Vietnamese Scene Text Detection and Recognition

Instuctor: Mai Tien Dung

TEAM MEMBERS

Bui viET DAT

- 20521162

Bui QUOC THINH

- 20520934

TABLE OF CONTENT

INTRODUCTION

More Details

APPLICATIONS

Scene Text Detection and Recognition

Information retrieval and automatic data entry in internet banking

License Plate Recognition at condos, parking lots, shopping malls

APPLICATIONS

Scene Text Detection and Recognition

Self-driving car to recognize traffic signs, addresses

Assisting visually impaired individuals

INPUT & OUTPUT

More Details

DETECTION

INPUT

- An Image
- Dataset that contains bounding boxes

OUTPUT

- The bounding boxes of the input image

RECOGNITION

INPUT

- An Image
- Dataset that contains labels

OUTPUT

- The label of the input image

APPROACH

More Details

PIPELINE

More Details

Pipeline

Detection

Perspective Transformation

Recognition

DATASET

More Details

Vintext

TRAIN DATA

- 1200 images

VALIDATION DATA

- 300 images

TEST DATA

- 500 images

FORMAT DATA

- x1, y1, x2, y2, x3, y3, x4, y4, TRANSCRIPT

VN_DICTIONARY.TXT

```
,113,494,187,343,160,PHÒNG
1,141,641,213,510,190,KHÁM
2,200,348,342,272,331,BS.
3,195,480,356,358,342,ĐOÀN
2,210,578,368,508,359,VĂN
7,230,703,384,593,370,HÙNG
8,153,687,162,671,159,NGỌC
0,156,700,164,687,161,PHU
1,150,701,155,671,148,###
2,165,701,171,671,165,###
```

DETECTION

SAST

A Single-Shot Arbitrarily-Shaped Text Detector based on Context Attended Multi-Task Learning

- **P Wang** et al.
- Proceedings of the 27th **ACM international** conference, 2019
- Effective in detecting **arbitrarily-shaped** text
- Robust in generalizing to **multilingual** scene text datasets

TEXT DETECTION RESULTS

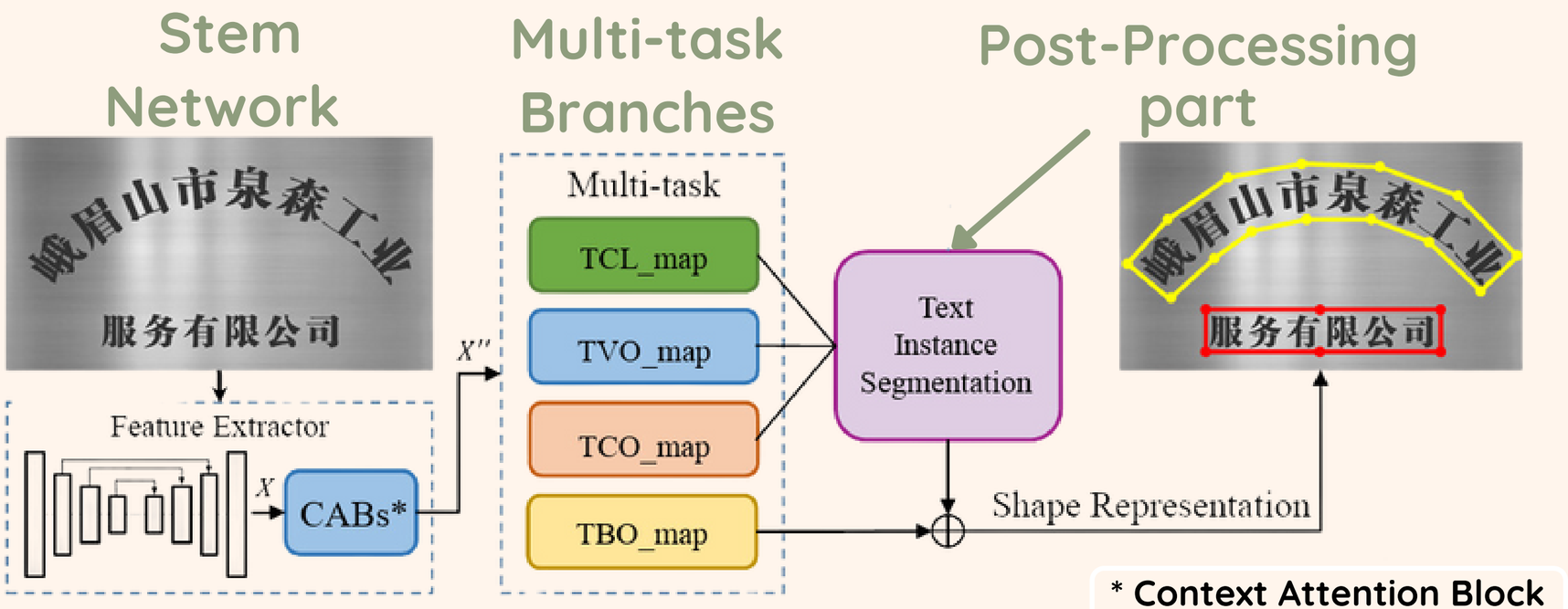
On the ICDAR2015 dataset

Model	Backbone	Precision	Recall	Hmean
EAST	ResNet50_vd	88.71%	81.36%	84.88%
EAST	MobileNetV3	78.20%	79.10%	78.65%
DB	ResNet50_vd	86.41%	78.72%	82.38%
DB	MobileNetV3	77.29%	73.08%	75.12%
SAST	ResNet50_vd	91.39%	83.77%	87.42%
PSE	ResNet50_vd	85.81%	79.53%	82.55%
PSE	MobileNetV3	82.20%	70.48%	75.89%
DB++	ResNet50	90.89%	82.66%	86.58%

- ICDAR2015: text appears in the scene without the user's prior action to change its appearance or positioning/quality
- The dataset covers a wide range of perspectives, e.g. wearable cameras

→ SAST achieves state-of-the-art results on the ICDAR2015 dataset

Pipeline


Stem Network
Multi-task Branches
Post-Processing part
峨眉山市泉森工业
服务有限公司
Feature Extractor
X
CABs*
X''
Multi-task
TCL_map
TVO_map
TCO_map
TBO_map
Text Instance Segmentation
Shape Representation
峨眉山市泉森工业
服务有限公司
* Context Attention Block

STEM NETWORK

- Network Backbone: **ResNet-50** with FPN
  - Pre-trained weight: **ImageNet**
- Serially stack two CABs behind

→ Capture rich contextual information

MULTI-BRANCHES

Predict four maps for each text region:

- **TCL (Text center line)**:
  - A segmentation map to **distinguish text** from non-text areas

- **TCO (Text center offset)**:
  - **Predicts the offset of the text region's center point from a reference location within the bounding box**

MULTI-BRANCHES

Predict four maps for each text region:

- **TVO (Text vertex offset):**
  - Predicts the **offset** of **each corner** of the text region (top-left, top-right, etc.) from a reference point within the bounding box
- **TBO (Text border offset):**
  - Predict the **offset** of the **text region's borders** from a reference point

Post-Processing

Point-to-quad assignment

1. For each pixel in the **TCL** map, the corresponding offset vector from the **TCO** map is used to identify a "**low-level center**" that the pixel likely belongs to

2. The **TVO** map regresses the four vertices of bounding quadrangle of text region directly (**high-level object knowledge**)

3. Combining **low-level center** and the **high-level object knowledge,** We can **group** each pixel in **TCL map** into different text instances

Post-Processing

Reconstruction

1. **Sample** an **adaptive number** of **points** in the **center line** of each text instance, calculate corresponding points in **upper** and **lower borders** with the help of **TBO map**

⟶ **Reconstruct the representation of arbitrarily-shaped scene text**

TRAIN

TRAIN

train

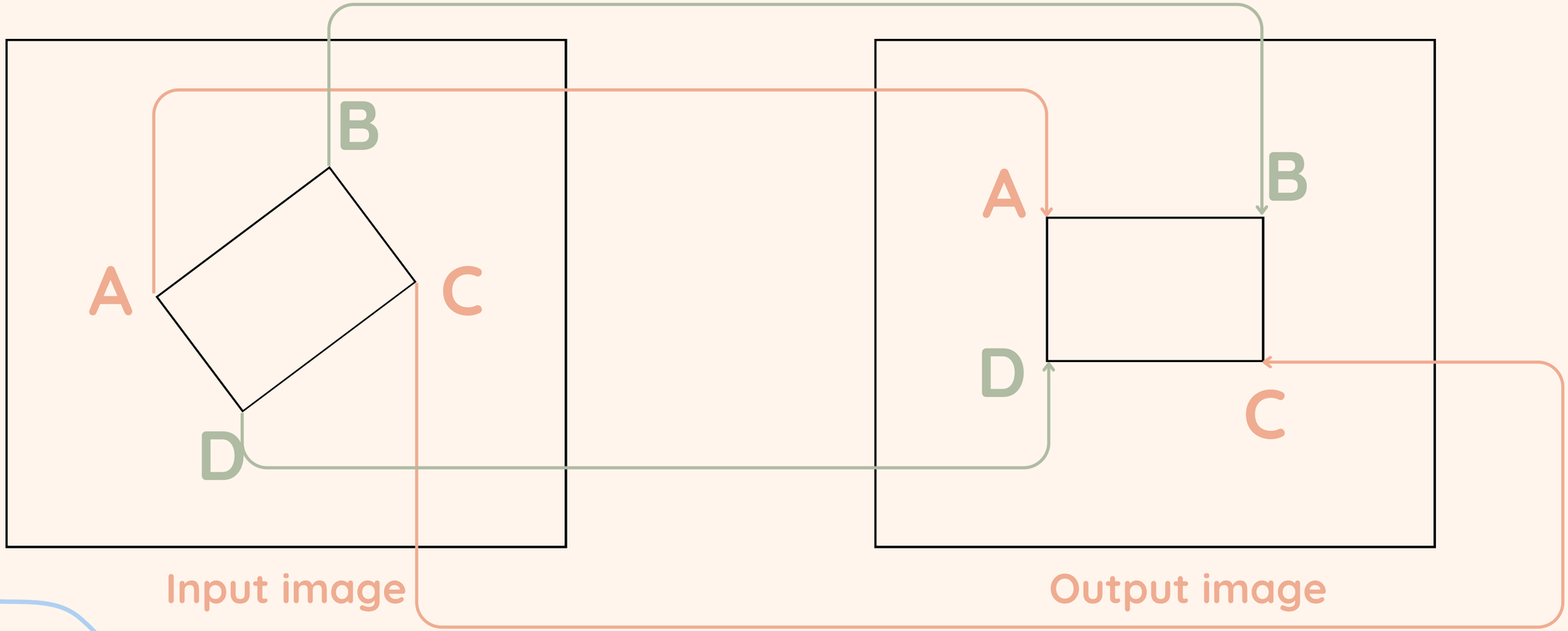
- Epoch: 100
- Dataset: Vintext

test

- Precision:0.87
- Recall:0.74
- F1 score:0.80

PERSPECTIVE TRANSFORMATION

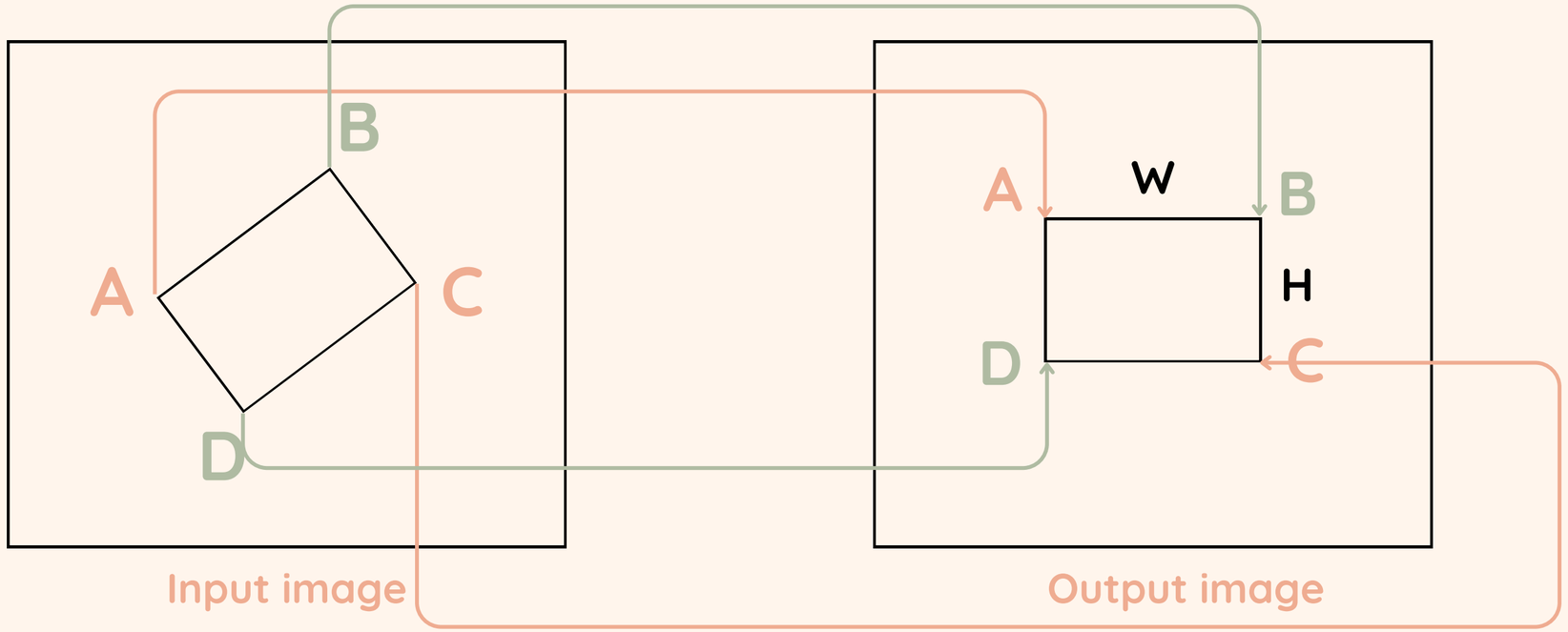
PERSPECTIVE TRANSFORMATION



PERSPECTIVE TRANSFORMATION

Scr points

A(xA,yA),
B(xB,yB),
C(xC,yC),
D(xD,yD)



Dst points

A (0, 0),
B (W, 0),
C (W, H),
D (0, H)

Transformation matrix

[a, b, c]
[d, e, f]
[g, h, 1]

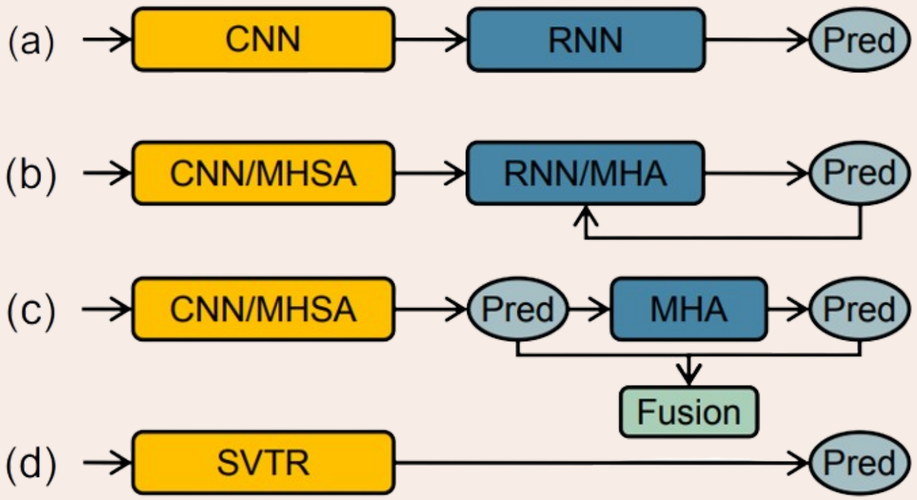
• Scr points = Dst points

RECOGNITION

OVERVIEW

- Extracting text from images in real-world scenarios (street signs; posters; etc.)
- Crucial for various applications: self-driving cars, document analysis, image retrieval, visual impairment support, etc.

STR APPROACHES



CNN-RNN based models

Encoder-Decoder models

Vision-Language models

Simplified models

CNN-RNN BASED MODELS

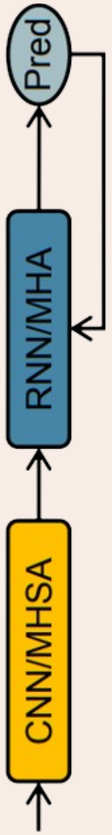
Advantages:

- Efficient: Fast training and inference.
- Commercial viability: Used in some STR products.

Limitations:

- Sensitive to text disturbances: occlusions, deformations, overlaps.
- Limited expressiveness: Sequential model might not capture full context of complex text.

ENCODER-DECODER MODELS

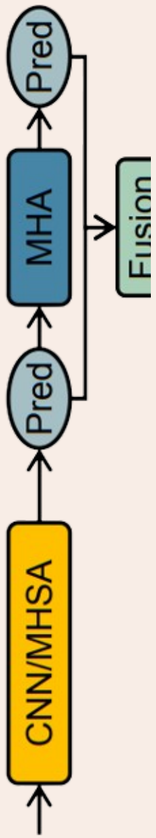


✅

- Achieve better accuracy than earlier approaches because they consider context information within the text sequence.

❌

- Slow inference speed: Predicting characters one by one.
- Complex pipeline: The two-stage architecture (encoder + decoder) can be complex to manage and optimize.



VISION-LANGUAGE MODELS

Pros:

- Higher Accuracy: Language knowledge boosts text understanding.
- Faster: Some methods allow parallel character prediction.

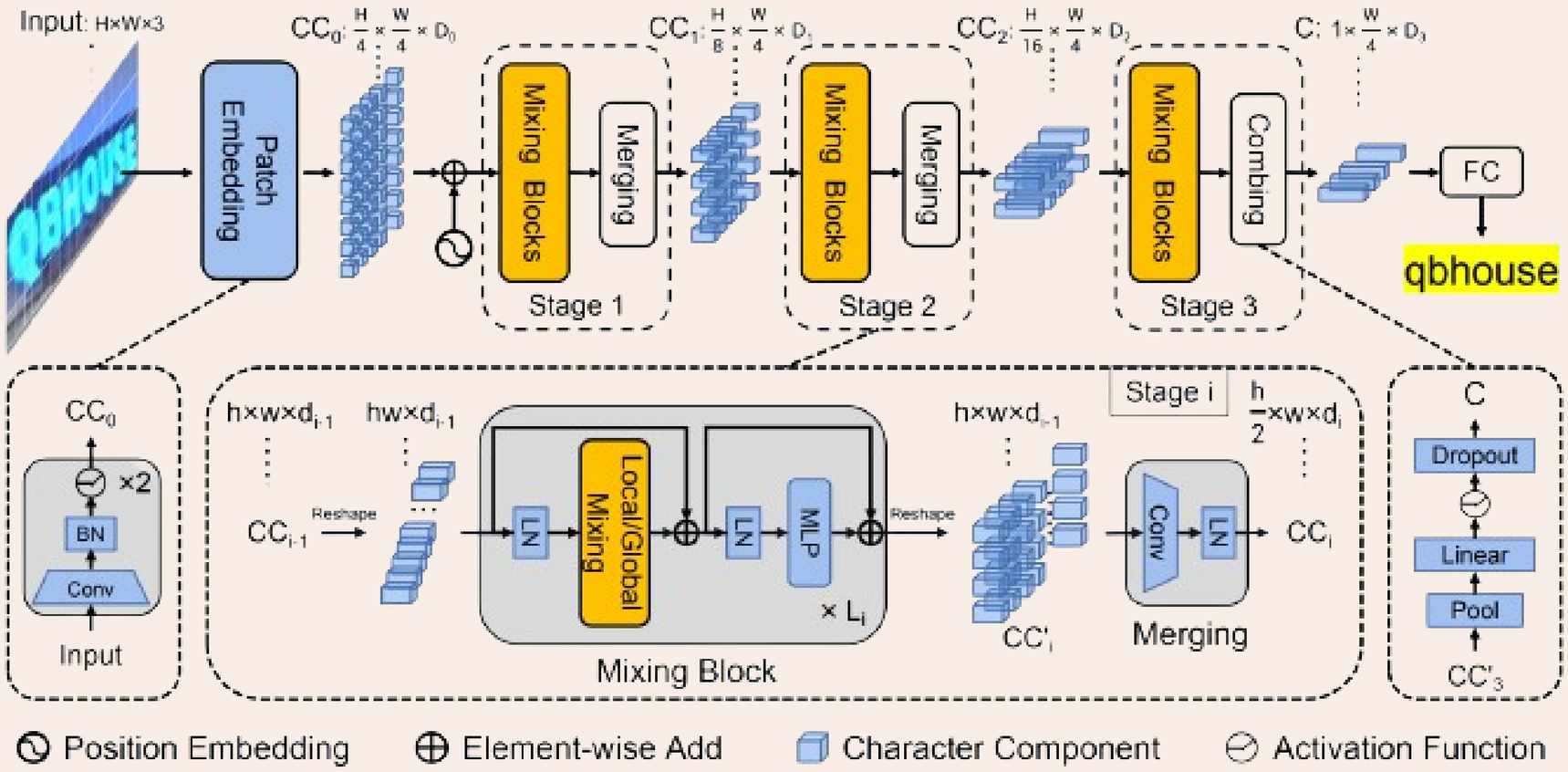
Cons:

- Complex: Integrating language knowledge makes recognition intricate.
- Resource-Heavy: Large models need significant computing power.

SVTR

Scene Text Recognition with a Single Visual Model

ARCHITECTURE



VARIANTS

Model	IC13 857	SVT	IIIT5k 3000	IC15 1811	SVTP	CUTE80	Avg_6	IC15 2077	IC13 1015	IC03 867	IC03 860	Avg_10	Chinese scene_test
SVTR Tiny	96.85	91.34	94.53	83.99	85.43	89.24	90.87	80.55	95.37	95.27	95.70	90.13	67.90
SVTR Small	95.92	93.04	95.03	84.70	87.91	92.01	91.63	82.72	94.88	96.08	96.28	91.02	69.00
SVTR Base	97.08	91.50	96.03	85.20	89.92	91.67	92.33	83.73	95.66	95.62	95.81	91.61	71.40
SVTR Large	97.20	91.65	96.30	86.58	88.37	95.14	92.82	84.54	96.35	96.54	96.74	92.24	72.10

Multiple architecture options: SVTR-L (large) for high accuracy, SVTR-T (tiny) for speed.

Impressive results on English and Chinese datasets:

- SVTR-L: competitive accuracy in English.
- SVTR-T: efficient and fast, ideal for real-time applications.

TRAIN

TRAIN

train

- Epoch: 120
- Dataset: 25794 cropped images

test

- Dataset: 7220 cropped images

DATA AUGMENTATION

- Geometry
- Deterioration
- Color Jitter

GEOMETRY

Random Rotation:

- Applies random rotations within a specified range (e.g., ±15 degrees).
- Maintains image center after rotation.

GEOMETRY

Random Affine:

- Rotation: within a specified range (e.g., ±15 degrees).
- Scaling: random factor between user-defined limits (scale=(0.5, 2.0)).
- Shearing: subtle tilting of the image (shear=(45, 15)).
- Translation: horizontal and vertical shifts (translate=(0.3, 0.3)).

GEOMETRY

Random Perspective:

- Applies random perspective distortions to images (distortion=0.5).

GEOMETRY

Benefits:

- Enhances model generalizability to unseen viewpoints, diverse image orientations and scales.
- Prepares models for real-world image variations.
- Enhances model robustness to minor distortions and misalignments.
- Can be combined with other augmentation techniques.

DETERIORATION

- Gaussian noise: Added with a standard deviation of 20.
- Motion blur: Applied with an angle of 6 degrees.
- Rescaling: 1/4 of original image.
- 0.25 probability of applying all effects (75% chance of original image).

DETERIORATION

Benefits:

- Introduces random intensity variations across the image, simulating sensor noise or low-light conditions.
- Blurs the image in a specific direction, mimicking camera movement or object motion.
- Shrinks the image, simulating situations where the object of interest is further away or captured with a lower resolution camera.

COLOR JITTER

- Brightness, Contrast, Saturation: Adjust image intensity and color balance (0.5).
- Hue Shift: Randomly change image colors (0.1).
- Frequency: Control how often color jitter is applied (probability 0.25).

COLOR JITTER

Benefits:

- Reduces dependence on specific color schemes.
- Enhances model performance in real-world scenarios with varying illumination and colors.

EVALUATION

Evaluation SVTR-T

Data Augmentation

- Acc: 0.794
- Norm_edit_dis: 0.882

No Data Augmentation

- Acc: 0.7605
- Norm_edit_dis: 0.865

Evaluation SVTR-L

Data Augmentation

- Acc: 0.6419
- Norm_edit_dis: 0.92

No Data Augmentation

- Acc: 0.54
- Norm_edit_dis: 0.73

DEMO

More Details

DAYTIME


SỐNG & LÀM VIỆC PHẢI CÓ LƯƠNG TÂM VÀ TRÁCH NHIỆM
ĐỪNG BAO GIỜ "EM TƯỞNG"
SỐNG & LÀM VIỆC PHẢI CÓ LƯƠNG TÂM VÀ TRÁCH NHIỆM
ĐỪNG BAO GIỜ "EM TƯỞNG"

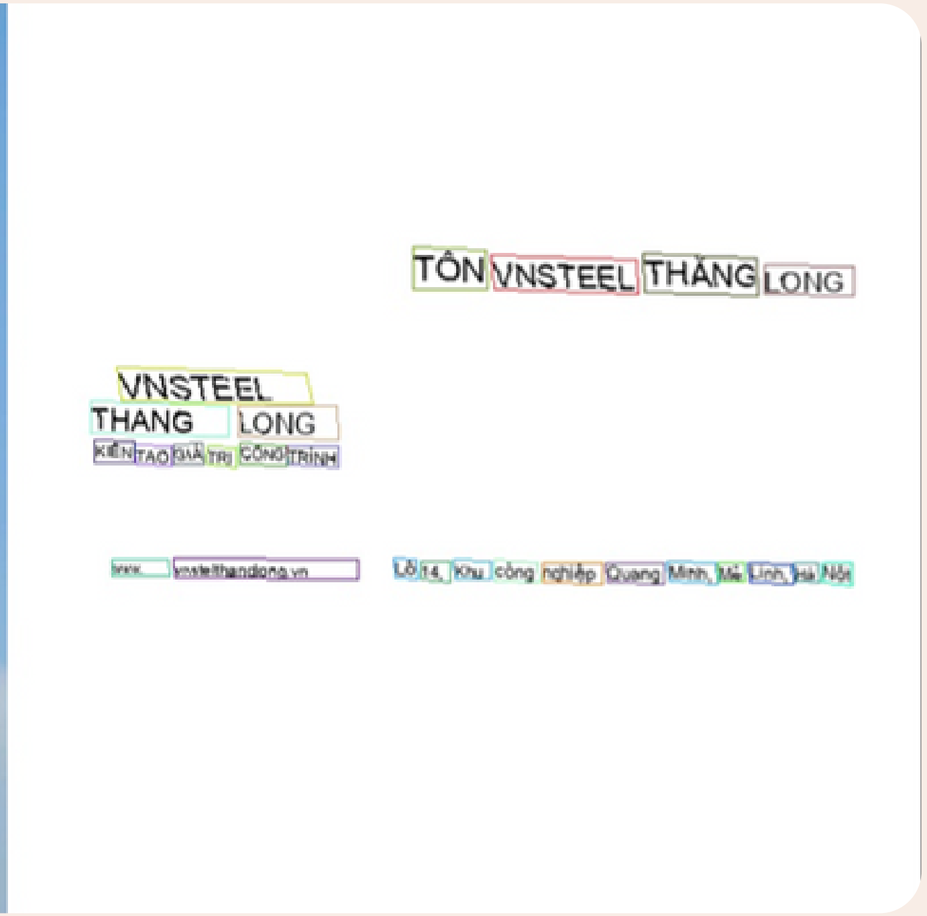
DAYTIME


TRƯỜNG TIỂU HỌC
NGUYỄN BỈNH KHIÊM

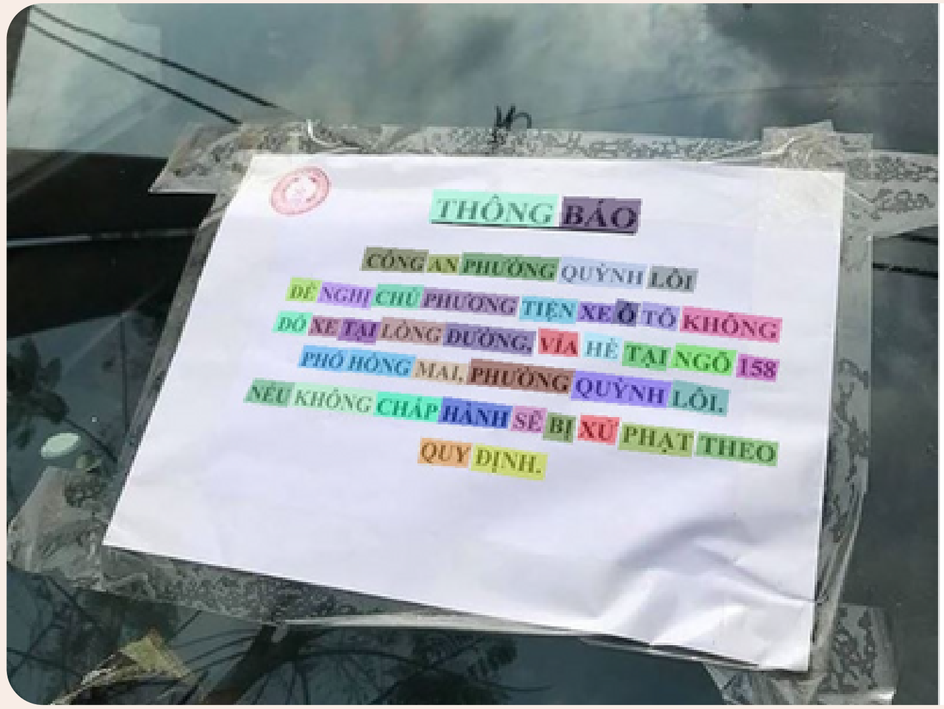

TRƯỜNG TIU HỌC
NGUYN BỈNH KHIÊM

DAYTIME


TÔN VNSTEEL THĂNG LONG
VNSTEEL
THANG LONG
KIẾN TẠO GIÁ TRỊ CÔNG TRÌNH
www.vnsteelthanglong.vn
Lô 14, Khu công nghiệp Quang Minh, Mê Linh, Hà Nội

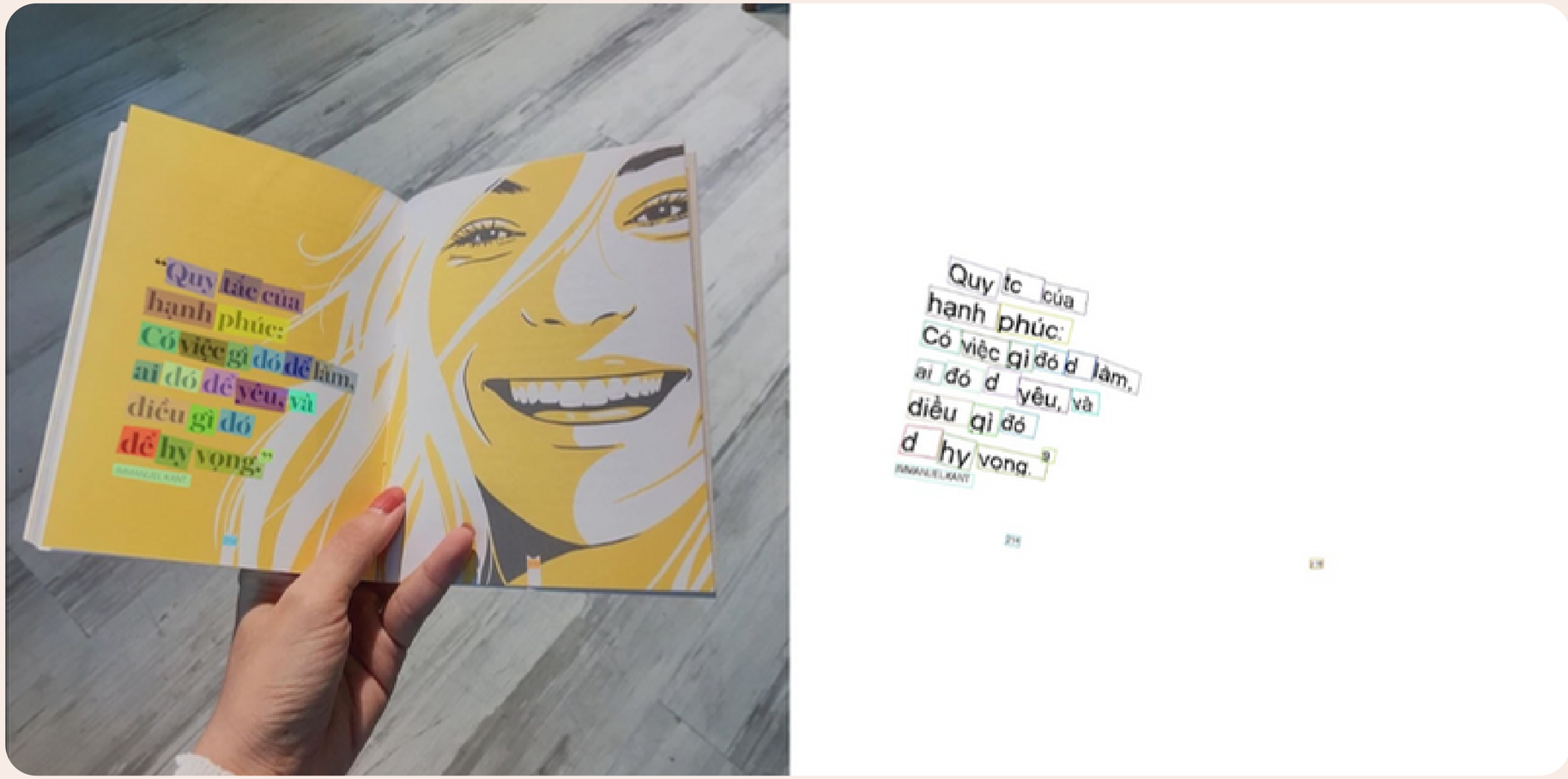

TÔN VNSTEEL THĂNG LONG
VNSTEEL
THANG LONG
KIẾN TẠO GIÁ TRỊ CÔNG TRÌNH
www. vnsteelthanglong.vn
Lô 14, Khu công nghiệp Quang Minh, Mê Linh, Hà Nội

DAYTIME



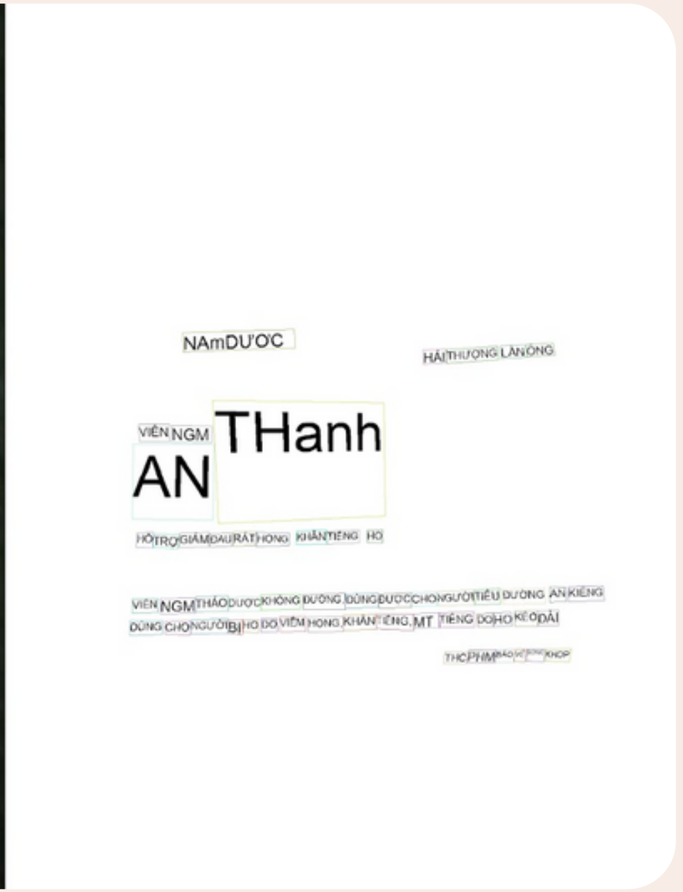
THÔNG BÁO
CÔNG AN PHƯỜNG QUỲNH LÔI
Đ NGHỊ CHỦ PHƯƠNG TIỆN XE Ô TÔ KHÔNG
Đ XE TẠI LÒNG ĐƯỜNG, VỈA HÈ TẠI NG 158
PHỐ HỒNG MAI, PHƯỜNG QUỲNH LÔI.
NẾU KHÔNG CHẤP HÀNH S BỊ XỬ PHẠT THEO
QUY ĐỊNH.

DAYTIME


"Quy tắc của
hạnh phúc:
Có việc gì đó để làm,
ai đó để yêu, và
điều gì đó
để hy vọng."
IMMANUEL KANT
Quy tc của
hạnh phúc:
Có việc gì đó đ làm,
ai đó đ yêu, và
điều gì đó
đ hy vọng.
IMMANUEL KANT

NIGHTTIME

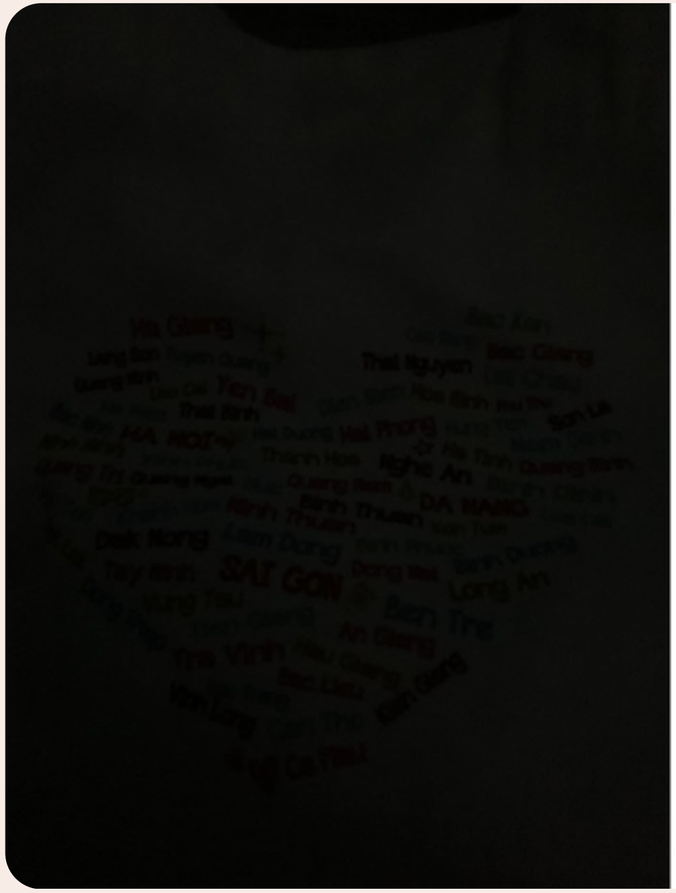

NAM DƯỢC
VIÊN NGẬM
anThanh


NAmDƯỢC
HẢI THƯỢNG LÃN ÔNG
VIÊN NGM
AN
THanh
HỖ TRỢ GIẢM ĐAU RÁT HỌNG KHẢN TIẾNG HO
VIÊN NGẬM THẢO DƯỢC KHÔNG ĐƯỜNG DÙNG ĐƯỢC CHO NGƯỜI TIỂU ĐƯỜNG AN KIÊNG
DÙNG CHO NGƯỜI BỊ HO ĐỜ VIÊM HỌNG KHẢN TIẾNG MẤT TIẾNG ĐỜ HO KÉO DÀI
THC PHM

NIGHTTIME

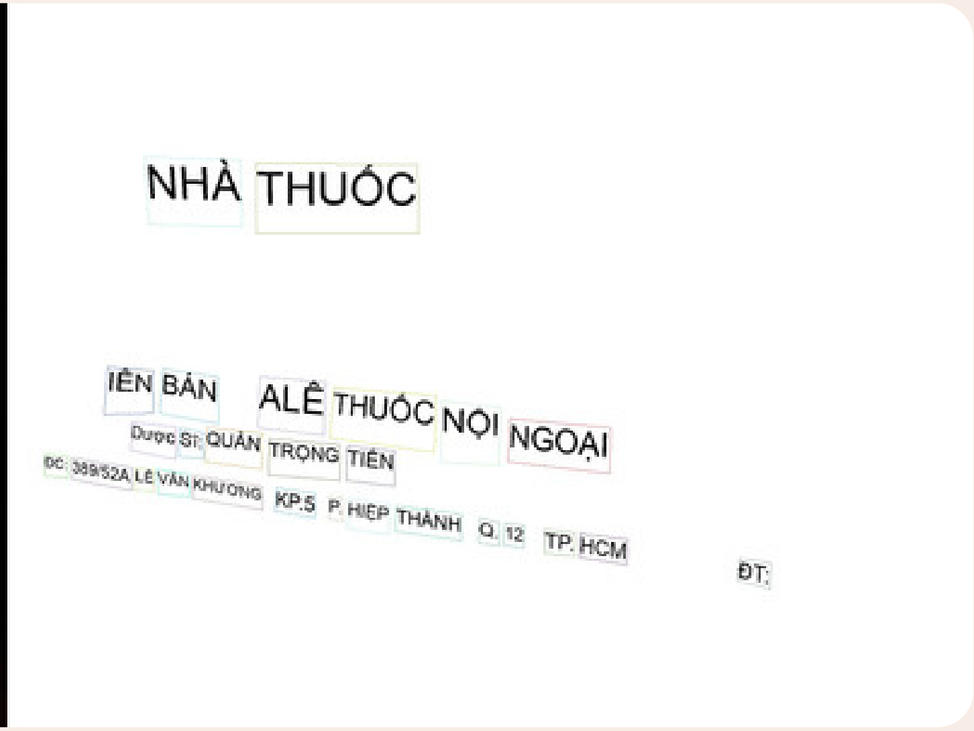
Hea Giang
Thai Nguyen
Bac Giang
Lang Son
Quong Ninh
La
CAI Yen Bai
Thai Binh
Hoa Binh
Sonla
Hai Phong
HA NOI
Binh
Nghe An
Quang Binh
Quang Tri Quang
DA NANG
Binh Thuan
Ninh Thuan
Lan Đang
Dak Nong
SAI GON
Dong
Long An
Tay Ninh
Ben Tre
Tra Vinh
Bac Lieu
Vinh Long
Ca Mau
Sunkis

NIGHTTIME

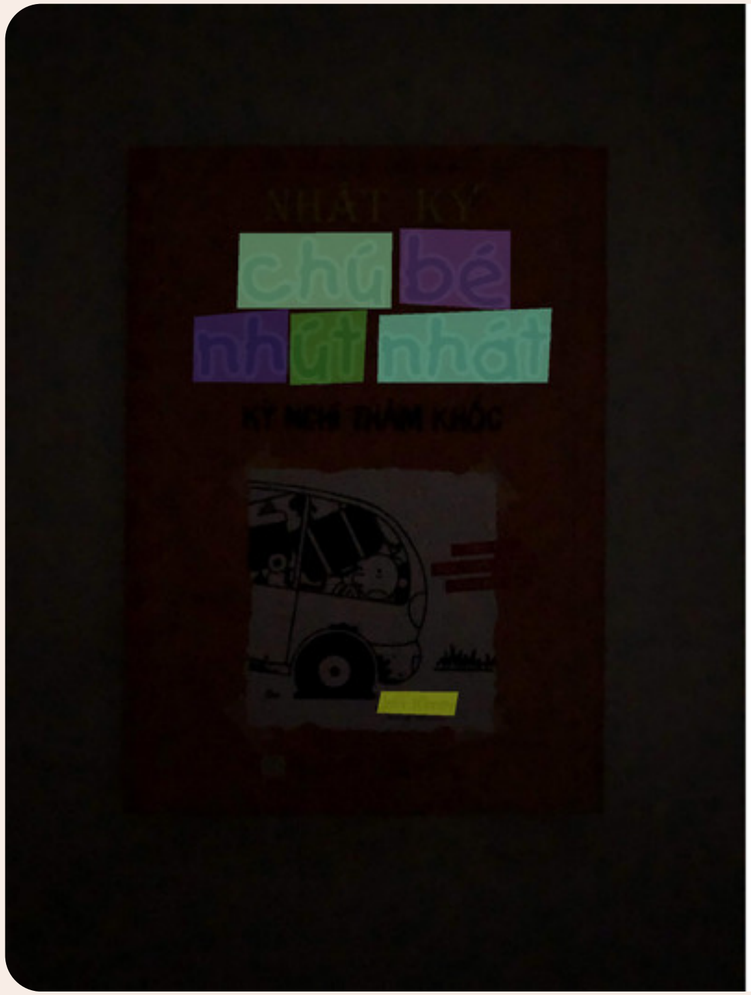


NIGHTTIME


ĐẠT CHUẨN GPP
NHẬP CÁC LOẠI


NHÀ THUỐC
BÁN LẺ THUỐC NỘI NGOẠI
Dược Sĩ: QUẢN TRỌNG TIẾN
ĐC: 389/52A LÊ VĂN KHƯƠNG KP.5 P. HIỆP THÀNH Q. 12 TP. HCM
ĐT:

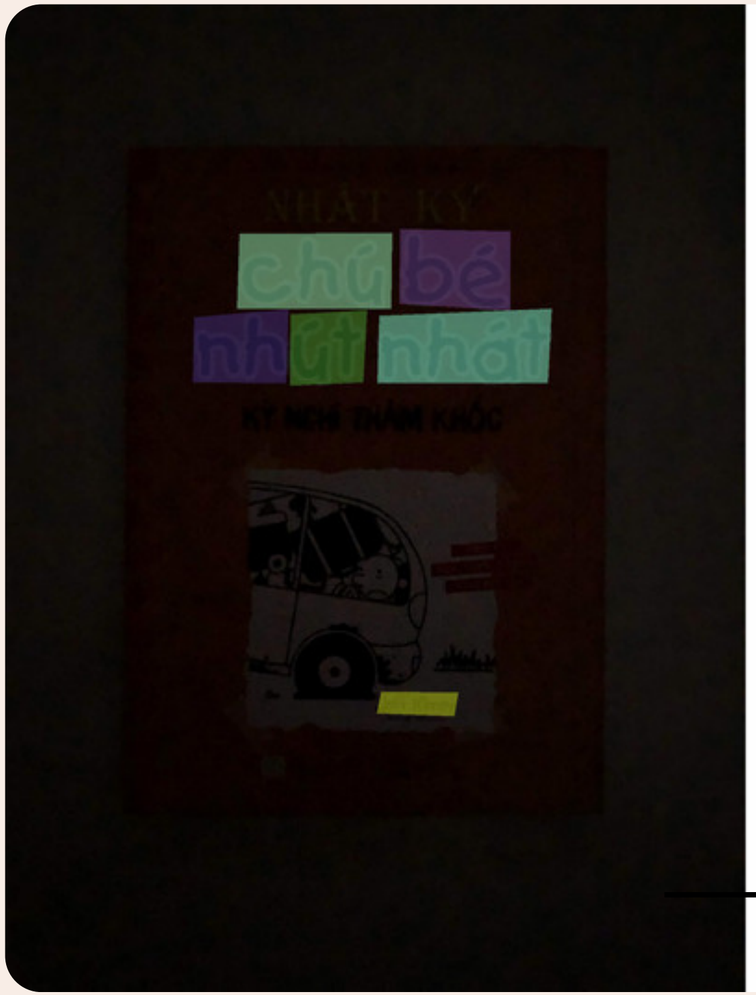
NIGHTTIME



Chý bé
nh Gt nht

ATKimey

NIGHTTIME



Chý bé
nh Gt nht

ATKimey

Low-light Image Enhancement

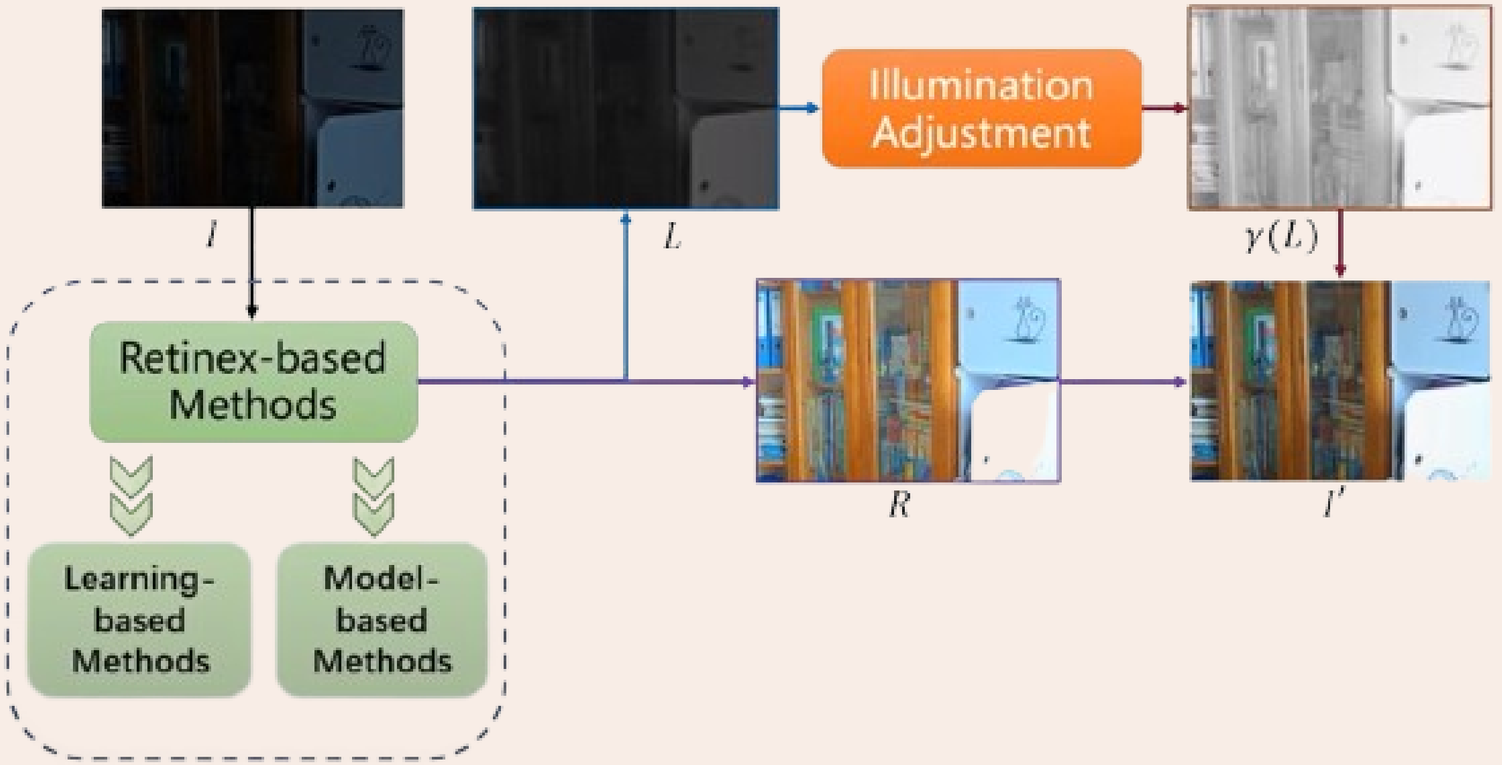
URETINEX-NET

Retinex-based methods

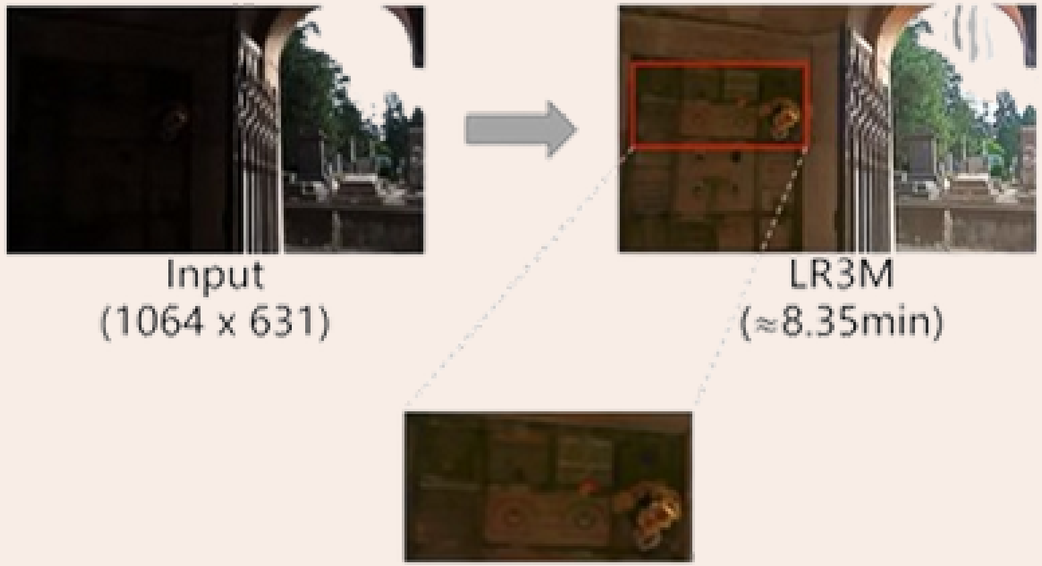
Modeling:

$$I = R \cdot L$$
$$I' = R \cdot \gamma(L)$$

I: input image
R: reflectance layer
L: illumination layer
$\gamma(\cdot)$: adjustment function

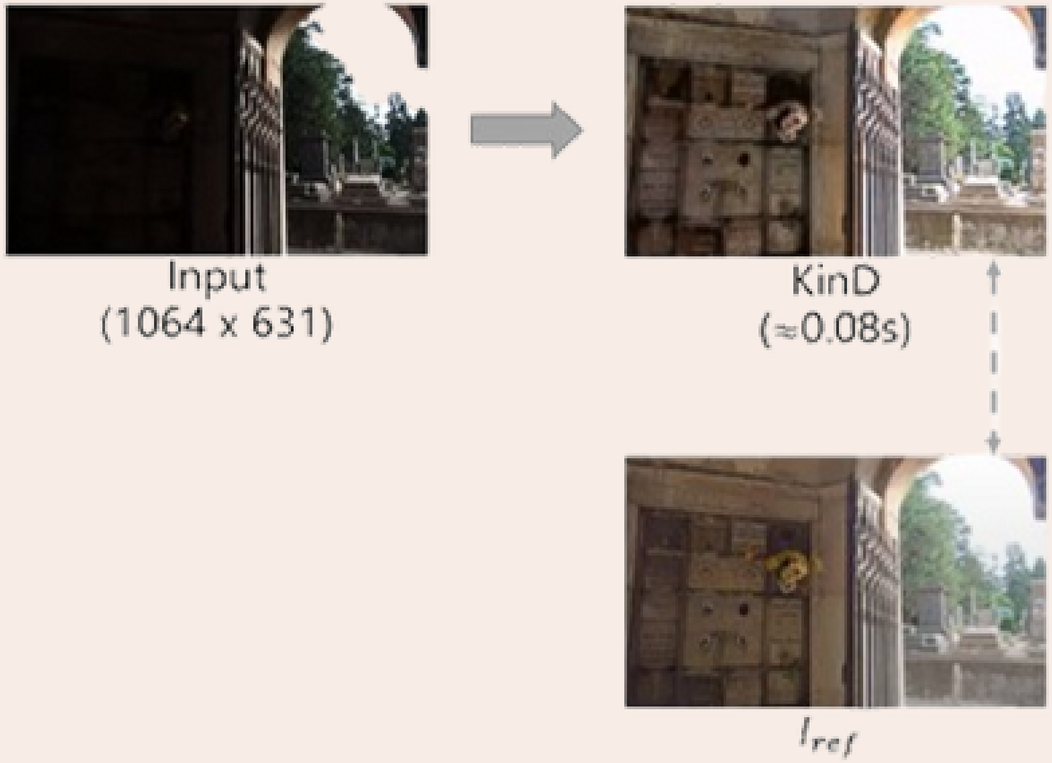


Retinex-based methods



- Hand-crafted features cannot be adaptive enough in various scenes

RETINEX-BASED METHODS



- Hand-crafted features cannot be adaptive enough in various scenes.
- Model-based method is time-consuming.
- Results in loss of details or unnatural look.

RETINEX-BASED METHODS

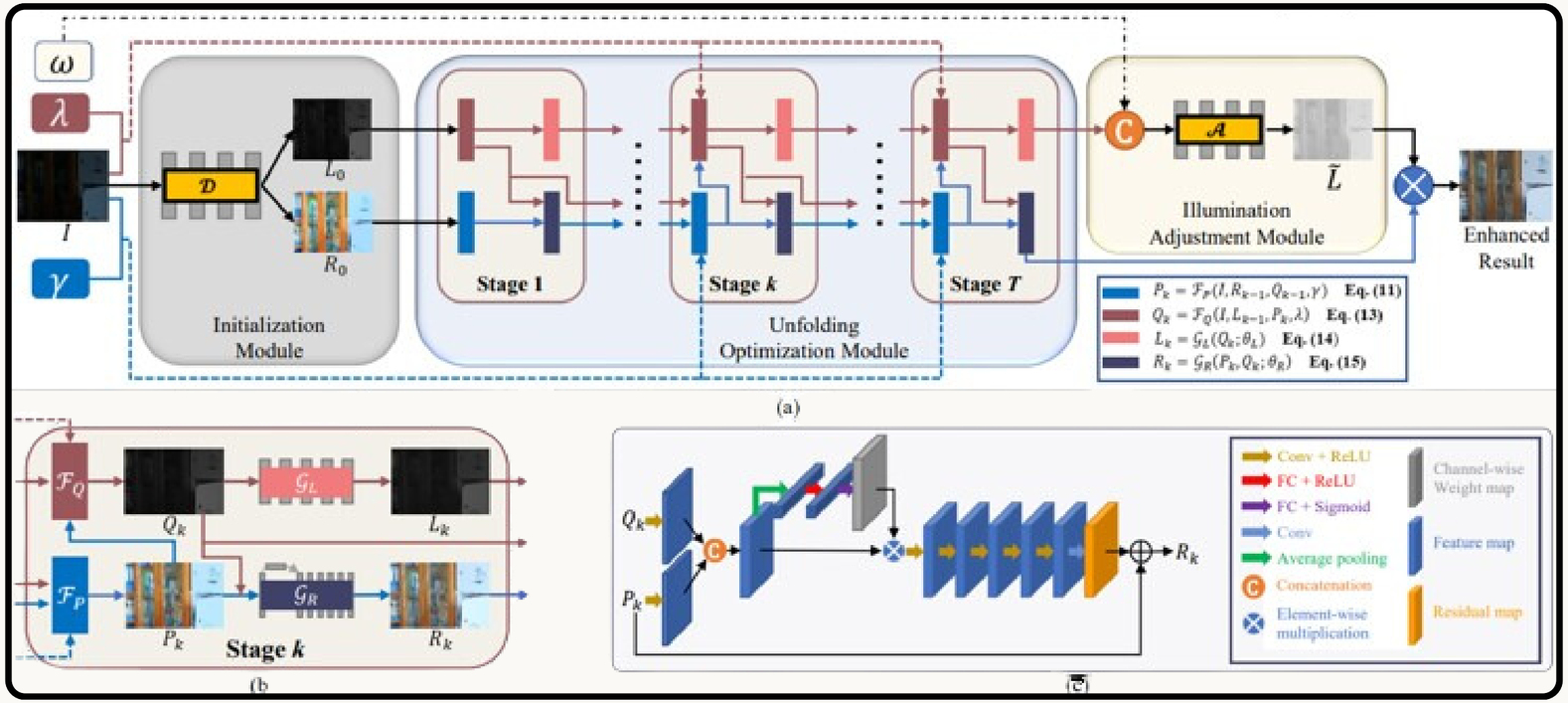
Based on the traditional model-based methods, deep unfolding network for low-light image enhancement, consisting of:

- Initialization module.
- Optimization module.
- Illumination adjustment module.

Unfold the optimization procedure into a deep network:

- Inherits the flexibility and interpretability from model-based methods.
- Leverages the powerful model ability of learning-based methods to adaptively fit data-dependent priors.

RETINEX-BASED METHODS



Pipeline

Low-light Image Enhancement

Detection

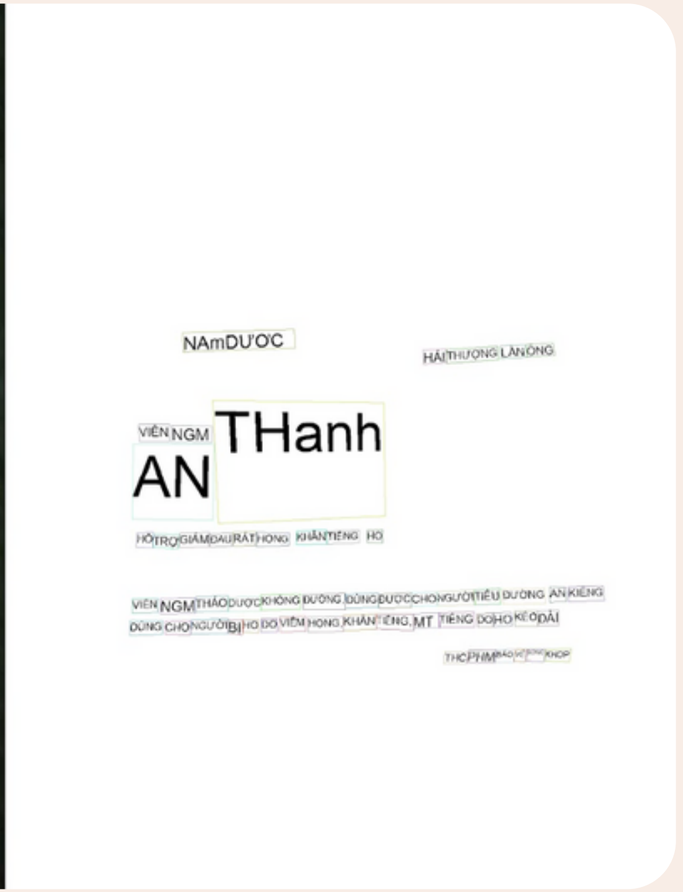
Perspective Transformation

Recognition

RESULTS

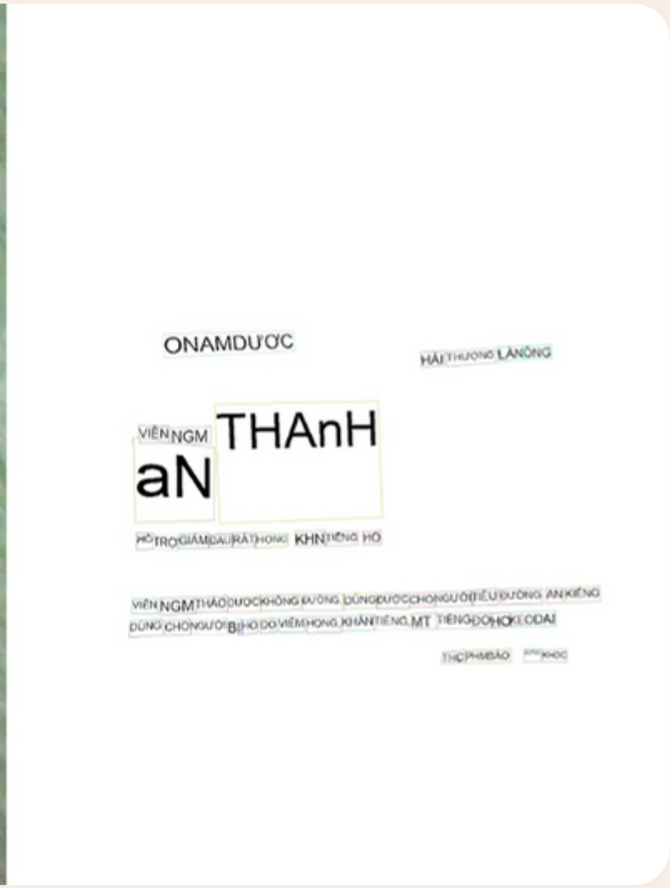
NIGHTTIME


NAM DƯỢC
VIÊN NGẬM
an Thanh

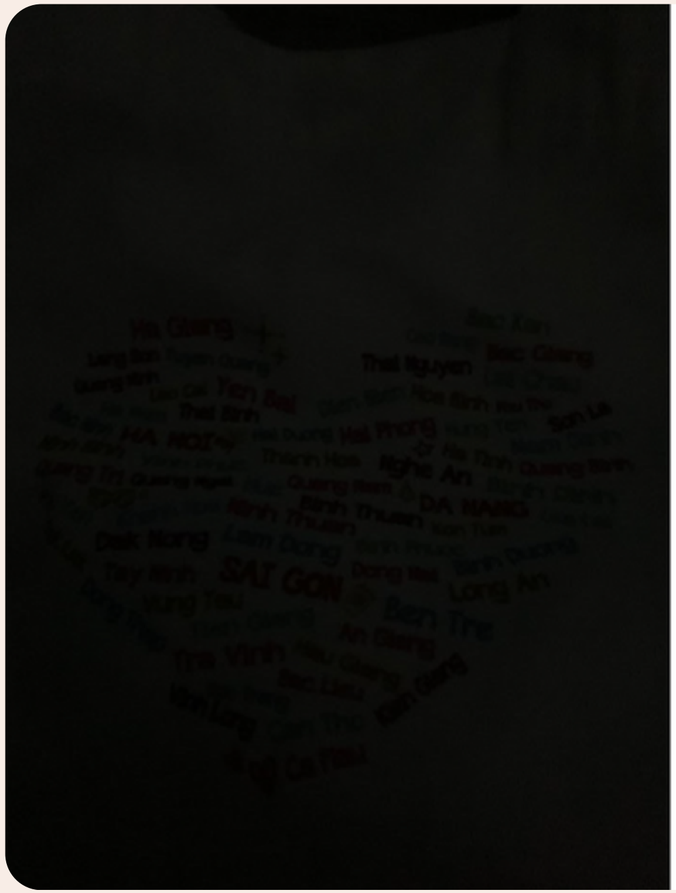

NAmDƯỢC
HẢI THƯỢNG LÃN ÔNG
VIÊN NGM
AN
THanh

NIGHTTIME


NAM DƯỢC
HẢI THƯỢNG LÃN ÔNG
VIÊN NGẬM
an Thanh
HỖ TRỢ GIẢM ĐAU RÁT HỌNG - KHẢN TIẾNG - HO
VIÊN NGẬM THẢO DƯỢC KHÔNG ĐƯỜNG, DÙNG ĐƯỢC CHO NGƯỜI TIỂU ĐƯỜNG, ĂN KIÊNG
DÙNG CHO NGƯỜI BỊ HO DO VIÊM HỌNG, KHẢN TIẾNG, MẤT TIẾNG DO HO KÉO DÀI


ONAMDƯỢC
HẢI THƯỢNG LÃN ÔNG
VIÊN NGẬM
THAnH
aN
HỖ TRỢ GIẢM ĐAU RÁT HỌNG KHẢN TIẾNG HO
VIÊN NGẬM THẢO DƯỢC KHÔNG ĐƯỜNG DÙNG ĐƯỢC CHO NGƯỜI TIỂU ĐƯỜNG ĂN KIÊNG
DÙNG CHO NGƯỜI BỊ HO DO VIÊM HỌNG KHẢN TIẾNG MẤT TIẾNG DO HO KÉO DÀI
THỰC PHẨM BẢO KHỎE

NIGHTTIME



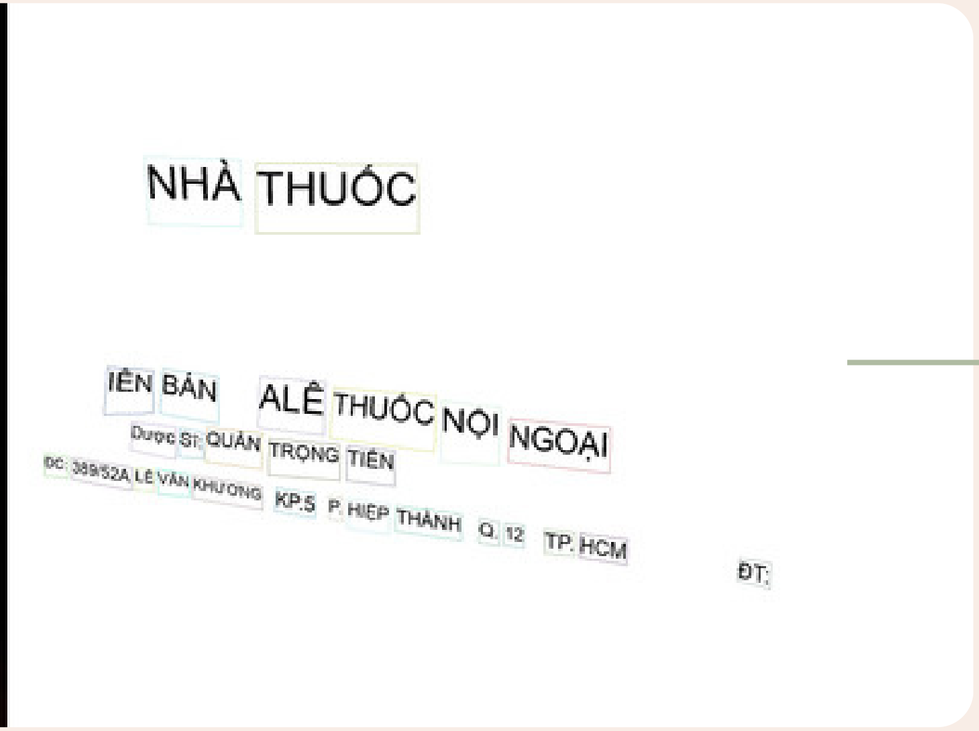
NIGHTTIME


Ha Giang
Bac Kan
Bac Giang
Thai Nguyen
Yen Bai
Phong
Nghe An
DA NANG
Dak Nong
SAT GON
Ben Tre
An Giang
Long An
Tra Vinh
Bac Lieu
Kien Giang
Ca Mau


Giong
Long Ion
QuangNinh
Thai
VE
Thel Binh
Bai
HOI
Duong Heal
Thành Hoa
Tint
An
Quan
Ninh
Binh
Thur
DA NANKG
Thuan
Than
NOI
ăm Dong
Tay anh
SAT GON
Donget
nng
Tai
Vinh
Vinnlng

NIGHTTIME


ĐẠT CHUẨN GPP
NHẬP CÁC LOẠI

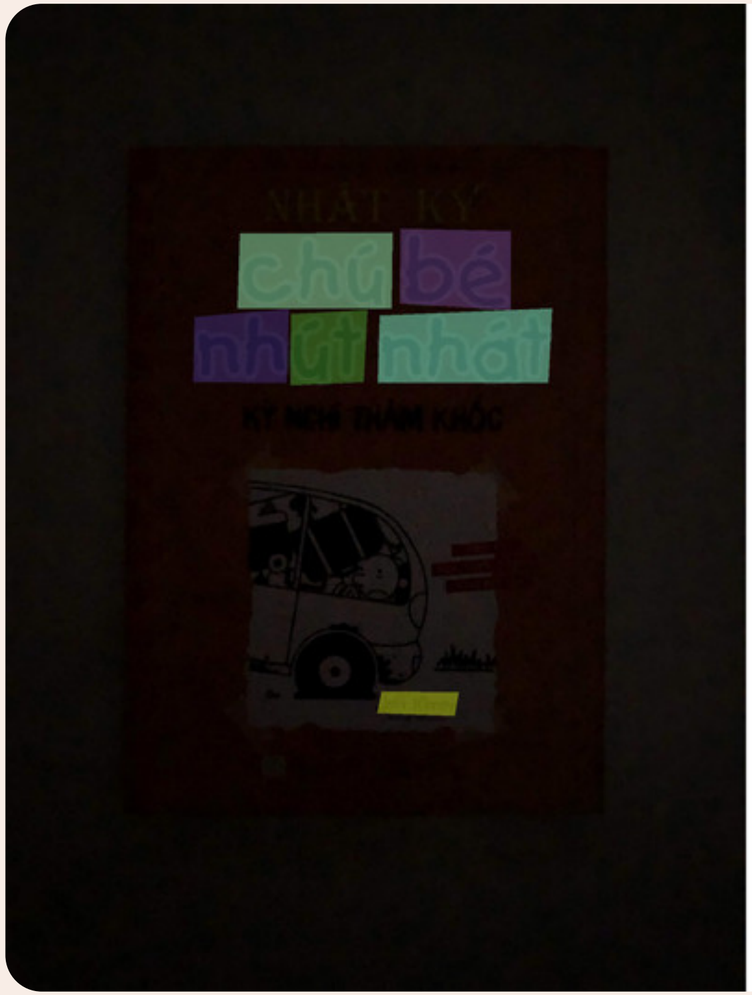

NHÀ THUỐC
LIÊN BÁN ALÊ THUỐC NỘI NGOẠI
Dược Sĩ: QUẢN TRỌNG TIẾN
ĐC: 389/52A LÊ VĂN KHƯƠNG KP.5 P. HIỆP THÀNH Q. 12 TP. HCM
ĐT:

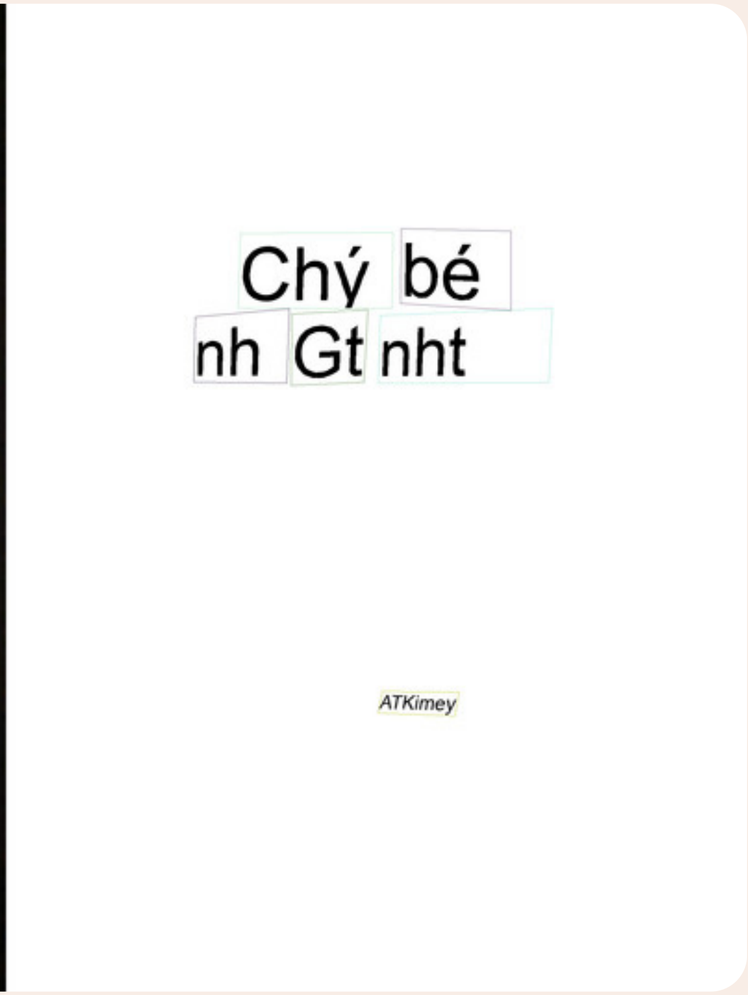
NIGHTTIME


NHÀ THUỐC ĐẠT CHUẨN GPP
AN THIỆN PHÚC
CHUYÊN BÁN SỈ & LẺ THUỐC NỘI NGOẠI NHẬP CÁC LOẠI
Dược Sĩ: QUẢN TRỌNG TIẾN
ĐC: 389/52A LÊ VĂN KHƯƠNG - KP.5 - P. HIỆP THÀNH - Q. 12 - TP. HCM
GPS: 41.8029877
ĐT: 0964 096 868

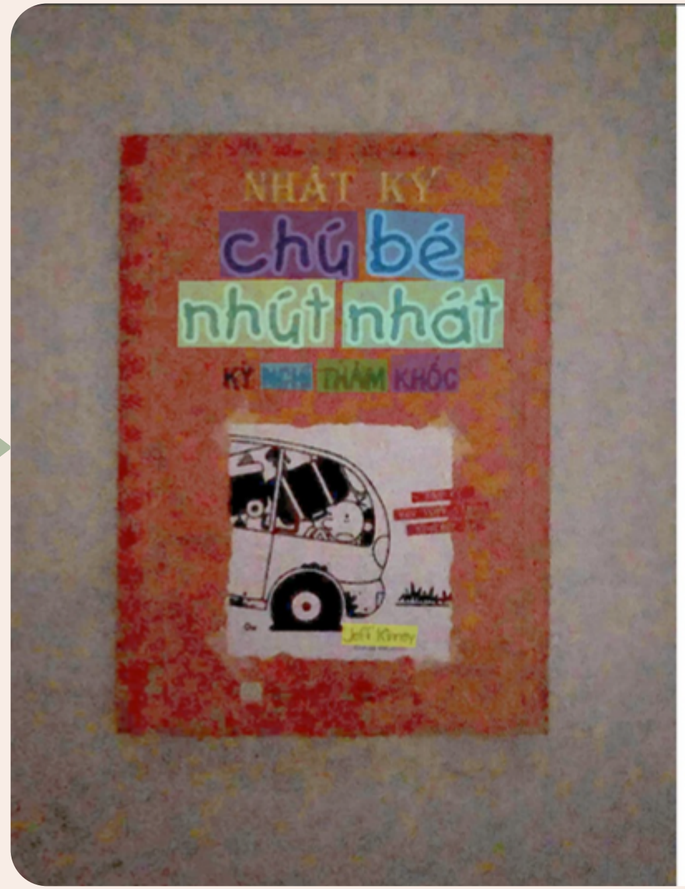

NHÀ THUỐC ĐẠT CHUN
CHUYÊN BÁN GÌ L THUỐC NỘI NGOẠI NHP CÁC LOẠI
Dược Sĩ: QUẢN TRỌNG TIÊN
ĐC: 389/52A LÊ VĂN KHƯƠNG KP.5 P. HIỆP THÀNH 12 TP. HCM
GPS: 41.8029877
ĐT: 0964 096868

NIGHTTIME


NHẬT KÝ
chú bé
nhút nhát
KỲ NIỆM ĐÁNH KHỐC


Chý bé
nh Gt nht
ATKimey

NIGHTTIME



Chý bé
nhýt nht
NCHÍ THM KHÓC

HTKirey

CHALLENGES

- Limited training data: Small amount of high-quality Vietnamese text data available, lead to overfitted and perform poorly on unseen data
- Font diversity: A variety of fonts, difficult for models to generalize to unseen fonts, leading to recognition errors
- Complex character set: Diacritics and tonal, challenging to distinguish and recognize accurately (ơ, â, ê,...)

CHALLENGES

- Image quality issues: Low resolution, noise, blur, and other distortions, difficult to extract accurate text information
- Occlusion and perspective distortion: Partially obscured by other objects or distorted by perspective
- Background complexity: A variety of backgrounds, make it difficult to distinguish between text and background noise

FUTURE WORKS

- Collect more Vietnamese Scene Text data.
- Text Data Generator on different Vietnamese fonts.
- Image Super Resolution: ERSGAN, HAT,...
- Reconstruct the image after low-light enhanced

REFERENCES

- A Single-Shot Arbitrarily-Shaped Text Detector based on Context Attended Multi-Task Learning
- SVTR: Scene Text Recognition with a Single Visual Model
- URetinex-Net: Retinex-based Deep Unfolding Network for Low-light Image Enhancement
- PaddleOCR

THANK YOU FOR LISTENING

aNY QUESTIONS?!?!