Ad-hoc Personalization of Offline Handwritten Text Recognition Using Style Transfer

Anonymous ACL submission

Abstract

Personalizing handwritten text recognition can significantly enhance recognition accuracy, but requires a substantial number of handwritten samples, which limits its practical relevance. In this work, we investigate the potential of style-transferred synthetic samples for ad-hoc personalization, i.e., adapting a generic handwriting recognition system to a new writer using only a few samples. We show that one-shot and few-shot generators are able to produce visually similar handwriting samples. However, our experiments also show that the style-transferred data has no measurable personalization effect. This finding holds in a fair comparison with the same amount of samples and when using larger quantities of synthetic data. It also holds when controlling for non-stylistic influences.

1 Introduction

Offline handwritten text recognition (HTR) is an important step in supporting tasks like grading handwritten university exams (Rowtula et al., 2019) or giving feedback in primary education (Laarmann-Quante, 2017). The person writing is usually known (e.g. exam written by a specific student; sentences attempted by a specific first grader), which means that the quality of HTR can be considerably improved by personalization (Kienzle and Chellapilla, 2006; Gold et al., 2021; Scius-Bertrand et al., 2023). However, the amount of handwritten samples required for this purpose is a major factor limiting practical usefulness. In this paper, we explore the idea to use the precious real handwriting samples to generate more style-transferred samples which can then be used to personalize the HTR system. While recent work has shown remarkable progress in one-shot and few-shot handwriting style transfer (Dai et al., 2024; Nikolaidou et al., 2025), it remains unclear if such data is useful for ad-hoc personalization.

We outline our approach in Figure 1. First, a large but generic dataset with real handwriting samples from many writers is used to train both a generic HTR model and a handwriting generator. The HTR model learns general handwriting recognition, while the generator builds a continuous latent space to synthesize new handwriting styles. Next, we take samples from unseen specific writers and create a style-transferred synthetic dataset which is then used for personalization. To investigate the impact of our approach, we conduct qualitative and quantitative analyses to compare style-transferred synthetic data with real handwriting samples and evaluate their impact on HTR system personalization.

Our findings reveal that while style-transferred synthetic data might look similar, it has no measurable personalization effect. We substantiate this conclusion through experiments that rule out other influences like data quantity, specific vocabulary, the complete lack of real data in personalization, and non-stylistic image properties. We make all our experiment code publicly available under *removed for anonymous review*.

2 Personalization of HTR Systems

Personalization of HTR systems, also known as writer adaptation, involves transforming a generic recognizer into a user-specific one to improve recognition rates (Kienzle and Chellapilla, 2006). This can either be done with real writer data or synthetic data.

2.1 Real writer data

Fine-Tuning If a sufficient number of transcribed samples from a specific writer is available, personalization can be achieved by fine-tuning a generic model. Gold et al. (2021) fine-tuned an HTR system with 528 writer-specific samples, reducing mean CER from 14.1% to 8.0% across 40 writers. Similarly, Kienzle and Chellapilla (2006) showed that character-level personalization with 2,000 writer-specific samples (20 per character)

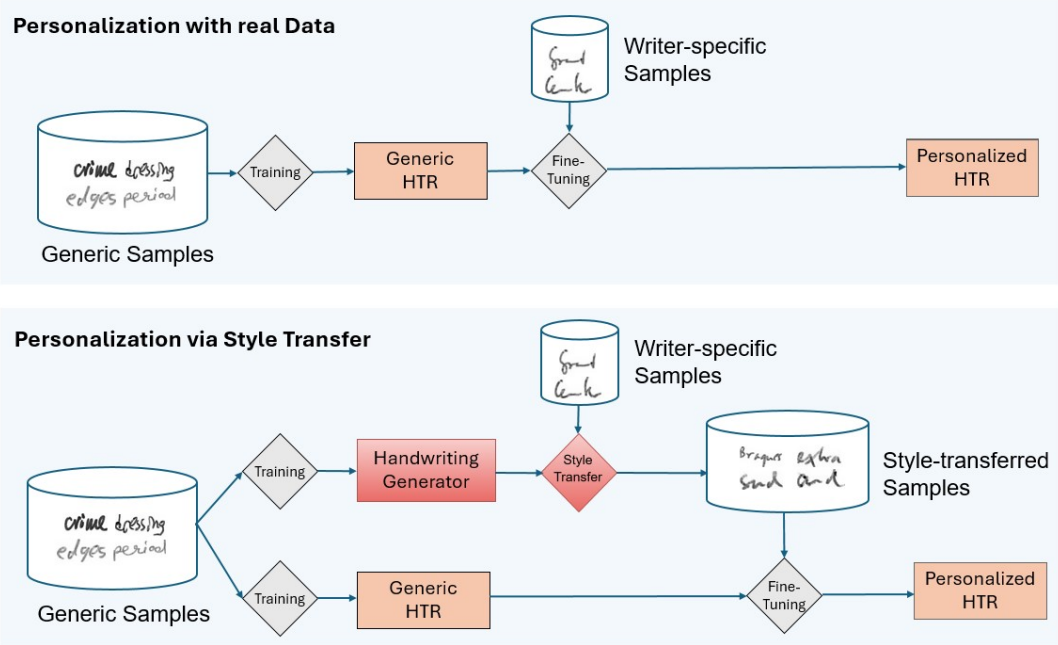


Figure 1: Personalization with real data vs. personalization via style transfer

reduced mean CER from 10.2% to 4.4% across 21 writers. Personalization also improved recognition of historical handwriting, reducing CER from around 10% to 4.9% for eight Irish writers (Ó Raghallaigh et al., 2022) and from 9.6% to 8.6% for 106 German writers (Scius-Bertrand et al., 2023). All these approaches require rather large quantities of transcribed handwriting samples for each writer.

Style Embeddings Instead of directly fine-tuning the model, the transcribed samples can also be used to learn style embeddings for each writer. Kohút et al. (2023) integrate these into the HTR system by conditioning an Adaptive Instance Normalization layer (Huang and Belongie, 2017), allowing the model to adapt feature processing to the writer's style. For known writers, this approach reduced CER by 9.2%. However, for unknown writers, neither selecting the closest existing embedding nor fine-tuning a new one proved effective; instead, standard fine-tuning outperformed both.

Weakly-supervised In real-world scenarios, such as exam settings, large amounts of personalized but unlabeled handwritten samples may be available enabling weakly-supervised approaches. Kang et al. (2019) propose domain adaptation from synthetic to real HTR, aligning the feature distributions of both domains. This reduces average CER from 24.3% to 18.3% on IAM validation data. Other approaches align feature distributions through deep transfer mapping with linear transformations (Yang et al., 2018), or focus on extracting writer-specific stylistic features (Wang and Du, 2022; Nosary et al., 2004).

While these approaches show improvements from incorporating untranscribed data, they perform much worse in terms of CER compared to using transcribed samples.

2.2 Synthetic Data

Fine Tuning – Data Augmentation Data augmentation can be beneficial for improving generic HTR in low-resource languages (Garces Arias et al., 2023; Koch et al., 2023). For personalization, results are rather negative. Jemni et al. (2022) apply augmentation techniques, including affine transformations and morphological distortions, to generate additional data for personalizing Arabic HTR. However, these methods improved only the generic model and hindered personalization. A general downside of data augmentation methods is that they rely on transforming existing data and cannot generate unseen words (Luo et al., 2022).

Style Transfer Style transfer is employed in current generators (Dai et al., 2024; Gan et al., 2022) to mimic a handwriting style derived from a small set of writer-specific samples to generate arbitrary new text. To the best of our knowledge, style trans-

fer has not been applied to the personalization of individual writers. Existing studies instead focus on improving generic HTR, either by generating synthetic data in the style of the target test set or by using unrelated writer styles, with results showing limited effectiveness when used alone.

Pippi et al. (2023b) show that transformer-based synthetic data in the target test style performs poorly, but adding only 1–5% real data yields substantial CER gains. Similarly, Dai et al. (2024) and Pippi et al. (2023a) applied diffusion- and transformer-based style transfer, where training on synthetic data only resulted in CERs of 11.7% and 11.9% on IAM, compared to 5–6% with real data. Likewise, Muth et al. (2024) showed that GAN-generated handwriting underperformed real samples, but pre-training reduced real data needs by 70–80% with comparable performance. Cascianelli et al. (2021) also found that halving real data and supplementing with font-based synthetic data achieved similar results on the Leopard dataset. Kang et al. (2022) investigate Spanish number recognition with GAN-based style transfer, showing that 160 real samples augmented with synthetic data outperform 298 real samples alone. Other GAN-based studies also found gains only with mixed real and synthetic data (Wang et al., 2025; Yeleussinov et al., 2023; Luo et al., 2022). Contrary to these findings, Ding et al. (2023a) found that diffusion-generated handwriting outperformed real samples on the IAM dataset (4.1% vs. 7.3% CER), with a combination further improving to 3.8%. Specialized methods such as handwriting generators based on a Y-Autoencoder combined with AdaIN also improved Japanese HTR, increasing F1 from 86.6% to 92.8% (Kitagawa et al., 2022). Overall the studies show a mixed picture. Style transfer may contribute to improving HTR systems and reducing reliance on real data. If successful for individual writers, it could enable unlimited user-specific data with few real samples, overcoming data limitations and manual transcription challenges.

3 Style Transfer

Style transfer allows us to generate synthetic handwriting samples by transferring a writer's style onto arbitrarily chosen text (see Figure 2). This technique allows for generating unlimited samples for personalization. In this section, we describe the style transfer methods and assess their quality through qualitative and quantitative analysis.

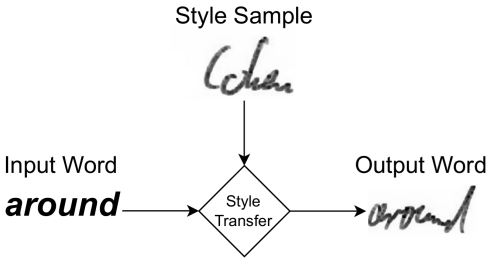


Figure 2: Style transfer example

3.1 Style Transfer Methods

One-shot (Gan et al., 2022) and few-shot (Kang et al., 2022) methods achieve style transfer for unknown styles by utilizing a learned continuous latent space, where each point corresponds to a known writer, enabling interpolation between styles. Earlier approaches have predominantly been using GAN-based methods (Gan et al., 2022; Luo et al., 2022; Kang et al., 2020) in both one-shot and few-shot settings. However, recent few-shot models leveraging encoder-decoder transformers aim to improve character-level style variation through cross-attention (Pippi et al., 2023a; Bhunia et al., 2021). Meanwhile, denoising diffusion-based models are becoming increasingly popular for one-shot handwriting generation (Dai et al., 2024; Nikolaidou et al., 2023; Ding et al., 2023b), as they yield higher image quality, broader distribution coverage, and more stable training (Dhariwal and Nichol, 2021).

When comparing both approaches, one-shot learning is more user-friendly, as few-shot learning is inconvenient and time-consuming due to its reliance on multiple samples (Dai et al., 2024). However, one-shot learning is likely more challenging, as it lacks robustness to input variations due to the absence of additional reference samples.

3.2 Data & Evaluation

We use the **GoBo dataset** (Gold et al., 2021) to evaluate style transfer quality. The dataset comprises English words from the following sources: random words from the Brown Corpus, pseudowords from the ARC Nonword Database, balanced CEDAR letter samples, and two domain-specific word lists, totaling 37,000 words written by 40 individuals, with an average of 926 words per writer.

Following prior work (e.g., Nikolaidou et al., 2025; Dai et al., 2024; Pippi et al., 2023a; Kang

et al., 2022), we use the **Fréchet Inception Distance** (FID), the standard metric for evaluating the similarity between real and synthetic handwriting (Heusel et al., 2017).

3.3 Style Transfer Experiments

We use the One-Shot Diffusion Mimicker (One-DM) and the Visual Archetypes-based Transformer (VATr), both trained on the IAM dataset, to generate personalized data with style transfer. These systems were selected for their state-of-the-art advancements in style transfer and their reported superiority in quality and adaptability over previous generators.

One-DM integrates high-frequency component analysis into a conditional diffusion model, using two parallel encoders to extract spatial and high-frequency style features and a style-content fusion module to perform style transfer from a single reference (Dai et al., 2024).

VATr combines a Transformer encoder-decoder with visual archetypes and leverages a pre-trained convolutional backbone on a large synthetic dataset for robust style extraction to perform style transfer from a few reference samples (Pippi et al., 2023a).

The generation of all synthetic datasets was carried out on Google Colab using an NVIDIA A100, requiring approximately 380 GPU-hours in total.

Quantitative Analysis To evaluate the dependency on references in one-shot and few-shot scenarios, we generated synthetic datasets with One-DM and VATr, based on the vocabulary of 40 writers from the GoBo dataset and randomly selected reference images for style transfer. We then evaluate this synthetic dataset against real handwritten data using the FID metric. The results (see Figure 3) show that One-DM's one-shot performance varies significantly with the chosen reference. A broad FID range indicates high intra-writer variation, making handwriting harder to map consistently in the latent space, whereas a narrow range suggests a more uniform handwriting with minimal deviations across references. Notably, the few-shot approach of VATr with 15 shots almost always surpasses the best one-shot cases, indicating that the use of multiple references simultaneously improves style transfer precision by reducing input variation effects. However, additional references beyond this point provide no further improvement, likely due to VATr's optimization for this sample size. Another observation is that FID values vary across

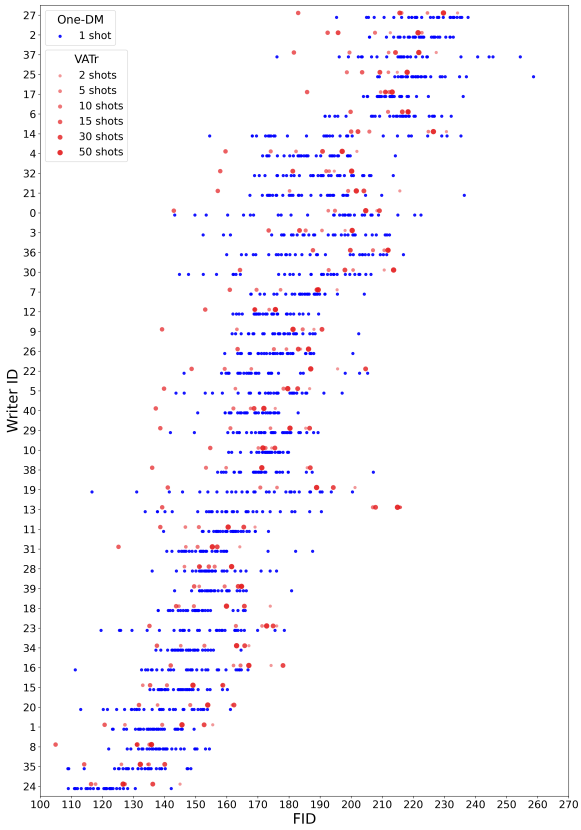


Figure 3: Style transfer quality (in terms of FID values) per writer. The spread of the blue one-shot dots shows the variance due to selection effects (it matters considerably, which sample is used). The red few-shot results show no clear trend for more shots improving results.

handwriting styles, suggesting that the latent space may not fully capture the diversity of handwriting styles, with some being underrepresented.

Qualitative Analysis Table 1 shows the writers with the highest and lowest FID ranges, highlighting their best and worst references for style transfer.

While lower FID values should generally reflect better style transfer, this is not always the case. For example for writer 0, the reference *loudds* (lowest FID) yielded a style transfer of the letter *s* in *discuss* less similar to the original than that generated with the reference *mircs* (highest FID). This discrepancy likely arises because the FID metric measures distributional similarity between synthetic and real datasets, rather than capturing stylistic differences in individual images. For generations with the lowest FID spreads, style transfer for *well* and *vocased* closely matches the original handwriting, regardless of the reference.[1]

[1] In the Appendix (Tables 4 and 5), we show additional synthetic samples from writers with lower/higher FID spreads.

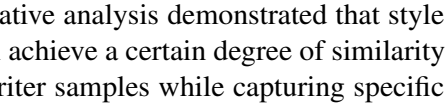
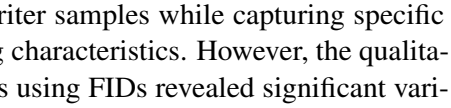
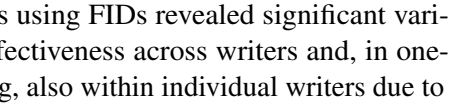
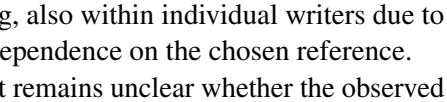
	Highest		Lowest	
Writer ID	**14**	**0**	**1**	**15**
FID Spread	154/235	143/222	123/149	135/160
best style				
worst style				
best transfer				
worst transfer				
real writer				

Table 1: Examples with highest and lowest FID spread

3.4 Discussion of Style Transfer Quality

Our quantitative analysis demonstrated that style transfer can achieve a certain degree of similarity with real writer samples while capturing specific handwriting characteristics. However, the qualitative analysis using FIDs revealed significant variations in effectiveness across writers and, in one-shot learning, also within individual writers due to the strong dependence on the chosen reference.

Overall, it remains unclear whether the observed similarity is sufficient for effective HTR personalization. We hypothesize that style-transferred synthetic data may contribute to reducing CERs but is likely less effective than real handwriting due to deviations from the true handwriting style.

Next, we begin personalizing HTR systems with real data to demonstrate its effectiveness. Then, through various experiments, we personalize with synthetic few-shot and one-shot data and compare the results with those obtained from real data.

4 Personalized Handwriting Recognition – Experimental Setup

We use the state-of-the-art **AttentionHTR**, an end-to-end system that leverages ResNet for feature extraction and bidirectional LSTMs for sequence modeling, incorporating a content-based attention mechanism as part of an encoder-decoder architecture (Kass and Vats, 2022).

We compute the standard evaluation metric **CER** (Character Error Rate) over three datasets (Table 2 gives an overview). In addition to the **GoBo** dataset (introduced in Section 3.2), we use the **IAM** dataset (Marti and Bunke, 2002) which is derived from the Lancaster-Oslo-Bergen Corpus and consists of 1,539 scanned handwritten English forms, penned by 657 different writers, with a total of 115,320 words.[2] We also use the **CVL-database** (Kleber et al., 2013) which contains seven handwritten texts (one in German, six in English) from 310 writers. Of these, 27 writers contributed seven texts each, while 283 writers contributed five texts.

Dataset	Words	Unique Words	Writers	∅ words per writer
CVL-en	84,514	253	310	273
GoBo	37,000	437	40	926
IAM	115,320	10,841	657	176

Table 2: Statistics of personalized handwriting datasets

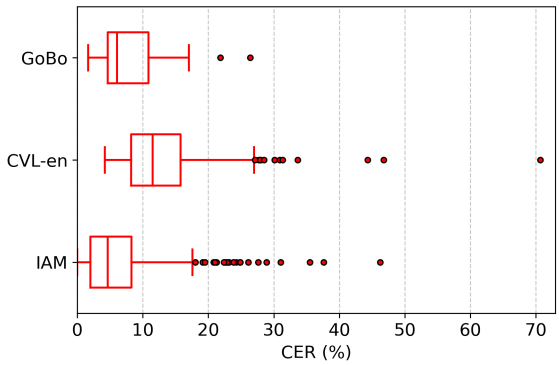


Figure 4: Distribution of CER for individual writers

For all personalization experiments, we use Google Colab with NVIDIA L4 and T4 GPUs, requiring approximately 300 GPU-hours.

Baseline Results We train AttentionHTR on IAM and validate using 5-fold cross-validation, achieving a mean CER of 3.22% ± 0.11%.[3] This value is reasonably close to the current state of the art to serve as the starting point for our personalization experiments. At the time of writing, PapersWithCode[4] lists 18 systems for this setup with CERs between 2.4 and 7.6%.

Writer Dependence Next, we evaluate how much recognition performance varies between writers. As Figure 4 shows: quite a lot.[5] For example, while the median writer in the IAM dataset experiences a system with a CER of 4.7% and half

[2] https://fki.tic.heia-fr.ch/databases/iam-handwriting-database

[3] For CER and standard deviation values over all writers for all folds see Table 6 in the Appendix.

[4] https://paperswithcode.com/sota/handwritten-text-recognition-on-iam

[5] Note however that the 70% outlier in CVL results from a transcription error. The author wrote entirely in uppercase, while transcriptions are case normalized.

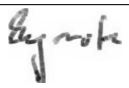
ID	Real	Generic	Personalized
17		Suymort	keynote
10		malnde	include
1		sundons	scholars
26		dissmubue	disputed

Table 3: Examples of HTR with and without personalization for writers with the highest CER

of the writers even less, some writers are faced with unacceptable CERs of 20 to 46%. Therefore, personalization has the potential to considerably improve performance for single writers with high CER values. In the next section, we experiment with such personalized models.

5 Personalization with Real Data

As we have shown in the previous section, CER values vary considerably between writers. Before we turn to style transfer in the next section, we now establish how much improvement is possible when fine-tuning with real samples from the GoBo dataset. This is at the same time a replication of the results from Gold et al. (2021), but with the use of AttentionHTR instead of an HTR approach based on a Convolutional Recurrent Neural Network (CRNN). For this purpose, we fine-tune the baseline model individually for each of the 40 writers, i.e. starting from the same pretrained model and creating 40 separately adapted models. At each step, 10 user-specific samples are added incrementally, reaching a total of 528 instances (maximum from the GoBo training data), and the performance is evaluated on the remaining 398 test samples. At each step, we trained the model for five epochs using a batch size of 10 and a learning rate of 0.001.

Figure 5 presents the results by Gold et al. (2021) (a screenshot from that paper) and ours (copying their diagram style for better comparison). The slope of the personalization curves is very similar, but our results start and arrive at much lower CER values due to the overall performance advantages of AttentionHTR compared to the CRNN used in (Gold et al., 2021). Our replication shows that personalization with a few hundred user-specific samples is feasible. In Table 3, we present examples of fully correct predictions after personalization for challenging writers with the highest CER.

To further assess the contribution of different word types to personalization, we repeated the experiment for each word group. The results show that all word groups except nonwords contributed to personalization.[6] As we are controlling for word overlap in the test data, this cannot be a vocabulary effect, but probably a result of forcing writers to write unusual character combinations leading to an 'impersonal' writing style.

Importance of Real Writer Data There is a theoretical possibility that CER values improve not because we are adding data for a specific writer, but in general due to fine-tuning with more data (as the vocabulary for each writer in the dataset is fixed). To test this, we rerun the personalization experiment using for each writer samples from a different writer for 'personalization'. The (bad) results in Figure 6 clearly show that the improvements can indeed be attributed to personalization with samples from a specific writer.

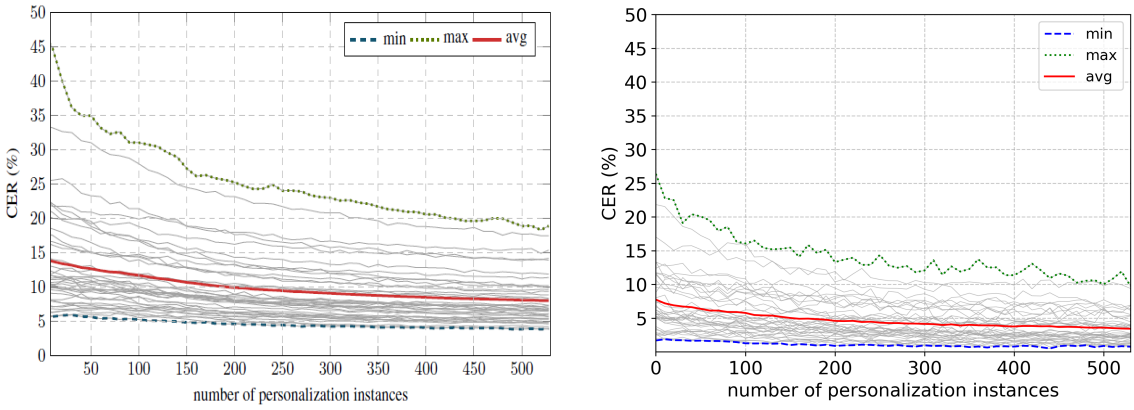
Even if we have shown that data from a random writer cannot stand in for another writer, maybe there are structurally similar handwriting styles where this is possible. We test this by personalizing the model for the writers with the highest CERs (as we can see the biggest effects here) using handwriting data from each of the other writers. In general, no improvements can be observed. However, there is one exception of two writers (with very similar writing styles) who can mutually improve each other's CERs to some extent (but less than with their own handwriting).[7] If similar writing styles can be mutually be used for personalization, maybe this is also true for synthetically generated 'similar' data.

6 Personalization via Style Transfer

We now repeat the personalization experiment from the previous section, but instead of real data we are using style transfer to create synthetic data as described in Section 3. First, we generate few-shot and one-shot synthetic datasets for each writer, using 15 random samples for few-shot and the best sample for one-shot (see Figure 3), while using the same words as in the previous experiments to ensure a fair comparison. Since the best one-shot

[6] See Figure 10 in Appendix C for detailed results.

[7] In Appendices G and H, we have provided examples for the writers with the highest and lowest CER, where these two writers (number 10 and 17) are also included. In Figure 11 we show the corresponding personalization graphs.



(a) Personalization of a CRNN by Gold et al. (2021)

(b) Personalization of AttentionHTR

Figure 5: HTR results with increasing samples used for personalization on the GoBo dataset

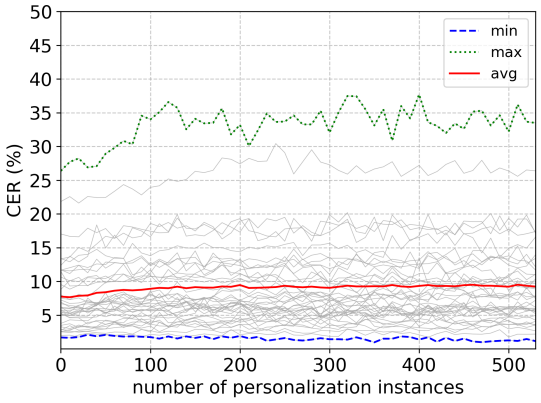


Figure 6: Failed 'personalization' when using samples from a different writer

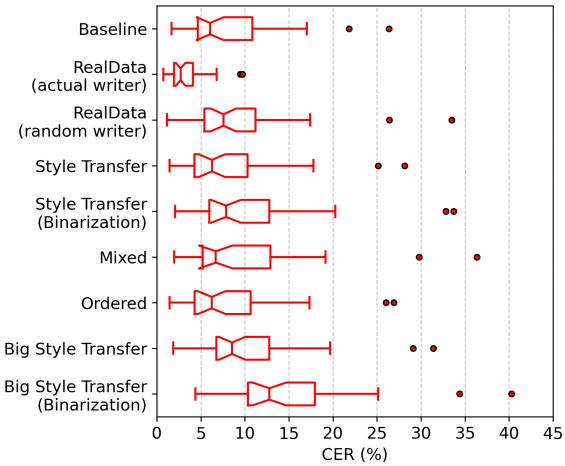


Figure 7: Overview of personalization results

sample is unknown in practice, this scenario is unrealistic. However, if unsuccessful, it would strongly indicate insufficient style transfer quality. We find that neither one-shot nor few-shot synthetic data improve recognition, and both yield similar outcomes, leading us to exclude one-shot from further experiments.

Figure 7 shows the results for personalization with few-shot style-transferred data under 'Style Transfer' and for comparison also shows the results for the (unpersonalized) baseline as well as using real data from the actual writer and from a random writer.[8] We find that style transfer performs on the level of the unpersonalized baseline, i.e. having no measurable effect.

Mixing real and synthetic Maybe synthetic style-transferred data alone is not sufficient, as previous work has successfully used a combination of real and synthetic data (Wang et al., 2025; Muth et al., 2024; Yeleussinov et al., 2023). We thus additionally personalize using a mixed dataset composed of 50% synthetic and 50% real samples. The results are also shown in Figure 7 under 'Mixed'. To better see the effect of synthetic data, we additionally examine the effect of initially personalizing with 250 real samples before continuing with synthetic data (shown as 'Ordered' in Figure 7). As can be seen in the personalization plot in Figure 8, the model starts to personalize well when being fed real data, but as soon as synthetic samples are added, the error rate rises again. After having seen only 50 synthetic samples, personalization results are back to original (bad) levels and then stay there. This indicates differences between synthetic and real data, as well as the sensitivity of the personalization process.

[8]The remaining entries in the figure are results from additional experiments described later in this paper.

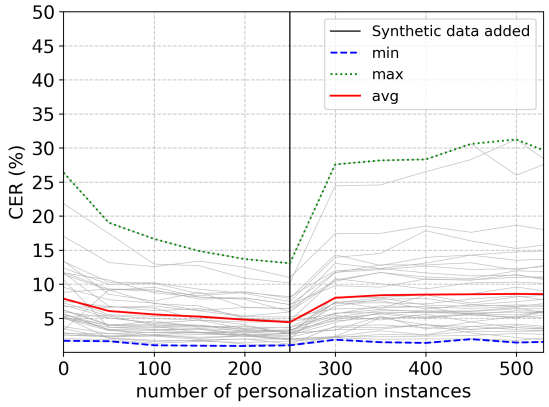


Figure 8: Setup 'Ordered' - Personalization with 250 real samples followed by 250 synthetic samples

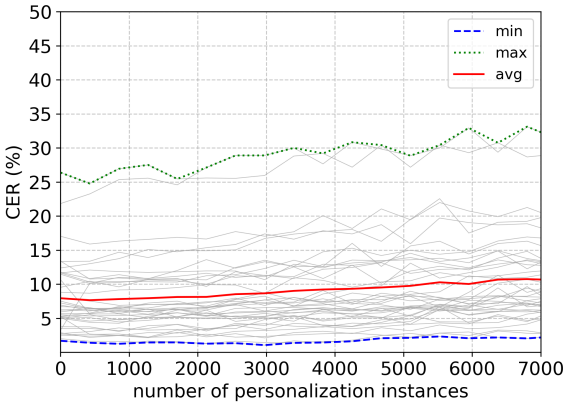


Figure 9: Setup 'Big Style Transfer' - Personalization using a much bigger dataset of 10,000 style-transferred samples (7000 training / 3000 validation)

More data Although our previous experiments showed that synthetic data does not contribute significantly to performance, we want to rule out the possibility that data quantity is the limiting factor. Therefore, we extended our experiments from 500 to 10,000 few-shot synthetic samples based on a dictionary of frequent English words[9], using 30% for validation to select the best model within 1–6 epochs (see Figure 9). The unchanged CERs further confirm our finding that style-transferred data is insufficient to achieve adequate personalization.

Non-stylistic influences Finally, we want to rule out non-stylistic influences (i.e. noise that humans do not pick up when visually inspecting style similarity, but machines might) in the synthetic data as a reason for failed personalization. To this end, we tested their impact by adding scanner streaks, compression noise, and blob noise to the real test data using OpenCV functions and re-evaluating performance (see Figures 12 and 13 in Appendix E). The results show that non-stylistic effects can negatively influence personalization, with blob noise having the strongest impact. Thus, noise introduced by the style generator might also interfere with personalization. To investigate this, we binarized the synthetic samples. The resulting images only consist of the black letter shapes on white background; any other noise from the generator will be removed. We re-ran the style transfer experiments in the replication setting ('Style Transfer (Binarization)') and with the extended amount of data ('Big Style Transfer (Binarization)') and show results again in Figure 7. In both settings, CER increased a bit, likely due to the loss of grayscale information through binarization. The more important result though is that personalization still fails, so we can rule out any non-stylistic influence.[10]

7 Conclusion

We started with the hypothesis that personalization of HTR system could be achieved through the use of style transfer. However, while synthetic handwriting often appears visually similar to real handwriting, this similarity was not sufficient in our experiments to enable successful personalization. We ruled out potential confounders, including data quantity, specific vocabulary, the absence of real data in personalization, and non-stylistic image properties, as explanations for this outcome.

This leaves two possibilities: either the synthetic handwriting must be made substantially more similar to the target writer, or the intra-writer variance must be represented more faithfully. We speculate that the latter is the dominant limitation, as individual samples already show a close resemblance to the target style, while 1–15 examples are unlikely to capture the full variability of a writer's handwriting.

Future research should focus more on understanding what makes handwriting styles similar and to identify suitable metrics for measuring this similarity. Currently, it remains uncertain whether existing style-transfer approaches adequately capture the writer-specific characteristics required for personalization.

[9] https://wortschatz.uni-leipzig.de/en/download/English

[10] For completeness, we also binarized the real data (see Figure 15 in Appendix F), which had only a marginal effect and thus cannot explain failed personalization on synthetic data.

Limitations

We only evaluate on one language (English), mainly due to the sparsity of suitable data for other languages. However, synthetic data generation was shown to also work for other languages (Dai et al., 2024), so the overall setup should be applicable as well - possibly with even more room for improvement starting from less well performing baselines. Another limitation is that our FID-based style transfer evaluation is restricted to 40 writers in the GoBo dataset. While this dataset was designed to be diverse (Gold et al., 2021), handwriting is highly individual, and some variations, especially outlier styles, may not be fully represented.

Ethical Considerations

Being able to synthesize handwriting from just a few samples may pose significant risks, as it can be exploited for fraudulent activities such as identity theft, forgery of signatures on legal documents, manipulation of handwritten records, and social engineering scams that deceive individuals by mimicking authentic handwriting. Storing writing samples is thus a potential security risk. Our research into ad-hoc personalization is a step towards solving this issue, as no writing samples (and derived style-transfer models) have to be stored if personalization from a few, directly obtained and then discarded, samples is possible.

References

Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Mubarak Shah. 2021. Handwriting transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1086–1094.

Silvia Cascianelli, Marcella Cornia, Lorenzo Baraldi, Maria Piazzi, Rosiana Schiuma, and Rita Cucchiara. 2021. *Learning to Read L'Infinito: Handwritten Text Recognition with Synthetic Training Data*, pages 340–350.

Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. 2024. One-DM: One-Shot Diffusion Mimicker for Handwritten Text Generation. In *Computer Vision – ECCV 2024*, pages 410–427, Cham. Springer Nature Switzerland.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc.

Haisong Ding, Bozhi Luan, Dongnan Gui, Kai Chen, and Qiang Huo. 2023a. Improving Handwritten OCR with Training Samples Generated by Glyph Conditional Denoising Diffusion Probabilistic Model. In *Document Analysis and Recognition - ICDAR 2023*, pages 20–37, Cham. Springer Nature Switzerland.

Haisong Ding, Bozhi Luan, Dongnan Gui, Kai Chen, and Qiang Huo. 2023b. Improving handwritten ocr with training samples generated by glyph conditional denoising diffusion probabilistic model. In *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part IV*, page 20–37, Berlin, Heidelberg. Springer-Verlag.

Jie Gan, Weifeng Wang, Ji Leng, and Xiaoxiang Gao. 2022. Higan+: Handwriting imitation gan with disentangled representations. *ACM Transactions on Graphics (TOG)*, 42(1):1–17.

Eduardo Garces Arias, Vihang Pai, Moritz Schöffel, Christian Heumann, and Matthias Aßenmacher. 2023. Automatic transcription of handwritten old occitan language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15416–15439, Singapore. Association for Computational Linguistics.

Christian Gold, Dario van den Boom, and Torsten Zesch. 2021. Personalizing Handwriting Recognition Systems with Limited User-Specific Samples. In *Document Analysis and Recognition – ICDAR 2021*, pages 413–428, Cham. Springer International Publishing.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519.

Sana Khamekhem Jemni, Sourour Ammar, and Yousri Kessentini. 2022. Domain and writer adaptation of offline Arabic handwriting recognition using deep neural networks. *Neural Computing and Applications*, 34(3):2055–2071.

Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. 2022. Content and style aware generation of text-line images for handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8846–8860.

Lei Kang, Marçal Rusiñol, Alicia Fornés, Pau Riba, and Mauricio Villegas. 2019. Unsupervised adaptation for synthetic-to-real handwritten word recognition. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3491–3500.

Lingfei Kang, Pau Riba, Y Wang, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. 2020. Ganwriting: Content-conditioned generation of styled handwritten word images. In *European Conference on Computer Vision (ECCV)*, volume 12368, pages 237–289. Springer.

Dmitrijs Kass and Ekta Vats. 2022. Attentionhtr: Handwritten text recognition based on attention encoder-decoder networks. In *Document Analysis Systems*, pages 507–522, Cham. Springer International Publishing.

Wolf Kienzle and Kumar Chellapilla. 2006. Personalized handwriting recognition via biased regularization. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 457–464. Association for Computing Machinery.

Tomoki Kitagawa, Chee Siang Leow, and Hiromitsu Nishizaki. 2022. Handwritten character generation using Y-autoencoder for character recognition model training. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7344–7351, Marseille, France. European Language Resources Association.

Florian Kleber, Stefan Fiel, Markus Diem, and Robert Sablatnig. 2013. CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting. In *2013 12th International Conference on Document Analysis and Recognition*, pages 560–564.

Philipp Koch, Gilary Vera Nuñez, Esteban Garces Arias, Christian Heumann, Matthias Schöffel, Alexander Häberlin, and Matthias Aßenmacher. 2023. A tailored handwritten-text-recognition system for medieval latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 103–110, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jan Kohút, Michal Hradiš, and Martin Kišš. 2023. Towards Writing Style Adaptation in Handwriting Recognition. In *Document Analysis and Recognition - ICDAR 2023*, pages 377–394, Cham. Springer Nature Switzerland.

Ronja Laarmann-Quante. 2017. Towards a tool for automatic spelling error analysis and feedback generation for freely written german texts produced by primary school children. In *Slate*.

Chenguang Luo, Y Zhu, L Jin, Z Li, and D Peng. 2022. Slogan: Handwriting style synthesis for arbitrary-length and out-of-vocabulary text. *IEEE Transactions on Neural Networks and Learning Systems*.

Urs-Viktor Marti and H. Bunke. 2002. The iam-database: An english sentence database for offline handwriting recognition. In *International Journal on Document Analysis and Recognition*, volume 5, pages 39–46.

Markus Muth, Marco Peer, Florian Kleber, and Robert Sablatnig. 2024. Maximizing data efficiency of htr models by synthetic text. In *Document Analysis Systems*, pages 295–311, Cham. Springer Nature Switzerland.

Konstantina Nikolaidou, George Retsinas, Vincent Christlein, Mathias Seuret, Giorgos Sfikas, Elisa Barney Smith, Hamam Mokayed, and Marcus Liwicki. 2023. Wordstylist: Styled verbatim handwritten text generation with latent diffusion models.

Konstantina Nikolaidou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. 2025. DiffusionPen: Towards Controlling the Style of Handwritten Text Generation. In *Computer Vision – ECCV 2024*, pages 417–434, Cham. Springer Nature Switzerland.

Ali Nosary, Laurent Heutte, and Thierry Paquet. 2004. Unsupervised writer adaptation applied to handwritten text recognition. *Pattern Recognition*, 37(2):385–388.

Brian Ó Raghallaigh, Andrea Palandri, and Críostóir Mac Cárthaigh. 2022. Handwritten text recognition (HTR) for Irish-language folklore. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 121–126, Marseille, France. European Language Resources Association.

Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. 2023a. Handwritten text generation from visual archetypes. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22458–22467.

Vittorio Pippi, Silvia Cascianelli, Christopher Kermorvant, and Rita Cucchiara. 2023b. How to Choose Pretrained Handwriting Recognition Models for Single Writer Fine-Tuning. In *IEEE International Conference on Document Analysis and Recognition*.

Vijay Rowtula, Subba Reddy Oota, and Jawahar C.V. 2019. Towards Automated Evaluation of Handwritten Assessments. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 426–433.

Anna Scius-Bertrand, Phillip Ströbel, Martin Volk, Tobias Hodel, and Andreas Fischer. 2023. The Bullinger Dataset: A Writer Adaptation Challenge. In *Document Analysis and Recognition - ICDAR 2023*, pages 397–410, Cham. Springer Nature Switzerland.

Yiming Wang, Hongxi Wei, Heng Wang, Shiwen Sun, and Chao He. 2025. GL-GAN: Perceiving and integrating global and local styles for handwritten text generation with mamba. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2434–2444, Abu Dhabi, UAE. Association for Computational Linguistics.

Zi-Rui Wang and Jun Du. 2022. Fast writer adaptation with style extractor network for handwritten text recognition. *Neural Networks*, 147:42–52.

Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, Jun Sun, and Cheng-Lin Liu. 2018. Deep transfer mapping for unsupervised writer adaptation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 151–156.

Arman Yeleussinov, Yedilkhan Amirgaliyev, and Lyailya Cherikbayeva. 2023. Improving ocr accuracy for kazakh handwriting recognition using gan models. *Applied Sciences*, 13(9).

A One-Shot and Few-Shot Examples

Style	Real				
Synthetic Data					
FID	0	**277**	**277**	**292**	**275**

Table 4: One-shot examples for writer ID 17

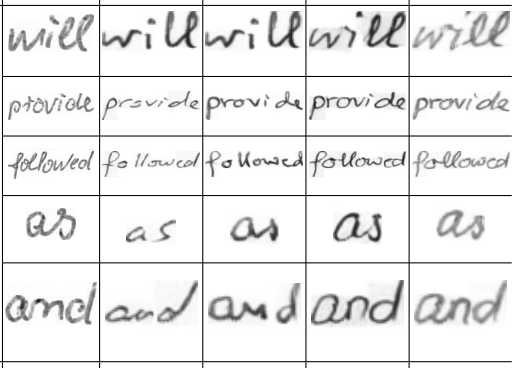
Shots	Real	2	5	10	15
Synthetic Data					
FID	0	**321**	**283**	**287**	**290**

Table 5: Few-shot examples for writer ID 19

Notably, in the one-shot examples, certain styles yield better results, such as the reference you, where the y in the word system is closer to that of writer 17. In the few-shot examples, it is noticeable that increasing shots improve similarity to the original. This is particularly evident in the style transfer for will, where the two final l differ with 2 shots but closely match the original with 15 shots.

B 5-Fold Cross-Validation

Fold	CER	σ
1	3.17	14.5
2	3.42	15.2
3	3.24	16.2
4	3.13	14.3
5	3.12	14.7
ø	3.22	

Table 6: Results of 5-Fold Cross-Validation on IAM

C Word Group Impact on Personalization

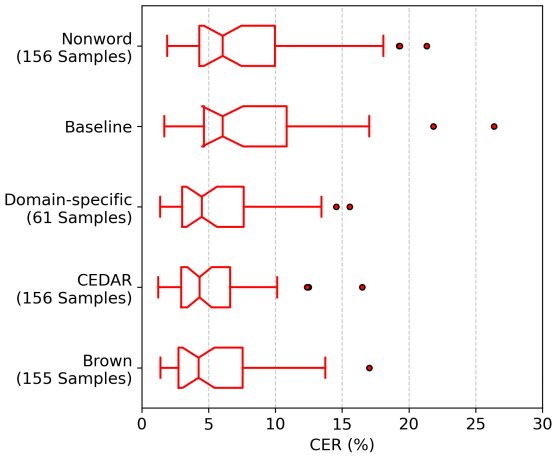
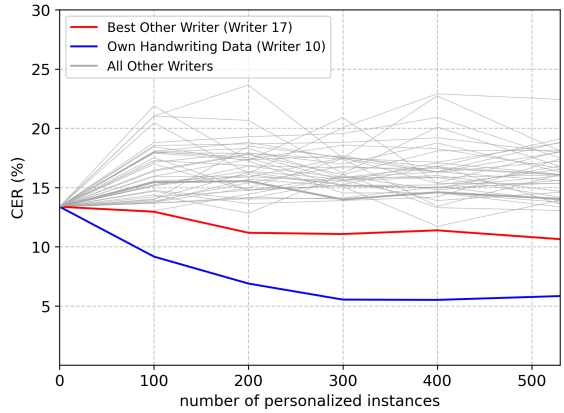


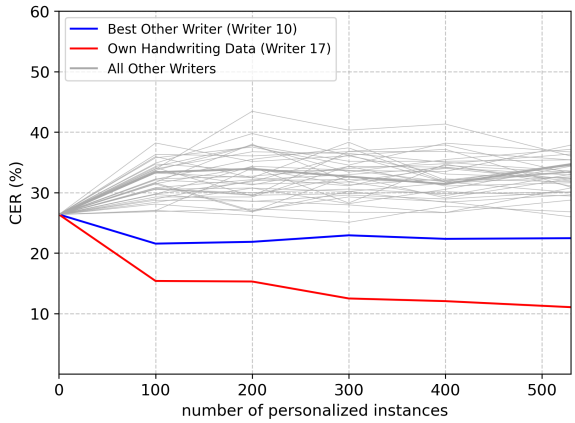
Figure 10: Personalization by training data

Nonwords did not contribute to personalization, consistent with Gold et al. (2021), who observed the same effect. They suggested this may result from the unnatural behavior of writers when writing uncommon words, who consciously considered their writing and sometimes switched between cursive and print, requiring the model to adapt to both styles.

D Personalizing with data from similar writers



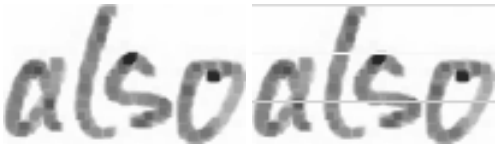
(a) Writer 10



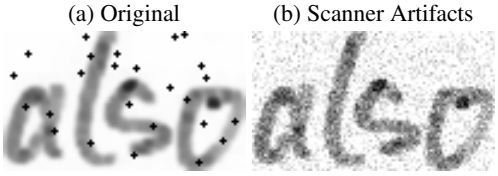
(b) Writer 17

Figure 11: Personalizing with data from other writers does not work in most cases, except for very similar handwriting styles that can mutually stand in for each other to some extent

E Impact of Noise and Degradation



(a) Original (b) Scanner Artifacts



(c) Blob Noise (d) Gaussian Noise

Figure 12: Examples of Noise and Degradation

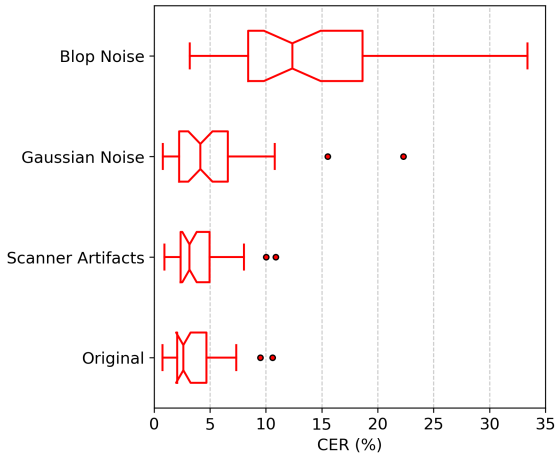


Figure 13: Impact of noise and degradation on real test data after personalization

F Impact of Binarization

(a) Original (b) Binarization

Figure 14: Comparison between original and binarized image

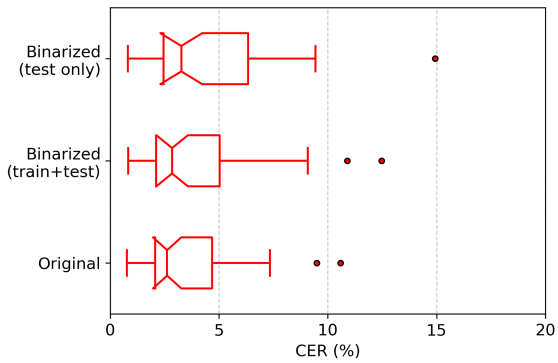


Figure 15: Personalization performance on binarized real data

Binarizing both training and test data has only a marginal effect on personalization performance. Training on grayscale but testing on binarized data performs slightly worse, as the grayscale information learned during training is absent at test time.

G Handwriting Examples of Writers with Highest CERs

Figure 16: Writers with Highest CERs (continued on next page)

Figure 16: Writers with Highest CERs

H Handwriting Examples of Writers with Lowest CERs

Figure 17: Writers with Lowest CERs (continued on next page)

Figure 17: Writers with Lowest CERs