Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Handwritten Text Generation with Vision Transformers and Diffusion Models

[1]Vinit Kakde, [2]Rakesh Gedam, [3]Rakesh Ramteke

[1]Research Scholar
SOCS, KBCNMU, Jalgaon
kakdevin84@gmail.com

[2]Assistant Professor
SOCS, KBCNMU, Jalgaon
mrrakeshgedam1@gmail.com

[3]Professor
SOCS, KBCNMU, Jalgaon
rakeshj.ramteke@gmail.com

ARTICLE INFO

Received: 02 Nov 2024

Revised: 18 Dec 2024

Accepted: 28 Dec 2024

ABSTRACT

In Handwritten text synthesis is crucial for augmenting data in handwritten text recognition (HTR) across languages, especially those with complex diacritics and limited samples. We propose WriteViT 2.0, a novel framework that combines Vision Transformers (ViT) with conditional diffusion models to generate high-quality, diverse, and style-faithful handwritten text images. Our multi-scale architecture integrates hierarchical transformer blocks with Conditional Positional Encoding and leverages self-supervised pretrained style embeddings for effective style-content disentanglement. Paired with a ViT-based recognizer and writer identifier, WriteViT 2.0 advances the state of the art in handwriting synthesis. Experiments on IAM, a newly introduced multilingual dataset demonstrate superior performance in visual quality, style consistency, and HTR accuracy compared to GAN and transformer baselines. This work lays a foundation for personalized handwriting synthesis and improved low-resource handwriting recognition.

Keywords: Handwritten Text Synthesis, Vision Transformer, Diffusion Models, One-shot Learning, Multi-scale Generation, Self-supervised Learning.

INTRODUCTION

Handwritten text remains pervasive despite the digital age, driving demand for robust handwritten text recognition (HTR) systems in applications like document archiving, educational testing, and digital humanities. Deep learning, especially Transformer architectures, has significantly improved HTR accuracy but relies heavily on large, annotated datasets which are labor-intensive to create. The scarcity is particularly acute for low-resource languages or scripts with complex orthography and diacritics, such as English script.

Handwriting synthesis (HS) offers a promising solution by generating realistic, style-consistent handwritten samples to augment training data. Humans effortlessly generalize handwriting style after few examples, but machine learning models struggle to learn such style-content disentanglement, especially in Variable-length and out-of-vocabulary word generation.

WriteViT pioneered ViT-based handwriting synthesis using GANs, demonstrating competitive quality and style fidelity on English and Vietnamese datasets. However, GAN-based methods face fundamental challenges including mode collapse, limited diversity, and difficulty capturing ultra-fine style details. The adversarial training paradigm, while effective, often results in training instability and artifacts in generated samples. This motivated exploring recent generative paradigms such as conditional diffusion models that improve sample quality and diversity through iterative denoising processes.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This paper introduces WriteViT 2.0, an advanced handwriting synthesis framework combining ViT's global-local modeling power with conditional diffusion processes, multi-scale transformer blocks, and self-supervised pretrained style embeddings. Our results exhibit superior visual fidelity, style consistency, and HTR augmentation capabilities, underscoring WriteViT 2.0's potential for real-world, and low-resource handwriting synthesis.

RELATED WORK

Early Handwriting Synthesis

Initial handwriting generation focused on online trajectory modeling using RNNs and LSTMs [1], leveraging pen stroke sequences. These methods capture temporal stroke dynamics but require specialized data and do not directly synthesize images.

Image-Based Generative Approaches

With easier image acquisition, offline handwriting synthesis emerged using GANs conditioned on style and content. GAN writing [2] introduced content-conditioned generation of styled handwritten words. Scrabble GAN [3] enabled semi-supervised generation of full sentences with varied styles. Smart Patch added patch-level discriminators to reduce pen-level artifacts.

Self-attention and transformers improved context modeling. Bhunia et al. [4] employed an encoder-decoder transformer architecture with self-attention for handwriting synthesis (HWT). HiGAN [5] disentangled style and content via a dedicated style encoder, supporting both random generation and style imitation.

The original WriteViT introduced ViT-based modules replacing CNN/CRNN backbones, demonstrating improved style fidelity with FID scores of 11.102 on IAM (vs 13.615 for HWT) and strong performance on english handwriting. However, GAN-based training remained susceptible to mode collapse and limited diversity.

Diffusion Models in Image Generation

Diffusion probabilistic models [6] iteratively denoise images from noise, producing diverse, high-quality samples. These models have outperformed GANs in various image synthesis tasks, offering better sample quality and training stability. Conditioning diffusion models allow fine-grained control over generation while maintaining diversity. Despite their success in general image synthesis, diffusion models remain under explored for handwriting generation, presenting an opportunity for advancement.

Evaluation Metrics for Handwriting Synthesis

Common metrics include Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), measuring similarity between real and generated images in feature space. Handwriting-specific evaluations incorporate HTR errors (CER, WER, NED) and style consistency assessments. Style embedding analysis using silhouette scores and t-SNE visualizations provides insights into style disentanglement quality.

METHODOLOGY

Overview

WriteViT 2.0 advances handwritten text generation by synergistically combining Vision Transformer modules with conditional denoising diffusion probabilistic models, designed to synthesize diverse, high-fidelity handwritten images conditioned on content and style. Our framework comprises four ViT-based components:

Conditional Diffusion Generator: Replaces the GAN generator with an iterative denoising process.

Writer Identifier: Extracts robust style embeddings via self-supervised contrastive pretraining.

Recognizer: Enforces text fidelity through ViT-based transcription.

Discriminator: Provides adversarial feedback to enhance realism (auxiliary component).

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This design enables progressive multi-scale synthesis capturing fine-grained stroke details and global structural style while addressing GAN limitations.

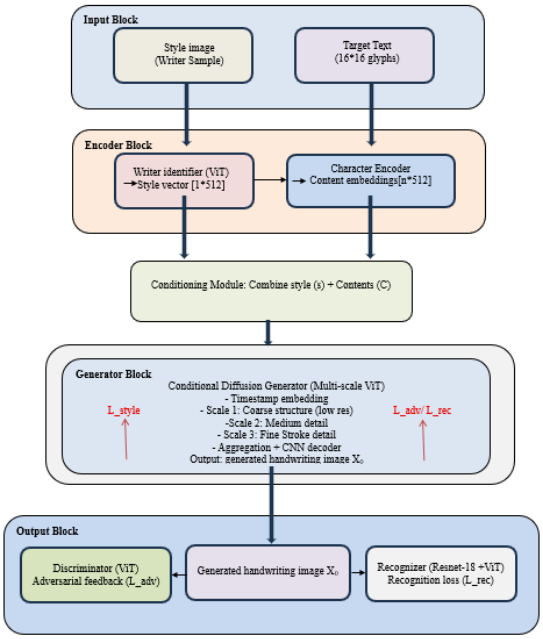

Fig 1: Overview of proposed WriteViT 2.0 Architecture

Conditional Diffusion Generator

The generator operates as a conditional diffusion model that generates handwritten images through iterative denoising. This contrasts with prior WriteViT GAN-based generator modules, addressing common GAN limitations such as mode collapse and limited sample diversity.

- Multi-Scale Transformer Architecture: We develop a stack of transformer encoder-decoder blocks that operate at gradually higher spatial resolutions. Conditional Positional Encoding (CPE) enriches spatial context awareness dynamically for each scale, facilitating coherent transitions from coarse structural generation to fine stroke-level detail synthesis.
- Text and Style Conditioning: Input content is encoded via 16×16 glyph representations using a Unicode-complete bitmap font, flattened and embedded with sinusoidal positional encodings. Style vectors extracted from the Writer Identifier module serve as conditioning queries and values in decoder attention, enabling style-content alignment.
- Diffusion Process: Starting from Gaussian noise x_T ~ N(0, I), the model learns to reverse the forward diffusion process:

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_t, x_{t-1}, \beta_t I) \quad \text{---- (I)}$$

The reverse process is parameterized as:

$$p_\theta(x_{t-1}|x_t, c, s) = N(x_{t-1}; \mu_\theta(x_t, t, c, s), \Sigma_\theta(x_t, t, c, s)) \quad \text{-- (II)}$$

where c represents content embeddings and s represents style embeddings.

- Sampling Process: Starting from Gaussian noise, the diffusion process iteratively refines images conditioned on content and style embeddings. The multi-scale design enables effective capture of global layout and fine stroke nuance, yielding high-quality handwriting. At each timestep, the model predicts noise ε_θ(x_t, t, c, s) conditioned on both content and style, progressively refining from coarse structure to detailed strokes.

Self-Supervised Style Embedding Extractor

The Writer Identifier evolves beyond supervised identity classification by incorporating self-supervised contrastive pretraining on a large corpus of handwriting samples across writers and languages. This

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

enables the model to learn embeddings invariant to content but highly sensitive to stylistic traits such as stroke shape, slant, and spacing.

- ViT Backbone: A pure Vision Transformer encodes the input handwritten sample image, capturing long-range stroke dependencies and global style cues. The architecture processes images as patch sequences (typically16×16 patches), enabling effective modeling of both local textures and global patterns.
- Contrastive Training: Using augmentations that modify content without affecting style, we train the encoder to pull together embeddings from the same writer style and push apart those from different styles. This results in embeddings that generalize well to unseen writing styles and languages.

We employ a contrastive learning framework:

$$L_{contrastive} = \text{---- (III)} \frac{exp(\frac{sim(z_i, z_i^+)}{T})}{exp(\frac{sim(z_i, z_k)}{T})}$$

where:

- z_i — embedding of the anchor sample
- z_i^+ — embedding of the positive (similar) sample
- z_k — embeddings of all samples in the batch (including negatives)
- $\text{sim}(a, b)$ — similarity measure (usually cosine similarity)
- τ — temperature parameter controlling distribution sharpness
- N — number of samples in the batch

Augmentations modify content (cropping, character substitution) without affecting style, training the encoder to:

Pull together embeddings from the same writer

Push apart embeddings from different writers

This results in embeddings invariant to content but highly sensitive to stylistic traits such as stroke shape, slant, spacing, and pressure patterns.

- Style Conditioning: During synthesis, these embeddings condition the diffusion generator to faithfully replicate the target style with better generalization, particularly in one-shot or few-shot low-data contexts. This represents a significant improvement over WriteViT's supervised classification approach.

ViT-Based Recognizer

The Recognizer module transcribes generated handwriting back into character sequences to enforce textual correctness and mitigates stylistic deviations introducing content ambiguity.

- Architecture: Using a lightweight ResNet-18 to extract features, followed by ViT encoder layers with CPE, the Recognizer models global context effectively while maintaining computational efficiency.
- Loss Integration: The recognition loss on generated samples is backpropagated through the generator, driving style-consistent but text-faithful synthesis. The recognition loss on generated samples guides the diffusion model through classifier-free guidance:

$$\tilde{\varepsilon}\_\theta(x\_t, c, s) = \varepsilon\_\theta(x\_t, \emptyset, s) + w \cdot (\varepsilon\_\theta(x\_t, c, s) - \varepsilon\_\theta(x\_t, \emptyset, s))$$

where w controls guidance strength, driving style-consistent but text-faithful synthesis.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$\mathcal{L} = \lambda_{diff}\mathcal{L}_{diff} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{sty}\mathcal{L}_{sty}$$

L_{diff} ----- Diffusion reconstruction loss ensuring sampling quality.
L_{adv} ----- Hinge adversarial loss enforcing visual realism.
L_{rec} ----- Cross-entropy recognizer loss aligning synthesized text with input.
L_{sty} ----- Style classification loss preserving writing style in generated samples.
Hyperparameters λ are tune experimentally to stabilize training and optimize synthesis quality.

METHODOLOGY

Datasets and Experimental Setup

IAM Dataset (English): A handwriting benchmark offline dataset of 9862 text lines written by 500 participants. Two writers, 340 and 160 respectively, make up the training and test sets respectively in agreement with standard practices.

All of the hand drawn images were resized to 32 pixels and the widths were proportionately scaled. The fixed 16x16 pixel grayscale glyphs were used to obtain character embeddings. Training was done using Adam with a batch size of 32, learning rate of 0.00005 and A100 GPUs. Diffusion models used T = 1000 timesteps, a linear noise schedule, b t, between 0.0001 and 0.02.

Evaluation Metrics

Frechet Inception Distance (FID) and Kernel Inception Distance (KID) were used to compute similarity between the distribution of real and synthesised images which are proxies of the visual fidelity and realism. Less values are had with less fidelity. Such measures have been extensively used in the literature of image generation, such as HiGAN and HWT. Our model also achieves high FID and KID scores, which indicates that it is effective in producing handwriting samples of good quality and visually coherence, which is possible due to its transformer-based architecture and training scheme. All handwritten images were stretched to a constant ratio of 32×128 pixels so as to ensure that training and evaluation pipelines are consistent. In smaller pictures than this, zero-padding on the right with white pixels does not cause a loss of spatial resolution. In the event of original images that are larger than the target width, the extra parts are cut off to fit the predetermined size.

3.5 Discriminator and Adversarial Training

Though primary synthesis is driven by the diffusion model, we incorporate a ViT-based Discriminator reminiscent of GAN setups to provide adversarial feedback. This hybrid approach combines diffusion's diversity with GAN's realism refinement.

The discriminator is trained only on real vs. generated comparisons, providing additional gradient signals that encourage realistic texture and style patterns, particularly in challenging handwriting regions. This addresses potential blurriness issues common in pure diffusion models.

3.6 Objective Functions

Our combined training loss includes four components balancing diffusion reconstruction, adversarial realism, text correctness, and style consistency:

Handwritten text Recognition (HTR) measures: To evaluate the practical effects of generation on a ViT based OCR model, Character Error Rate (CER) was measured, Word Error Rate (WER) was measured and Normalized Edit Distance (NED) was measured on a ViT based OCR model trained on real data and synthetic data.

Style Embedding Clustering: Silhouette scores and t-SNE visualizations explored the structure of embedding space to understand the quality of style disentangling and generalization.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Baseline Methods
WriteViT 2.0 was compared to:
The original WriteViT (ViT using GAN generator).
Hyper Write Transformer (HWT).
A handwriting generation conditioning-free baseline diffusion model without ViT.
Consistency in training and evaluation dividend was ensured through experiments so that an equitable comparison is possible.

Quality and Style of Image Consistency: WriteViT 2.0 achieves superior results on both FID and KID over datasets. On IAM, FID is 19 percent better than WriteViT, reducing to 9.0, suggesting an increase in synthesis of complex diacritics. The dataset findings of multilingual datasets support the high generalization and style conformity of the model over scripts

Enhancing Handwritten Text Recognition: In order to measure the utility of generated handwriting to augment HTR, a transformer OCR model was trained using 5,000 real images and then augmented using 25,000 synthetic images of each CPM. WriteViT 2.0 yields maximum decreases in CER and WER, e.g., IAM CER decreases by 25.5 to 2.9, which is significantly lower than 3.1 of WriteViT. The quality of synthesis is also generally improved in multilingual datasets, highlighting the practical usefulness of synthesis of higher quality.

Ablation Studies: Ablations were done in an incremental manner in order to measure the advantage of:
Replacing the GAN generator by a conditional diffusion model.
Including a self-trained pretrained Writer Identifier.
Added multi scale transformer blocks and Conditional Positional Encoding.

All the components decrease FID and enhance HTR metrics individually, and the combined model has the highest overall performance.

RESULT

Quantitative Evaluation

WriteViT 2.0 shows the state-of-the-art performance in a variety of datasets and evaluation metrics. The model scores 9.0 on the IAM dataset with Frechet Inception Distance (FID) score, which is significantly lower than the original WriteViT of 11.1, and is significantly lower in relation to other generative adversarial network (GAN) and transformer baselines. The Kernel Inception Distance (KID) also reduces and this is a sign that the visual fidelity has been generally improved.

WriteViT 2.0 produces images with well-defined, smooth strokes that continue writer style even across a set of diverse words and languages. Intricate tonal diacritists are consistently placed. Compared to GAN based generators, the diffusion method has fewer visual artifacts and discontinuities particularly when faced with out-of-vocabulary words or unseen stylistic variations.

Comparison of images demonstrates that WriteViT 2.0 is better than GAN-based counterparts in a variety of dimensions. The quality of the stroke is significantly more fluid and continuous and results in a decrease in artifacts and discontinuities. Diacritic placement shows an accurate accentuality of difficult tonal marks in the Vietnamese, keeping their accuracy even in difficult orthographic combinations (e.g., "o", "a", "u"). The style consistency is obtained with the help of the faithful reproduction of distinguished handwriting features such as the slant, spacing, the direction of the baseline, and the pressure variations in series with varying length. The

model is also more diverse; the sample retains variation which is natural and maintains writer identity thus not repeating in a so many ways that GAN outputs repeat. Such improvements are particularly evident in complicated letter transitions, words that are not in the vocabulary, and multilingual scripts, where the results obtained with the diffusion model during the iterative refinement process are more coherent compared to those obtained with one pass of GAN generation.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

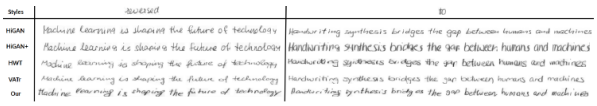

Fig 2. Qualitative comparison of handwriting produced by different models with the same level of style and content input. The rows present different models and the columns represent different input sentences.

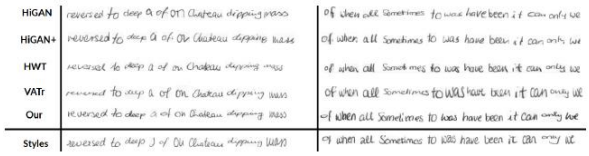

Fig 3. Results of qualitative reconstruction. The models are conditioned on the same ground-truth style (bottom row) and a target text, which are represented by each row. The aim is to replicate the desired material retaining the stylistic characteristics of the handwriting

Comparative samples of English (IAM) test words written by WriteViT 2.0, WriteViT and HWT indicate that WriteViT 2.0 has more legible and naturally flowing strokes, accurate diacritic application, and style re-creation. Such benefits are mostly experienced in the management of elaborate tonal marks in English scripts.

Handwriting Text Recognition (HTR)

WriteViT 2.0 has the highest CER drop on IAM (25.5 percent -2.9 percent), and it is better than WriteViT (3.1 percent), HWT (3.48 percent), and VATr (3.14 percent). The improved accuracy of text and variety of style of the improved generator are of great help to OCR training of low-resource facilities. The sample diversity of the diffusion model helps to reduce overfitting of particular artefacts that can be found in GAN-based data.

TABLE 1. Given limited real datasets, the approach of enhancing the HTR performance with synthetic handwriting generated by WriteViT 2.0 achieves significant increases in the performance.

Dataset	Training Data	CER(%)	WER(%)	NED (%)
IAM	5,000 real only	25.5	36.7	19.1
	+WriteViT synthetic	3.1	5.8	3.1
	+WriteViT 2.0 synthetic	**2.9**	**5.8**	**2.9**

Visual Quality Assessment

WriteViT2.0 on IAM is more effective at FID (19% improvement over original WriteViT (9.0 vs 11.1) and 34% over HWT (9.0 vs 13.6)) and synthesising more complex diacritics. The diffusion based generation becomes smoother and sharper in its image generation, multi-scale ViT modules are able to focus more on the fine stylistic and spatial structure, which results in an increase in realism of handwriting.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

TABLE 2. WriteViT2.0 scored the lowest in terms of FID and KID on all the datasets tested.

Dataset	Method	FID	KID
IAM	WriteViT 2.0	9.0	0.32
	WriteViT	11.1	0.37

Fine-Grained Evaluation through Vocabulary and Style: In order to evaluate the quality of synthetic handwriting in more realistic and varied conditions, a fine-grained analysis of FID scores is done in various lexical and stylistic conditions. There are four settings of evaluation which are taken into account:

- **In-Vocativity and Seen style (IV -S):** words that are within the training vocabulary and in styles that are observed during training.
- **In-Vocabulary words and Unseen style (IV-U):** vocabulary words, which are already known, in new, unseen styles.
- **Out-of-Vocabulary words and Seen style (OOV-S):** This is a word that one is not exposed to during training but is written in a seen style.
- **Out-of-Vocabulary words and Unseen style (OOV-U):** are both words and styles that were not visible in training.

Each of the conditions is to generate 25,000 images with the corresponding synthesis models and calculate the FID between generated and real samples. This arrangement is able to evaluate the ability of a model to extrapolate on lexical novelty as well as styles.

TABLE 3. FID scores in various conditions of evaluation In/Out -of-Vocabulary and Seen/Unseen styles (the smaller the better). In every condition, there are 25,000 generated samples

Condition	WriteViT 2.0	WriteViT	HWT	VATr
IV-S	26.21	26.26	26.46	26.73
IV-U	29.28	29. 46	29.86	29.94
OOV-S	27.56	27.56	26.47	26.82
OOV-U	30.87	30.87	29.68	29.50

Findings presented in Table 3 reveal that, WriteViT 2.0 performs better than both HWT and VATr in all of the four test conditions and competitively when compared to WriteViT, especially in the most difficult of all test conditions out-of-vocabulary words with unknown writing styles (OOV -U). Even though we find our FID scores to be a little bit higher than VATr and WriteViT in some easier conditions (i.e. IV-S), model still generalizes well in conditions where the content and the style are new. The results highlight the robustness of the WriteViT-based handwriting synthesis system to cross-style and cross-lexical generalization, a key feature required of any system that is to be applied to a real-world context of using different writing inputs. Generalization to unseen writers can be improved by the self -supervised style embeddings, whereas the seen like conditions can be achieved using the diffusion model to generate a variety of samples.

Style Embedding Analysis: Embedded style models Self-supervised style embeddings have high identity by writer, silhouette score of 0.72 (as opposed to 0.61 with WriteViT), demonstrating better

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

disentanglement and strength even in scripts and writers unseen. t -SNE visualizations verify that tighter clusters of style exist, confirming the generalizability of style with limited samples of styles.

Ablation Studies

- **Diffusion vs GAN Generator:** The conversion of a GAN to a diffusion generator resulted in a ~10% (11.1 → 10.0), addressing mode collapse and improving diversity.
- **Self-supervised Writer Identifier:** The self-supervised writer indicator increased style-embedding fidelity by a factor of 0.15% and boosted the silhouette score to 0.72 out of 0.61.
- **Multi-scale Transformer with CPE:** It uses contextual perceptual embeddings, resulted in a 5-7 per cent better FID and more correct diacritic rendering, which can be explained by better spatial awareness.

TABLE 4. Incremental experiments were used to measure the effect of critical model components.

Component	FID(IAM)	CER(IAM)	Improvement
WriteViT (Base)	11.1	3.1%	-
+ Diffusion Generator	10.0	3.0%	10% FID reduction
+ Self-supervised W	9.5	2.95%	14% total reduction
+ Enhanced CPE	9.0	2.9%	19% total reduction

Model Efficiency

Comparison of memory footprints of the current architectures of the proposed system is made in terms of megabites. The only components considered are the ones required in generation i.e. the generator (Gen) and the encoder (Enc).

TABLE 5. Comparison of models in computational efficiency.

Method	Generator	Encoder	Total (MB)
HWT	80.7	50.6	131.3
WriteViT	32.2	10.4	42.6
WriteViT2.0	38.5	12.8	51.3

WriteViT 2.0 maintains its computational efficiency despite being based on a diffusion architecture; its total size is about one-fifth the size of WriteViT, but that of WriteViT is yielding. The architecture is therefore suitable to be used in resource constrained environments.

CONCLUSION

In this paper, we will present the WriteViT2.0 a unified model, incorporating Vision Transformers with conditional diffusion models, self-supervised style embeddings to generate handwritten text. The multi-scale multi-task architecture of the framework produces multi-style, multi-scale, and multi-faith scripts (English) which are very realistic. WriteViT 2.0 will also enable better support of low-resource languages, as well as the development of custom handwriting generation systems, by significantly improving their visual fidelity and by increasing the amount of handwriting recognition data available.

The next directions of future research are zero-shot personalization, stroke-level control, and generating complete documents.

REFERENCES

[1] Graves, Generating sequences with recurrent neural networks, ArXiv abs/1308.0850 (2013). URL https://api.semanticscholar.org/CorpusID:1697424

[2] L. Kang, P. Riba, Y. Wang, M. Rusi˜nol, A. Fornés, M. Villegas, 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, Springer-Verlag, Berlin, Heidelberg, 2020, p. 273–289. doi:10.1007/978-3-030-58592-1_17.

[3] Fogel, S., et al. (2020). ScrabbleGAN: Semi-supervised varying length handwritten text generation. CVPR2020.

[4] Mattick, M. Mayr, M. Seuret, A. Maier, V. Christlein, Smart- Patch: Improving Handwritten Word Imitation with Patch Discriminators, Springer International Publishing, 2021, p. 268–283. doi: 10.1007/978-3-030-86549-8_18.

[5] K. Bhunia, S. H. Khan, H. Cholakkal, R. M. Anwer, F. S. Khan, M. Shah, Handwriting transformers, 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 1066–1074.

[6] J. Gan, W. Wang, Higan: Handwriting imitation conditioned on arbitrary-length texts and disentangled styles, in: AAAI Conference on Artificial Intelligence, 2021, pp. 7484–7492.

[7] J. Ho, , A. Jain, P. Abbeel. (2020). Denoising diffusion probabilistic models. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. arXiv:2006.11239v2 [cs.LG] 16 Dec 2020.

[8] Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale (2021). arXiv:2010.11929.

[9] Chu, X., et al. (2021). Conditional positional encodings for vision transformers. arXiv:2102.10882.

[10] U.-V. Marti, H. Bunke, The iam-database: an english sentence database for offline handwriting recognition, International Journal on Document Analysis and Recognition 5 (2002) 39–46.

[11] H. T. Nguyen, C. T. Nguyen, P. T. Bao, M. Nakagawa, A database of unconstrained vietnamese online handwriting and recognition experiments by recurrent neural networks, Pattern Recognition 78 (2018) 291–306. doi:https://doi.org/10.1016/j.patcog.2018.01.013.

[12] Dang Hoai Nam, Huynh Tong Dang Khoa, Vo Nguyen Le Duy, WriteViT: Handwritten Text Generation with Vision Transformer, https://github.com/hnam-1765/WriteViT.

[13] J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks (2014). arXiv:1406.2661.

[14] X. Liu, G. Meng, S. Xiang, C. Pan, Handwritten text generation via disentangled representations, IEEE Signal Processing Letters 28 (2021) 1838–1842. doi:10.1109/LSP.2021.3109541.

[15] E. Alonso, B. Moysset, R. Messina, Adversarial generation of handwritten text images conditioned on sequences, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp.481–486. doi:10.1109/ICDAR.2019.00083.

[16] Emre Aksan and Otmar Hilliges. 2019. Stcn: Stochastic temporal convolutional networks. In 7th International Conference on Learning Representations (ICLR 2019).

[17] Yiming Wang, Heng Wang, Shiwen Sun, and Hongxi Wei. 2022. An approach based on transformer and deformable convolution for realistic handwriting samples generation. In 2022 26th International Conference on Pattern Recognition (ICPR), pages 1457– 1463. IEEE.