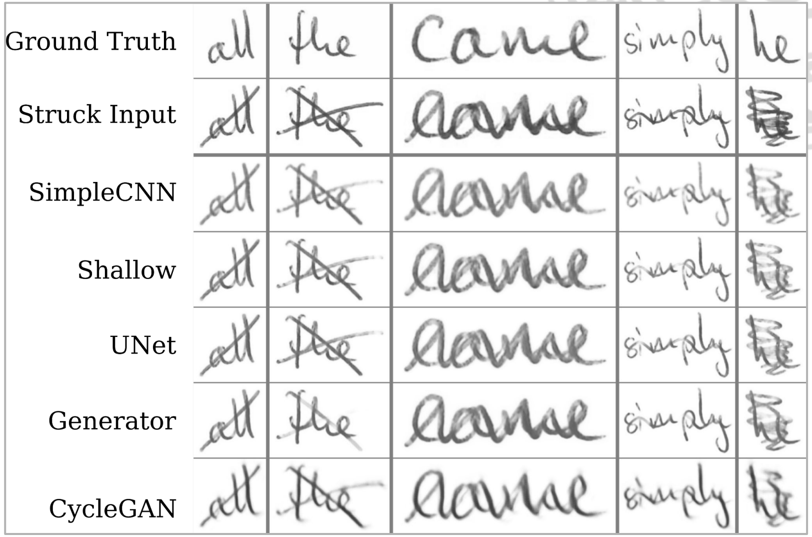


Fig.: Lemon-picked examples. Mean greyscale images, averaged over output from 30 separate models each.

UPPSALA UNIVERSITET

Conclusions

1. paired approaches outperform attribute-guided CycleGAN
    - general cleaning performance
    - model size
2. writer-specific models outperform multi-writer in domain
    - future work: impact of finetuning from multi to single-writer
3. varying difficulty of stroke types
    - most challenging: zigzag, wave and scratch
    - future work: focus efforts on challenging strokes

Contact:
raphaela.heil@it.uu.se
ekta.vats@abm.uu.se
anders.hast@it.uu.se

For links to our code and datasets, please see Appendix A of our paper!

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

This work is partially supported by Riksbankens Jubileumsfond (RJ) (Dnr P19-0103:1)

UPPSALA UNIVERSITET