Khmer Optical Character Recognition Using a Squeeze-and-Excitation Transformer Network

1. Dataset

Synthetic data were generated using Pillow, a collection of Khmer fonts, and a Khmer text corpus. Each generated image contains text of varying lengths and heights to simulate the diversity of real-world printed text. The applied data augmentation techniques include erosion, noise addition, text thinning and thickening, blurring, perspective distortion, erosion, and rotation.
A total of 100,000 synthetic document-text images were produced, each incorporating different augmentations and font styles.

For scene-text data, we used SynthTIGER in combination with the Stanford Background Dataset to generate realistic backgrounds. Minor augmentations such as rotation, blur, and noise were also applied, resulting in another 100,000 synthetic scene-text images.

The dataset consists of 200,000 synthetic images spanning both document-text and scene-text samples. As shown in Fig. 1, most textline contain roughly 50 characters, while the longest sequences approach 200 characters.

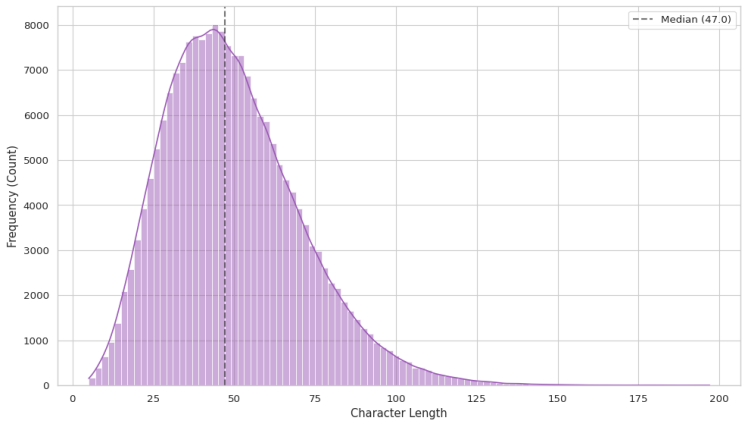


Figure 1: Character Length Distribution of Synthetic Train Dataset

1.1 Khmer OCR Benchmark Dataset (KHOB)

The KHOB dataset serves as a standardized, open-source benchmark for Khmer OCR engines. While the images possess relatively clean backgrounds, they suffer from quality degradation due to compression artifacts. Furthermore, the text lines in this dataset have significantly smaller heights compared to our training data. Consequently, resizing and chunking during inference introduce distortions that present a challenge for OCR performance.

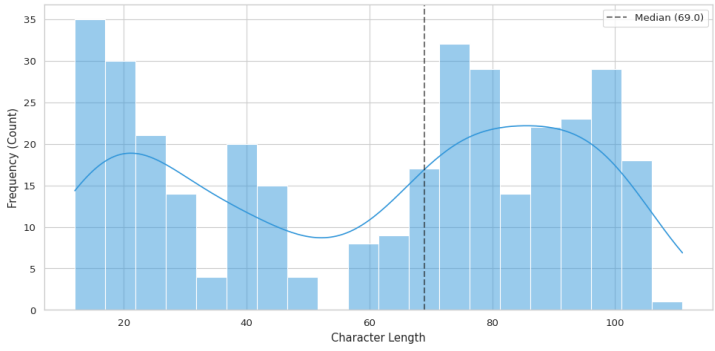


Figure 2: Character Length Distribution of KHOB Dataset

1.2 Legal Documents

This dataset consists of 227 images of legal documents, including birth certificates, identification cards, and academic diplomas. The documents contain textual elements such as dates, personal names, places of birth, and identification numbers. The images exhibit varying degrees of degradation, illumination conditions, and perspective distortion. Text line lengths vary across samples, with most text lines comprising approximately 10 characters.

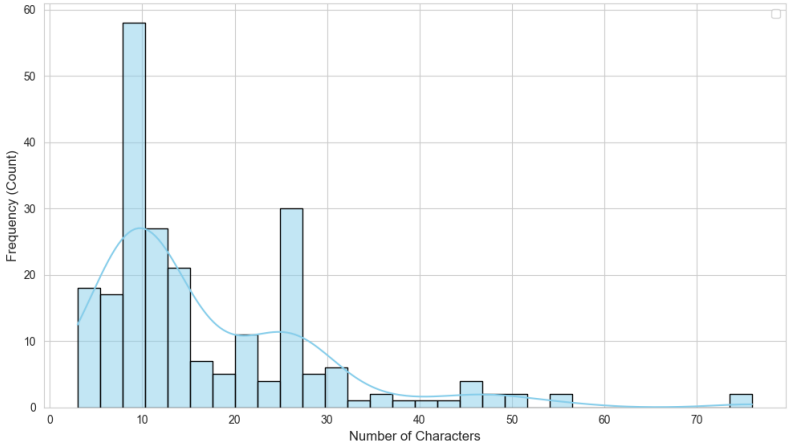


Figure 3: Character Length Distribution of Legal Documents Dataset

1.3 Khmer Printed Words

This dataset contains 136,117 images of Khmer words rendered in 10 different fonts. The images are generally clean and represent short words. This dataset will test the model's robustness against short text, since it was primarily trained on long text-line images.

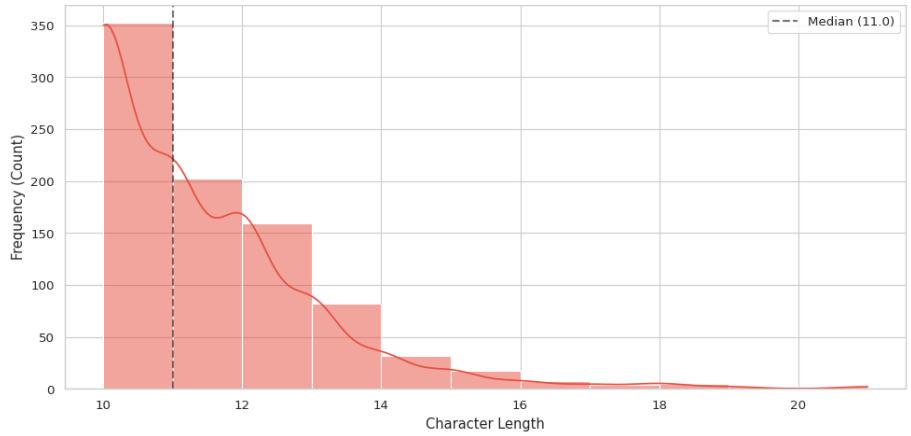


Figure 4: Character Length Distribution of Printed Words Dataset

TABLE 1: List of the used datasets, type and sizes.

Dataset	**Type**	**Training**	**Evaluation**
Doc & ST	Synthetic	200,000	---
KHOB	Real	---	325
Legal Documents	Real	---	220
Printed Word	Synthetic		1,000

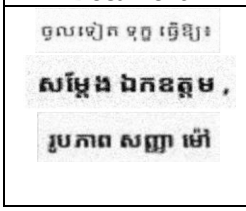
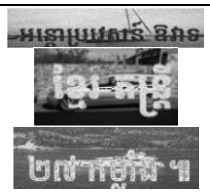
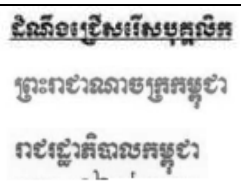
Synthetic Document	Synthetic Scene Text	KHOB	Legal Documents	Printed Word
				

Figure 5: Some preview images of the dataset

2. Methodology

2.1 Data Annotation

To further evaluate the model and assess its robustness on real-world data, the Legal Documents dataset was manually annotated. To support efficient and consistent annotation, we developed a lightweight custom annotation tool using the Tkinter framework, as illustrated in Fig. 6. The tool allows annotators to assign corresponding transcriptions, and save them to a designated folder.

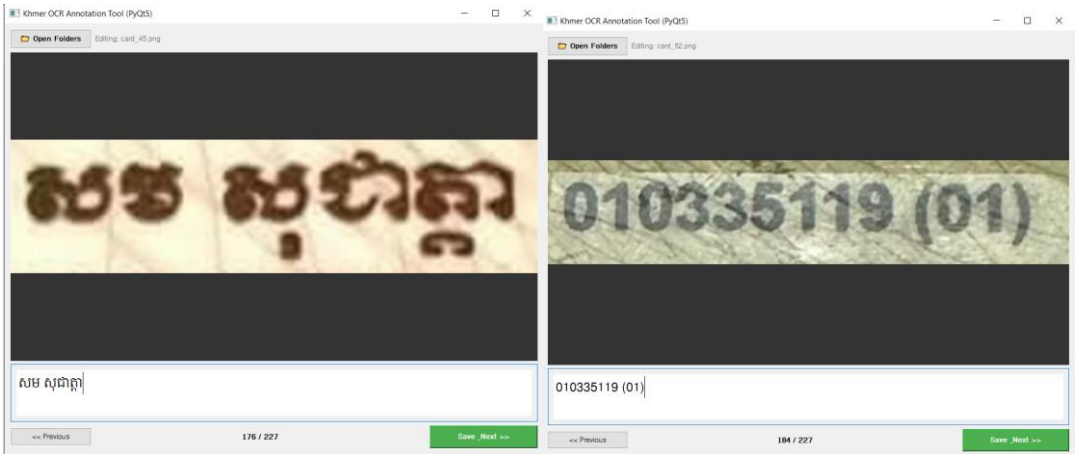


Figure 6: Annotation Tools

2.2 Image Chunking and Merging Method

To handle variable-length text lines efficiently, each image is padded at both the beginning and end so that its width becomes divisible by 100. The image is then split into overlapping chunks with an overlap margin of 16 pixels, and the height of each chunk is resized to 48 pixels. Only chunks with a width greater than or equal to (minimum chunk ratio × chunk width) are retained.

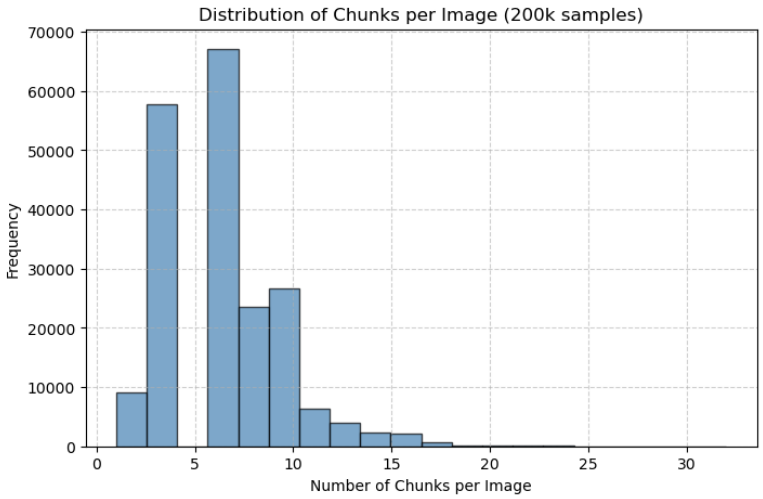


Figure 7: Chunk Distribution Per Image

2.3 Model Architecture

To address the challenges of Khmer OCR, we investigated two architectural approaches. We first established a baseline performance using a standard CNN-Transformer architecture, and subsequently introduced our proposed Squeeze-and-Excitation Transformer Network architecture, which incorporates sequence-aware attention mechanisms and recurrent smoothing to enhance recognition accuracy.

2.3.1 Baseline Architecture

To establish a performance benchmark, we implemented a standard CNN-Transformer architecture based on Buoy et al, (2023). This baseline model consists of five key modules:

- CNN Module
- Patch Module
- Transformer Encoder
- Merging Module
- Transformer Decoder

The CNN module is a modified VGG, and ResNet architecture that takes a grayscale input image of size 48×100 pixels and outputs a feature map with 512 channels, a height of 2 pixels, and a width of 32 pixels.

The resulting feature map is then processed by the Patch Encoder, which projects the spatial features into a 384-dimensional embedding space, resulting in a sequence of 32 visual token per chunk. Positional encoding is applied to preserve the spatial order of the patches, and the sequence of embeddings is then passed through the Transformer Encoder to capture contextual relationships among visual tokens.

Next, the Merging Module aggregates contextual features from all chunks of the same text line, combining their encoded representations into a unified sequence. This merged representation is then passed to the Transformer Decoder, which generates Khmer characters sequentially.

Before going through the model, the input image undergoes the chunking process described earlier. The features from multiple chunks (e.g., Chunk 1 and Chunk 2) are processed independently by the

CNN and Transformer Encoder, and then merged before decoding. By merging the chunk-level contextual features before character decoding, instead of decoding characters independently per chunk, the Transformer Decoder can leverage contextual information from all previously decoded characters across chunks, thereby improving recognition accuracy for long or complex text sequences.

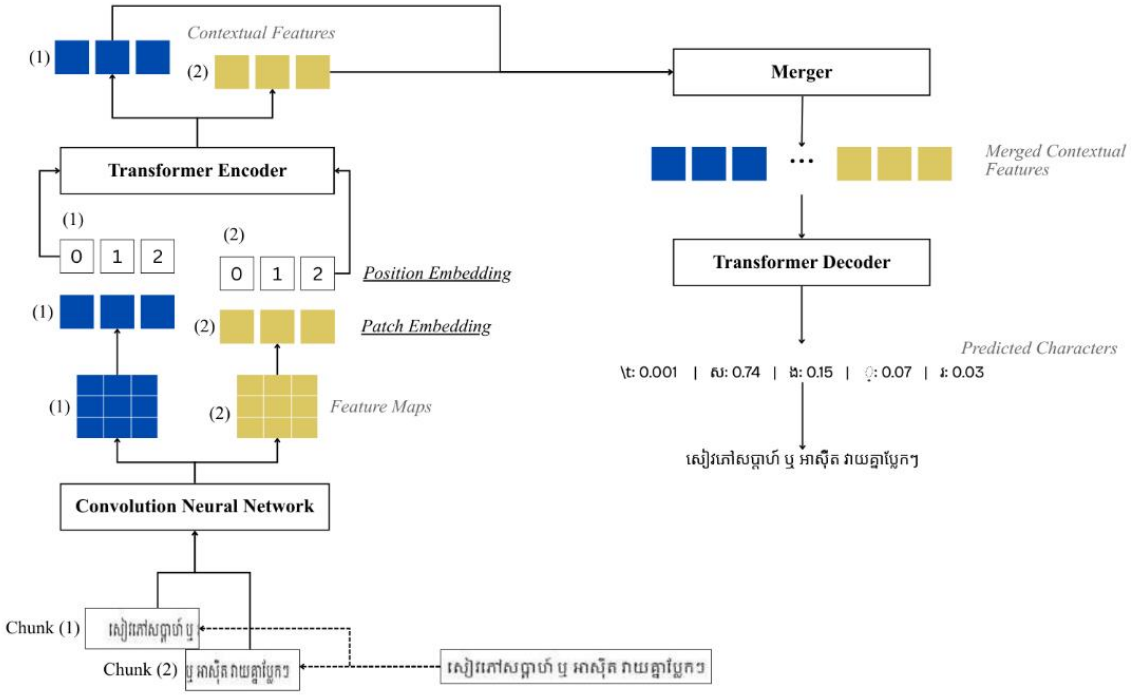


Figure 8: Baseline Transformer-based Encoder-Decoder architecture

2.3.2 The Proposed Text Recognition Architecture

To increase the accuracy of the baseline approach, we propose the Squeeze-and-Excitation Transformer Network. This architecture integrates two critical modules to improve feature representation and sequence continuity: the embedding of specialized Squeeze-and-Excitation (SE) blocks within the CNN backbone, and the insertion of a Bidirectional Long Short-Term Memory (BiLSTM) layer acting as a context smoother bridging the encoder and decoder.

The architecture is organized into six sequential modules: The Squeeze-and-Excitation Network, Patching Module, Transformer Encoder, Merging Module, BiLSTM Context Smoother, and Transformer Decoder. This configuration allows the system to leverage the parallel processing capabilities of the Transformer Encoder for individual image chunks, while subsequently utilizing the BiLSTM to effectively resolve boundary discontinuities and restore global context before the final character generation phase.

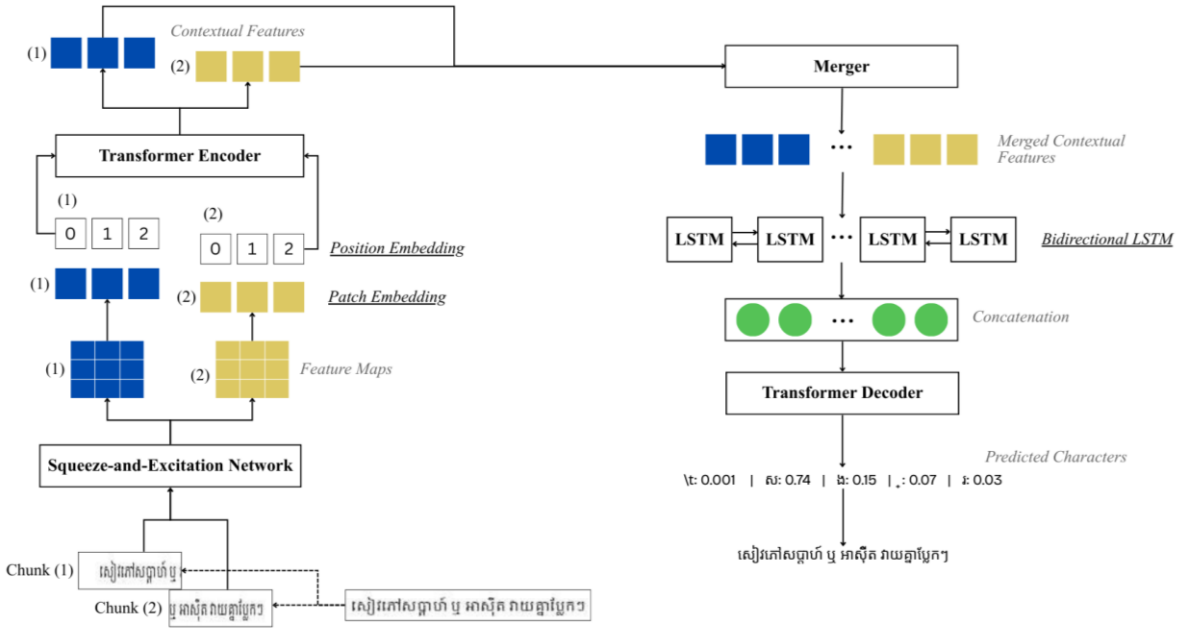


Figure 9: The Proposed Squeeze-and-Excitation Transformer-based Encoder-Decoder architecture

A. Squeeze-and-Excitation Network

The feature extraction module is built upon a modified VGG architecture tailored for text recognition. Standard Convolutional Neural Networks (CNNs) rely on local receptive fields (typically 3x3 kernels) to extract features. While effective at detecting local edges and textures, standard convolutions treat all feature channels as equally important and often lack the global context required to distinguish between complex Khmer characters that share similar sub-parts.

To address this, we integrate a 1D Squeeze-and-Excitation (SE) blocks after convolutional blocks 3, 4, and 5. This integration recalibrates the feature maps by explicitly modeling the interdependencies between channels, but with a critical modification for OCR tasks.

In a standard SE block, global average pooling is used to compress the entire spatial dimension (HxW) into a single vector. However, determining the importance of a channel based on the entire image creates a loss of positional information, which is detrimental for text recognition where the horizontal axis (W) represents the temporal sequence.

Our SE mechanism modifies the standard Squeeze-and-Excitation process to operate in tandem with the CNN by altering the pooling strategy to respect the sequential nature of text. Given an input feature map from the preceding convolutional block, the module applies average pooling strictly along the vertical height dimension. This operation generates a sequence of channel statistics ensuring the network captures the global vertical context of a character while rigorously preserving its horizontal position in the text line. This 1D feature vector is then processed through two 1D convolutional layers arranged in a bottleneck structure to learn a set of adaptive channel-wise weights. These weights function as a dynamic attention mechanism, determining which feature maps are most relevant at each specific step of the sequence, and are finally applied to the original feature map via element-wise multiplication to selectively emphasize informative features.

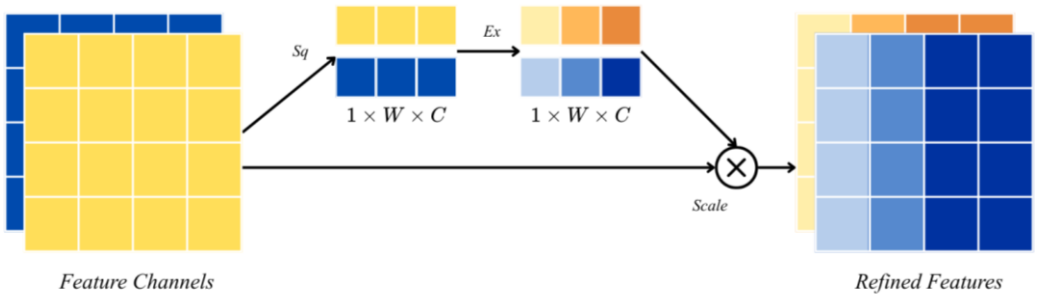


Figure 10: The proposed 1D Squeeze-and-Excitation (1D-SE) module. Unlike standard SE blocks, it utilizes vertical pooling to preserve the width dimension, maintaining spatial alignment with the text sequence

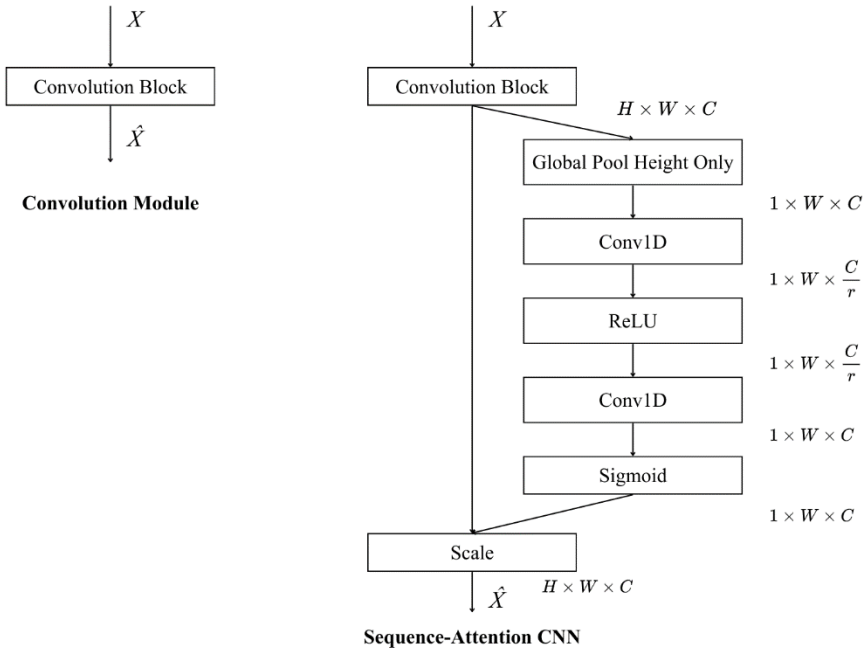


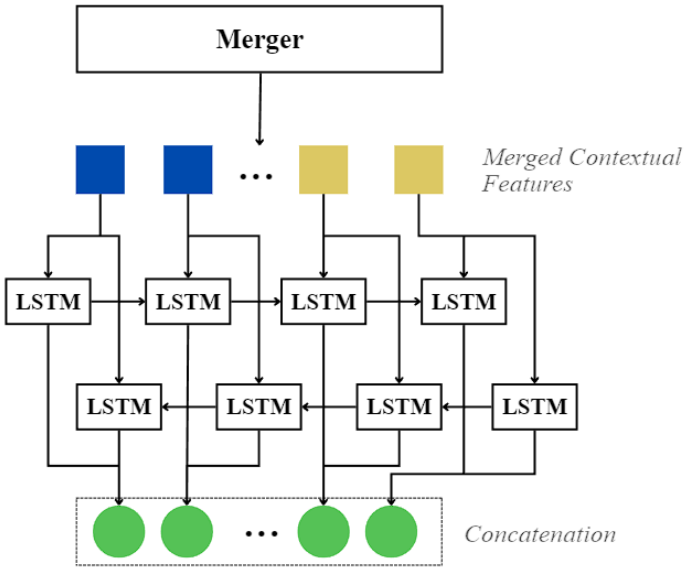
Figure 11: The Convolutional Neural Network module before (left) and after (right) adding the 1D-SE module. 1D-SE module is only added to layer 3, 4 and 5 of the CNN module because these deeper layers encode abstract, high-level semantic features where channel recalibration is most effective

B. Bidirectional Long Short-Term Memory Context Smoother

The second critical component addresses a structural limitation of the "Chunk-and-Merge" architecture. In the baseline model, input images are sliced into chunks (e.g., Chunk A and Chunk B) which are processed independently by the Transformer Encoder. Consequently, the encoder generates representations for tokens at the end of Chunk A without any knowledge of the tokens at the start of Chunk B. When these are concatenated, it results in boundary discontinuities that can confuse the decoder, leading to missing or duplicated characters at chunk seams.

To mitigate this, we insert a Bidirectional Long Short-Term Memory (BiLSTM) layer effectively acting as a "Context Smoother" between the Merging Module and the Decoder.

By fusing the forward and backward hidden states, every visual token is enriched with context from the entire text line. This effectively smooths the transitions between chunks, transforming the disjointed local features into a coherent, global sequence prior to decoding.



3. Training Configuration and Fine-tuning

The model was trained for 100 epochs using the Adam optimizer with cross-entropy loss. Rather than employing a fixed learning rate, a staged cyclic learning rate schedule was adopted. The learning rate was initially set to 1e-4 for the first 15 epochs to promote rapid convergence, followed by cyclic learning between 1e-4 and 1e-5 for the next 15 epochs to improve training stability. For the remaining epochs, the learning rate was cycled between 1e-5 and 1e-6 to enable fine-grained optimization. At each epoch, 50,000 images were randomly sampled and augmented for training.

4. Results and Analysis

To assess the proposed architecture, we conducted evaluations using a set of real-world test samples comprising low-resolution printed documents, documents under various illumination and degradation, and isolated printed words. Although our model was trained entirely on synthetic data, these tests aimed to verify its robustness in real-world scenarios. We benchmarked performance against Tesseract-OCR (an industry standard) and two deep learning baselines: VGG-Transformer and ResNet-Transformer.

The quantitative results summarized in Table 2 demonstrate that the proposed Squeeze-and-Excitation Transformer Network consistently achieves the highest accuracy across most datasets. On long text lines from the KHOB dataset, our model achieved the best performance with a Character Error Rate (CER) of 4.79%, improving upon the VGG-Transformer baseline (5.07%) by leveraging the BiLSTM Context Smoother to effectively resolve chunk boundary discontinuities. Furthermore, on the degradation-heavy Legal Documents dataset, the architecture achieved a CER of 9.13%, significantly outperforming the VGG-Transformer (10.27%) and Tesseract-OCR (24.30%), a result we attribute to the Sequence-Aware SE blocks' capacity to suppress background noise and emphasize character strokes. Although the ResNet-Transformer maintained a slight edge on short, isolated words (2.80%), our model (3.44%) still outperformed its direct VGG-based predecessor (3.61%), suggesting that our architecture provides the optimal trade-off by delivering state-of-the-art performance in continuous, complex text recognition scenarios where global context is critical.

TABLE 2: Character Error Rate (CER in %) results on the KHOB, and Printed Word

Model	**KHOB**	**Legal Documents**	**Printed Word**
Tesseract-OCR	9.36	24.30	8.02
VGG-Transformer	5.07	10.27	3.61
ResNet-Transformer	5.85	11.57	**2.80**

SeqSE-CRNN-Transformer	**4.79**	**9.13**	3.44	

TABLE 3: Failure cases on KHOB, Legal Document, and Printed Word dataset

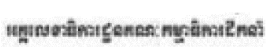
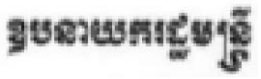
Failure Cases					
Images					
Ground-Truth	អគ្គលេខាធិការដ្ឋនគណៈកម្មាធិការជីកនាំ	និងកែសម្រួលសមាសភាពរាជរដ្ឋាភិបាលនៃព្រះរាជាណាចក្រកម្ពុជា	ឧបនាយករដ្ឋមន្ត្រី	180818125	
Proposed Model	អគ្គរលេខ:ជិក្សារដ្ឋឧករណៈកម្មាធិការដើរទាំ	និងកែសប្រួលសមាសភាពរាជរដ្ឋាភិបាលខៃព្រះរាជាណាចក្រកម្ពុជា	ឧបនាយករដ្ឋមន្ត្រី	180818125	
VGG-Tr	អគ្គរលេខ:និការដ្ឋនគណៈកម្មាឧការដើកនាំ	និងកែសម្រួលសមាសភាពរាជរដ្ឋាភិបាលខៃព្រះរាជាណាចក្រកម្ពុជា	ឧបនាយករដ្ឋមន្ត្រី	180818125	
ResNet-Tr	ក្នេរលខ:ជិការដ្ឋន#ណៈកម្មជិការជីកនាំ	និងកែសម្រួលសមាសរភាពរាជរដ្ឋាភិបាលខៃព្រះរាជាណាចក្រកម្ពុជា	ឧបនាយករដ្ឋមន្ត្រី	180818125	
Tesserract	.«ក្លរលទានិការដ្ឋទកណៈកម្មាទិការជីកទាំ	និងកែសម្រួលសមាសភាពរាជរន្ធាភិបាលនៃទ្រៈរាជាណាចក្រកម្ពុជា	ទូបនាយករដ្ឋមន្ត្រី	180818125	

TABLE 4: Example of proposed, and baseline model compared with the ground truth. Errors in the predictions are highlighted in red.

Successful Cases					
Images					
Ground-Truth	រាជរដ្ឋាភិបាលកម្ពុជា	រដ្ឋបាលរាជធានីភ្នំពេញ	ភូមិរបោះអង្កាញ់ ឃុំព្រៃកឯង	ស្រុកដង្កោ ខេត្តកណ្តាល	020728787 (01)
Proposed Model	រាជរដ្ឋាភិបាលកម្ពុជា	រដ្ឋបាលរាជធានីភ្នំពេញ	ភូមិរបោះអង្កាញ់ ឃុំព្រៃកឯង	ស្រុកដង្កោ ខេត្តកណ្តាល	020728787 (01)
VGG-Tr	រាបរដ្ឋាភិបាលកម្ពុជា	រដ្ឋបាលរាជ្ជធានីភ្នំពេញ	ភូមិរបោះអង្កាញ់ ឃុំព្រៃកឯង	ស្រុកដណ្តោ ខេត្តកណ្តាល	0207287(01)

ResNet-Tr	រាជរដ្ឋាភិភាលកម្ពុជា	រដ្ឋបាលរាជ្ជធានីភ្នំពេញ	ភូមិរបោះអង្គាញ់ ឃុំព្រែកឯង	ស្រុកដណ្តោ ខេត្តកណ្តាល	0207287 (01)
Tesserract	រាជរដ្ឋាគិធាលកម្ពុជា	រដ្ឋចាលរាជ្ជធានីភ្នំពេញ	ភូមិរបោះអង្គាញ់ ឃុំព្រែកឯង	ម	020728787 (01)

Successful Cases					
Images					
Ground-Truth	រាជរដ្ឋាភិបាលកម្ពុជា	រដ្ឋបាលរាជធានីភ្នំពេញ	ភូមិរបោះអង្គាញ់ ឃុំព្រែកឯង	ស្រុកដង្កោ ខេត្តកណ្តាល	020728787 (01)
Proposed Model	រាជរដ្ឋាភិបាលកម្ពុជា	រដ្ឋបាលរាជធានីភ្នំពេញ	ភូមិរបោះអង្គាញ់ ឃុំព្រែកឯង	ស្រុកដង្កោ ខេត្តកណ្តាល	020728787 (01)
VGG-Tr	រាបរដ្ឋាភិបាលកម្ពុជា	រដ្ឋបាលរាជ្ជធានីភ្នំពេញ	ភូមិរបោះអង្គាញ់ ឃុំព្រែកឯង	ស្រុកដណ្តោ ខេត្តកណ្តាល	0207287(01)
ResNet-Tr	រាជរដ្ឋាភិភាលកម្ពុជា	រដ្ឋបាលរាជ្ជធានីភ្នំពេញ	ភូមិរបោះអង្គាញ់ ឃុំព្រែកឯង	ស្រុកដណ្តោ ខេត្តកណ្តាល	0207287(01)
Tesseract	រាជរដ្ឋាគិធាលកម្ពុជា	រដ្ឋចាលរាជ្ជធានីភ្នំពេញ	ភូមិរបោះអង្គាញ់ ឃុំព្រែកឯង	ម	020728787 (01)

References

1. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**
   Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al.
   ICLR 2021.
   arXiv:2010.11929

2. **TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models**
   Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, Furu Wei.
   AAAI 2023.
   arXiv:2109.10282

3. **Toward a Low-Resource Non-Latin-Complete Baseline: An Exploration of Khmer Optical Character Recognition**
   R. Buoy, M. Iwamura, S. Srun and K. Kise.
   IEEE Access, vol. 11, pp. 128044-128060, 2023.
   DOI: 10.1109/ACCESS.2023.3332361

4. **Balraj98.** (2018). *Stanford background dataset* [Data set].
   Kaggle. https://www.kaggle.com/datasets/balraj98/stanford-background-dataset

5. **EKYC Solutions.** (2022). *Khmer OCR benchmark dataset (KHOB)* [Data set].
   GitHub. https://github.com/EKYCSolutions/khmer-ocr-benchmark-dataset

6. **Em, H., Valy, D., Gosselin, B., & Kong, P.** (2024). *Khmer text recognition dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/emhengly/khmer-text-recognition-dataset

7. **Squeeze-and-Excitation Networks**
   Jie Hu, Li Shen, and Gang Sun.
   CVPR 2018.
   arXiv:1709.01507

8. **Bidirectional Recurrent Neural Networks**
   Mike Schuster and Kuldip K. Paliwal.
   IEEE Transactions on Signal Processing, 1997.
   DOI: 10.1109/78.650093