SyntheticHTR: Handwritten Text Image Synthesis based on Latent Diffusion Models

Liam Tabibzadeh
Department of Information Technology
Uppsala University
Uppsala, Sweden
liam.tab7@gmail.com

Alex Kangas
Department of Information Technology
Uppsala University
Uppsala, Sweden
alex.kangas.5644@student.uu.se

***Abstract*—The primary challenge in handwritten text recognition (HTR) arises from the vast diversity of human writing styles, leading to the importance of highly varied training data in HTR systems. Additionally, annotating such data often requires specialized knowledge, especially for historical documents, which makes the process expensive. With the emergence of generative machine learning models, particularly the advent of novel latent diffusion models following the promising works with Generative Adversarial Networks (GANs), presents new opportunities for data synthesis in HTR. As a result, this work introduces SyntheticHTR framework that employs a latent diffusion model to synthesize three widely-used HTR benchmark datasets, both attempting to replicate the original images and to extrapolate words with respect to writer styles. This work then introduces ways to evaluate the quality of the synthesized images, using the uncertainty estimation inherent in the prediction of state-of-the-art HTR system. The results of the experiments indicate that the model is capable of producing high-quality synthetic images, although its performance is somewhat less successful for out-of-vocabulary words. The pre-trained models will be released on Github, along with the highest-quality selected synthetic datasets and the source code. We hope that these will be a valuable resource to the HTR community.**

I. Introduction

The digitization of handwritten text documents has increased in the last decades, and with the advent of deep neural networks, the possibility to extract the textual content inside handwritten documents has attracted great attention. This has the potential to aid in preservation of historical documents from deterioration. As a result, research within this has increased substantially. Moreover, with the introduction of transformers, Handwritten text recognition (HTR) models have attempted to integrate them into the neural network architecture, thus further improving the performance of these models. A critical challenge with HTR models, however, is their dependence on extensive training datasets. This issue is particularly acute in the context of historical text documents, where available data is often scarce and restricted to only a few writers. As a result, there has been an increasing interest in exploring generative models for image synthesis. These models hold promise for augmenting existing datasets and thereby enriching the training material available for HTR models, potentially enhancing their accuracy and robustness.

Due to the above, we identified the need to increase the available training data by making use of the latest research in generative machine learning models. As a starting point, we improve upon the latent diffusion model introduced in WordStylist [1] to synthesize three common HTR benchmark datasets, and perform extensive testing to identify the best subsets of those synthetic datasets.

The main contributions of this work are as follows. Firstly, this work introduces SyntheticHTR framework that leverages latent diffusion models for handwritten text synthesis, both attempting to replicate the original images and to extrapolate words with respect to writer styles. Secondly, methods to evaluate the quality of the synthesized images are studied using the uncertainty estimation inherent in the prediction of state-of-the-art HTR systems. Thirdly, we make available three pre-trained models on common HTR benchmark datasets that can be utilized to synthesize images. Lastly, we fully synthesize four datasets that are made available to the community that can be used as training data in HTR models.

II. Related Work

HTR systems are designed to extract textual content from handwritten documents [2], and their development has been significantly improved by the advent of deep neural networks. Popular approaches include Recurrent Neural Networks [3] and attention-based sequence to sequence models [4]–[6]. Moreover, the recent integration of transformers into HTR models, such as TrOCR [7], has further boosted their performance, setting new benchmarks in the HTR field. TrOCR utilizes an image transformer encoder to extract features from input images, which are then interpreted by a text transformer decoder to form a wordpiece sequence.

Despite these advancements, the challenges of data scarcity in this domain have directed the research towards image synthesis, recently enabled by advancements in generative machine learning models. Initially the focus was on Generative Adversarial Networks (GANs), as demonstrated by models like ScrabbleGAN [8] which consists of a generator network G that synthesizes the images, a discriminator D that serves as an evaluator of the synthetic images, and a text recognition network R that ensures that the text inside the images are readable. The approach is therefore to independently generate each letter with overlapping patches to preserve style coherence, controlled by a consistent noise vector.

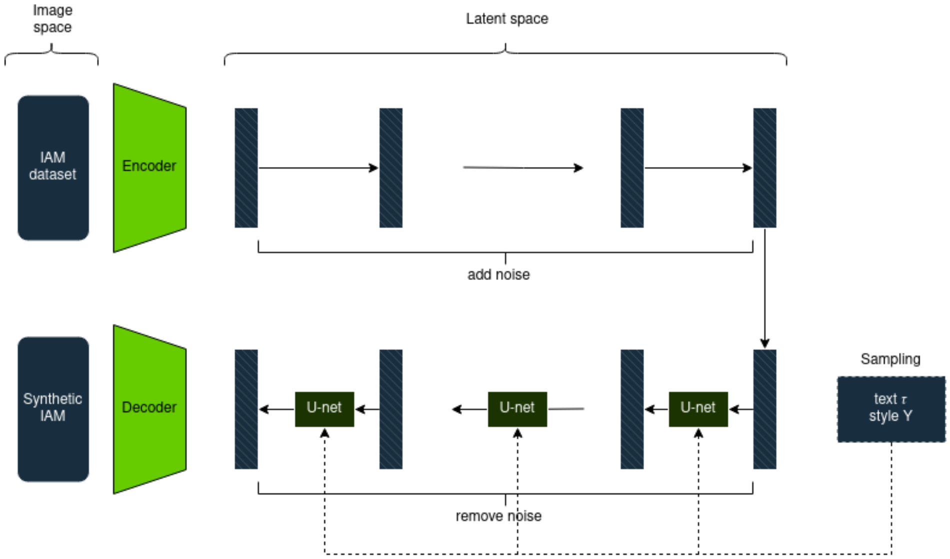

Fig. 1. SyntheticHTR model architecture.

While ScrabbleGAN has shown effectiveness, the field has evolved with the introduction of Latent Diffusion Models (LDMs). LDMs utilize techniques to transform images into a latent space before synthesizing them, which simplifies the generation process while preserving essential features. This work improves upon the Latent Diffusion model WordStylist introduced in [1], that has proven to be effective for image synthesis.

III. Methodology

This section presents the overall architecture of the proposed SyntheticHTR framework for handwritten image synthesis, the benchmark datasets used in the study and the quality assessment techniques.

A. SyntheticHTR Architecture

Figure 1 depicts the architecture of the WordStylist model, which is used in this work as a starting point to synthesize handwriting images. There are two main models inside this architecture; a Variational Autoencoder (VAE) which is responsible for encoding images into a lower dimensional latent space, and then decoding the message back into the image space. The VAE uses a probabilistic Bayesian framework, as it is defined by:

$$p_\theta(x|z) = \mathcal{N}(x; \mu_\theta(z), \sigma_\theta^2 I_D)$$

$$p_\theta(z) = \mathcal{N}(x; 0, I_M)$$

where $p_\theta(x|z)$ is the likelihood and $p_\theta(z)$ is the prior, both assumed to be Gaussian distributed. Due to the intractability of the posterior, VAEs approximate the posterior by maximizing the Evidence Lower Bound (ELBO) which is a lower bound of the log-likelihood of the data. The ELBO is given by the reconstruction error and the Kullback–Leibler (KL) divergence between the approximate posterior and the prior, where the latter acts as a regularizer. Therefore, the VAE effectively learns the data distribution whilst not memorizing the data, leading to a latent space that can synthesize new images beyond the training data.

The second major block in this architecture is the latent diffusion model, which acts on the latent space learnt by the VAE, as opposed to the original image. This reduces the computational burden of the diffusion process, due to the lower dimensional latent space. Starting with the encoded message, the diffusion model gradually adds noise to the message until it consists of complete noise. Then the reverse process is started to remove noise to reconstruct the original message. In this work, The VAE is a pre-trained Autoencoder KL model from the HuggingFace repository [1] which substantially reduces the computational burden in training. Therefore, the training only updates the parameters of the diffusion process, whose reverse process is modelled using a U-net.

With a fully trained diffusion model, sampling is initiated by providing a text string τ and a style condition Y, that instantiate a noise removing process from a noisy image that is then passed to the VAE decoder which outputs a realistic looking image. In our work, we have decided to set the number of timesteps T in the diffusion process to 1000 and employ a linearly increasing noise schedule over the range of 1×10^{-4} to 0.02.

B. Datasets Studied

Three widely recognized HTR benchmark datasets were chosen for our study, with example images from each of these datasets presented in Table I. Below follows descriptions of each dataset.

[1] https://huggingface.co/stabilityai/stable-diffusion-2-1

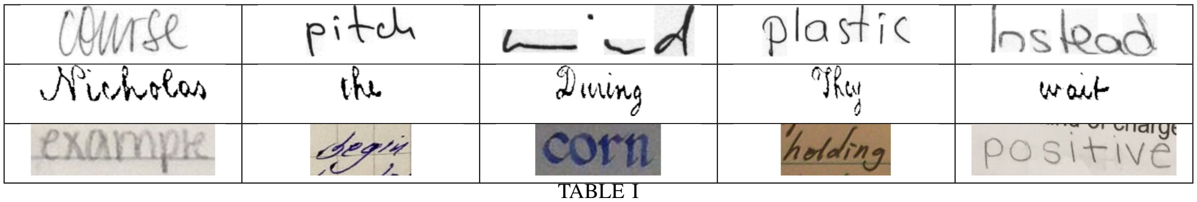

TABLE I

SAMPLE IMAGES FROM THE DATASETS, DEMONSTRATING THE VARIETY OF WRITER STYLES INHERENT IN EACH DATASET. FIRST ROW: IAM DATASET. SECOND ROW: GEORGE WASHINGTON DATASET. THIRD ROW: IMGUR5K DATASET.

IAM dataset: It comprises of $115,375$ handwritten word images from 657 writers, and was utilized in the work [9]. Specifically, the Aachen training set was selected to train the model and generate a corresponding synthetic dataset. A key preprocessing step involved filtering the data to include only words ranging from 2 to 10 characters in length, resulting in a refined dataset of $44,412$ word-level images from 339 authors.

IMGUR5K Dataset: This dataset comprises $8,177$ page images sourced from imgur.com, further segmented into $230,573$ word-level images contributed by over $5,000$ authors [10]. For our research, a specific subset of IMGUR5K was selected based on certain criteria: page images containing more than 150 words were excluded, as were words that fell outside the 2–10 character range or did not use the upper and lower case of the Latin alphabet. This resulted in a refined subset comprising 9,166 word-level images from 51 page images. Due to the absence of authorship metadata, we assume that each page image was written by a different individual, estimating the writer styles for this subset to be around 51.

George Washington (GW) dataset: Originating from the George Washington papers housed at the Library of Congress, this dataset initially included $4,894$ word-level images authored by two writers [11]. By applying similar criteria for word length and character set, the dataset was narrowed to $4,506$ word-level images.

C. Synthetic Datasets Quality Assessment

Fréchet-Inception Distance (FID): The FID is used to assess image quality in generative models, focusing on how similar two sets of images are. It compares feature distributions from images using the Inception v3 network, rather than just looking at each pixel. Mathematically, it measures the difference between two sets of images by calculating the distance between their feature distributions, modeled as multivariate Gaussian distributions:

$$\begin{aligned} FID = ||\mu_1 - \mu_2||_2^2 + tr(\Sigma_1 + \Sigma_2 \\ -2 \cdot \sqrt{(}\Sigma_1 \Sigma_2)) \end{aligned} \tag{1}$$

Here, (μ_1, Σ_1) and (μ_2, Σ_2) are the parameters of the real and generated datasets. The FID's role in this work is to provide an objective measure of the similarity between the feature distributions of our generated images and those in the original dataset.

Normalized Mean Absolute Error (NMAE): The Mean Absolute Error (MAE) is a metric for comparing two images, calculated as the average of the absolute differences between their corresponding pixels. For images A and B it is defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |A(i) - B(i)| \tag{2}$$

where i iterates over all pixels in these images. MAE thus provides a direct measure of how different two images are at the pixel level. In this research, MAE is normalized between 0 and 1 for better interpretability of the pixel-wise difference between the synthetic and the original image. As a result, NMAE values close to 0 indicate low differences, and values close to 1 indicate high differences.

AttentionHTR Confidence Score (CS): The AttentionHTR model [4] is a HTR model that uses transfer learning from the scene text recognition domain, using a pre-trained STR benchmark model [12], and further fine-tuning on IAM and IMGUR5K datasets. The model's last layer includes a softmax function that outputs a probability distribution over the character set for each prediction. The main advantage of using AttentionHTR model is that the final model is trained on handwriting from thousands of authors, with varying image conditions, in order to aid generalization in the real-world.

The value of the maximum of this probability distribution is referred to as the AttentionHTR Confidence Score (CS), and it intuitively reflects how certain the model is of its prediction. Although it may not fully align with human judgement regarding the quality of an image, it nonetheless offers an insight into the complexity of the recognition task at hand and thereby attempts to measure the practical value that the synthetic images bring to the model.

IV. EXPERIMENTAL RESULTS

This section presents the overall procedure of the experiments conducted in this work and presents the results obtained.

A. Experimental Design

Our experimental evaluation involved a structured pipeline applied to three handwritten text datasets: IAM, IMGUR5K, and the GW dataset. For each dataset, a dedicated model was trained to capture the distinct handwriting styles and word

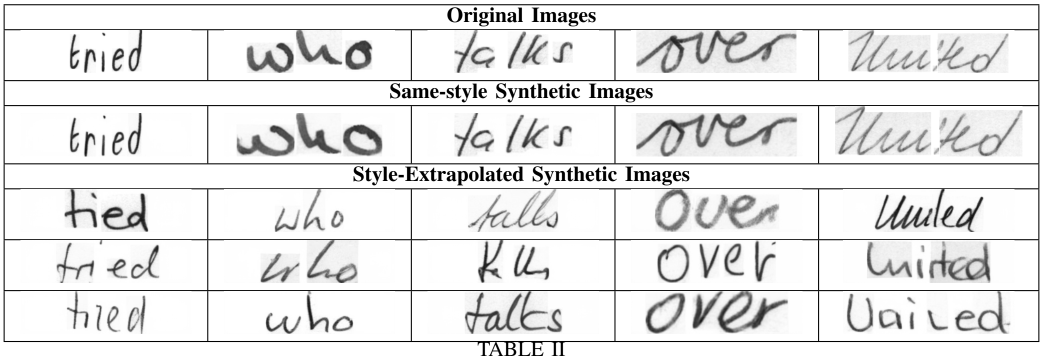
Original Images
Same-style Synthetic Images
Style-Extrapolated Synthetic Images

TABLE II
COMPARISON OF ORIGINAL AND SYNTHETIC HANDWRITING SAMPLES FROM THE IAM DATASET

formations. The training consists of $1,000$ epochs, with a batch size of 224 on an Nvidia A40 GPU. Optimization during training used the *AdamW* optimizer with a learning rate of 10^{-4}.

The trained models were then used to generate synthetic images of the same styles and words of the original datasets. Additionally, to evaluate the model's style extrapolation capabilities, we created an **out-of-vocabulary (OOV) dataset** from the IAM-trained model. This involved generating 50 words for each writer style that were not in the writer's original set but existed in other writers' vocabularies.

The quality of the synthetic datasets were evaluated using the confidence score and NMAE metrics, which subsequently acted as a selection method for refining the datasets by keeping the highest 50% or 90% performing images with respect to each of the metrics. This procedure leads to smaller subsets of the original synthetic datasets that consist of the highest quality images, as only they are selected to remain in this subset.

In order to calculate the AttentionHTR confidence scores for the synthetic IAM and IMGUR5K datasets, we used the publicly available AttentionHTR models pre-trained on these respective datasets. In the absence of a corresponding pre-trained model for the George Washington dataset, we fine-tuned a new AttentionHTR model on this dataset and used it to calculate the confidence scores on each image within the synthesized dataset.

Finally, the FID score was calculated to compare the synthetic datasets with their corresponding real datasets across all subsets, thus making the comparison fair. Since there does not exist the same word and style combination for the synthetic out-of-vocabulary dataset, we compared this dataset with the complete IAM dataset.

B. Results

Table II highlights a sequence of images for comparison, where the top row contains the original handwriting from the IAM dataset, followed by synthetic reproductions in the corresponding writer's style, and below are synthetic versions in different styles that did not originally include the given word. The synthetic images maintain stylistic fidelity to the originals, and the extrapolated styles demonstrate the model's ability to generate diverse handwriting variations, despite occasional character misrepresentations.

Figure 2 presents the distribution of NMAE and AttentionHTR Confidence Scores (CS) across synthetic datasets. The histograms for each dataset show the frequency of scores with summary statistics provided in each plot's upper left. The Synthetic GW dataset shows a high mean CS and a low NMAE, suggesting that all images are of high quality. Meanwhile, Synthetic IAM and IMGUR5K datasets show a wider range of scores but indicate satisfactory metrics.

Table III presents synthetic images from the IAM dataset, categorized based on their quality as assessed by NMAE and CS metrics. The arrows signify the relative quality of each image: a downward arrow (↓) indicates lower quality, a bidirectional arrow (↔) denotes average quality, and an upward arrow (↑) represents higher quality compared to other images in the dataset according to the respective metric. The CS metric is segmented into three categories based on specific thresholds: for $CS < 0.8$ the label is $\downarrow CS$; for $0.8 < CS < 0.99$ it is $\leftrightarrow CS$; and for $CS > 0.99$ the label is $\uparrow CS$. Similarly, for NMAE, the categories are defined as follows: $NMAE > 0.02$ is labelled as $\downarrow NMAE$, $0.02 < NMAE < 0.15$ as $\leftrightarrow NMAE$, and $NMAE > 0.15$ as $\leftrightarrow NMAE$. These thresholds were selected to achieve a balanced distribution of images across each category.

It can be seen that the CS metric is associated with the clarity and legibility of individual characters in the images. This is evident in cases where cursive handwriting styles lead to less distinct characters. In contrast, the NMAE seem to be solely related with the quality of image reconstruction, which aligns with the definition of this metric. Consequently, it can therefore be concluded that the NMAE metric assesses the ability of our model to interpolate the training data by reconstructing the images, whilst the CS metric quantifies the model's ability to generate legible text and whether the produced images contain clear and readable words, regardless of whether the images are similar to the training data. In this view, these two metrics measure distinct dimensions of the model.

Table IV illustrates the performance metrics for the four synthetic datasets: IAM, Out-of-Vocabulary IAM (IAM-OOV),

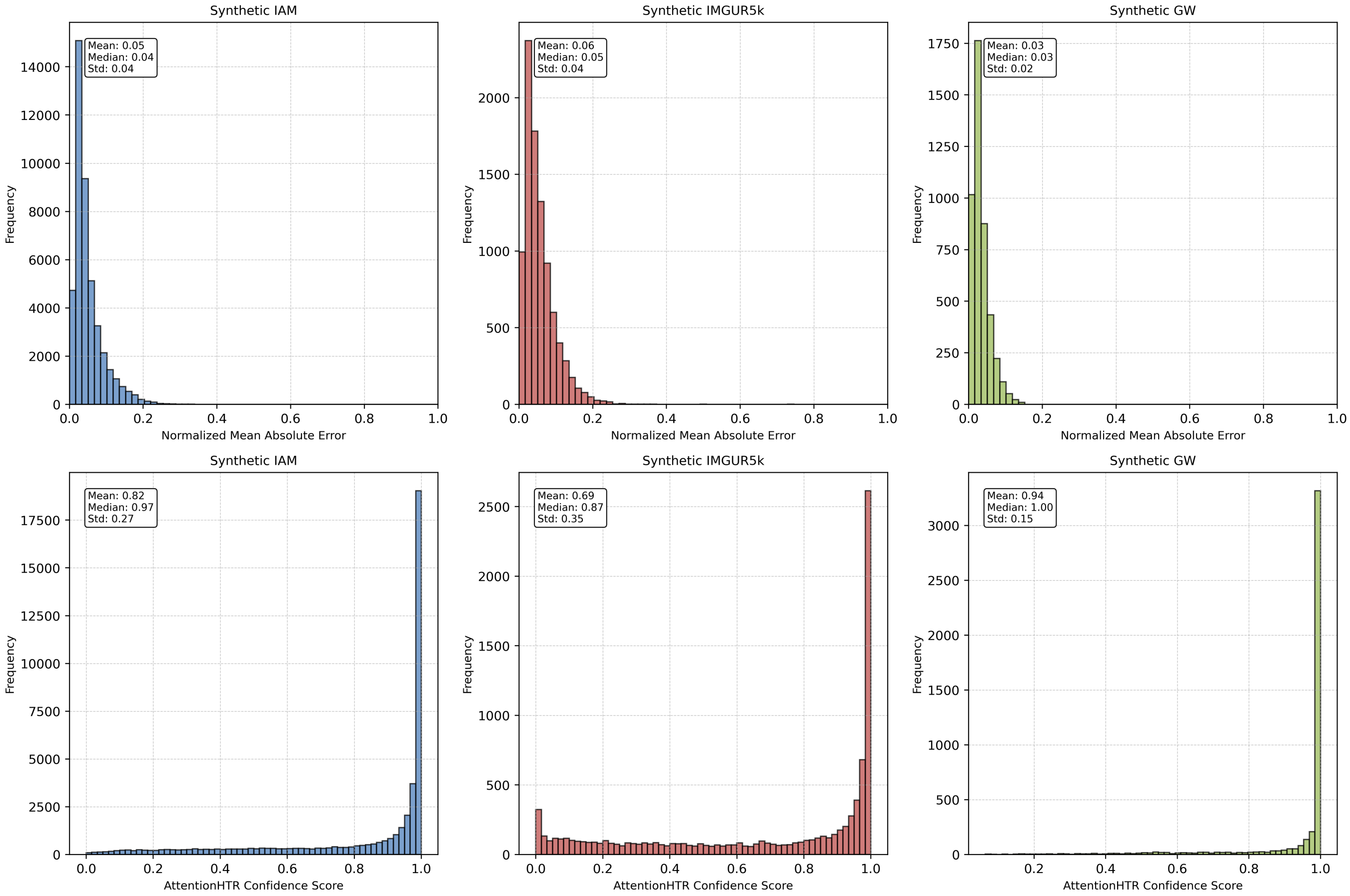

Fig. 2. Histograms of Normalized Mean Absolute Error (NMAE) and AttentionHTR confidence scores for synthetic datasets

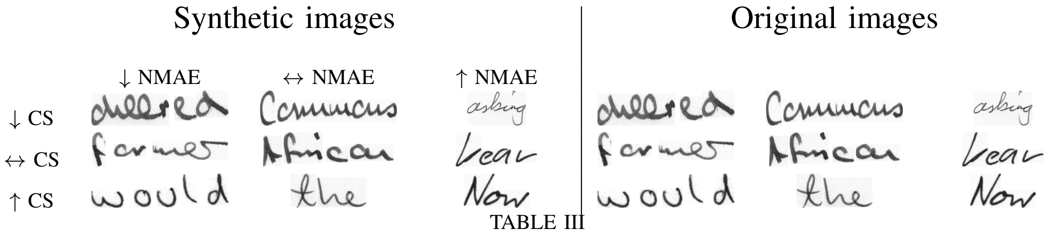
Synthetic images
Original images
↓ NMAE
↔ NMAE
↑ NMAE
↓ CS
↔ CS
↑ CS

TABLE III

EVALUATION OF SYNTHETIC AND ORIGINAL IAM IMAGES BASED ON NMAE AND CONFIDENCE SCORE (CS) METRICS. ARROWS INDICATE IMAGE QUALITY: DOWNWARD (↓) FOR LOWER, HORIZONTAL (↔) FOR AVERAGE, AND UPWARD (↑) FOR HIGHER QUALITY. THE TABLE'S RIGHT SIDE PRESENTS ORIGINAL IMAGES FOR COMPARISON.

George Washington dataset (GW), and IMGUR5K, including their respective subsets. The datasets and subsets are evaluated on key metrics including FID, Mean Confidence Score (Mean CS), and NMAE. We identify the complete synthetic datasets as 100%, and their subsets are labelled with the corresponding percentages of the full dataset, 50% or 90%. The subset selection criteria are marked as "CS" if it is based on the Confidence Score and "NMAE" if it is based on the Normalized Mean Absolute Error.

In comparing the four categories of datasets, we find that subsets from the IAM dataset generally show the lowest FID, indicating a significant performance improvement over the full IAM dataset. This difference is not as noticeable in other datasets. It is worth highlighting the performance of the synthetic GW datasets, which score well on the AttentionHTR Mean CS and NMAE. Also, among the complete synthetic datasets, GW stands out as the highest performer, followed by IAM. For FID, subsets from all datasets, including IMGUR5K, GW, and Out-of-Vocabulary IAM (IAM-OOV), show performance levels comparable to their respective full datasets, indicating that all synthetic images are of high quality. Additionally, the most effective subset within each dataset category aligns with the metric used for selection, pointing to a direct correlation between the method of selection

Dataset	Number of Images	FID	Mean CS	NMAE
100%-Synthetic IAM	44,412	12.78	0.82	0.05
CS **90%**-Selected Synthetic IAM	39,971	5.17	0.89	0.05
NMAE **90%**-Selected Synthetic IAM	39,971	4.73	0.81	0.04
CS **50%**-Selected Synthetic IAM	22,206	5.97	0.99	0.06
NMAE **50%**-Selected Synthetic IAM	22,206	**4.52**	0.76	0.02
100%-Synthetic IMGUR5K	9,166	32.61	0.69	0.06
CS **90%**-Selected Synthetic IMGUR5K	8,250	**31.92**	0.76	0.06
NMAE **90%**-Selected Synthetic IMGUR5K	8,250	32.86	0.69	0.05
CS **50%**-Selected Synthetic IMGUR5K	4,583	32.43	0.97	0.06
NMAE **50%**-Selected Synthetic IMGUR5K	4,583	32.36	0.70	0.03
100%-Synthetic GW	4,506	**9.05**	0.94	0.03
CS **90%**-Selected Synthetic GW	4,056	9.41	0.99	0.03
NMAE **90%**-Selected Synthetic GW	4,056	9.41	0.95	0.03
CS **50%**-Selected Synthetic GW	2,253	11.68	1.00	0.03
NMAE **50%**-Selected Synthetic GW	2,253	10.83	0.95	0.02
100%-Synthetic IAM-OOV	16,581	14.65	0.45	*NA*
CS **90%**-Selected Synthetic IAM-OOV	14,923	14.06	0.49	*NA*
NMAE **90%**-Selected Synthetic IAM-OOV	14,923	*NA*	*NA*	*NA*
CS **50%**-Selected Synthetic IAM-OOV	8,291	**13.03**	0.71	*NA*
NMAE **50%**-Selected Synthetic IAM-OOV	8,291	*NA*	*NA*	*NA*

TABLE IV

Comparative analysis of synthetic handwriting datasets: IAM, IMGUR5K, George Washington (GW), and Out-of-Vocabulary IAM (IAM-OOV), along with their respective subsets. The datasets and subsets are evaluated on key metrics: FID, Mean Confidence Score (Mean CS), and NMAE. Selection criteria for subsets are based on either CS or NMAE scores, with the best FID performances highlighted in bold. 'NA' indicates data not available.

and the overall performance of each dataset.

Furthermore, as can be observed in Table IV, the 100% synthetic IAM dataset achieved an FID score of **12.78**, which is significantly lower than the FID score of **22.74** for the synthetic IAM dataset achieved by the original WordStylist paper [1]. For clarity, the lower the FID, the better the synthetic dataset. Our improvement upon the results from the original WordStylist model may be attributed to the newer pre-trained Stable Diffusion model we used from the HuggingFace.

V. Discussion

The challenge of HTR primarily stems from a lack of sufficient data. Therefore, this work has aimed to tackle this issue by employing latent diffusion models to generate synthetic datasets. Specifically, we utilized a latent diffusion model architecture to generate synthetic versions of well-known handwritten text recognition (HTR) benchmark datasets, including the IAM dataset, a portion of the IMGUR5K dataset, and the GW dataset. To make sure that these synthetic datasets are of high quality, we experimented with refining the datasets, by selecting images based on their performance in terms of NMAE and the AttentionHTR Confidence Score (or CS metric for short). This process ensured that only the synthetic images with the best scores in these metrics were retained for each dataset. For all datasets and subsets thereof, we evaluated the quality using FID, NMAE and CS metric.

Based on Table III, we reasoned that the CS and NMAE metrics quantify different dimensions of the generative ability of our model. We concluded that the CS metric evaluates the legibility of individual characters in images, with its effectiveness declining for images with cursive handwriting styles which tend to obscure the clarity of individual characters. Conversely, the NMAE metric focuses exclusively on the quality of image reconstruction, aligning with its intended purpose. These metrics, therefore, serve distinct evaluative purposes; while NMAE measures the model's capability to interpolate and reconstruct training images, CS assesses the model's ability to generate legible text, irrespective of their resemblance to training images.

Considering the problem which our model is trying to solve is that of data scarcity in the field of HTR, our model's capacity to generate images beyond the existing data distribution emerges as a critical dimension. The CS metric's focus on assessing text legibility in generated images irrespective of the training images makes it particularly relevant for HTR applications, where clarity and readability are paramount. Consequently, the results in Table IV should be interpreted with an emphasis on the CS metric, since in a broader scope, which is that of our model's ability to generate data to solve the HTR data scarcity problem, then we believe that this goal is best quantified through this metric.

The performance of our synthetic datasets appears to be linked to the complexity of each original dataset. Here, 'complexity' refers to the variety of writer styles present. For example, the GW dataset, which includes handwriting from only two writers, generally shows better performance on our evaluation metrics compared to datasets with a wider range of writer styles. This trend suggests that datasets with fewer writer styles, like the GW dataset, are easier for our latent diffusion models to learn the data distribution effectively. Therefore, the relative simplicity of the GW dataset in terms of writer style diversity seems to contribute to its higher scores in NMAE and CS. This observation indicates a correlation between the diversity of handwriting styles in a dataset and the resulting quality of the synthetic dataset generated.

VI. Conclusion

This work presented SyntheticHTR framework based on latent diffusion models for generating realistic yet synthetic handwritten text images to help address the training data scarcity issues with the current HTR methods. The experimental results demonstrate the effectiveness of the proposed method on the benchmark datasets, validated using the metrics: FID, NMAE and AttentionHTR Confidence Score. In summary, we have successfully synthesized four word-images datasets and introduced novel ways to assess the quality of synthetic images, thereby contributing to the field with training data and an empirical analysis on evaluation methods of the quality of synthetic data. The results, pre-trained models and the best synthesized datasets across each dataset category will be made available to the research community on Github.

Limitations

While our model has demonstrated its capability to synthesize datasets effectively, there are still areas that require further exploration. For example, the variation in style extrapolation quality, as shown in Table II, suggests the need for additional experimentation in this aspect. Moreover, our experiments are restricted to grayscale images, leaving the model's effectiveness on colored images untested. Additionally, any variety or lack thereof in the original datasets is inherent in our synthetic datasets, possibly misrepresenting some writer styles. Lastly, the impact of using these synthesized datasets on the performance of HTR models has not been evaluated in this work. Therefore, future research should focus on integrating these synthesized datasets into an HTR model to determine their practical value in real-world HTR applications.

Ethics Statement

Our research aims to enable knowledge extraction from historical collections. The research does not pose any risk or societal harm. To enable future re-use, transparency, and dissemination of knowledge to the public, we will make our source code, synthetic datasets and the pre-trained models available to the research community. Model pre-training is also an effort towards minimizing the environmental impact of the machine learning model training, as opposed to training from scratch for future research.

Acknowledgment

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

[1] K. Nikolaidou, G. Retsinas, V. Christlein, M. Seuret, G. Sfikas, E. B. Smith, H. Mokayed, and M. Liwicki, "Wordstylist: Styled verbatim handwritten text generation with latent diffusion models," 2023.

[2] J. Nockels, P. Gooding, S. Ames, and M. Terras, "Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of transkribus in published research," *Archival Science*, vol. 22, no. 3, pp. 367–392, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s10502-022-09397-0

[3] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 2, pp. 211–224, 2011. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5871643?casa_token=6mKdi5tiFioAAAAA:g-phUz9H_GfylmWxbmh-MCX8jkgQwIoAUbsGmXZtRblLjmLjevvYGaKLld22gTbs_pSviNcANw

[4] D. Kass and E. Vats, "Attentionhtr: Handwritten text recognition based on attention encoder-decoder networks," in *Document Analysis Systems*, ser. Lecture Notes in Computer Science, S. Uchida, E. Barney, and V. Eglin, Eds., vol. 13237. Springer, Cham, 2022.

[5] T. Bluche, J. Louradour, and R. Messina, "Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention," in *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 1050–1055. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8270105?casa_token=3H2s_ETxgtMAAAAA:E_4Vw8ANLo5nKwt-qrJsG9a3cUeCB4ZRqznuqu61Yy_QHCT7BrToV_KEA8c4vuVKwrLIsdbzjw

[6] L. Kang, J. I. Toledo, P. Riba, M. Villegas, A. Fornés, and M. Rusinol, "Convolve, attend and spell: An attention-based sequence-to-sequence model for handwritten word recognition," in *Pattern Recognition: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings 40*. Springer, 2019, pp. 459–472. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-12939-2_32

[7] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "Trocr: Transformer-based optical character recognition with pre-trained models," 2022.

[8] S. Fogel, H. Averbuch-Elor, S. Cohen, S. Mazor, and R. Litman, "Scrabblegan: Semi-supervised varying length handwritten text generation," *arXiv preprint arXiv:2003.10557*, Mar 2020. [Online]. Available: https://arxiv.org/abs/2003.10557

[9] U.-V. Marti and H. Bunke, "The iam-database: an english sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, vol. 5, pp. 39–46, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:29622813

[10] P. Krishnan, R. Kovvuri, G. Pang, B. Vassilev, and T. Hassner, "TextStyleBrush: Transfer of Text Aesthetics from a Single Example," 2021.

[11] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character hmms," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 934–942, 2012, special Issue on Awards from ICPR 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865511002820

[12] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in *International Conference on Computer Vision (ICCV)*, 2019. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/papers/Baek_What_Is_Wrong_With_Scene_Text_Recognition_Model_Comparisons_Dataset_ICCV_2019_paper.pdf