





A Transformer Based Handwriting Recognition System Jointly Using Online and Offline Features

Ayush Lodh¹, Ritabrata Chakraborty^{1,2}, Shivakumara Palaiahnakote³,
and Umapada Pal¹

¹ Indian Statistical Institute, India
{ayushlodh26, ritabrata04}@gmail.com
umapada@isical.ac.in

² Manipal University Jaipur, India

³ University of Salford, United Kingdom
s.palaiahnakote@salford.ac.uk

Abstract. We posit that handwriting recognition benefits from complementary cues carried by the rasterized complex glyph and the pen’s trajectory, yet most systems exploit only one modality. We introduce an end-to-end network that performs early fusion of offline images and online stroke data within a shared latent space. A patch encoder converts the grayscale crop into fixed-length visual tokens, while a lightweight transformer embeds the (x, y, pen) sequence. Learnable latent queries attend jointly to both token streams, yielding context-enhanced stroke embeddings that are pooled and decoded under a cross-entropy loss objective. Because integration occurs before any high-level classification, temporal cues reinforce each other during representation learning, producing stronger writer independence. Comprehensive experiments on IAMOnDB, and VNON-DB demonstrate that our approach achieves state-of-the-art accuracy, exceeding previous bests by up to 1%. Our study also shows adaptation of this pipeline with gesturification on the ISI-Air dataset. Our code can be found here.

Keywords: Handwritten text recognition · Transformer · Online–offline fusion.

1 Introduction

Handwritten Text Recognition (HTR) is a foundational task in the broader domain of handwriting analysis [1], with widespread applications in document digitization, education technology, and human-computer interaction. While many advances have been made in recognizing isolated characters using deep learning, most existing models are trained on datasets such as UNIPEN [14], comprising neatly segmented, individual characters written in isolation. However, this is far removed from real-world scenarios, where characters look different when

* Ayush Lodh and Ritabrata Chakraborty contributed equally to this work.

** Work done during internship at Indian Statistical Institute.

embedded within a word, influenced by neighboring characters, and written in a variety of natural handwriting styles.

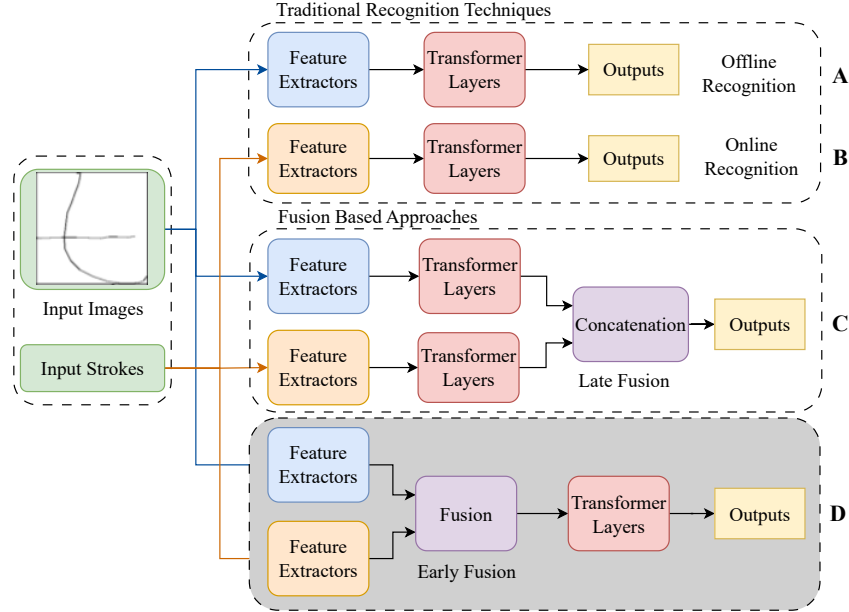


Fig. 1: An overview of input modalities and their architectures for handwritten text recognition. A: Image-only input, B: Stroke-only input, C: Late Fusion based dual input, D: Early Fusion based dual input (Ours).

Early handwritten-character recognition systems were designed around a single modality. Offline pipelines (Fig. 1-A) treat a scanned glyph as an image and rely on convolutional or transformer backbones to map pixels to character sequences [24]. Conversely, online pipelines (Fig. 1-B) discard appearance altogether, interpreting only the pen-tip trajectory captured by the digitiser [31]. Each view is incomplete as images lose temporal order, while stroke traces omit shading, pen pressure, and context such as ligatures or serifs. To mitigate this limitation, bimodal fusion emerged in a multiscale network [37] that concatenates high-level image and stroke embeddings before classification (late fusion; Fig. 1-C). Such strategies boost robustness, yet they still process both streams independently until the penultimate layer. As a result, misaligned timelines, redundant features, and modality-specific noise remain unaddressed, and the network cannot learn shared primitives—e.g., a cusp that appears as both a sharp curvature in trajectory space and a dark pixel cluster in image space.

We posit that early fusion (Fig. 1-D) unlocks deeper complementarity. **By projecting raw visual patches and stroke tokens into a shared la-**

tent space and allowing cross-attention before any task-specific transformer layers, the model can utilise the interaction between how a character looks in pixel space and how it is traced through the pen vectors. We utilize IAMOn-DB [27] and VNOn-DB [30] datasets adapted for character-level recognition to design our experimental study for this hypothesis. We also display recognition results on the ISI-Air [33] dataset, showing strong results to air-writing data.

- i. We explore a novel direction in the field of handwritten text recognition, leveraging the relationship of character images and the pen stroke vectors used in tracing them.
- ii. We propose HATCharClassifier, a first of its kind handwritten text recognizer that provides robust character recognition results through early fusion of multiple inputs (image + stroke).
- iii. We benchmark our model as state-of-the-art for multiple multimodal handwritten text datasets such as IAMOn-DB [27] and VNOn-DB [30], along with air writing datasets like ISI-Air [33].

The rest of the paper is structured as follows: Section 2 explores related literature on early classical methods, deep learning and transformer based methods, and fusion based models for handwritten text recognition. Section 3 explores our proposed framework HATCharClassifier. Section 4 describes experiment design, datasets used and metrics we employ to show the performance of our method. Section 5 delves into quantitative and qualitative results for the proposed framework. Section 6 provides a discussion on the importance of such a method with regards to current research and its caveats, and finally, Section 7 concludes the paper.

2 Related Work

2.1 Early Methods for Handwriting Recognition

Statistical and classical machine learning approaches have played a pivotal role in advancing handwritten text recognition in the early years of the field. Hidden Markov Models (HMMs) emerged as the dominant framework for modeling sequential handwriting, particularly excelling in cursive script recognition due to their ability to perform implicit segmentation and probabilistic modeling of character sequences [7,32,29,5]. Support Vector Machines (SVMs) gained traction for isolated character recognition by leveraging margin-based classification and kernel methods, with adaptations such as dynamic time warping kernels enabling their application to online handwriting sequences [3]. To capture richer dependencies, discriminative models like Conditional Random Fields (CRFs) were also explored, for example, [10] showed that CRFs can outperform HMMs on whole-word recognition tasks. These classical methods typically operate on carefully engineered features. For instance, Bai and Huo [4] extract 8-directional histogram features from pen trajectories for online Chinese character recognition. Likewise,

many systems convert raw strokes into offline images or other local descriptors to feed into the sequence model. Alongside these, k-Nearest Neighbors (k-NN) and shallow neural networks like multilayer perceptrons served as competitive baselines for digit and character recognition, especially on benchmarks such as UNIPEN [14] and CEDAR [16]. These classifiers were highly reliant on carefully engineered features such as zoning, projection histograms, contour profiles, and geometric descriptors—extracted from normalized and preprocessed handwriting samples [18,32]. Together, these classical methods laid the algorithmic foundation for contemporary handwritten text recognition systems.

2.2 Deep Learning Based Methods

With the advent of deep learning [22], end-to-end neural models became standard for HTR. [34] introduced the CRNN architecture, combining CNN feature extraction with bidirectional RNN sequence modeling. Bi-directional LSTM (BiLSTM) networks [17] or gated RNNs then capture long-range context in the stroke/image sequence. [12] demonstrate that a BLSTM trained with Connectionist Temporal Classification (CTC) loss [11] can significantly outperform traditional HMM baselines on unconstrained handwriting recognition. Encoder-decoder models with attention have also been applied for lexicon-free transcription. More recently, fully Transformer-based OCR models have appeared. For example, [23] propose TrOCR, which uses a pre-trained Vision Transformer encoder and text Transformer decoder, yielding state-of-the-art results on handwritten text recognition benchmarks. These works paved the way towards performance-maximising models suitable for multilingual and multidomain use-cases.

2.3 Recent Attention-Based Methods

With the advent of attention mechanisms [36], transformer architectures have begun to be applied to handwriting recognition. [24] adapted the Vision Transformer (ViT) [9] for line-level text recognition. Their HTR-VT model uses a CNN for feature extraction and employs sharpness-aware minimization, achieving competitive accuracy on standard HTR datasets like IAMOn-DB [27]. [20] introduce Character Queries, a transformer decoder where each character is represented by a learned query vector; this approach excels at segmenting on-line strokes into characters given a known transcription. C-TST [8], a two-stream model using a 1D convolution + Transformer branch for temporal stroke features and a Vision Transformer for spatial image features; fusing both streams yields high accuracy on Chinese benchmarks. [25] utilized the Swin Transformer as the encoder to extract image features, focusing on Chinese character characteristics. These Transformer-based methods complement and often improve upon earlier RNN and CNN-based systems.

2.4 Bimodal Fusion Methods

Multimodal (image + stroke) fusion has been widely studied to improve robustness. Most recent methods employ late-fusion multi-stream architectures: sepa-

rate encoders process pen trajectories and images, and their outputs are merged. For example, [37] propose a multi-scale bimodal fusion network that combines features from both streams using Transformers, achieving state-of-the-art accuracy on IAMOn-DB (e.g. 4.7% CER). Similarly, Bhunia et al. [6] fuse online trajectory features with rendered images for Indic script recognition. While these late-fusion models yield high accuracy, they incur extra complexity due to separate image rendering and fusion modules. We identify this as a **domain gap**: training stroke and image encoders independently can limit joint feature learning. To address this, we introduce an *early fusion* strategy that jointly embeds stroke and image information from the outset.

3 Methodology

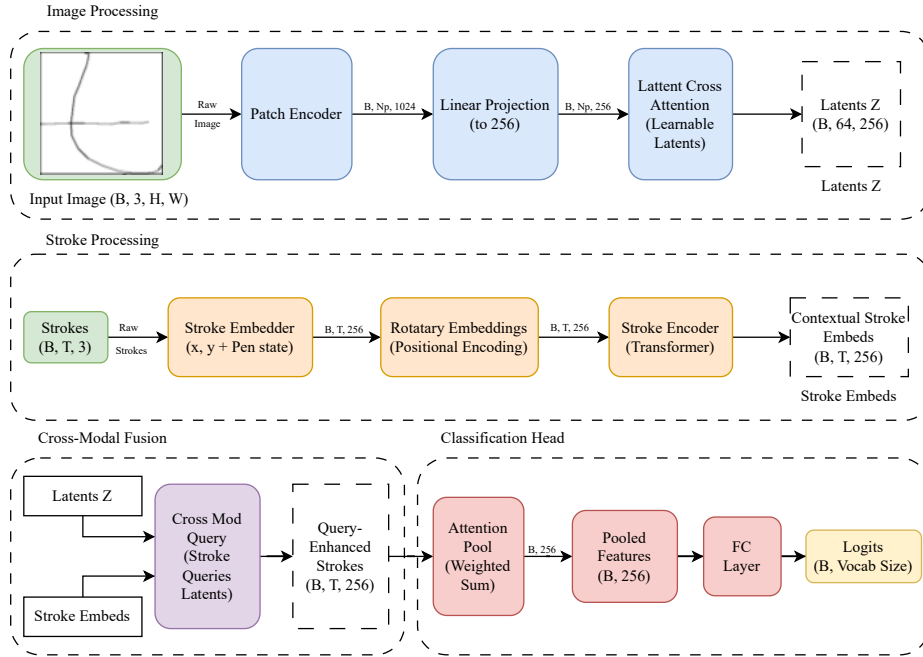


Fig. 2: Our proposed pipeline.

Notation. Let $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{S}_i, y_i)\}_{i=1}^M$ with $\mathbf{I}_i \in \mathbb{R}^{H \times W \times C}$ (gray or RGB image), $\mathbf{S}_i = [(x_t, y_t, p_t)]_{t=1}^{T_i} \in \mathbb{R}^{T_i \times 3}$ (online stroke sequence, $p_t \in \{0, 1\}$ pen state), and $y_i \in \{1, \dots, V\}$. HAT converts either modality—or both—into a common d -dimensional token space and classifies with a linear head (Fig. 2).

Image Patch Encoder. We first (bi-linearly) resize \mathbf{I} to 224×224 and, when necessary, replicate its single channel to obtain $C=3$. A pretrained Swin-B [26] backbone outputs the last-stage feature map $\mathbf{F} \in \mathbb{R}^{7 \times 7 \times 1024}$. Flattening the spatial axes and projecting with $\mathbf{W}_p \in \mathbb{R}^{1024 \times d}$ gives

$$\mathbf{E}_p = \text{reshape}(\mathbf{F}) \mathbf{W}_p \in \mathbb{R}^{N \times d}, \quad N = 7 \times 7. \quad (1)$$

Latent Cross-Attention. Following Perceiver-IO [19], we introduce L learnable *latent tokens* $\mathbf{Z}^{(0)} \in \mathbb{R}^{L \times d}$. For layers $\ell = 0, \dots, L_{\text{img}} - 1$

$$\tilde{\mathbf{Z}}^{(\ell)} = \text{MHA}(\mathbf{Q}=\mathbf{Z}^{(\ell)}, \mathbf{K}=\mathbf{E}_p, \mathbf{V}=\mathbf{E}_p), \quad (2a)$$

$$\mathbf{Z}^{(\ell+1)} = \text{TransformerLayer}(\mathbf{Z}^{(\ell)} + \tilde{\mathbf{Z}}^{(\ell)}). \quad (2b)$$

Layer-norm on the final state yields $\mathbf{Z} = \text{LN}(\mathbf{Z}^{(L_{\text{img}})})$.

Stroke Encoder. Each point (x_t, y_t, p_t) is embedded by concatenating raw coordinates with a pen-state lookup $\mathbf{E}_{\text{pen}} \in \mathbb{R}^{2 \times d/8}$:

$$\mathbf{s}_t = [x_t, y_t, \mathbf{E}_{\text{pen}}[p_t]] \in \mathbb{R}^{2+d/8}. \quad (3)$$

After a point-wise projection, BatchNorm, and Dropout we obtain

$$\mathbf{E}_s = [\mathbf{s}_t]_{t=1}^T \mathbf{W}_s \in \mathbb{R}^{T \times d}, \quad \mathbf{W}_s \in \mathbb{R}^{(2+d/8) \times d}. \quad (4)$$

Rotary positional encoding. Splitting every token into even/odd parts and rotating with angle $\phi_t = t\Theta$ (see [35]) gives $\hat{\mathbf{E}}_s \in \mathbb{R}^{T \times d}$.

Temporal transformer. An N_{stk} -layer Transformer processes the sequence

$$\mathbf{H} = \text{TransformerEncoder}(\hat{\mathbf{E}}_s) \in \mathbb{R}^{T \times d}, \quad (5)$$

which is refined via a residual 2-layer MLP:

$$\mathbf{E}_{\text{stroke}} = \mathbf{H} + \text{MLP}(\text{LN}(\mathbf{H})). \quad (6)$$

Cross-Modal Querying. With both modalities present, stroke tokens query the latent image tokens:

$$\tilde{\mathbf{E}}_{\text{stk}} = \text{MHA}(\mathbf{Q}=\mathbf{E}_{\text{stroke}}, \mathbf{K}=\mathbf{Z}, \mathbf{V}=\mathbf{Z}), \quad (7a)$$

$$\mathbf{E}_{\text{cross}} = \text{TransformerLayer}(\mathbf{E}_{\text{stroke}} + \tilde{\mathbf{E}}_{\text{stk}}). \quad (7b)$$

If images (resp. strokes) are missing we set $\mathbf{T} = \mathbf{Z}$ (resp. $\mathbf{T} = \mathbf{E}_{\text{stroke}}$).

Attention Pooling & Classification. Given token matrix $\mathbf{T} \in \mathbb{R}^{N_t \times d}$ ($N_t = L$ or T), scalar importances

$$\alpha_i = \frac{\exp(\mathbf{w}_2^\top \tanh(\mathbf{W}_1 \mathbf{t}_i))}{\sum_{j=1}^{N_t} \exp(\mathbf{w}_2^\top \tanh(\mathbf{W}_1 \mathbf{t}_j))}, \quad \mathbf{g} = \sum_{i=1}^{N_t} \alpha_i \mathbf{t}_i, \quad (8)$$

are computed with $\mathbf{W}_1 \in \mathbb{R}^{d \times d}$, $\mathbf{w}_2 \in \mathbb{R}^d$. The classifier is

$$\mathbf{o} = \mathbf{W}_c \mathbf{g} + \mathbf{b}_c \in \mathbb{R}^V, \quad (9)$$

and we minimise cross-entropy

$$\mathcal{L} = -\frac{1}{M} \sum_{i=1}^M \log \frac{\exp(o_{i,y_i})}{\sum_{v=1}^V \exp(o_{i,v})}. \quad (10)$$

Algorithm 1 HAT (refer to Eq. (1)–(10) as aforementioned)

Require: mini-batch $\{(\mathbf{I}_i, \mathbf{S}_i, y_i)\}_{i=1}^B$, mode $m \in \{\text{IMAGE}, \text{STROKE}, \text{BOTH}\}$

```

1 for  $i \leftarrow 1$  to  $B$  do
2   if  $m \neq \text{STROKE}$  then                                ▷ image branch
3      $\mathbf{E}_p \leftarrow \text{IMAGEPATCHENCODER}(\mathbf{I}_i)$               ▷ Eq. (1)
4      $\mathbf{Z} \leftarrow \text{LATENTCROSSATTN}(\mathbf{E}_p)$                   ▷ Eqs. (2a)–(2b)
5   end if
6   if  $m \neq \text{IMAGE}$  then                                ▷ stroke branch: Sec. 3
7      $\mathbf{E}_{\text{stk}} \leftarrow \text{STROKEENCODER}(\mathbf{S}_i)$               ▷ Eqs. (4)–(6)
8   end if
9   Select  $\mathbf{T}$ 
10    •  $m = \text{IMAGE} \rightarrow \mathbf{T} = \mathbf{Z}$ 
11    •  $m = \text{STROKE} \rightarrow \mathbf{T} = \mathbf{E}_{\text{stk}}$ 
12  if  $m = \text{BOTH}$  then                                ▷ hybrid: Sec. 3
13     $\mathbf{T} \leftarrow \text{CROSSMODALQUERY}(\mathbf{E}_{\text{stk}}, \mathbf{Z})$           ▷ Eq. (7b)
14  end if
15   $\mathbf{g} \leftarrow \text{ATTENTIONPOOL}(\mathbf{T})$                       ▷ Eq. (8)
16   $\mathbf{o}_i \leftarrow \mathbf{W}_c \mathbf{g} + \mathbf{b}_c$                           ▷ Eq. (9)
17 end for
18  $\mathcal{L} \leftarrow \text{CROSSENTROPY}(\{\mathbf{o}_i\}, \{y_i\})$           ▷ Eq. (10)
19 Back-propagate  $\nabla \mathcal{L}$ ; update parameters

```

In summary, HAT maps images (via a frozen Swin_B) and online strokes (via a rotary-encoded transformer) to a shared d -dimensional token space. Stroke tokens optionally query image tokens through a single cross-modal attention layer, after which attentional pooling yields a global vector for linear classification with cross-entropy. The same lightweight architecture handles image-only, stroke-only, and hybrid inputs without altering parameters.

4 Experiments

4.1 Datasets

Table 1: Dataset statistics after pre-processing

Dataset	Train Set	Val. Set	Test Set	# Cls.
IAMOn-DB [27]	72,508	18,954	21,455	84
VNOn-DB [30]	197,140	57,763	77,389	145
ISI-Air [33]	10,000	2,000	-	10

Numbers = character samples.

We evaluate on three on-line handwriting corpora, each trained and tested in isolation. Qualitative examples are given in Fig. 3, 4, 5, while Table 1 (left) reports the final number of character instances per split together with the number of target classes.

The **IAMOn-DB** [27] dataset is a widely used benchmark for online handwriting recognition, particularly focused on English cursive script. It contains handwritten text samples collected from 221 writers using a stylus on a tablet,

capturing the temporal sequence of pen strokes along with their spatial coordinates. IAMOn-DB supports writer-independent and writer-dependent evaluation protocols and is frequently used for training and evaluating sequence models like HMMs and RNNs. Its high-quality online handwriting data has made it a standard in evaluating temporal modeling capabilities in handwriting recognition systems.

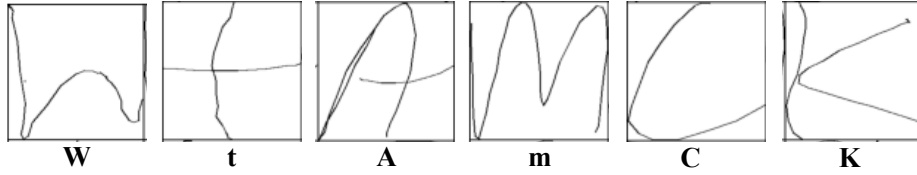


Fig. 3: Some examples from IAMOn-DB dataset.

The VNOn-DB (Vietnamese Online Handwriting Database) [30] is a large-scale dataset designed to support research in Vietnamese online handwriting recognition. It comprises pen trajectory data collected from over 200 writers, covering all 134 Vietnamese characters including diacritics. Each character is annotated with stroke order and pen-up/pen-down events, providing rich temporal and spatial information. VNOn-DB presents challenges specific to the Vietnamese language, such as compound characters and tonal marks, making it a valuable resource for evaluating script-specific handwriting models. The dataset has been used to benchmark both character-level and word-level recognition tasks in low-resource language settings.

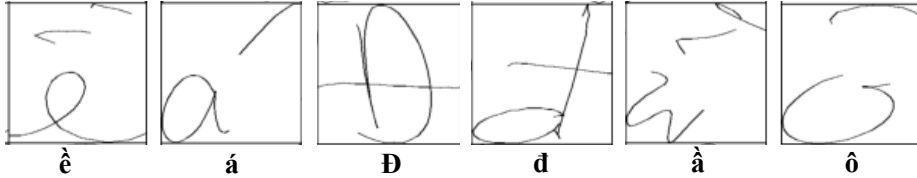


Fig. 4: Some examples from VNOn-DB dataset.

The ISI-AIR dataset[33] is a publicly available corpus designed for research in mid-air handwriting recognition using motion capture. Collected at the Indian Statistical Institute (ISI), the dataset comprises 3D hand trajectory recordings of English digits captured using a webcam. Unlike traditional handwriting, ISI-AIR features freehand air gestures without physical contact, introducing challenges such as higher spatial variance, motion blur, and absence of surface constraints.

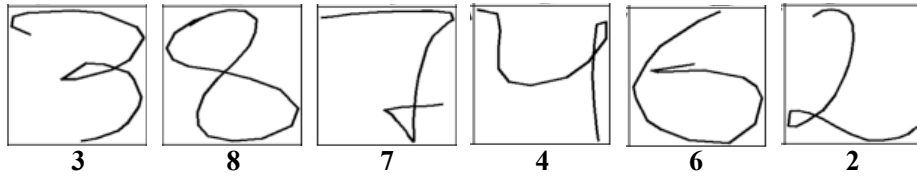


Fig. 5: Some examples from ISI-Air dataset.

4.2 Implementation Details

All experiments are conducted on a single NVIDIA RTX A5000 (24 GB) GPU using the PyTorch [2] framework. Training proceeds with the AdamW [28] optimizer (initial learning rate 1×10^{-4} , $\beta_1=0.9$, $\beta_2=0.999$, weight-decay 0.01). The learning rate follows a cosine-annealing schedule that decays to zero, and gradients are clipped to an ℓ_2 -norm of 1.0. Classification uses cross-entropy with label-smoothing 0.1, while dropout ($p=0.1$) and batch-normalization regularize the stroke pathway.

We evaluate with overall accuracy $Acc = \frac{\sum_c TP_c}{N}$ and the class-balanced macro variants of precision, recall and F_1 : for each class c we compute $P_c = \frac{TP_c}{TP_c + FP_c}$, $R_c = \frac{TP_c}{TP_c + FN_c}$ and $F_{1,c} = \frac{2TP_c}{2TP_c + FP_c + FN_c}$, then average them to obtain $\bar{P} = \frac{1}{C} \sum_c P_c$, $\bar{R} = \frac{1}{C} \sum_c R_c$ and $\bar{F}_1 = \frac{1}{C} \sum_c F_{1,c}$. All scores are reported in percentage.

5 Results

Table 2: Comparison across datasets. Highest values per metric are highlighted. I: Image, S: Stroke, D: Dual

Architecture	Mode	Acc. (%)	Precision (%)	Recall (%)	F1 (%)
IAM Dataset [27]					
HTR-VT [24]	I	95.3	94.9	94.7	94.8
HAT	I	91.5	90.8	88.0	89.4
LSTM [13]	S	90.7	89.3	90.0	88.6
HAT	S	89.5	90.2	85.0	87.5
OLHTR [37]	D	95.3	95.1	94.6	93.8
HAT	D	96.4	94.0	92.5	93.7
VNOn-DB Dataset [30]					
CNN-LSTM [21]	I	95.3	95.0	94.2	94.6
HAT	I	92.6	91.5	90.7	91.1
HAT	S	72.1	71.0	68.5	69.7
HAT	D	95.8	95.5	95.0	95.2
ISI-Air Dataset [33]					
HAT	I	99.5	98.7	99.1	98.4
HAT	S	98.1	97.3	97.4	98.0
RNN-LSTM [33]	D	98.7	98.6	98.5	98.6
HAT	D	99.8	99.5	98.2	98.7

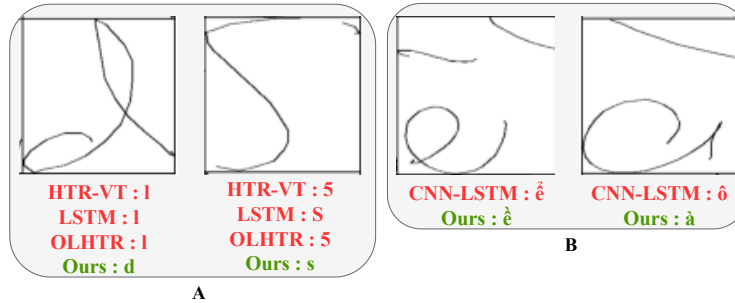


Fig. 6: Qualitative results (A): IAM-OnDB, (B): Vn-OnDB. Red denotes incorrect recognition, our methods show correctly recognized class in green for all cases shown here.

Table 2 displays the performance of comparable models in literature against our proposed recognizer. For the **IAMOn-DB** dataset, we note a mean 1.5% improvement in accuracy across all modes (image-only, stroke-only and dual input). OLHTR [37] exceeds for dual input marginally ($< 1\%$) in precision, recall and F-1 score. We observe an interesting 3.8% gain in accuracy for image-only mode, highlighting the robustness and utility of our framework even when both inputs are not available.

For **VNOn-DB** and we provide their first dual-input benchmarking results. For VNOn-DB, we note an accuracy of 95.8% for dual input, as opposed to 72.1% on stroke-only mode. This further reinforces our claim, because although we assume the per-character stroke richness and vocabulary size (145) of the dataset would be much higher than IAMOn-DB, simply relying on stroke information leads to confusion in predictions. Combining image inputs and correlating offline characteristics with the strokes leads to robust identification.

We notice a similar trend in accuracy for the **ISI-Air** dataset, with our dual-input model achieving a 1.7% increase from stroke-only input mode, and an overall 1.1% increase compared to the traditional RNN-LSTM architecture previously reported in literature [33].

We display some qualitative recognition results in Figure 6. It is interesting to note that even though character isolation removes the context of the word it is taken from, our model provides correct results for slightly varying characters even when other paradigms fail.

5.1 Ablations

We provide insights into various settings of feature extraction by swapping out the patch encoder backbone shown in Table 3. Swin-B 224 [26] displays the best accuracy when used in training; by using frozen weights, the accuracy drops notably by 10%. At its lightest setting (ResNet-18) [15] we get an accuracy of 79.13% highlighting the impact of our chosen Swin-B [26] backbone. Training

Table 4: Comparison of convergence with other models on IAMOn-DB [27]

Models	Acc. (%)	#Params (M)	Conv. Epochs
LSTM [13]	90.07	—	150
HAT (Strokes)	89.50	4.3	46
HTR-VT [24]	95.30	53.5	1 165
HAT (Images)	91.50	88.6	4
OLHTR [37]	95.30	—	—
HAT (Fusion)	96.42	94.4	3

“#Params” denotes total model parameters; “—” = not reported in the paper.

We benchmark the proposed HAT variants against representative on-line/off-line recognisers. We observe LSTM and recent transformer-based systems require hundreds to thousands of epochs. In contrast, the efficient-fusion HAT converges in just **three** epochs while retaining competitive accuracy (see Table 4).

cost is considerably alleviated since our model converges at the 4th epoch itself for the chosen setting.

Table 3: Ablation study with different image feature extractors.

Feature Extractor	Status	Acc. (%)	#Params (M)	Conv. Epochs
ResNet-18[15]	❄	79.13	—	15th
ResNet-18[15]	🔥	82.95	25.04	8th
ResNet-34[15]	❄	76.31	—	16th
ResNet-34[15]	🔥	80.23	35.14	9th
ViT (Base)[9]	❄	82.42	—	9th
ViT (Base)[9]	🔥	86.14	93.47	4th
Swin-B 224[26]	❄	82.33	—	22nd
Swin-B 224[26]	🔥	92.42	94.48	4th

All results are measured on the IAMOn-DB for dual inputs. #Params refers to the parameters of the complete model, including both the stroke and the image branch.

❄ denotes frozen and 🔥 denotes trainable parameters

6 Discussion

Limitations. The current system recognises isolated glyphs, assumes pre-segmented inputs, and depends on a training visual backbone of 90 M parameters. Performance has not been audited across writer demographics or fine-grained stroke disorders, and segmentation errors in IAMOn-DB and VNOOn-DB introduce small but uncorrected label noise. We also notice some failure cases for our model in Figure 7.

We benchmark two fusion strategies for our HAT architecture on IAMOn-DB. Early and mid-level fusions are contrasted with the transformer-based OLHTR baseline. Numerical results are summarised in Table 5.

Table 5: Fusion-level comparison of our models on IAMOn-DB [27].

Model Variant	Fusion	Acc. (%)
HAT	Early	96.40
HAT	Middle	92.10
OLHTR [37]	Late	95.30

“Acc.” = accuracy.

Table 6: Robustness of the dual-trained model to modality dropout (IAMOn-DB). Δ is the absolute accuracy drop relative to the full-input baseline.

Train \rightarrow Test	Acc. (%)	Δ (%)
Dual \rightarrow Dual	92.4	0.0
Dual \rightarrow Image-only	88.1	-4.3
Dual \rightarrow Stroke-only	85.7	-6.7

We train a single model with dual modalities and then evaluate it under three conditions: (i) the full input, (ii) image stream only and (iii) stroke stream only. The network gracefully degrades when a modality is absent at test time. A modest drop of 4–7% shows the model remains fairly robust to real-world sensor failure.

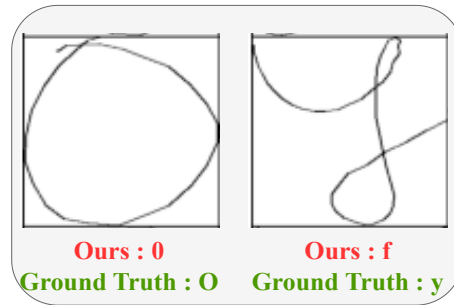


Fig. 7: Failure cases for our model.

Future Work. Future work lies in curation of a large level multilingual word level and line level dataset to extend the usage of this framework to real applications. There is also potential in providing for manually segmented character information that might help in better benchmarking results.

Ethical Considerations. Pen trajectories may act as a biometric, enabling unintended writer-identification. Some security concerns might stem from learned correlations between motor patterns through stroke and subsequent formed character images, which warrants careful auditing of software that uses this framework.

7 Conclusion

We propose HATCharClassifier, a novel framework that utilises early fusion utilising online and offline input modalities for handwritten text recognition. Our method displays the utility of capturing the correlation between the two modalities using cross-modal querying, leading to more robust recognition across multiple datasets as discussed above. Our approach opens a new direction in this field, motivating future work for word-level and line-level air-writing dataset collection and benchmarking. We notice some limitations such as minor sparse inconsistencies in character stroke segmentation used in preprocessing for the IAMOn-DB and VNon-DB datasets from word-level, and heavier parameter and floating-point operations (FLOPs) than other existent methods using dual input. We believe this leaves potential for further improvement in future work.

References

1. AlKendi, W., Gechter, F., Heyberger, L., Guyeux, C.: Advancements and challenges in handwritten text recognition: A comprehensive survey. *Journal of Imaging* **10**(1) (2024). <https://doi.org/10.3390/jimaging10010018>, <https://www.mdpi.com/2313-433X/10/1/18>
2. Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C.K., Maher, B., Pan, Y., Puhersch, C., Reso, M., Saroufim, M., Siraichi, M.Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., Chintala, S.: PyTorch 2: Faster Machine Learning Through Dynamic Python Byte-code Transformation and Graph Compilation. In: *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ASPLOS '24, Association for Computing Machinery (Apr 2024). <https://doi.org/10.1145/3620665.3640366>, <https://docs.pytorch.org/assets/pytorch2-2.pdf>
3. Bahlmann, C., Haasdonk, B., Burkhardt, H.: Online handwriting recognition with support vector machines—a kernel approach. In: *Proceedings eighth international workshop on frontiers in handwriting recognition*. pp. 49–54. IEEE (2002)
4. Bai, Z.L., Huo, Q.: A study on the use of 8-directional features for online handwritten chinese character recognition. In: *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. pp. 262–266. IEEE (2005)
5. Bengio, Y., LeCun, Y., Nohl, C., Burges, C.: Lerec: A nn/hmm hybrid for on-line handwriting recognition. *Neural computation* **7**(6), 1289–1303 (1995)
6. Bhunia, A.K., Mukherjee, S., Sain, A., Bhunia, A.K., Roy, P.P., Pal, U.: Indic handwritten script identification using offline-online multi-modal deep network. *Information Fusion* **57**, 1–14 (2020)
7. Bozinovic, R.M., Srihari, S.N.: Off-line cursive script word recognition. *IEEE Transactions on pattern analysis and machine intelligence* **11**(1), 68–83 (1989)
8. Chen, Y., Zheng, H., Li, Y., Ouyang, W., Zhu, J.: Online handwritten chinese character recognition based on 1-d convolution and two-streams transformers. *IEEE Transactions on Multimedia* **26**, 5769–5781 (2023)

9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Feng, S., Manmatha, R., McCallum, A.: Exploring the use of conditional random field models and hmms for historical handwritten document recognition. In: Second International Conference on Document Image Analysis for Libraries (DIAL'06). pp. 8–pp. IEEE (2006)
11. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376 (2006)
12. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence* **31**(5), 855–868 (2008)
13. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems* **28**(10), 2222–2232 (2016)
14. Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., Janet, S.: Unipen project of on-line data exchange and recognizer benchmarks. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5). vol. 2, pp. 29–33. IEEE (1994)
15. 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)
16. www.cedar.buffalo.edu/ilt/research.html
17. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
18. Impedovo, D., Pirlo, G.: Zoning methods for handwritten character recognition: A survey. *Pattern Recognition* **47**(3), 969–981 (2014)
19. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)
20. Jungo, M., Wolf, B., Maksai, A., Musat, C., Fischer, A.: Character queries: A transformer-based approach to on-line handwritten character segmentation. In: International Conference on Document Analysis and Recognition. pp. 98–114. Springer (2023)
21. Le, A.D., Nguyen, H.T., Nakagawa, M.: An end-to-end recognition system for unconstrained vietnamese handwriting. *SN Computer Science* **1**(1), 7 (2020)
22. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436–444 (2015)
23. 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)
24. Li, Y., Chen, D., Tang, T., Shen, X.: Htr-vt: Handwritten text recognition with vision transformer. *Pattern Recognition* **158**, 110967 (2025)
25. Li, Z., Zhao, H., Nishizaki, H., Leow, C.S., Shen, X.: Chinese character recognition based on swin transformer-encoder. *Digital Signal Processing* **161**, 105080 (2025)

26. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
27. Liwicki, M., Bunke, H.: Iam-ondb-an on-line english sentence database acquired from handwritten text on a whiteboard. In: Eighth International Conference on Document Analysis and Recognition (ICDAR'05). pp. 956–961. IEEE (2005)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), <https://openreview.net/forum?id=Bkg6RiCqY7>
29. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition* **5**, 39–46 (2002)
30. Nguyen, H.T., Nguyen, C.T., Nakagawa, M.: Icfhr 2018–competition on vietnamese online handwritten text recognition using hands-vnondb (voht2018). In: 2018 16th International conference on frontiers in handwriting recognition (ICFHR). pp. 494–499. IEEE (2018)
31. Ott, F., Rügamer, D., Heublein, L., Bischl, B., Mutschler, C.: Auxiliary cross-modal representation learning with triplet loss functions for online handwriting recognition. *IEEE Access* **11**, 94148–94172 (2023). <https://doi.org/10.1109/ACCESS.2023.3310819>
32. Plamondon, R., Srihari, S.N.: Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence* **22**(1), 63–84 (2000)
33. Rahman, A., Roy, P., Pal, U.: Air writing: Recognizing multi-digit numeral string traced in air using rnn-lstm architecture. *SN Computer Science* **2**(1), 20 (2021)
34. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence* **39**(11), 2298–2304 (2016)
35. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
37. Xu, Z., Chen, Z., Wu, Y., Li, H., Lv, W., Jin, L., Wang, Q.: A multi-scale bi-modal fusion network for robust and accurate online handwriting recognition. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6460–6464. IEEE (2024)