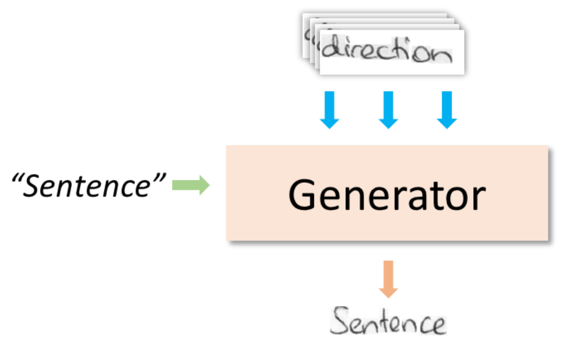
Handwriting Transformer

Ankan Kumar Bhunia[1] Salman Khan[1,2] Hisham Cholakkal[1] Rao Muhammad Anwer[1] Fahad Shahbaz Khan[1,3] Mubarak Shah[4]
[1]Mohamed bin Zayed University of AI, UAE [2]Australian National University, Australia [3]Linköping University, Sweden [4]University of Central Florida, USA

2021 ICCV OCTOBER 11-17 VIRTUAL

Problem formulation

We are given (a) set of handwritten word images as few-shot calligraphic **style examples** of one writer, (b) **query text** from an unconstrained set of vocabulary, our model strives to generate handwritten images with the same text in the writing style of the given writer.

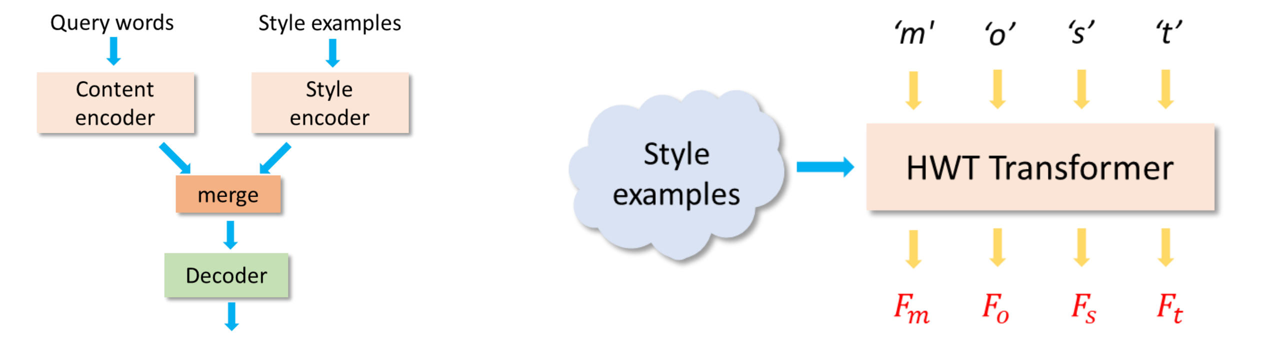


Motivation

Limitation of existing frameworks

We distinguish the main architectural constraint that impede the quality of handwritten text image generation in the existing GAN-based methods [1,2].

- Separate processing of style and content: In these models, both Style and content are loosely connected as their representative features are processed separately by their respective encoders and then later concatenated.
- Global and Local style imitation: While such a scheme enables entanglement between style and content at the word-level, it does not explicitly enforce style-content entanglement at the character-level. *As a result, they struggle to accurately imitate local styles such as character shapes or ligatures.*



(1) Existing approaches (2) Our proposed approach

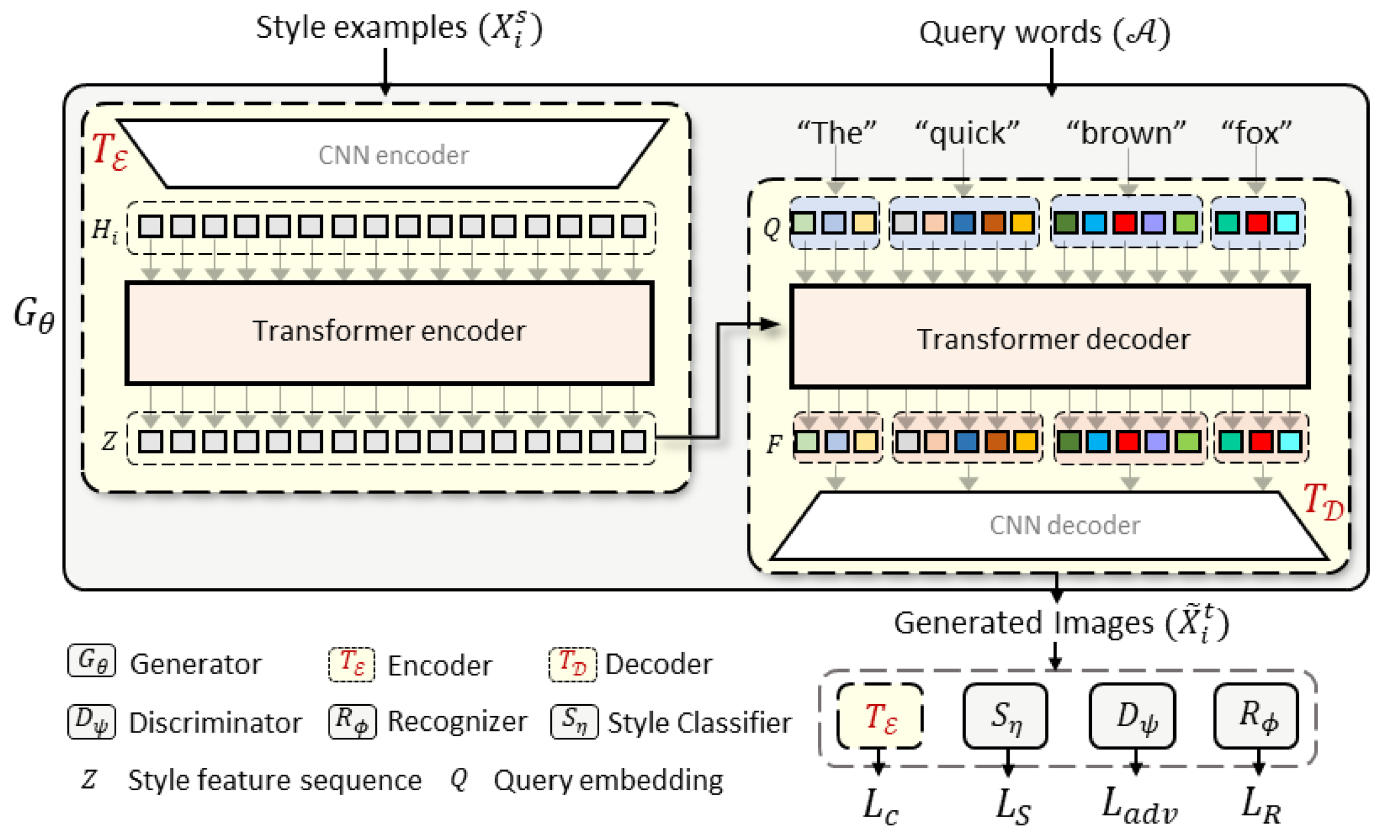
Why Transformer-based Design?

We propose a transformer based design model (HWT).

- Our proposed HWT imitates the style of a writer for a given query content through *self- and encoder-decoder attention* that emphasizes relevant self attentive style features with respect to each character in that query.
- This enables us to *(a) capture style-content entanglement at the character-level,* and *(b) model both the global as well as local style features for a given calligraphic style.*
- Further, such a tight integration between style and content leads to a *cohesive architecture design.*

Methodology

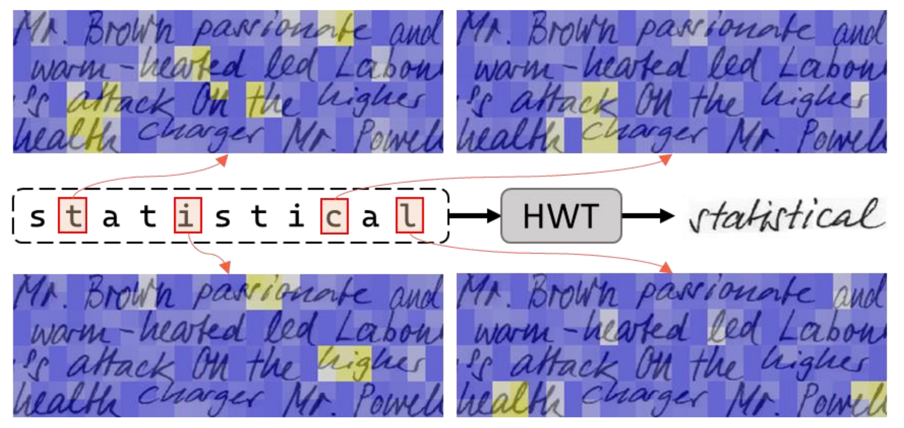
- Our proposed generative model (G_θ) comprises an encoder-decoder Transformer network.



Our training algorithm follows the traditional GAN paradigm,

- where a discriminator network (D_ψ) is employed to ensure realistic generation of handwriting styles,
- A recognizer network (R_ϕ) aids in textual content preservation,
- A writer style classifier (S_η) ensures satisfactory transfer of the calligraphic styles.
- In addition, we use cycle loss. that ensures the original style feature sequence can be reconstructed from the generated image.

Visualization of Attention maps



The attention maps are computed for each character in the query word (*statistical*) which are then mapped to spatial regions in the given example style images.

Experiments

Quantitative analysis of style imitation

	IV-S ↓	IV-U ↓	OOV-S ↓	OOV-U ↓
GANwriting [1]	120.07	124.30	125.87	130.68
Davis *et al* [2]	118.56	128.75	127.11	136.67
HWT (Ours)	**106.97**	**108.84**	**109.45**	**114.10**

Our HWT performs favorably in all four settings: In-Vocabulary words and seen style (**IV-S**), In Vocabulary words and unseen style (**IV-U**), Out of vocabulary content and seen style (**OOV-S**) and Out of vocabulary content and unseen style (**OOV-U**).

Quantitative analysis of Handwritten Text Generation

	FID ↓	GS ↓
ScrabbleGAN [3]	20.72	2.56×10^{-2}
Davis *et al* [2]	20.65	4.88×10^{-2}
HWT (Ours)	**19.40**	$\mathbf{1.01 \times 10^{-2}}$

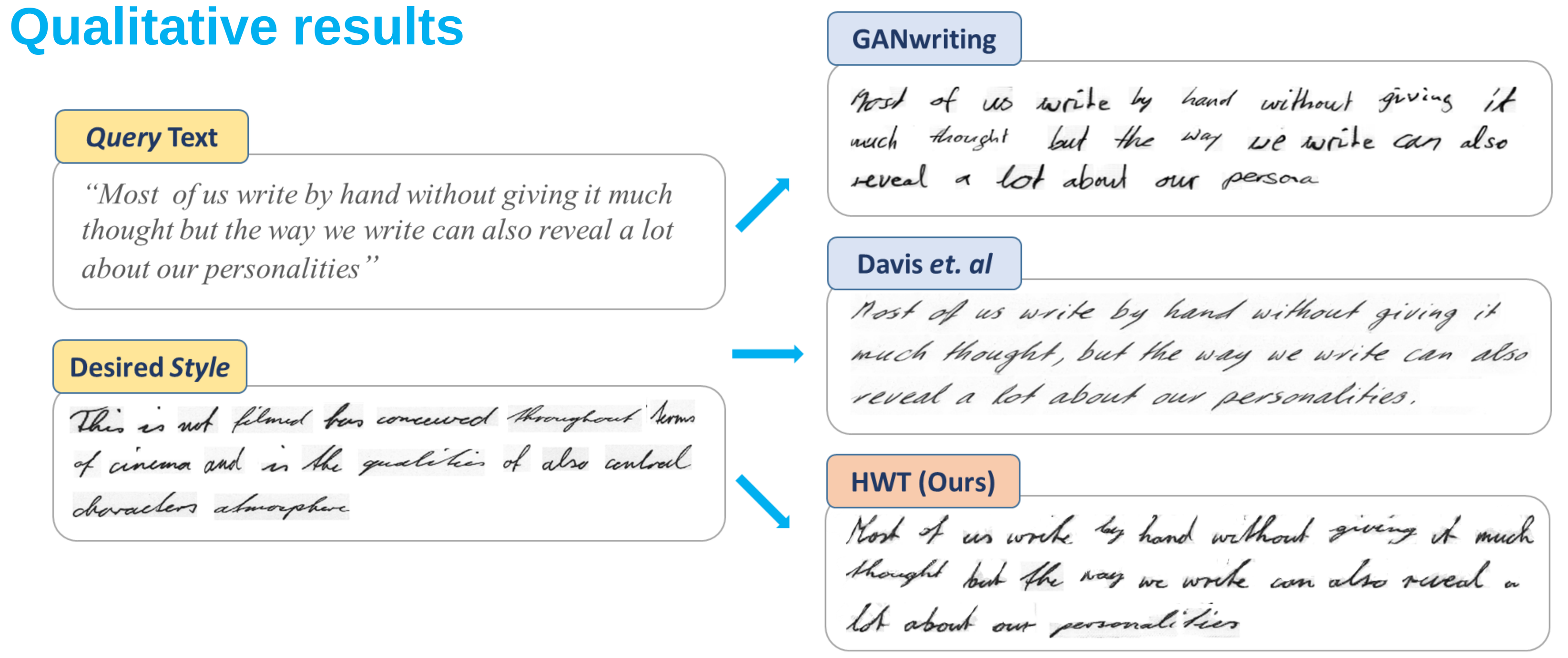
We evaluate the quality of the text image generated by our HWT following the same evaluation settings as used in ScrabbleGAN. Our HWT performs favorably against these methods in terms of both FID and GS score.

Handwritten Text Recognition (HTR)

Method	Training Data			CVL(%)		CVLoov(%)	
	GAN	CVL	IAM	WER	CER	WER	CER
—	✗	✓	✓	29.41	13.13	37.63	17.16
HiGAN [4]	✓	✓	✓	28.91	12.54	37.06	16.67
ScrabbleGAN [3]	✓	✓	✓	28.68	12.13	37.10	16.73
HWT (Ours)	✓	✓	✓	**27.81**	**11.84**	**36.47**	**15.95**

We utilize our generated samples for training HTR model to validate if the generated images can help improve text recognition performance.

Qualitative results



Conclusion

Qualitative, quantitative and human-based evaluations show that our HWT produces realistic styled handwritten text images with varying length and any desired writing style.

[1] Kang *et.al*, Ganwriting: Content conditioned generation of styled handwritten word images. In ECCV, 2020.
[2] Davis *et al.*, Text and style conditioned gan for generation of offline handwriting lines. BMVC, 2020.
[3] Fogel *et al*, Scrabblegan: semi-supervised varying length handwritten text generation. In CVPR, 2020.
[4] Gan *et. al.* HiGAN: Handwriting Imitation Conditioned on Arbitrary-Length Texts and Disentangled Styles. In AAAI, 2021