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. 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.

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. To investigate this, 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 style-transferred synthetic data lacks sufficient similarity to real handwriting for effective HTR personalization, highlighting the need for further research to better capture and replicate individual writing styles. 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).

2.1 Real writer data

Style Embeddings Kohút et al. (2023) proposed a model that integrates individual handwriting styles into the HTR system by learning writer embeddings during training. These embeddings condition an Adaptive Instance Normalization (AdaIN) layer, allowing the model to adapt feature processing to the writer's style. For known writers, this approach reduced Character Error Rate (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.

This study highlights the challenge of generalizing handwriting with learned model parameters, which leads to overfitting due to its individual nature. Explicit fine-tuning could provide a more effective solution.

Fine-Tuning – Transcribed Samples If a sufficient number of transcribed samples from a specific writer are available, personalization can be achieved by fine-tuning a generic model with these samples. Gold et al. (2021) demonstrate the effectiveness of this method by fine-tuning a HTR system with 528 writer-specific samples, resulting in a mean CER reduction from 14.1% to 8.0%

across 40 writers. Notably, the writer with the highest initial CER of 47.2% benefited the most, achieving a reduction to 18.9%. Similarly, Kienzle and Chellapilla (2006) showed that personalizing a HTR system at the character level with 2,000 user-specific samples (20 per character) reduced the mean CER from 10.2% to 4.4% across 21 writers. Scius-Bertrand et al. (2023) were also able to demonstrate an improvement through personalization for challenging historical handwriting, achieving a mean CER reduction from 9.6% to 8.6% across 106 writers.

All these approaches require rather large quantities of transcribed handwriting samples for each writer.

Fine-Tuning – Untranscribed Samples In real-world scenarios, such as exam settings, large amounts of handwritten samples may be available without transcriptions. Kang et al. (2019) propose an adversarial domain adaptation approach that fine-tunes a model initially trained on synthetic data by aligning synthetic and real features using a Discriminator and a Gradient Reversal Layer, minimizing the domain gap and enhancing its ability to recognize new handwriting styles. This adaptation reduces average CER from 24.3% to 18.3% on IAM validation data, with a mean of 135 words per writer.

Beyond this work, other unsupervised adaptation methods have also demonstrated improvements, such as Deep Transfer Mapping, which aligns feature distributions through linear transformations (Yang et al., 2018), a Style Extractor Network, which captures writer-specific features using a CNN-GRU architecture (Wang and Du, 2022), and the use of writer invariants to iteratively refine character models by analyzing allograph patterns (Nosary et al., 2004).

While these approaches show improvements, they perform much worse in terms of CER.

2.2 Synthetic Data

Fine Tuning – Data Augmentation Synthetic data has been used to personalize HTR systems. Jemni et al. (2022) applied 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 by altering the writer's style and limiting adaptation to individual characteristics. Moreover, they rely on existing data and cannot generate entirely new words or characters, reducing diversity (Luo et al., 2022).

Fine Tuning – Style Transfer Style transfer is employed in modern 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.

Most previous work on style transfer aims to improve generic HTR models rather than personalization, with most studies indicating its limited effectiveness when used alone. For instance, Dai et al. (2024) and Pippi et al. (2023a) used diffusion- and transformer-based models for style transfer, showing that HTR models trained solely on synthetic data achieved CERs of 11.7% and 11.9% on IAM test data, while real data training reached approximately 5–6%. Similarly, Muth et al. (2024) found that GAN-generated handwriting performed significantly worse than real samples, but pre-training reduced reliance on real data by 70–80% while maintaining comparable performance. Likewise, Pippi et al. (2023b) show that transformer-based handwriting imitation for historical HTR improved only when real samples were included. Kang et al. (2022) investigate Spanish number recognition with GAN-based style transfer, achieving slightly better results by combining 160 real samples with synthetic data than using 298 real samples alone. Contrary to these findings, Ding et al. (2023a) reported that training exclusively on diffusion-generated handwriting achieved better CERs, reaching 4.1% on IAM test data compared to 7.3% with real samples, while a combination of both further improved the result to 3.8%.

The studies show that style transfer can 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 Research Hypothesis

From previous research on HTR personalization, we conclude that recent advances in style transfer show the biggest potential for improved personalization. We thus explore whether synthetic data generated through style transfer can enable 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

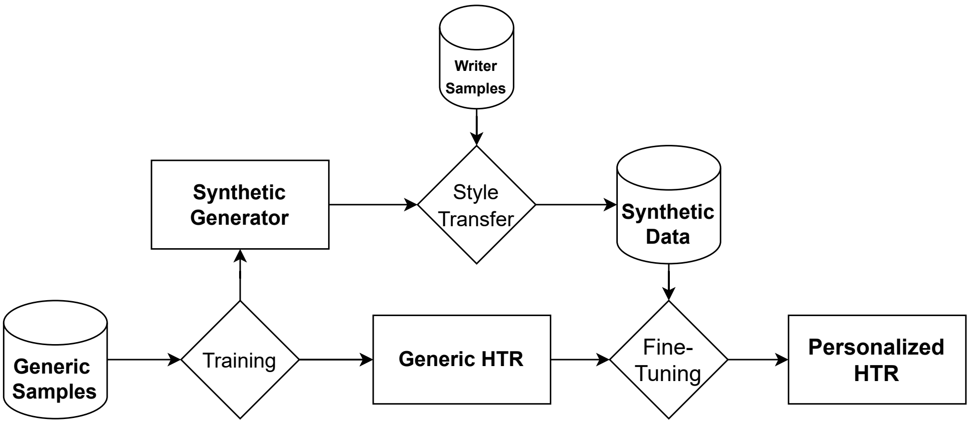


Figure 1: Personalization with style-transferred data

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 specific writers and apply style transfer along with predefined text content to create a synthetic dataset. This dataset is then used for personalization, and we evaluate whether the adaptation was successful.

4 Style Transfer

Style transfer allows us to generate synthetic handwriting samples by transferring a writer's style onto arbitrarily chosen text. This technique enables the generation of a virtually unlimited number of samples, which can be used for personalization. In this section, we describe the style transfer methods and assess their quality through both qualitative and quantitative analysis.

4.1 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 recent studies indicate that 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.

4.2 Data & Evaluation

We use the **GoBo dataset** (Gold et al., 2021) to evaluate style transfer quality. The dataset comprises 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.

We use the **Fréchet Inception Distance** (FID), to evaluate the alignment between style-transferred data and ground truth data (Heusel et al., 2017):

$$\text{FID} = \|\mu_s - \mu_r\|_2^2 + \text{Tr}(C_s + C_r - 2(C_s C_r)^{1/2}) \quad (1)$$

Here, m_s and C_s represent the mean and covariance matrix of the synthetic data, and m_r and C_r correspond to those of the real data.

4.3 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).

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 random samples for style transfer. We then evaluate this synthetic dataset against real handwritten data using the FID metric. The results (see Figure 2) 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 handwriting styles, suggesting that the latent space may not fully capture the diversity of handwriting styles, with some being underrepresented.

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

These examples show that for writers with a low FID range, style transfer yields more consistent handwriting, while a high FID range leads to greater inconsistencies. Notably, for writer 0, the

FID Spread	**Highest**		**Lowest**	
Writer ID	**14**	**0**	**1**	**15**
Best Style	called	leudds	sukwoorts	shrournds
FID	154	143	123	135
Worst Style	sleighphths	mircs	fruiols	knournds
FID	235	222	149	160
Best Transfer	present	discuss	voacsed	Well
Worst Transfer	present	discuss	voacsed	Well
Real	present	disuss	voacsed	well

Table 1: Writers with Highest and Lowest FID Spread

reference leudds, which yielded the lowest FID, produced a less accurate style transfer of the letter s in discuss than the reference mirc, which had the 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.

To further evaluate handwriting generators qualitatively, we generate synthetic data using samples from writers with higher and lower FID spreads and compare them to the original handwritten references (see Appendix, Tables 4 and 5). 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.

4.4 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.

Yet, whether this serves a reliable criterion for successful personalization remains uncertain, as FID measures distributional similarity but may not fully capture the unique characteristics of hand-

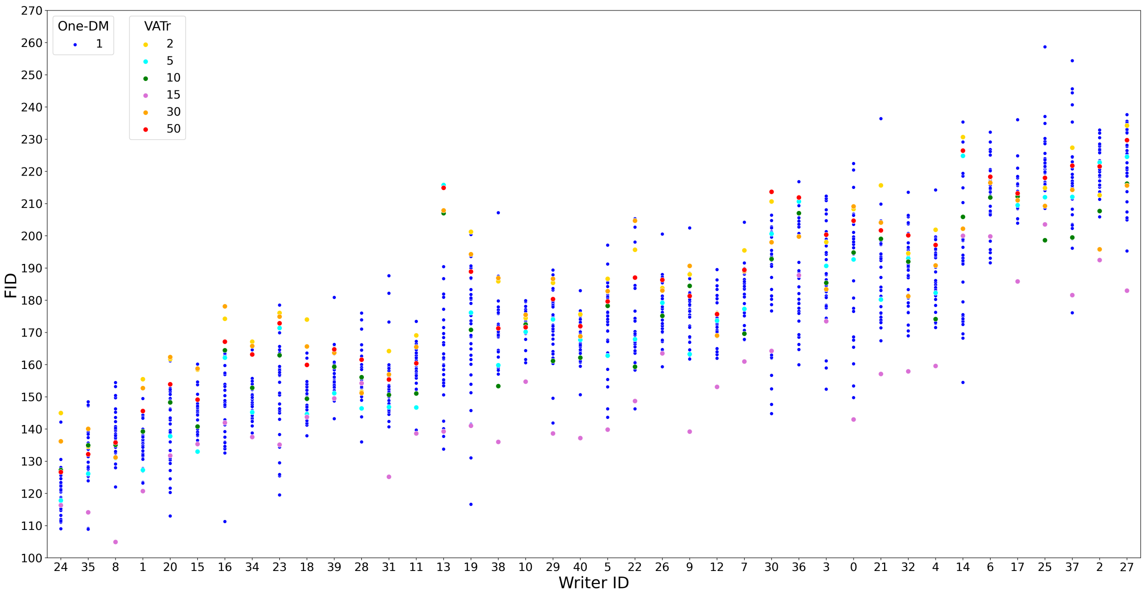


Figure 2: Style transfer quality (in terms of FID values) per writer

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: Dataset statistics

writing. However, the FID values offered valuable insights, particularly regarding the strong dependence on the chosen reference in one-shot learning. 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 will begin personalizing HTR systems with real data to demonstrate its effectiveness. Then, through various experiments, we will personalize with synthetic few-shot and one-shot data and compare the results with those obtained from real data.

5 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 use the standard evaluation metric **Character Error Rate** and three datasets (Table 2 gives an overview). In addition to the GoBo dataset (introduced in Section 4.2), 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.[1] 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.

Baseline Results We first train AttentionHTR on the IAM dataset, using a 70/15/15 split for training, validation, and testing. This yields an in-domain CER of 4.5%. This value is reasonably close to the current state of the art. The 18 systems listed on PapersWithCode[2] for this setup, yield CER values in the range between 2.4 and 7.6%.

Writer Dependence Next, we evaluate how much recognition performance varies between writers. As Figure 3 shows: quite a lot.[3] For example, while the median writer in the IAM dataset experiences a system with a CER of 4.7% and half

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

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

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

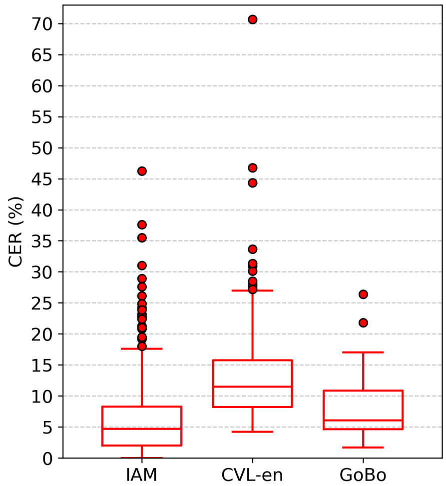


Figure 3: Distribution of CER for individual writers

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.

6 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 for each writer by incrementally adding 10 user-specific samples at each step, reaching a total of 528 instances and evaluate the performance 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 4 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 learning 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

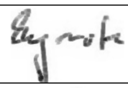
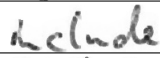
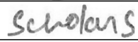
ID	Real	Generic	Personalized
17		Suymort	keynote
10		malnde	include
1		sundons	scholars
26		dissmubue	disputed

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

fully correct predictions after personalization for writers with the highest CER and Figure 10 (see Appendix B) shows the contribution of each word group to personalization.

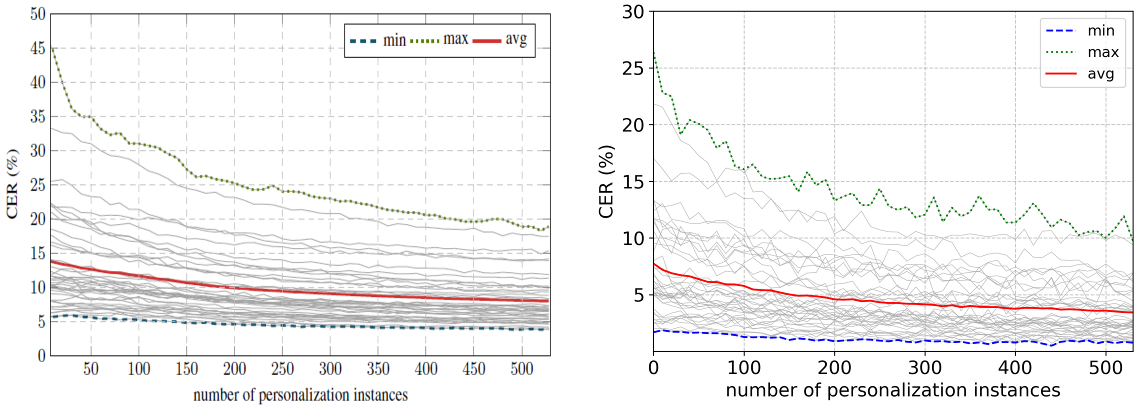
Importance of Real Writer Data There is a theoretical possibility that CER values improve not because we are adding data for a specific writer, but in general due to fine-tuning with more data (as the vocabulary for each writer in the dataset is fixed). To test this, we rerun the personalization experiment using for each writer samples from a different writer for 'personalization'. The (bad) results in Figure 5 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, in Figure 6 we show one exception (writers 10 and 17) who can mutually improve each other's CERs. Interestingly, the writing styles of these two writers are indeed highly similar.[4] If similar writing styles can be mutually be used for personalization, maybe similar generated data can also be used.

7 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 4. First, we generate few-shot and one-shot synthetic datasets for each writer, using 15 random samples for few-shot and the best

[4] In Appendices C and D, we have provided examples for the writers with the highest and lowest CER, where these two writers are also included.



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

(b) Personalization of AttentionHTR

Figure 4: HTR results with increasing samples used for personalization on the GoBo dataset

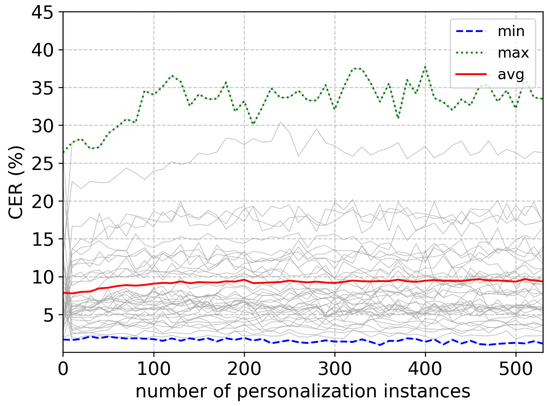


Figure 5: Failed ‘personalization’ when using samples from a different writer

one-shot sample for one-shot (see Figure 2), while using the same words as in the previous experiments to ensure a fair comparison. Since the best one-shot sample is unknown in practice, this scenario remains 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 compares personalization with few-shot style-transferred data to the baseline and to personalization with real data from random writers and the actual writer. Notably, the CERs with style-transferred data are only slightly lower than those with real data from a random writer, suggesting an inconsistent resemblance to real handwriting.

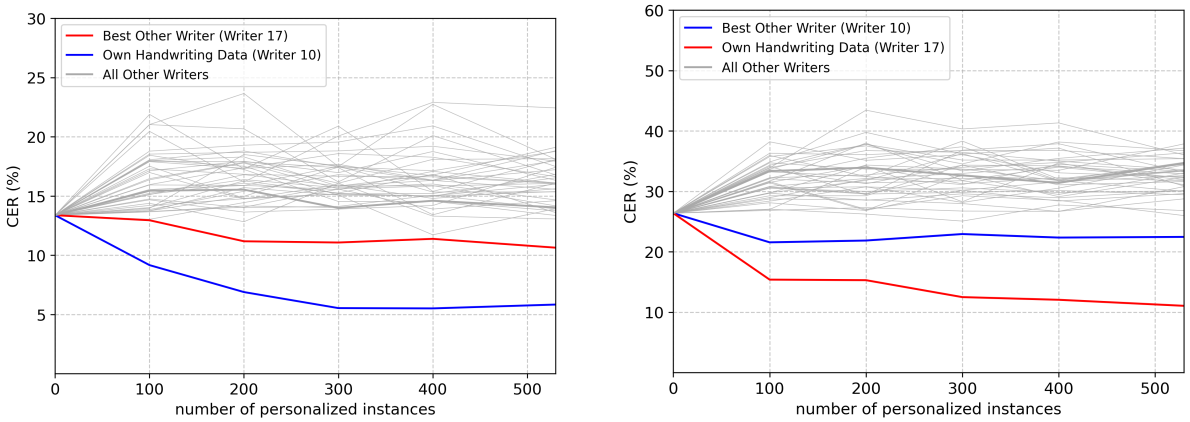
Mixing real and synthetic In previous work, using both real and synthetic data was successfully used, thus we next personalize using a mixed dataset composed of 50 % synthetic and 50 % real samples. The results are also shown in Figure 7. Notably, the CERs at 250 samples (≈ 47 % real data) without synthetic data in Figure 4b are significantly lower than those in the mixed synthetic dataset in Figure 7. Hence, we examine the effect of initially personalizing with 250 real samples before continuing with synthetic data (see Figure 8). After integrating the first 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.

More data Although our previous experiments showed that synthetic data do 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[5], using 30% for validation to select the best model within 1–6 epochs (see Figure 9). The unchanged CERs further confirm that the handwriting generator's style transfer is insufficient to achieve adequate similarity to real data.

8 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.

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



(a) Writer 10

(b) Writer 17

Figure 6: 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

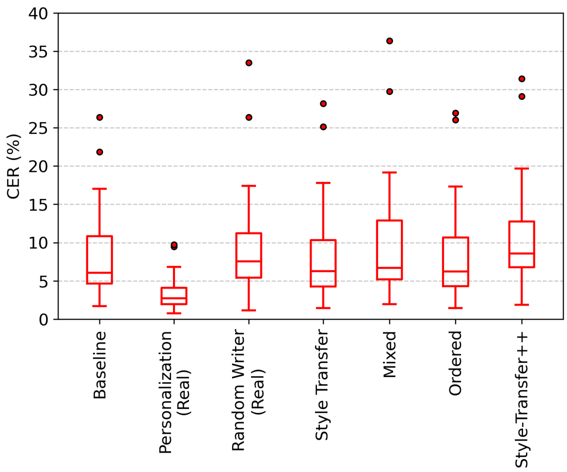


Figure 7: Overview of personalization results

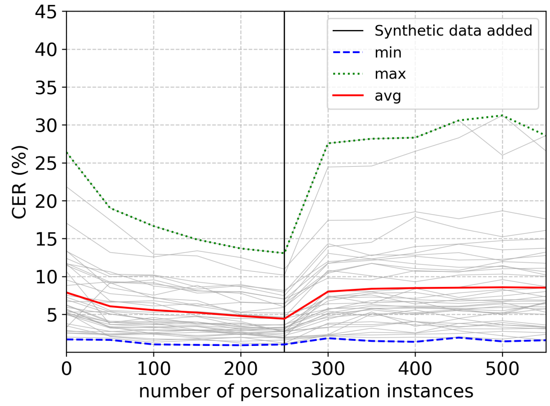


Figure 8: Personalization using mixed ordered datasets

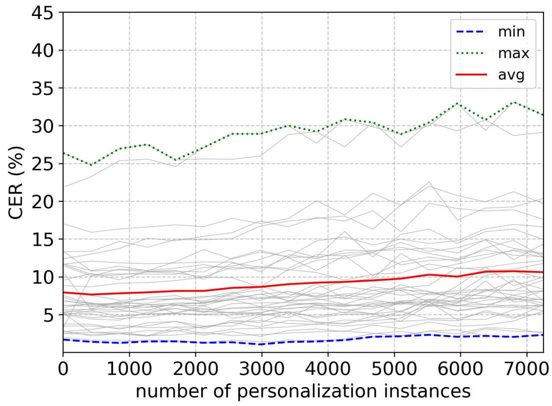


Figure 9: Personalization using 7000 synthetic training and 3000 validation samples of frequent English words

Even when using real data from writers with similar handwriting, personalization resulted in only a slight improvement in CER. This implies that synthetic handwriting must exhibit a much higher degree of similarity than what can be generated with currently available style transfer methods. In particular, it remains unclear which specific features of handwriting are essential for ensuring successful personalization.

Future Work Future research should focus more on investigating whether style-transferred data can effectively reduce CER in a writer-dependent manner, rather than just improving generic models. To address this, it is essential to understand when synthetic data aligns closely enough with real handwriting, and which specific features of handwriting must be considered in the evaluation process. To better evaluate this, further research is needed to identify suitable metrics for measuring handwriting similarity, as it remains uncertain whether existing approaches adequately capture the writer-specific characteristics required for personalization.

Limitations

We only evaluate on one language (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 a potential security risk. Our research into ad-hoc personalization is a step towards solving this issue, as no writing samples have to be stored if personalization from a single, directly obtained and then discarded, sample 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.

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.

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.

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.

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.

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.

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*.

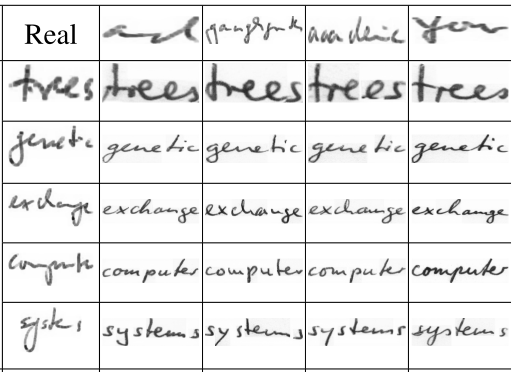
Vijay Rowtula, Subba Reddy Oota, and Jawahar C.V. 2019. Towards Automated Evaluation of Handwritten Assessments. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 426–433.

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.

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

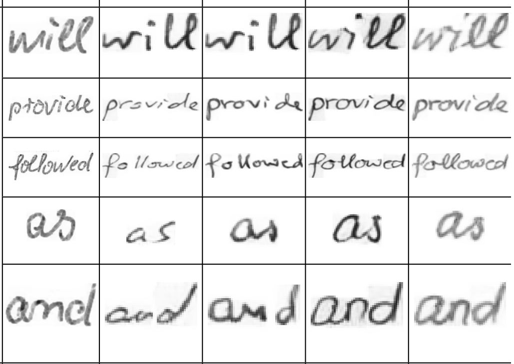
Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, Jun Sun, and Cheng-Lin Liu. 2018. Deep transfer mapping for unsupervised writer adaptation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 151–156.

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

B Word Group Impact on Personalization

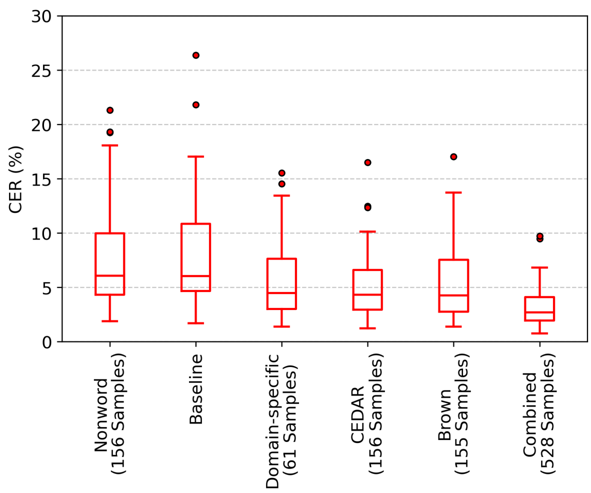


Figure 10: Personalization by word group

C Handwriting Examples of Writers with Highest CERs

Figure 11: Writers with Highest CERs (continued on next page)

Figure 11: Writers with Highest CERs

D Handwriting Examples of Writers with Lowest CERs

Figure 12: Writers with Lowest CERs (continued on next page)

Figure 12: Writers with Lowest CERs