StylusAI: Stylistic Adaptation for Robust German Handwritten Text Generation

Nauman Riaz (✉)[1,2,0009−0000−1416−3220], Saifullah Saifullah[1,2,0000−0003−3098−2458], Stefan Agne[1,3,0000−0002−9697−4285], Andreas Dengel[1,2,0000−0002−6100−8255], and Sheraz Ahmed[1,3,0000−0002−4239−6520]

[1] Smart Data and Knowledge Services (SDS), German Research Center for Artificial Intelligence GmbH (DFKI), Trippstadter Straße 122 67663 Kaiserslautern
{firstname.lastname}@dfki.de

[2] Department of Computer Science, RPTU Kaiserslautern-Landau, Erwin-Schrödinger-Straße 52, 67663 Kaiserslautern, Germany

[3] DeepReader GmbH, 67663 Kaiserlautern, Germany

Abstract. In this study, we introduce StylusAI, a novel architecture leveraging diffusion models in the domain of handwriting style generation. StylusAI is specifically designed to adapt and integrate the stylistic nuances of one language's handwriting into another, particularly focusing on blending English handwriting styles into the context of the German writing system. This approach enables the generation of German text in English handwriting styles and German handwriting styles into English, enriching machine-generated handwriting diversity while ensuring that the generated text remains legible across both languages. To support the development and evaluation of StylusAI, we present the 'Deutscher Handschriften-Datensatz' (DHSD), a comprehensive dataset encompassing 37 distinct handwriting styles within the German language. This dataset provides a fundamental resource for training and benchmarking in the realm of handwritten text generation. Our results demonstrate that StylusAI not only introduces a new method for style adaptation in handwritten text generation but also surpasses existing models in generating handwriting samples that improve both text quality and stylistic fidelity, evidenced by its performance on the IAM database and our newly proposed DHSD. Thus, StylusAI represents a significant advancement in the field of handwriting style generation, offering promising avenues for future research and applications in cross-linguistic style adaptation for languages with similar scripts.

Keywords: Handwriting Generation · Diffusion Models· Handwriting Text Recognition · Transformers

1 Introduction

Despite significant technological advancements in our society, the use of traditional handwritten text remains widely popular for documenting data, making

the task of handwritten text recognition (HTR) critically important for automated document processing. However, extensive data complexity in handwritten texts, such as varying writing styles and languages [7,14], low-quality images [1,16,24], and lighting variations [16,24,23], makes HTR a challenging task. While recent Deep Learning (DL)-based systems have shown promising potential for improvement [15,25,18], these models are mostly data-driven, especially transformer models, known for requiring extensive data for optimal performance. On the other hand, manually gathering and annotating handwritten text is an extremely labor-intensive and time-consuming task, requiring significant human effort [7,14].

Due to the aforementioned challenges in gathering handwritten text data, augmenting these datasets through image synthesis is seen as a popular alternative [6,2]. Numerous synthesis and data augmentation approaches have been proposed in recent years [2,17,21,26], leading to significant improvements in the performance of existing HTR models [15,22]. These include the previous state-of-the-art (SotA) approaches based on Generative Adversarial Network (GANs) [13,12,20,4] and the recent SotA approaches based on Diffusion Model (DMs) [17,21,31], both of which typically synthesize handwritten text images by conditioning the generation process on the target text. While the most recent DM-based synthesis methods [21,31] have significantly improved over the previous SotA approaches [13,12,20,4], there is still a significant lack of research that explores style adaptation between similarly written languages. This research gap is particularly important, especially for enhancing the synthesis of handwritten text for languages with limited resources. Specifically, resource-constrained languages could benefit from adopting diverse styles from well-resourced languages that share similar scripts. In this paper, we aim to investigate this possibility for the German language by exploring style adaptation from English to German.

Despite the similarity between the German and English languages in their written forms, there exist specific German characters (as shown in Fig. 1) such as **ä**, **ö**, **ü**, and **ß**, that do not exist in the English vocabulary. In addition, there is a huge scarcity of publicly available German handwritten text datasets, making it difficult to achieve sufficient diversity when synthesizing German handwritten text using existing data-driven text synthesis methods. While one may train existing approaches on publicly available English language datasets, such as the

Fig. 1: Few images showcasing a selection of written German words, with a particular focus on the unique characters found in the German alphabet, including umlauts and the eszett.

IAM Handwriting Database [19], to generate text for overlapping characters between the two languages (English and German), such an approach will naturally fail to generate text for out-of-vocabulary characters, such as those mentioned above. In this work, therefore, our main focus is to explore the possibility of adapting writing styles from large-scale English language datasets to the out-of-vocabulary German characters, so that small-scale German language datasets may then be augmented to produce diverse, style-rich datasets. We achieve this by designing a conditional diffusion model, the generation process of which is guided not only through text and writer style but also by an additional synthetic printed text image. This addition of a visual representation of the target text allows us to model the problem as an image-to-image translation task, which helps improve style adaptation compared to existing text-only conditional approaches.

The main contributions of this paper are three-fold:

1. We present the 'Deutscher Handschriften-Datensatz' (DHSD), a German handwriting dataset that comprises 37 distinct handwriting styles.
2. We propose StylusAI, an architecture based on diffusion models for handwritten text generation, for effective style adaptation incorporating the stylistic elements present in English handwriting into the context of the German writing system, creating a fusion that maintains legibility and coherence for both languages. This allows German to be generated in English writer styles and vice versa leading to diverse handwriting style generation.
3. Furthermore, we show that the proposed model also outperforms previously suggested models on datasets such as IAM and our new DHSD in terms of producing handwriting samples with superior text and stylistic quality.

The rest of the paper is structured into the following main sections. In Section 2, a summary of related work in the field of handwritten text generation is presented. Section 3 explores the proposed technique for the task at hand. Section 5 details the experimental setup, including preprocessing steps and implementation details. Section 6 presents the findings and their interpretation. Finally, Section 7 concludes the study and outlines future research directions.

2   Related Work

Handwritten text generation holds immense significance in the field of document analysis and recognition and has been widely explored in the past decade [6,13,20,4,31]. The earliest attempts in this domain involved using Recurrent Neural Networks (RNNs) for online handwriting generation [6]. In particular, Graves [6] proposed using Long Short-Term Memory (LSTM) networks [10] to predict text-conditioned real-valued data sequences for synthesizing handwritten text in an online fashion.

More recently, the field has greatly shifted its attention towards offline text generation using generative approaches [13,20,31]. Kang *et al.* [13] proposed GANwriting for handwritten text generation, a conditional Generative Adversarial Network (GAN) [5] that incorporated handwriting image samples of the

writers for style information alongside the target text conditioning to guide the generation process. While the approach achieved great success in generating realistic handwritten text images at the word level (later also extended for sentence-level generation [12]), allowing for a controlled generation of previously unseen text sequences with various writer styles, it greatly suffered from unrealistic pen-level artifacts introduced in the generated images. This issue was addressed in a follow-up work, SmartPatch [20], where an additional patch-level discriminator loss was employed to improve the generation quality. Fogel *et al.* [4] proposed ScrabbleGAN, where a GAN was trained in a semi-supervised fashion to generate comprehensive handwritten text sentences across various styles and content. In a different direction, Bhunia *et al.* [2] proposed an encoder-decoder transformer architecture that was trained to learn not only the long and short-range contextual relationships but also the style-content relationship through self-attention [30].

The recent success of diffusion models in natural image synthesis [9,26,28] has also sparked an interest in using them for the task of handwritten text generation. Luhman *et al.* [17] recently proposed a conditional diffusion model for generating handwritten text sequences in an online fashion. In particular, their approach uses diffusion to generate real-valued pen stroke sequences and introduces style and textual conditioning through embeddings generated by a pre-trained MobileNetV2 model [11]. It is important to highlight that while their approach utilized diffusion, it focused solely on online generation, which contrasts with the offline generation focus of this work. In another recent work, Nikolaidou *et al.* [21] proposed WordStylist, a UNet [27]-based text-to-image latent diffusion model [26] for offline generation of handwritten text. In particular, they utilized a transformer-based architecture to generate character-level text embeddings and incorporated a cross-attention mechanism into the UNet model to condition the generation process on the target text and style embeddings. It is worth mentioning that although WordStylist [21] demonstrated exceptional performance in generating realistic handwritten text sequences, significantly outperforming the previous state-of-the-art [13,20,2], it was not trained to achieve cross-language style adaptation, which is the main focus of this work. Later, we demonstrate through results that WordStylist [21] struggles to adapt styles from English to German when queried to generate out-of-vocabulary German characters in the style of English writers, despite being trained on a dataset that includes both English and German texts.

3   Background

3.1   Diffusion Models

Diffusion models represent a class of generative models utilizing Markov chains to systematically introduce noise, thus obscuring the data's original structure. These models are tasked with learning how to invert this process, aiming to restore the data to its initial form. Their design is influenced by principles found in thermodynamics [29], and these models have become increasingly prominent in image synthesis for their capability to produce high-fidelity images. The Diffusion

Model consists of two main phases: the forward process, which involves the diffusion of a sample, and the reverse process, which involves denoising [9].

Forward Diffusion In the forward diffusion process, an initial sample, denoted by x_0 is obtained from a distribution $q(x_0)$. This sample is subsequently perturbed with Gaussian noise to produce a latent variable x_1. The procedure of introducing noise and creating subsequent latent variables $(x_2, x_3, \ldots, x_T)$ is repeated until it reaches a predetermined hyperparameter T. Mathematically, the relationship between the latent variables is expressed as follows:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad q(x_t|x_{t-1}) \sim \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \quad (1)$$

where we have $\beta_i \in [0,1]$ for all $i \in [1,T]$. $\beta_1, \beta_2, ..., \beta_T$ form a noise variance schedule and dictate the quantity of noise added at every timestep. In the concluding step, assuming a sufficiently large T and an appropriate noise schedule, we will have x_T that closely resembles a sample drawn from pure Gaussian noise.

Reverse Diffusion During the denoising phase, a U-Net architecture is employed to iteratively mitigate noise emanating from a normal distribution, with the objective of restoring the original dataset. The process of image generation relies on a sequential sampling technique. It starts by generating a sample from $q(x_T)$. Then, a sample is taken from the distribution at the previous timestep, conditional on the value of x_T, and this procedure is repeated in reverse order until reaching x_0.

The noise reduction across reverse timesteps is guided by the transition:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \quad p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$
$$(2)$$

The network is trained by minimizing the variational lower bound between the posterior of the forward process and the joint distribution of the reverse process, represented as p_θ.

The training loss is defined as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x},t,\epsilon \sim \mathcal{N}(0,1)} \left[||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||_2^2 \right] \quad (3)$$

where this loss quantifies the discrepancy between the true noise ϵ and the noise predicted by the network ϵ_θ.

4   StylusAI: The Proposed Approach

Inspired by the work of Brooks *et al.* [3], instead of relying solely on simple text conditioning as done in previous work [21], we enhance our method by incorporating synthetic printed text image conditioning during the diffusion process. We
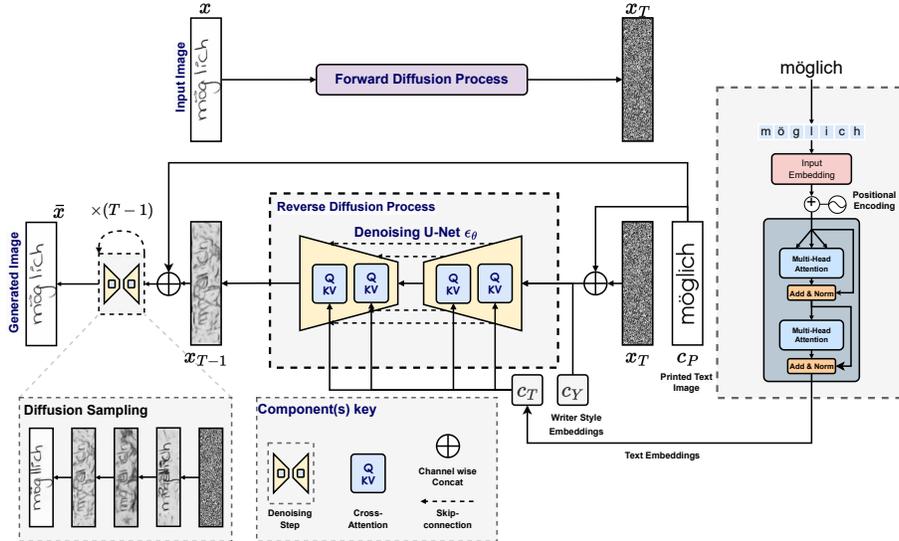
Fig. 2: Overview of the Proposed Architecture.

approach the problem of style adaptation by modeling a diffusion process whose goal is to generate a handwritten text image from a printed text image, adhering to a specific style and text. The printed text image provides details on the appearance of the characters, and the diffusion process meticulously generates those characters in the targeted writer's style. This innovative approach allows for better style adaptation of German characters in English writer styles and vice versa. Given a handwriting image x, the diffusion process aims to add noise to it, producing a noisy image x_t where the noise level increases over timesteps $t \in T$. We propose a U-Net based network ϵ_θ that predicts the noise added to x_t given the text conditioning c_T, writer style conditioning c_Y, and printed text image conditioning c_P. The following diffusion objective is minimized to train the network:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}, \mathbf{c}_T, \mathbf{c}_Y, \mathbf{c}_P, \epsilon \sim \mathcal{N}(0,1)} \left[||\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}_T, \mathbf{c}_Y, \mathbf{c}_P)||_2^2 \right] \tag{4}$$

StylusAI leverages synthetically generated printed text images, character-level text, and writer-style embeddings to facilitate controlled handwritten text generation. It employs a standard Transformer Encoder [30] for text conditioning. For predicting noise, StylusAI utilizes a U-Net [27] based model as illustrated in the Figure 2. The addition of printed text image in the conditioning allows for better style adaptation between similarly written languages as evidenced from results described in Section 6. The details of the architecture are given below.

4.1   Transformer Encoder and Style Conditioning

We use character-level tokenization for text conditioning and pass the tokens through two stacks of transformer encoder layers, which utilize multi-head self-attention layers to generate text embeddings. The attention mechanism is defined as $\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$ which gives a weighted sum of character embedding representation. These embeddings are used for conditioning via cross-attention with features from U-Net layers. The writer style condition c_Y is processed through an embedding layer and subsequently added to the time step embedding as shown in Figure 2.

4.2   Forward Diffusion and Training

We aim to train a diffusion model $p_\theta(x \,|\, c_T, c_Y, c_P)$, given the input text c_T, writer style c_Y, and a synthetic printed text image c_P. The content of the input text c_T, which is to be conditioned on, matches that of the printed text image. We sample the timesteps $t \in T$ from a uniform distribution which are encoded using a sinusoidal position embedding. Noise is incrementally added to the original image x. A noise scheduler incrementally increases the level of noise from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ over $T = 1000$ timesteps. To predict the noise ϵ, we utilize a U-Net architecture that comprises Residual Blocks [8] and intermediate cross-attention Transformer blocks [26]. The cross-attention blocks are employed to attend to the text conditioning. The network accepts as input the noisy image, the corresponding timestep, and the specified conditions c_Y, c_T, and c_P (i.e., the printed text image concatenated channel-wise with the noisy image). The training goal is to minimize the discrepancy between the noise prediction of the network and the actual noise in the image mathematically described in Equation 4.

4.3   Reverse Diffusion: Denoising

For the generation process, we employ the learned diffusion model for reverse diffusion. We feed a noisy random sample x_T into the learned network ϵ_θ along with a channel-wise concatenation of the printed text image c_P, text condition c_T, and style condition c_Y to predict the noise. The predicted noise is progressively removed at each timestep, repeating this process from T to $t = 0$, until we obtain our handwritten text image sample. We use the DDPM algorithm for the denoising process. The denoising process is graphically shown in Figure 2.

5   Experimental Setup

5.1   Datasets

We employ two different datasets, the IAM dataset [19] and the proposed DHSD dataset. The specifics of the mentioned dataset including the splits used for training are given below:

Fig. 3: These images provide a glimpse into the diversity of IAM dataset, showcasing a variety of unique handwriting styles

IAM: The IAM Handwriting Database is a comprehensive resource designed for use in handwriting recognition and related research. It encompasses a collection of handwriting samples provided by 657 different writers, ensuring a diverse representation of handwriting styles. In total, the database contains 1,539 pages of scanned text, which further breaks down into more granular elements including 5,685 isolated and labeled sentences. Researchers can also access 13,353 isolated and labeled text lines for studying specific handwriting characteristics at the line level. At the most detailed level, the database features a substantial compilation of 115,320 isolated and labeled words, offering an in-depth opportunity for word recognition analysis. These words were extracted from the scanned pages using an automatic segmentation method, which was subsequently followed by a thorough manual verification process to ensure the accuracy and reliability of the segmentation and labeling. We use labeled word images in our experiments and employ the Aachen splits, similar to [21], for training purposes. Figure 3 displays sample images of handwritten words from the IAM database.

Deutscher Handschriften-Datensatz (DHSD): We propose a novel German handwriting dataset, which has been populated with contributions from 37 individuals, each providing an average of 150 words. To guarantee the representation of the entire German alphabet within the dataset, we carefully selected words to ensure that each one includes at least one letter from the German alphabet. The dataset comprises words that are names of cities and streets in Germany, all extracted from the OpenStreetMap (OSM) database. The dataset is split into training and testing subsets, with 80% allocated for training and 20% for testing. Sample images of the handwritten words from our proposed dataset are displayed in Figure 4.

IAM + DHSD For the purpose of evaluating our model's ability to adapt German characters in different English writing styles, we integrated the IAM dataset with our newly proposed DHSD. Specifically, we compiled data from 37 English writers within the IAM dataset and paired it with data from 37

Fig. 4: These images provide a glimpse into the diversity of our proposed DHSD, showcasing a variety of unique handwriting styles.

German writers from our proposed dataset. This merged dataset was employed for training our model. The top 37 writers who contributed the most writing samples were selected from the IAM dataset with an average of 326 words per writer. The experimentation on this combined dataset is conducted differently compared to the datasets mentioned above. The outcomes of this integration are elaborated upon in Section 6.

5.2 Experimental Setup

Our experimental setup is largely based on the one described by [21], with modifications to accommodate our newly proposed DHSD dataset. We evaluate the model's performance by assessing the accuracy of both handwriting text recognition and writer style recognition on the handwritten text images produced by the generative model. We intentionally avoid using the Fréchet Inception Distance (FID) metric. Although FID is a common metric for assessing generative models, it might not be suitable for tasks that deal with image types substantially different from the natural images seen in ImageNet, which was used to train the underlying FID network. This discrepancy between the types of images can undermine the reliability of the evaluation process. However, modifying the metric by fine-tuning the FID network on a dataset of document images goes beyond the scope of this paper.

5.3 Implementation Details

Our proposed architecture is implemented using a U-Net model, which includes five encoder and decoder blocks. The initial 2 blocks and the last block in the encoder are ResNet layers with downsampling. We use one Resnet layer per each U-Net block. The remaining two blocks feature cross-attention layers with 4 heads for Multi-Head attention. Conversely, the U-Net's decoder segments mirror this configuration by employing upsampling, following the same sequence in reverse. During the denoising process, we append a synthetically generated printed text image to the noisy image at each timestep. This printed text image,

which corresponds to the text designated for generation, is appended along the channel dimension. The timestep and writer-style embeddings are both kept at a dimension of 256. The proposed architecture was trained using four A100-40GB GPUs, with a collective batch size of 256. All text images were resized to 256 x 64 pixels in grayscale, maintaining the aspect ratio. We trained and evaluated the architecture across three different dataset configurations as detailed in Subsection 5.1.

6   Results and Analysis

We conduct our assessment following the evaluation framework outlined in [21], spanning two dimensions: textual quality and style quality for IAM and DHSD. For textual quality, we create synthetic training sets (generated by the models) like the training splits of the dataset configurations IAM and DHSD as detailed in Section 5. We use these sets to train a handwriting recognition model [25], which is then evaluated on the original test sets to determine the model's performance on individual datasets. Meanwhile, style quality is gauged by training a conventional CNN on IAM and DHSD to classify writers' styles. The effectiveness of this approach is tested on the generated handwriting samples.

For the assesment of style adaptation we also use textual quality and style quality but the experimentation is done a bit differently. The evaluation process differs in this case, as we utilize a combined dataset comprising both the IAM and DHSD datasets for training the generative model. Subsequently, the trained model is employed to generate a modified version of the combined dataset, where the styles are reversed. This entails generating German words in the style of 37 English writers with a one-to-one correspondence between words and writers.

6.1   Handwriting Text Recognition (HTR)

We evaluate text quality using a handwriting text recognition model (HTR) proposed by [25], which we train on synthetically generated training sets of IAM and DHSD. This evaluation approach is similar to what is proposed by the authors of WordStylist [21]. These sets were synthesized utilizing previous methods such as GANwriting, Smart Patch, Word Stylist, and our proposed approach for comparison. The Handwritten Text Recognition (HTR) model, once trained, undergoes evaluation on the original test splits from both datasets using the Character Error Rate (CER) as a measure of text quality. A lower CER signifies that the HTR model has effectively generalized to the test set, suggesting that the synthetically generated words used in training are legible and closely resemble the original test set distribution. This similarity is the intended result of a good handwritten text generation model. The evaluation results for text quality on the IAM dataset are presented in Table 1, and on the DHSD dataset in Table 2. Our proposed model outperforms previous existing models in both cases.

Table 1: The HTR results for the IAM dataset using the Character Error Rate (CER), where a lower rate signifies improved performance.

Training Data	CER(%)
Real IAM	4.57 ± 0.07
GANwriting IAM	35.21 ± 0.23
SmartPatch IAM	30.25 ± 0.45
WordStylist IAM	8.50 ± 0.12
StylusAI IAM (Ours)	7.82 ± 0.09
Real IAM + WordStylist	4.42 ± 0.08
Real IAM + StylusAI IAM (Ours)	$\mathbf{3.85 \pm 0.09}$

Table 2: The HTR results for the DHSD dataset using the Character Error Rate (CER), where a lower rate signifies improved performance.

Training Data	CER(%)
Real DHSD	11.13 ± 0.07
GANwriting DHSD	42.31 ± 0.13
SmartPatch DHSD	37.47 ± 0.25
WordStylist DHSD	14.58 ± 0.12
StylusAI DHSD (Ours)	11.57 ± 0.09
Real DHSD + WordStylist DHSD	11.62 ± 0.08
Real DHSD + StylusAI DHSD (Ours)	$\mathbf{9.01 \pm 0.08}$

The authors of WordStylist propose this method of evaluating textual quality, which we have chosen to use as-is for IAM and DHSD datasets. However, we recognize that it may not accurately reflect the generative capabilities of a generative model. Generating the same training set on which the model was originally trained could result in good-quality images, but this may be due to overfitting. To effectively assess the extent of style adaptation of the generative model we follow a different approach.

HTR for style adaptation To assess the effectiveness of style adaptation in text quality, we employ the combined IAM+DHSD dataset to train the generative model. Once the model is trained, we use it to generate the DHSD training split in the styles of 37 English writers found in the combined dataset. This newly created synthetic training data (Eng-DHSD) is then used to train the HTR model and it does not depict the original training distribution as German words are not present in the style of English writers during the training of the generative model. The performance of the HTR model is evaluated on the test split of the DHSD dataset. A lower CER would indicate a more successful style

Table 3: The HTR results after training on Eng-DHSD dataset and testing on DHSD testing split using the Character Error Rate(CER), where a lower rate signifies improved performance.

Training Data	CER(%)
Real DHSD	11.13 ± 0.07
WordStylist Eng-DHSD	45.80 ± 0.12
StylusAI Eng-DHSD (ours)	30.86 ± 0.07
Real DHSD + WordStylist Eng-DHSD	12.67 ± 0.08
Real DHSD + StylusAI Eng-DHSD (Ours)	$\mathbf{10.24 \pm 0.09}$

adaptation. Only WordStylist and StylusAI are considered for comparison because GANwriting and SmartPatch did not yield promising results for this task. The comparison suggests that our model outperforms WordStylist as shown in Table 3.

Generating German characters in the style of English writers is a difficult task for a generative model that has been trained on handwritten text images without examples of German characters written in the style of English writers. Consequently, when it comes to generating German characters in the style of English writers, the model produces inconsistent results. The consistency of these generations varies among different English writer styles. Please refer to Figure 5 for visual examples.

Fig. 5: German text samples generated by StylusAI, emulating different IAM writer styles, where good generations are those in which the German characters have been better adapted to the English writer styles compared to average and poor generations.

Despite the inconsistencies, StylusAI consistently produces better results compared to WordStylist across various writing styles. The incorporation of a printed text image, along with character embeddings, during the denoising process provides additional guidance on the basic style of individual characters, resulting in more consistent outcomes. Figure 6 provides a comparison between generations of WordStylist and StylusAI (ours).
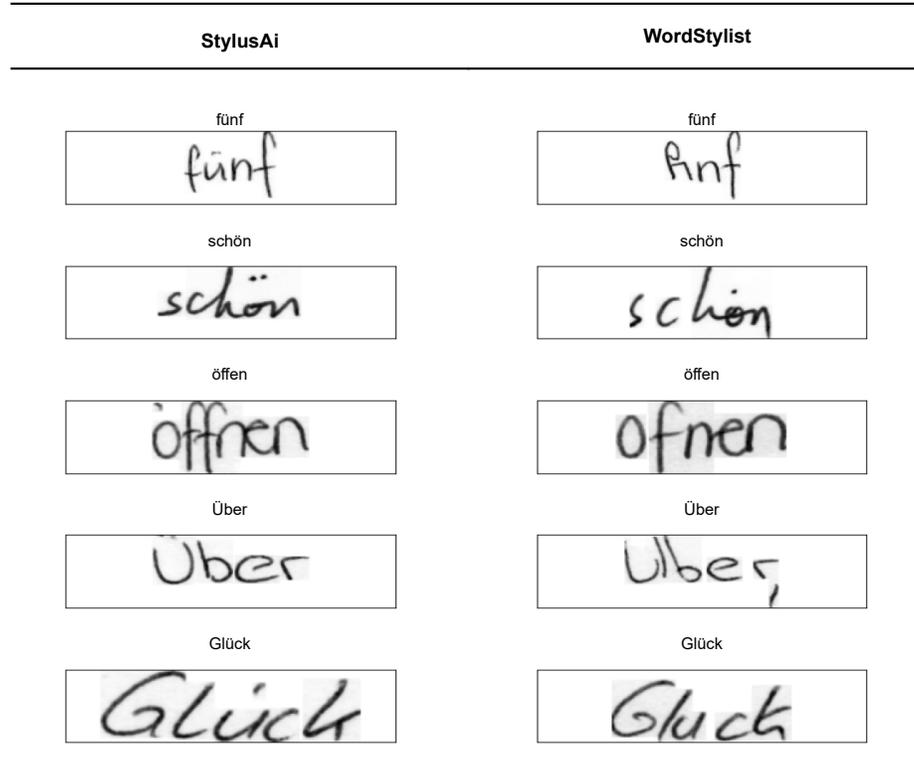

Fig. 6: Comparison between German word generations from StylusAi and WordStylist in IAM English writer styles. StylusAI demonstrates superior quality in adapting German characters to English writer styles.

6.2   Writing Style Analysis

To assess the model's effectiveness in capturing the diverse writing styles of different writers, we followed an evaluation methodology similar to the one outlined in [21]. To evaluate the generated styles, we fine-tuned a ResNet-18 CNN [8], originally pre-trained on ImageNet, utilizing the IAM and DHSD databases for

writer classification tasks. Following this, we utilized the datasets produced by the four generative approaches as test sets and presented the resulting accuracy. For analysis of writer style classification on style adaptation task, we utilize the same pattern by training on IAM+DHSD and testing on the Eng-DHSD synthesized set from the trained generative models. The evaluation for writer classification is shown in Table 4. The results show that StylusAI is able to adapt the style better while also producing fewer errors while generating German characters.

Table 4: Comparison of the accuracy of a ResNet18 model trained for writer identification on IAM, DHSD, and IAM+DHSH datasets and tested on the generated IAM, DHSD and Eng-DHSD datasets.

Method	IAM (%)	DHSD (%)	Eng-DHSD (%)
GANwriting	4.81	6.72	-
SmartPatch	4.09	7.28	-
WordStylist	70.67	73.50	62.38
StylusAI (Ours)	**75.25**	**75.02**	**66.79**

7   Conclusion and Future Work

This research introduces the 'Deutscher Handschriften-Datensatz'(DHSD), a comprehensive dataset encompassing a wide array of German handwriting styles, laying the groundwork for novel applications in German handwriting analysis and generation. Leveraging this dataset, we developed StylusAI, a state-of-the-art architecture premised on diffusion models, tailored for the intricate task of handwriting style adaptation. StylusAI represents a significant stride forward, that combines stylistic elements prevalent in English handwriting with those inherent to the German writing system. This amalgamation not only preserves but enhances the legibility and stylistic cohesion across both languages, promoting a seamless generation of diverse handwriting styles. Our extensive evaluations demonstrate that StylusAI not only achieves but surpasses the performance benchmarks of existing models in the realm of handwritten text generation. Its capabilities are evident when assessed on both the newly curated DHSD and the established IAM datasets, where it consistently generates handwriting samples of superior text and stylistic quality.

This paper signifies the immense potential of employing diffusion models in the context of cross-linguistic handwriting synthesis between similarly written languages, paving the way for advancements in the field of handwritten text generation. One interesting future direction includes the exploration of other similarly written languages to enhance handwritten text generation and consequently handwriting text recognition systems.

References

1. Aradillas, J., Murillo-Fuentes, J., Olmos, P.: Boosting offline handwritten text recognition in historical documents with few labeled lines. IEEE Access **PP**, 1–1 (05 2021). https://doi.org/10.1109/ACCESS.2021.3082689
2. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Khan, F.S., Shah, M.: Handwriting transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1086–1094 (2021)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
4. Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., Litman, R.: Scrabblegan: Semi-supervised varying length handwritten text generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4324–4333 (2020)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
6. Graves, A.: Generating sequences with recurrent neural networks. ArXiv **abs/1308.0850** (2013), https://api.semanticscholar.org/CorpusID:1697424
7. Grosicki, E., El-Abed, H.: Icdar 2011 - french handwriting recognition competition. In: 2011 International Conference on Document Analysis and Recognition. pp. 1459–1463 (2011). https://doi.org/10.1109/ICDAR.2011.290
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10.1162/neco.1997.9.8.1735
11. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017)
12. Kang, L., Riba, P., Rusinol, M., Fornes, A., Villegas, M.: Content and style aware generation of text-line images for handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 8846–8860 (2021)
13. Kang, L., Riba, P., Wang, Y., Rusinol, M., Fornés, A., Villegas, M.: Ganwriting: content-conditioned generation of styled handwritten word images. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16. pp. 273–289. Springer (2020)
14. Kleber, F., Fiel, S., Diem, M., Sablatnig, R.: Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 560–564 (2013). https://doi.org/10.1109/ICDAR.2013.117

15. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 13094–13102 (2023)
16. Lins, R.D.: Nabuco - two decades of document processing in latin america. J. Univers. Comput. Sci. **17**(1), 151–161 (2011), http://dblp.uni-trier.de/db/journals/jucs/jucs17.html#Lins11a
17. Luhman, T., Luhman, E.: Diffusion models for handwriting generation. ArXiv **abs/2011.06704** (2020), https://api.semanticscholar.org/CorpusID:226955899
18. Maqsood, A., Riaz, N., Ul-Hasan, A., Shafait, F.: A unified architecture for urdu printed and handwritten text recognition. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) Document Analysis and Recognition - ICDAR 2023. pp. 116–130. Springer Nature Switzerland, Cham (2023)
19. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for off-line handwriting recognition. International Journal on Document Analysis and Recognition **5**, 39–46 (2002)
20. Mattick, A., Mayr, M., Seuret, M., Maier, A., Christlein, V.: Smartpatch: Improving handwritten word imitation with patch discriminators. In: International Conference on Document Analysis and Recognition. pp. 268–283. Springer (2021)
21. Nikolaidou, K., Retsinas, G., Christlein, V., Seuret, M., Sfikas, G., Smith, E.B., Mokayed, H., Liwicki, M.: Wordstylist: Styled verbatim handwritten text generation with latent diffusion models. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) Document Analysis and Recognition - ICDAR 2023. pp. 384–401. Springer Nature Switzerland, Cham (2023)
22. Pippi, V., Cascianelli, S., Cucchiara, R.: Handwritten text generation from visual archetypes (2023)
23. Pratikakis, I., Gatos, B., Ntirogiannis, K.: Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012). In: ICFHR. pp. 817–822. IEEE Computer Society (2012), http://dblp.uni-trier.de/db/conf/icfhr/icfhr2012.html#PratikakisGN12
24. Pratikakis, I., Zagoris, K., Kaddas, P., Gatos, B.: Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018). In: ICFHR. pp. 489–493. IEEE Computer Society (2018), http://dblp.uni-trier.de/db/conf/icfhr/icfhr2018.html#PratikakisZKG18
25. Riaz, N., Arbab, H., Maqsood, A., Nasir, K., Ul-Hasan, A., Shafait, F.: Convtransformer architecture for unconstrained off-line urdu handwriting recognition. International Journal on Document Analysis and Recognition (IJDAR) **25**(4), 373–384 (2022)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models pp. 10684–10695 (2022)
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
28. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 36479–36494. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf

29. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2256–2265. PMLR, Lille, France (07–09 Jul 2015), https://proceedings.mlr.press/v37/sohl-dickstein15.html
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Zhu, Y., Li, Z., Wang, T., He, M., Yao, C.: Conditional text image generation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14235–14245 (June 2023)