Project

End-to-End Khmer OCR with Font Awareness

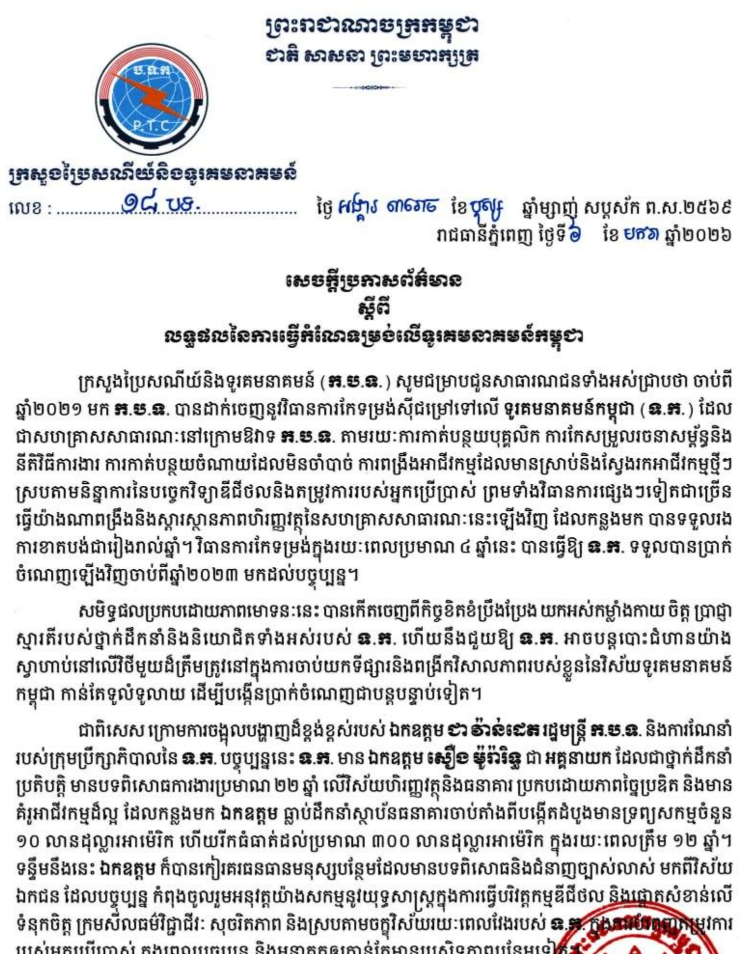
Abstract

Converting Khmer documents into editable formats like Markdown or HTML is challenged by the frequent use of multiple font styles within a single line—specifically Moul (for titles and honorifics) and Regular scripts. Current OCR solutions often lose this semantic formatting during extraction. We propose a unified Khmer OCR architecture that performs simultaneous character recognition and font classification using a shared CNN backbone. By branching a font classification head from the primary feature extractor, the model identifies styles (Regular, Moul, Bold, etc.) without additional computational overhead. This approach enables the generation of rich-text documents that preserve the visual hierarchy and semantic intent of the original Khmer source.

Introduction

The digital transformation of Khmer printed documents into structured formats (PDF to Docx/Markdown) requires more than simple character accuracy. A defining characteristic of Khmer typography is the functional use of distinct font styles. For instance, **Moul** fonts are traditionally reserved for headings, names of high-ranking officials, and sacred texts, while Regular fonts are used for body content.

Standard OCR pipelines typically treat font variation as "noise" to be normalized, resulting in a loss of the document's original structure and emphasis. To bridge this gap, we introduce a multi-task learning architecture that utilizes a shared CNN backbone for both sequence recognition and font identification.



ព្រះរាជាណាចក្រកម្ពុជា
ជាតិ សាសនា ព្រះមហាក្សត្រ

ក្រសួងប្រៃសណីយ៍និងទូរគមនាគមន៍

លេខ : ១៨ បទ. ថ្ងៃ អង្គារ ៣រោច ខែបុស្ស ឆ្នាំម្សាញ់ សប្តស័ក ព.ស.២៥៦៩
រាជធានីភ្នំពេញ ថ្ងៃទី៦ ខែ មករា ឆ្នាំ២០២៦

សេចក្តីប្រកាសព័ត៌មាន
ស្តីពី
លទ្ធផលនៃការធ្វើកំណែទម្រង់លើទូរគមនាគមន៍កម្ពុជា

ក្រសួងប្រៃសណីយ៍និងទូរគមនាគមន៍ (**ក.ប.ទ.**) សូមជម្រាបជូនសាធារណជនទាំងអស់ជ្រាបថា ចាប់ពីឆ្នាំ២០២១ មក **ក.ប.ទ.** បានដាក់ចេញនូវវិធានការកែទម្រង់ស៊ីជម្រៅទៅលើ **ទូរគមនាគមន៍កម្ពុជា** (**ទ.ក.**) ដែលជាសហគ្រាសសាធារណៈនៅក្រោមឱវាទ **ក.ប.ទ.** តាមរយៈការកាត់បន្ថយបុគ្គលិក ការកែសម្រួលរចនាសម្ព័ន្ធនិងនីតិវិធីការងារ ការកាត់បន្ថយចំណាយដែលមិនចាំបាច់ ការពង្រឹងអាជីវកម្មដែលមានស្រាប់និងស្វែងរកអាជីវកម្មថ្មីៗស្របតាមនិន្នាការនៃបច្ចេកវិទ្យាឌីជីថលនិងតម្រូវការរបស់អ្នកប្រើប្រាស់ ព្រមទាំងវិធានការផ្សេងៗទៀតជាច្រើន ធ្វើយ៉ាងណាពង្រឹងនិងស្តារស្ថានភាពហិរញ្ញវត្ថុនៃសហគ្រាសសាធារណៈនេះឡើងវិញ ដែលកន្លងមក បានទទួលរងការខាតបង់ជារៀងរាល់ឆ្នាំ។ វិធានការកែទម្រង់ក្នុងរយៈពេលប្រមាណ ៤ ឆ្នាំនេះ បានធ្វើឱ្យ **ទ.ក.** ទទួលបានប្រាក់ចំណេញឡើងវិញចាប់ពីឆ្នាំ២០២៣ មកដល់បច្ចុប្បន្ន។

សមិទ្ធផលប្រកបដោយភាពមោទនៈនេះ បានកើតចេញពីកិច្ចខិតខំប្រឹងប្រែង យកអស់កម្លាំងកាយ ចិត្ត ប្រាជ្ញាស្មារតីរបស់ថ្នាក់ដឹកនាំនិងនិយោជិតទាំងអស់របស់ **ទ.ក.** ហើយនឹងជួយឱ្យ **ទ.ក.** អាចបន្តបោះជំហានយ៉ាងស្វាហាប់នៅលើវិថីមួយដ៏ត្រឹមត្រូវនៅក្នុងការចាប់យកទីផ្សារនិងពង្រីកវិសាលភាពរបស់ខ្លួននៃវិស័យទូរគមនាគមន៍កម្ពុជា កាន់តែទូលំទូលាយ ដើម្បីបង្កើនប្រាក់ចំណេញជាបន្តបន្ទាប់ទៀត។

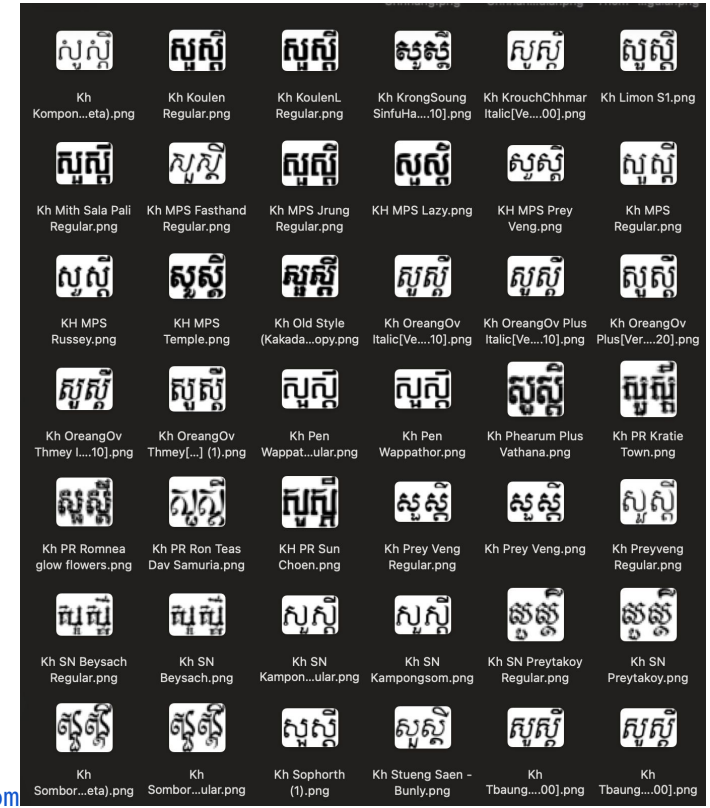
ជាពិសេស ក្រោមការចង្អុលបង្ហាញដ៏ខ្ពង់ខ្ពស់របស់ ឯកឧត្តម **ជា វ៉ាន់ដេត** រដ្ឋមន្ត្រី **ក.ប.ទ.** និងការណែនាំរបស់ក្រុមប្រឹក្សាភិបាលនៃ **ទ.ក.** បច្ចុប្បន្ននេះ **ទ.ក.** មាន ឯកឧត្តម **សឿន ម៉ូរ៉ាវិទ្ធ** ជា អគ្គនាយក ដែលជាថ្នាក់ដឹកនាំប្រតិបត្តិ មានបទពិសោធការងារប្រមាណ ២២ ឆ្នាំ លើវិស័យហិរញ្ញវត្ថុនិងធនាគារ ប្រកបដោយភាពច្នៃប្រឌិត និងមានគំរូអាជីវកម្មដ៏ល្អ ដែលកន្លងមក ឯកឧត្តម ធ្លាប់ដឹកនាំស្ថាប័នធនាគារចាប់តាំងពីបង្កើតដំបូងមានទ្រព្យសកម្មចំនួន ១០ លានដុល្លារអាមេរិក ហើយរីកធំធាត់ដល់ប្រមាណ ៣០០ លានដុល្លារអាមេរិក ក្នុងរយៈពេលត្រឹម ១២ ឆ្នាំ។ ទន្ទឹមនឹងនេះ ឯកឧត្តម ក៏បានកៀរគរធនធានមនុស្សបន្ថែមដែលមានបទពិសោធនិងជំនាញច្បាស់លាស់ មកពីវិស័យឯកជន ដែលបច្ចុប្បន្ន កំពុងចូលរួមអនុវត្តយ៉ាងសកម្មនូវយុទ្ធសាស្ត្រក្នុងការធ្វើបរិវត្តកម្មឌីជីថល និងផ្តោតសំខាន់លើទំនុកចិត្ត ក្រមសីលធម៌វិជ្ជាជីវៈ សុចរិតភាព និងស្របតាមចក្ខុវិស័យរយៈពេលវែងរបស់ **ទ.ក.** ក្នុងការបំពេញតម្រូវការ

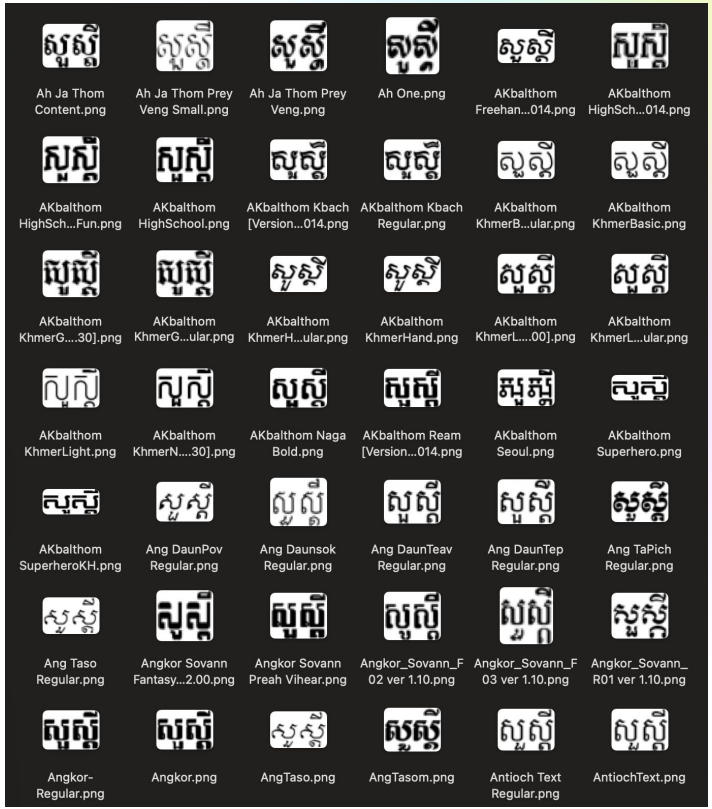
ព្រះរាជាណាចក្រកម្ពុជា
ជាតិ សាសនា ព្រះមហាក្សត្រ
15.95
ក្រសួងប្រៃសណីយ៍និងទូរគមនាគមន៍
១៨ us.
លេខ :
ថ្ងៃ អង្គារ 6 ខែ ហ្គុស្ស ឆ្នាំម្សាញ់ សប្តស័ក ព.ស.២៥៦៩
រាជធានីភ្នំពេញ ថ្ងៃទី ៦ ខែ មករា ឆ្នាំ២០២៦
សេចក្តីប្រកាសព័ត៌មាន
ស្តីពី
លទ្ធផលនៃការធ្វើកំណែទម្រង់លើទូរគមនាគមន៍កម្ពុជា
ក្រសួងប្រៃសណីយ៍និងទូរគមនាគមន៍ (ក.ប.ទ.)
សូមជម្រាបជូនសាធារណជនទាំងអស់ជ្រាបថា ចាប់ពី
ឆ្នាំ២០២១ មក ក.ប.ទ. បានដាក់ចេញនូវវិធានការកែទម្រង់ស៊ីជម្រៅទៅលើ
ទូរគមនាគមន៍កម្ពុជា (ទ.ក.) ដែល
ជាសហគ្រាសសាធារណៈនៅក្រោមឱវាទ ក.ប.ទ. តាមរយៈការកាត់បន្ថយបុគ្គលិក
ការកែសម្រួលរចនាសម្ព័ន្ធនិង
នីតិវិធីការងារ ការកាត់បន្ថយចំណាយដែលមិនចាំបាច់
ការពង្រឹងអាជីវកម្មដែលមានស្រាប់និងស្វែងរកអាជីវកម្មថ្មីៗ
ស្របតាមនិន្នាការនៃបច្ចេកវិទ្យាឌីជីថលនិងតម្រូវការរបស់អ្នកប្រើប្រាស់
ព្រមទាំងវិធានការផ្សេងៗទៀតជាច្រើន
ធ្វើយ៉ាងណាព
ង្រឹងនិងស្តារស្ថានភាពហិរញ្ញវត្ថុនៃសហគ្រាសសាធារណៈនេះឡើងវិញ ដែលកន្លងម
ការខាតបង់

Dataset

To support the development of a robust Khmer Optical Character Recognition (OCR) system with Automatic Font Recognition (AFR), a large-scale synthetic text-line dataset was constructed. Due to the limited availability of publicly annotated Khmer OCR corpora, the dataset was generated using an automated rendering pipeline designed to capture both linguistic and typographic diversity inherent in the Khmer script. The total images are 2M+ sentences in Khmer lanugage.

Dataset


Kh Kompon...eta).png
Kh Koulen Regular.png
Kh KoulenL Regular.png
Kh KrongSoung SinfuHa....10].png
Kh KrouchChhmar Italic[Ve....00].png
Kh Limon S1.png
Kh Mith Sala Pali Regular.png
Kh MPS Fasthand Regular.png
Kh MPS Jrung Regular.png
KH MPS Lazy.png
KH MPS Prey Veng.png
Kh MPS Regular.png
KH MPS Russey.png
KH MPS Temple.png
Kh Old Style (Kakada...opy.png
Kh OreangOv Italic[Ve....10].png
Kh OreangOv Plus Italic[Ve....10].png
Kh OreangOv Plus[Ver....20].png
Kh OreangOv Thmey I....10].png
Kh OreangOv Thmey[...] (1).png
Kh Pen Wappat...ular.png
Kh Pen Wappathor.png
Kh Phearum Plus Vathana.png
Kh PR Kratie Town.png
Kh PR Romnea glow flowers.png
Kh PR Ron Teas Dav Samuria.png
KH PR Sun Choen.png
Kh Prey Veng Regular.png
Kh Prey Veng.png
Kh Preyveng Regular.png
Kh SN Beysach Regular.png
Kh SN Beysach.png
Kh SN Kampon...ular.png
Kh SN Kampongsom.png
Kh SN Preytakoy Regular.png
Kh SN Preytakoy.png
Kh Sombor...eta).png
Kh Sombor...ular.png
Kh Sophorth (1).png
Kh Stueng Saen - Bunly.png
Kh Tbaung....00].png
Kh Tbaung....00].png


Ah Ja Thom Content.png
Ah Ja Thom Prey Veng Small.png
Ah Ja Thom Prey Veng.png
Ah One.png
AKbalthom Freehan...014.png
AKbalthom HighSch...014.png
AKbalthom HighSch...Fun.png
AKbalthom HighSchool.png
AKbalthom Kbach [Version...014.png
AKbalthom Kbach Regular.png
AKbalthom KhmerB...ular.png
AKbalthom KhmerBasic.png
AKbalthom KhmerG....30].png
AKbalthom KhmerG...ular.png
AKbalthom KhmerH...ular.png
AKbalthom KhmerHand.png
AKbalthom KhmerL....00].png
AKbalthom KhmerL...ular.png
AKbalthom KhmerLight.png
AKbalthom KhmerN....30].png
AKbalthom Naga Bold.png
AKbalthom Ream [Version...014.png
AKbalthom Seoul.png
AKbalthom Superhero.png
AKbalthom SuperheroKH.png
Ang DaunPov Regular.png
Ang Daunsok Regular.png
Ang DaunTeav Regular.png
Ang DaunTep Regular.png
Ang TaPich Regular.png
Ang Taso Regular.png
Angkor Sovann Fantasy...2.00.png
Angkor Sovann Preah Vihear.png
Angkor_Sovann_F 02 ver 1.10.png
Angkor_Sovann_F 03 ver 1.10.png
Angkor_Sovann_ R01 ver 1.10.png
Angkor-Regular.png
Angkor.png
AngTaso.png
AngTasom.png
Antioch Text Regular.png
AntiochText.png

Architecture & Configuration

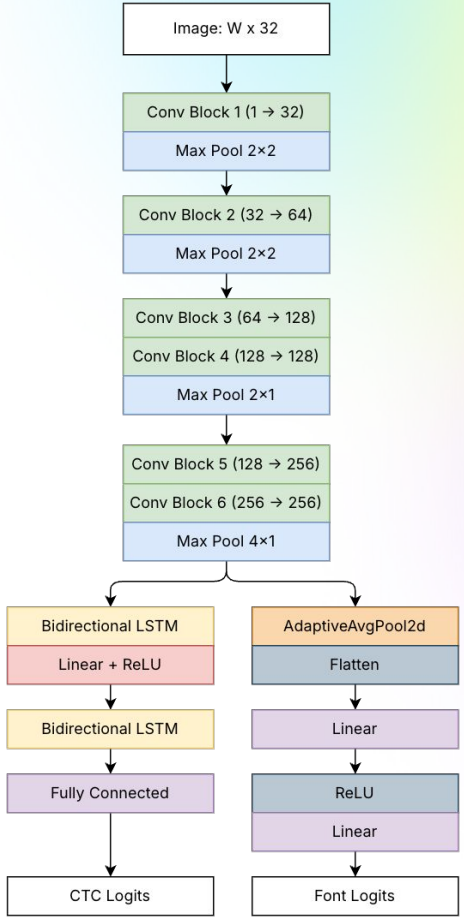
The backbone consists of the 6 convolution blocks followed two Bidirectional LSTMs for OCR task and AdaptiveAvgPool2d followed by two linear layers and ReLU for fonts classification head.

Optimizer: **AdamW**

Learning Rate: **0.0001**

OCR Loss Function: **CTCLoss**

Font Classify Loss Function: **Cross Entropy Loss**



Data Processing / Tokenizer

Category	Tokens
Special	[PAD], [SOS], [EOS], [UNK]
Consonants	ក, ខ, គ, ឃ, ង, ច, ឆ, ជ, ឈ, ញ, ដ, ឋ, ឌ, ឍ, ណ, ត, ថ, ទ, ធ, ន, ប, ផ, ព, ភ, ម, យ, រ, ល, វ, ឝ, ឞ, ស, ហ, ឡ
Independent Vowels	អ, ឣ, ឤ, ឥ, ឦ, ឧ, ឩ, ឪ, ឫ, ឬ, ឭ, ឮ, ឯ, ឰ, ឱ, ឲ, ឳ

Dependent Vowels	◌ា, ◌ិ, ◌ី, ◌ឹ, ◌ឺ, ◌ុ, ◌ូ, ◌ួ, ◌ើ, ◌ឿ, ◌ៀ, ◌េ, ◌ែ, ◌ៃ, ◌ោ, ◌ៅ
Diacritics & Marks	◌ំ, ◌ះ, ◌ៈ, ◌៉, ◌៊, ◌់, ◌៌, ◌៍, ◌៎, ◌៏, ◌័, ◌៑, ◌្
Punctuation	។, ៕, ៖, ៗ, ។ល។, ៎, ◌៑
Numerals	០, ១, ២, ៣, ៤, ៥, ៦, ៧, ៨, ៩

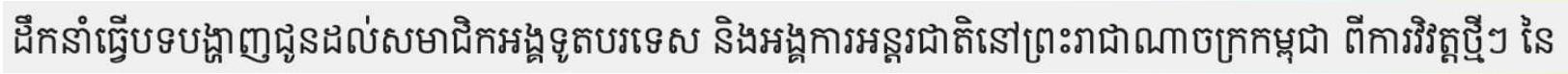
OCR Evaluation

The model achieved **CER** of **0.0051** on the self-split validation set with the validation loss of **0.0167.**

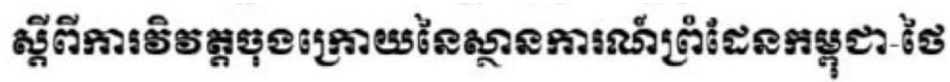
Font Classification Evaluation

Font	Precision	Recall	F1-Score
Regular	0.8112	0.8690	0.8391
Italic	0.7691	0.8136	0.7908
Bold	0.8517	0.7933	0.8215
BoldItalic	0.7911	0.7623	0.7764
Moul	0.8182	0.9020	0.8580
MoulLight	0.9034	0.5325	0.6701

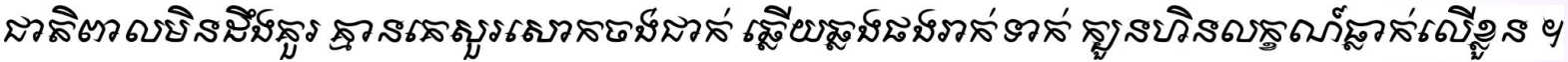
Result



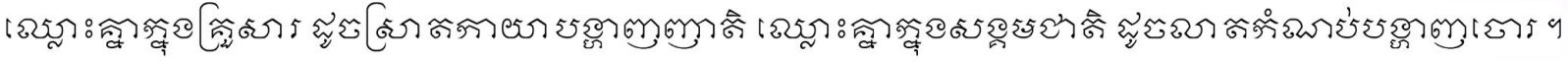
Predicted Text: **ដឹកនាំធ្វើបទបង្ហាញជូនដល់សមាជិកអង្គទូតបរទេសនិងអង្គការអន្តរជាតិនៅព្រះរាជាណាចក្រកម្ពុជាពីការវិវត្តថ្មីៗនៃ**
Font: **Regular (93%)**



Predicted Text: **ស្តីពីការវិវត្តចុងក្រោយនៃស្ថានការណ៍ព្រំដែនកម្ពុជាថៃ**
Font: **Moul (97%)**



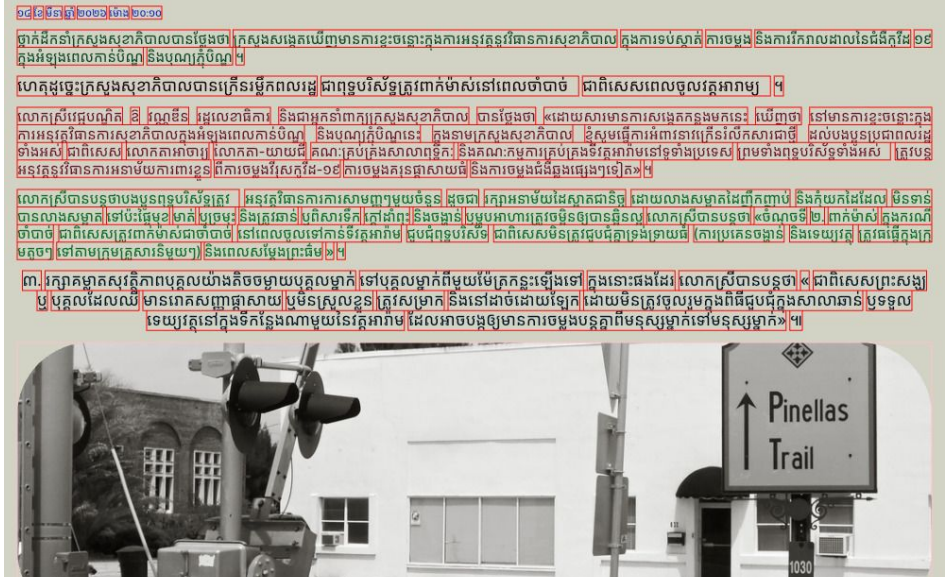
Predicted Text: **ជាតិពាលមិនដឹងគួរគ្មានគេសួរសោកចង់ជាក់ឆ្លើយឆ្លងផងរាក់ទាក់ក្បួនហិនលក្ខណ៍ធ្លាក់លើខ្លួន។**
Font: **Italic (69%)**



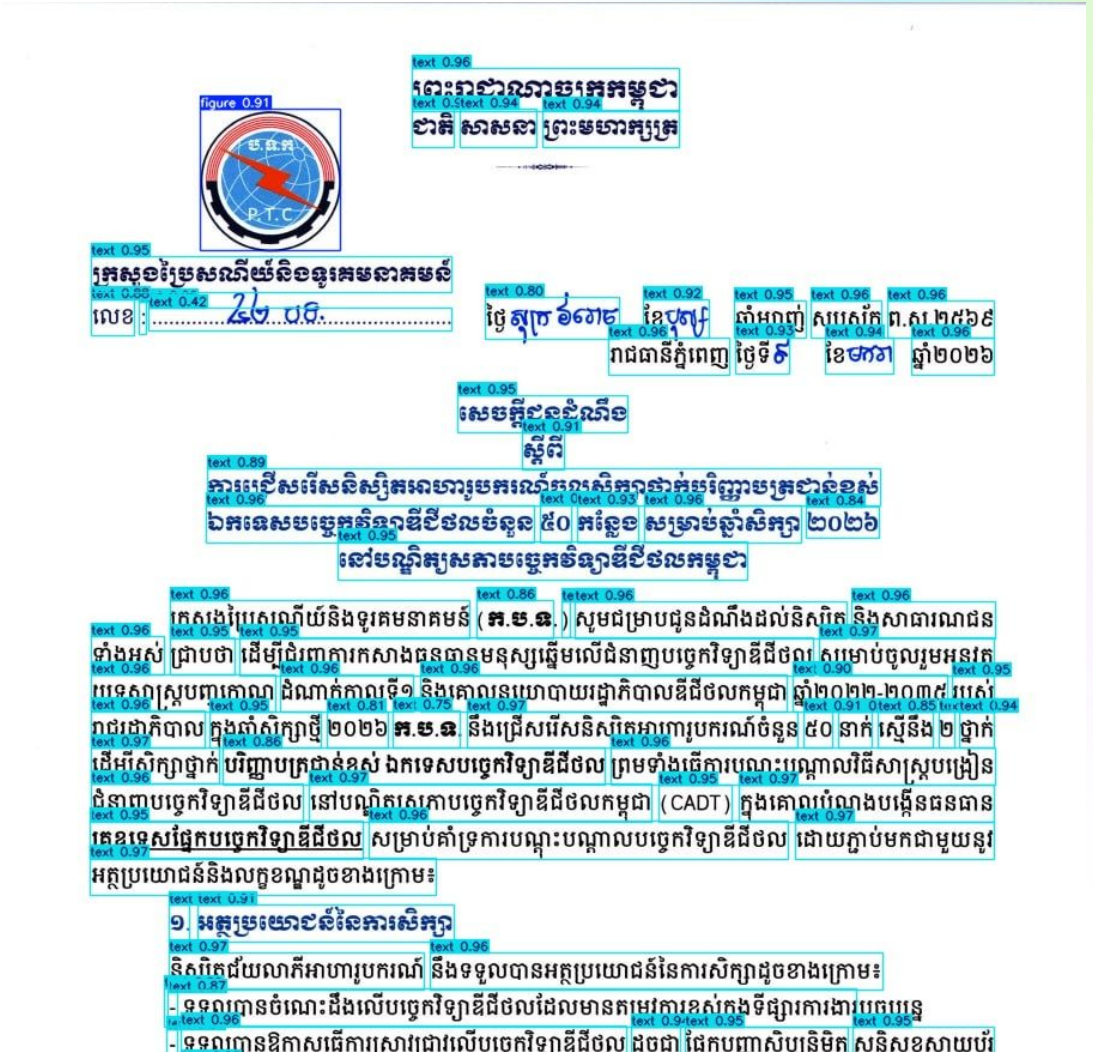
Predicted Text: **ឈ្លោះគ្នាក្នុងគ្រួសារដូចស្រោតកាយាបង្ហាញញាតិឈ្លោះគ្នាក្នុងសង្គមជាតិដូចលាតកំណប់បង្ហាញចោរ។**
Font: **MoulLight (73%)**

Text & Image Detection

We trained a YOLOv11 based text and image detection with the 5000+ Synthetic Documents generated by **Sone.js** **(https://github.com/seanghay/sone)**



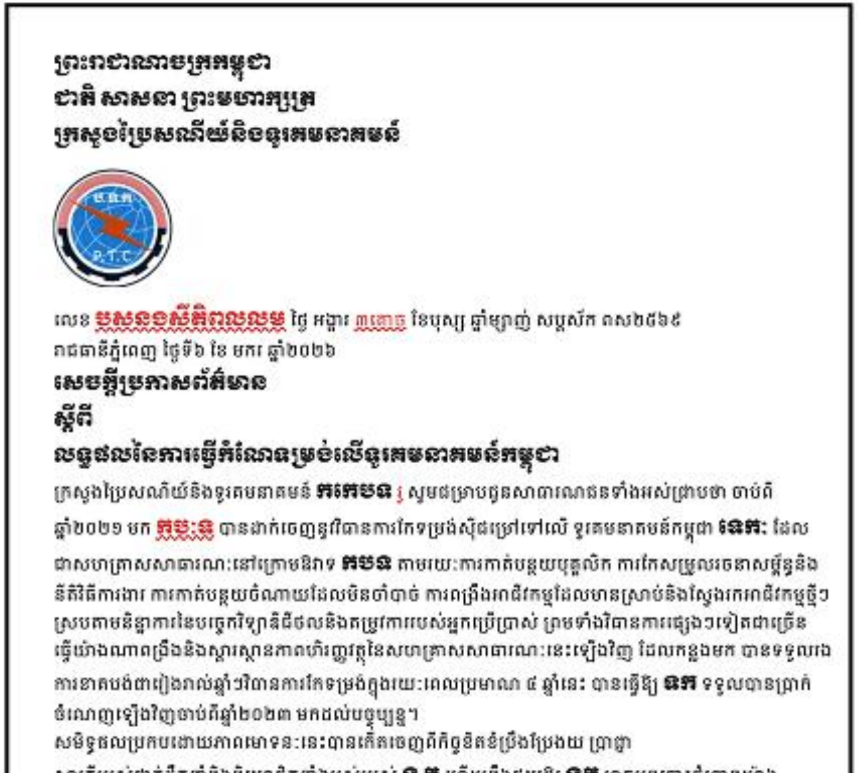
Detection Result



Full OCR

Image

Prediction



Full OCR

Image

Prediction

រាជធានីភ្នំពេញ ៖ ដើមឈើមានអាយុកាលជាច្រើនឆ្នាំមួយដើម អត់ស្គាល់ថា ជាដើមអ្វីនោះទេ ត្រូវបានជនមិនស្គាល់
អត្តសញ្ញាណលួចកាប់ចោល បង្កការភ្ញាក់ផ្អើលដល់អាជ្ញាធរចុះទៅពិនិត្យ និងឃាត់មនុស្ស ៣នាក់យកទៅសួរនាំ
កាលពីយប់ថ្ងៃទី ៨ ខែកុម្ភៈ ២០២៦ នៅចំណុចផ្លូវ ៥៧ កែងផ្លូវ ៣៣៨ ក្នុងសង្កាត់បឹងកេងកងទី ១ ខណ្ឌបឹងកេងកង រាជធានីភ្នំពេញ ។
តាមប្រភពបានអោយដឹងថា ដើមឈើមួយដើមនេះ មានអាយុកាលជាច្រើនឆ្នាំមកហើយ ជាសម្បត្តិរបស់រដ្ឋផងដែរ លុះនៅយប់ថ្ងៃកើតហេតុ ជនមិនស្គាល់អត្តសញ្ញាណបានធ្វើសកម្មភាពកាប់ដើមឈើនោះ មិនដឹងថាបំរុងយកទៅណានោះទេ រហូតដល់មានអ្នករាយការណ៍ជូនអាជ្ញាធរមូលដ្ឋានចុះទៅពិនិត្យ និងឃាត់មនុស្ស៣នាក់យកទៅសួរនាំ ។
ជុំវិញករណីខាងលើនេះ ខាងមន្ត្រីពាក់ព័ន្ធ មិនទាន់បានអោយដឹងថា អ្នកកាប់ដើមឈើរបស់រដ្ឋទាំងនេះ យកទៅធ្វើអ្វី
អោយពិតប្រាកដនៅឡើយទេ ។