

NHẬN DẠNG KÝ TỰ QUANG  
HỌC (OCR) CHO CHỮ NÔM  
TRONG TÀI LIỆU TIẾNG Việt

---

Presented by  
Nhóm 1

GV.Tran Tan Thanh



# Giới thiệu

Nhận dạng kí tự quang học (Optical Character Recognition - OCR) là một lĩnh vực quan trọng trong xử lý hình ảnh và trí tuệ nhân tạo, nhằm chuyển đổi văn bản từ hình ảnh sang định dạng kỹ thuật số có thể chỉnh sửa và tìm kiếm được.



# Lý do chọn OCR cho chữ Nôm trong tài liệu tiếng Việt

- Chữ Nôm, hệ chữ viết truyền thống của Việt Nam, mang giá trị lịch sử và văn hóa sâu sắc, là kho tàng tri thức cần được khai thác và bảo tồn. Việc phát triển OCR cho chữ Nôm trong tài liệu tiếng Việt có ý nghĩa quan trọng
- Hiện nay, việc dịch thuật chữ Nôm sang tiếng Việt hiện đại (Quốc Ngữ) gặp nhiều khó khăn do thiếu chuyên gia và quy trình thủ công tốn thời gian. Công nghệ OCR có thể đẩy nhanh quá trình này, tạo cầu nối giữa bảo tồn di sản và khả năng tiếp cận tri thức

# Tham khảo nghiên cứu về OCR Hán-Nôm

- 1. OCR\_chu\_nom

- Tác giả: Trần Hoàng Quân, Nguyễn Hoàng Anh Kiệt, Võ Trương Trung Chánh
- Tổng quan: Dự án sinh viên về OCR Chữ Nôm.
- Đặc điểm chính:
- Phát hiện văn bản xử lí trên từng từ.

- 2. NomNaOCR

- Tác giả: Nguyễn Đức Duy Anh, hướng dẫn bởi TS. Đỗ Trọng Hợp
- Tổng quan: Sử dụng Học sâu để số hóa tài liệu viết tay tiếng Việt cổ.
- Đặc điểm chính:
- Phát hiện văn bản xử lí theo chuỗi.





# Thủ vĩ ngâm

首尾吟

Nguyễn Du

谷城南孔蔑間  
奴揲倅少堆告  
昆隊遁揚埃眷  
埜馭檻少几經  
儻蹀狹回坤且猋  
茹涓趣庶礙援勵  
朝官拯沛隱拯沛  
谷城南均蔑間

Góc thành Nam. lều một gian.  
No nước uống. thiếu cơm ăn.  
Con đói trốn. đường ai quyến.  
Bà ngựa gầy. thiếu kẻ chăn.  
Ao bơi hép hòi khôn thả cá.  
Nhà quen thú thừa ngại nuôi vắn.  
Triều quan chẳng phải. ẩn chẳng phải.  
Góc thành Nam. lều một gian.



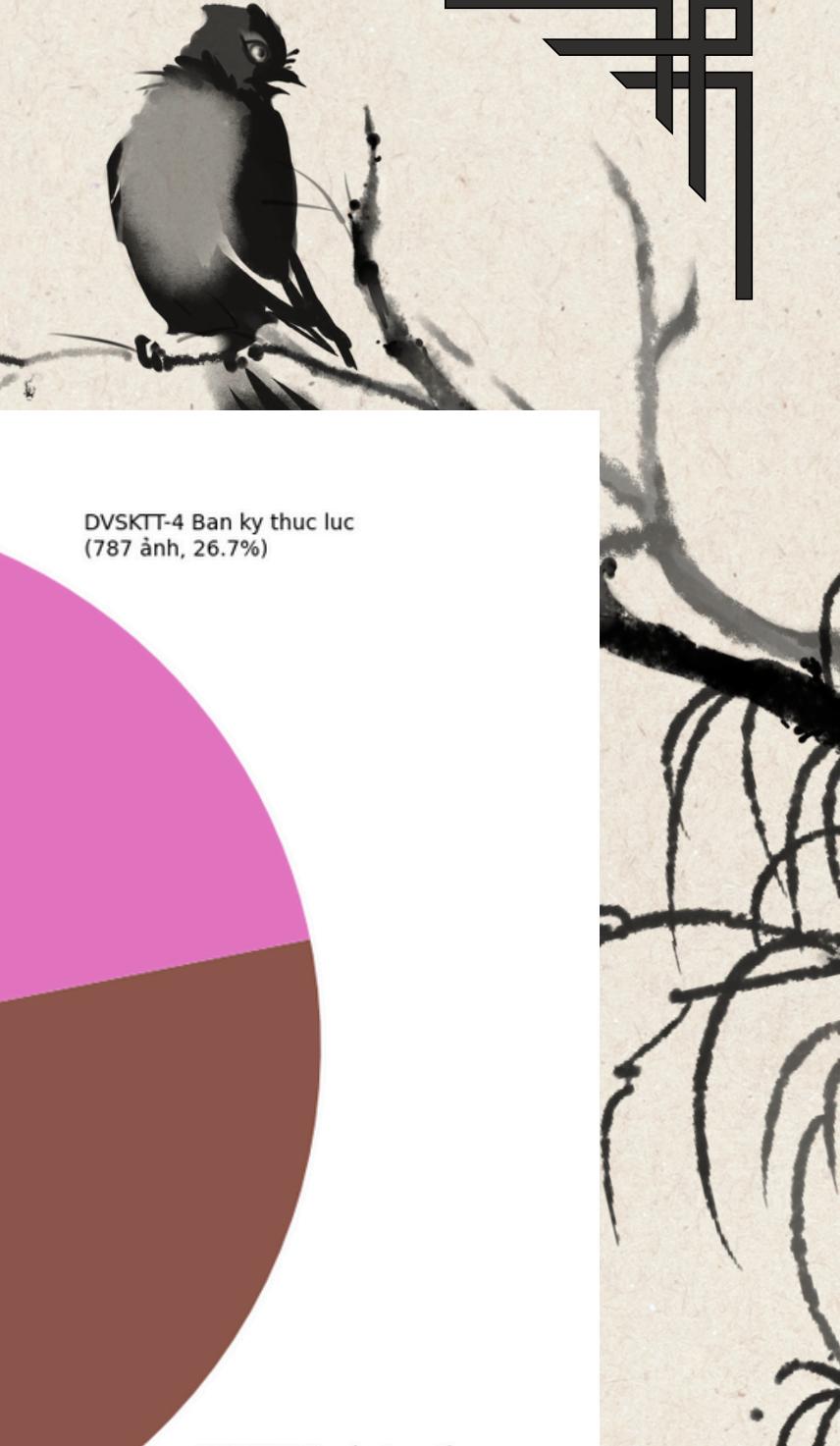


# Data

- Bộ dữ liệu: Sử dụng IHR-NomDB (<https://morphoboid.labri.fr/ihr-nom.html>) và NomNaCCR(<https://www.kaggle.com/datasets/quandang/nomnaocr>), bộ dữ liệu chứa các tài liệu chữ Nôm cũ và bị xuống cấp.



# Dữ liệu thô



## Tục biên tự [ 12 trang ]

### Tách câu và Phiên âm

大越史記續編序 . [1a\*1\*1]

Đại Việt sử kí tục biên tự.

國之有史尚矣 . [1a\*2\*1]

Quốc chi hữu sử thượng hĩ.

我越歷代史記先正黎文休潘孚先作  
之於前吳士連武瓊述之於後 . [1a\*2\*7]

Ngã Việt, lịch đại sử kí tiên chính Lê Văn Hưu, Phan  
Phu Tiên tác chí ư tiền, Ngô Sĩ Liên, Vũ Quỳnh thuật  
chi ư hậu.

其間事蹟之詳畧政治之得失莫不悉  
備於記載之中 . [1a\*4\*8]

Ki gian sự tích chí tường lược, chính trị chí đắc thất,  
mặc bất tất bị ư kí tài chí trung.

但未行鋟梓更手傳筆因循抄錄不能  
無陶陰帝廟之疑 . [1a\*6\*3]

Đãn vị hành tẩm tử, cánh thủ truyền bút, nhân tuân  
sao lục, bất năng vô đào âm để hổ chí nghi.

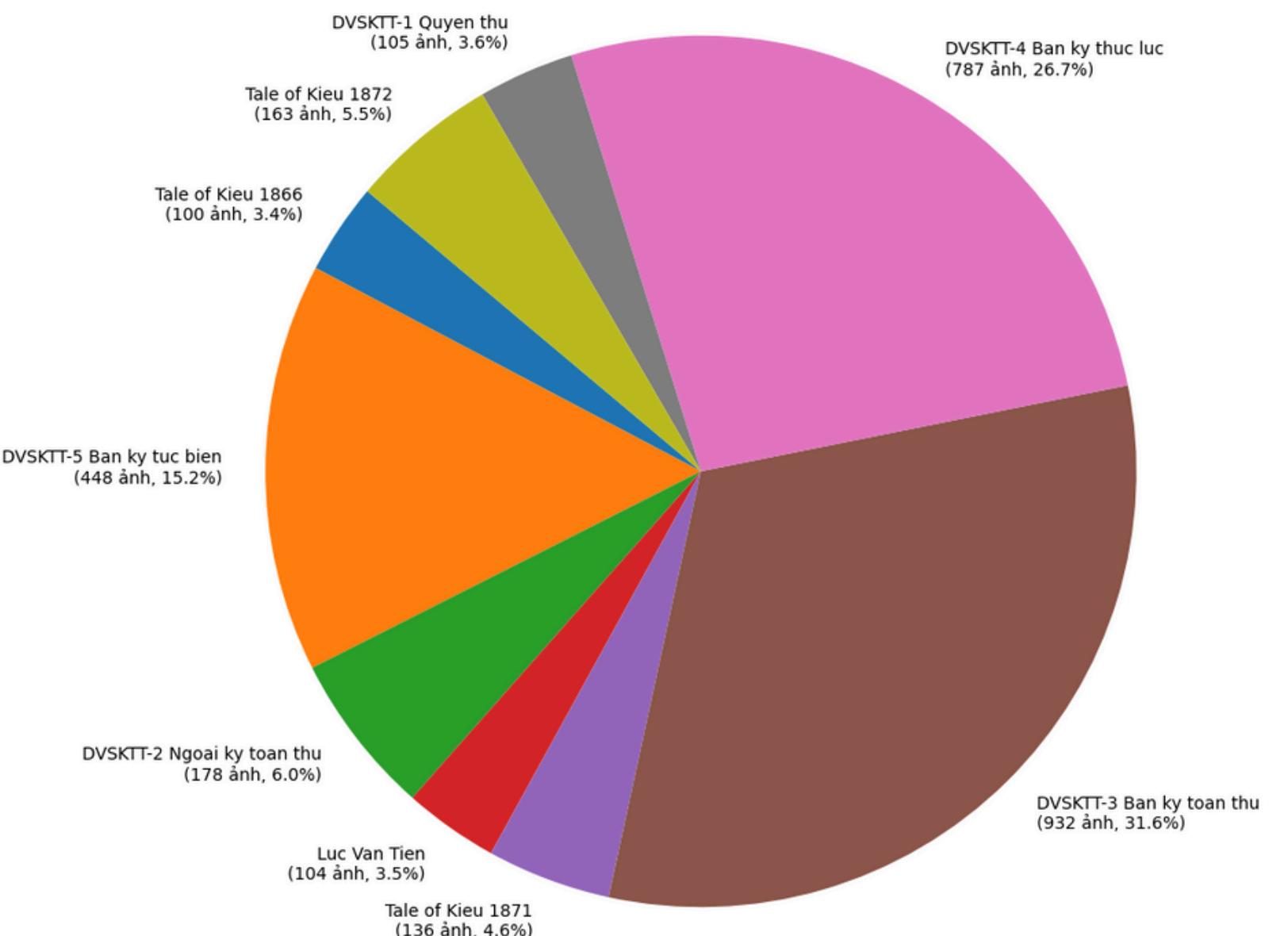
迨至 . [1a\*7\*12]

Đãi chí ...



Trang: 1a

Tổng số ảnh: 2953 trang



# Đặc điểm tài liệu Hán Nôm

慕辭劬揆駄些  
旣戈沒局波橒  
選之彼嗇斯豐  
稿貢客拱燭烟  
浪蔚茄靖朝明  
固茹員外戶王  
沒踰金珂素娥  
放骨格雪精神  
雲祐莊重恪鴻  
翹強色稍漫淥

笄才笄命窖哭怙饑  
仍調鼈覽也笏疸惡  
垂撐涓貝鴈紅打檻  
風情固錄辟傳史撐  
眾方榜湧岱京凭鑛  
家資擬拱常又塲中  
王兌兕笄綏潤儒家  
翠翹兕姊姪笄翠雲  
沒馱沒騷笄困賤苦  
還輸涪還雪讓牟襄  
揭皮才色吏笄分欣

Cấu trúc thường thấy  
trong các tác phẩm thơ lục  
bát

○ 卷之四	屬吳晉宋齊梁紀	起丁未至庚申	附趙燭
前李紀	起辛酉至丁		
士王紀	起丁卯至丙		
士王	在位四十年		

Cấu trúc bất định  
trong Đại Việt Sử kí  
Toàn Thư



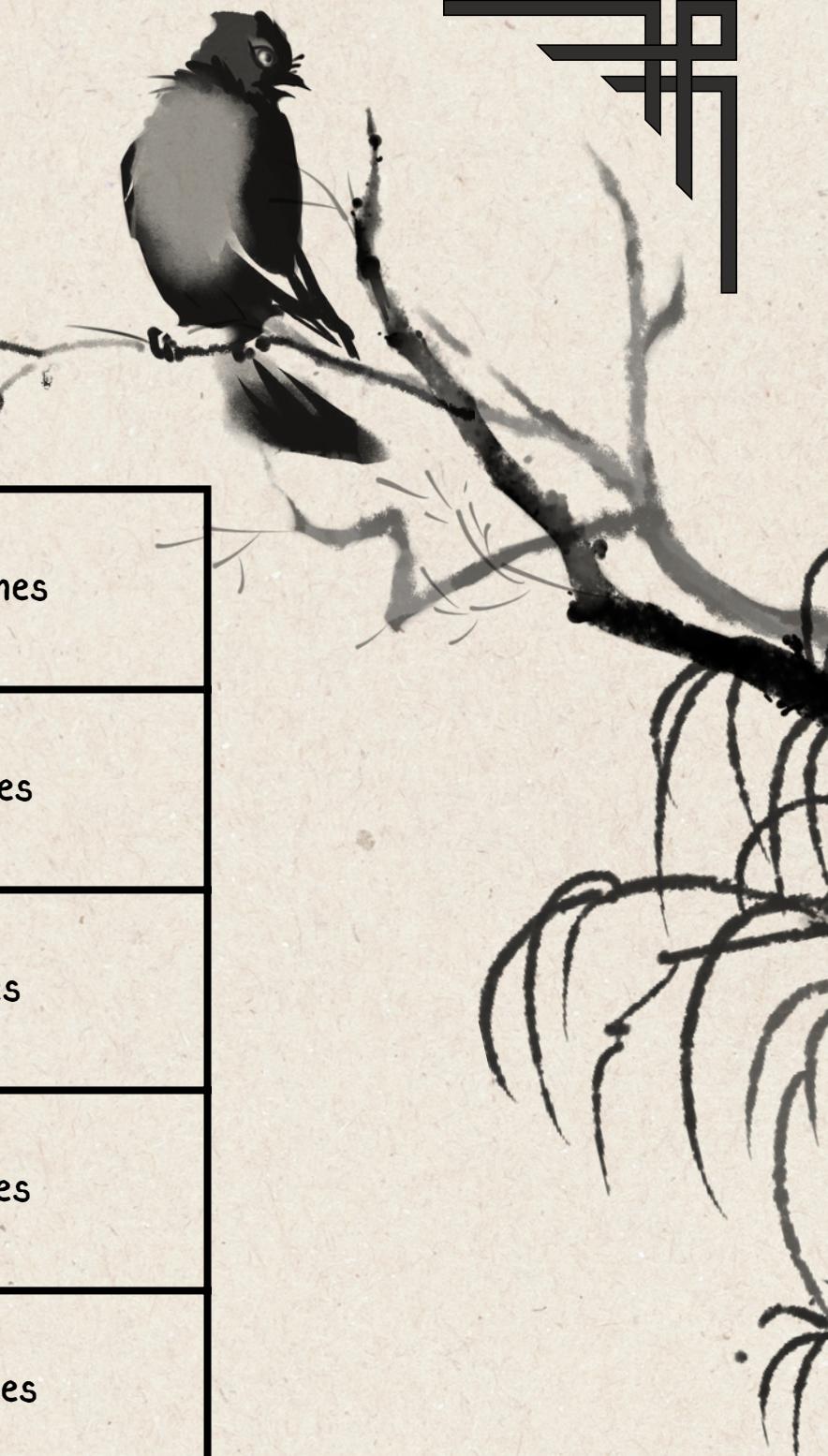
# Bộ dữ liệu cho training

## 1. Đối với task Detection

Training	2359 pages
Validate	357 pages
Test poem	45 pages
Test prose	192 pages
Tổng	2953 pages

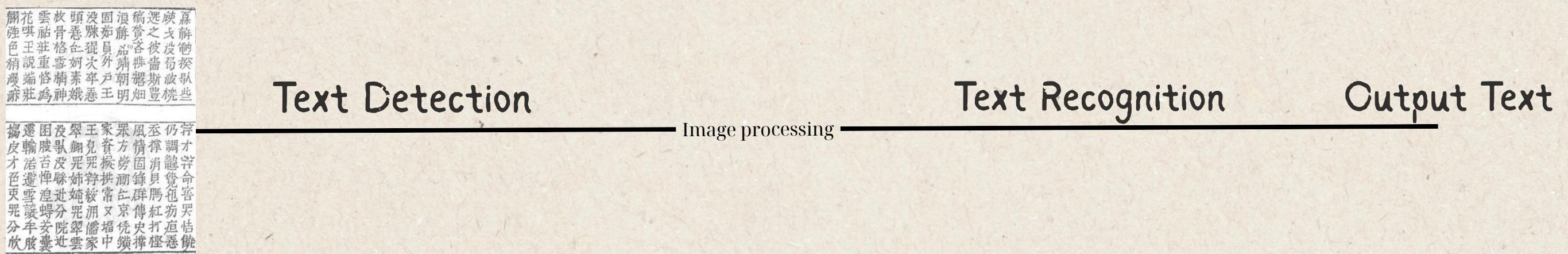
## 2. Đối với task Recognition

Training	30654 patches
Validate	4256 patches
Test poem	860 patches
Test prose	2548 patches
Tổng	38318 patches



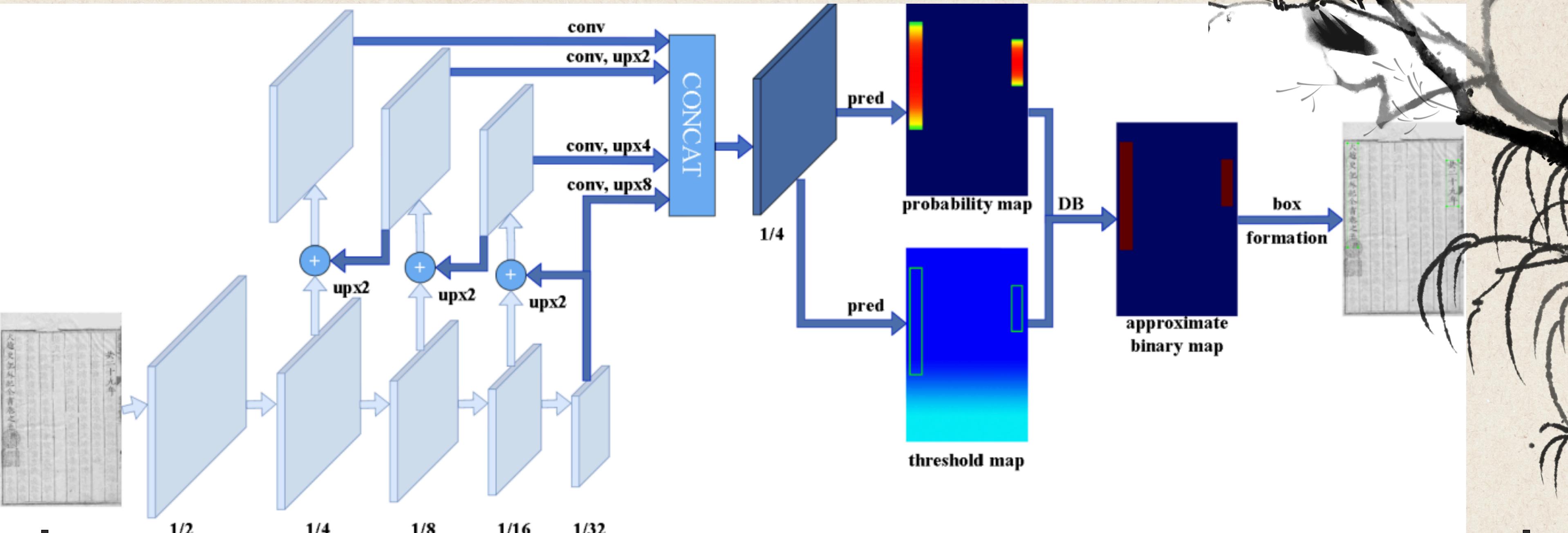
# Pipeline

Nhận thấy sự phát triển của Paddle CCR nên quyết định xây dựng dựa trên bộ công cụ này



# Text Detection

Phương pháp Segmentation-based với DBNet

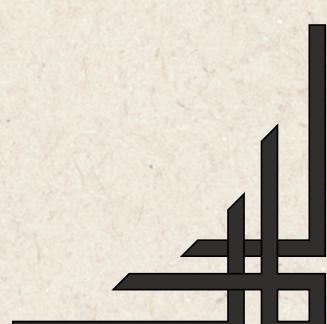


Thiết lập	Thông số
Epoch	15
lr	0.001
warm-up	2
resize	960x960
Post-process	
Thresh	0.3
Box thresh	0.5
Max candidates	1000
Upclip ratio	2.5

# Text Detection

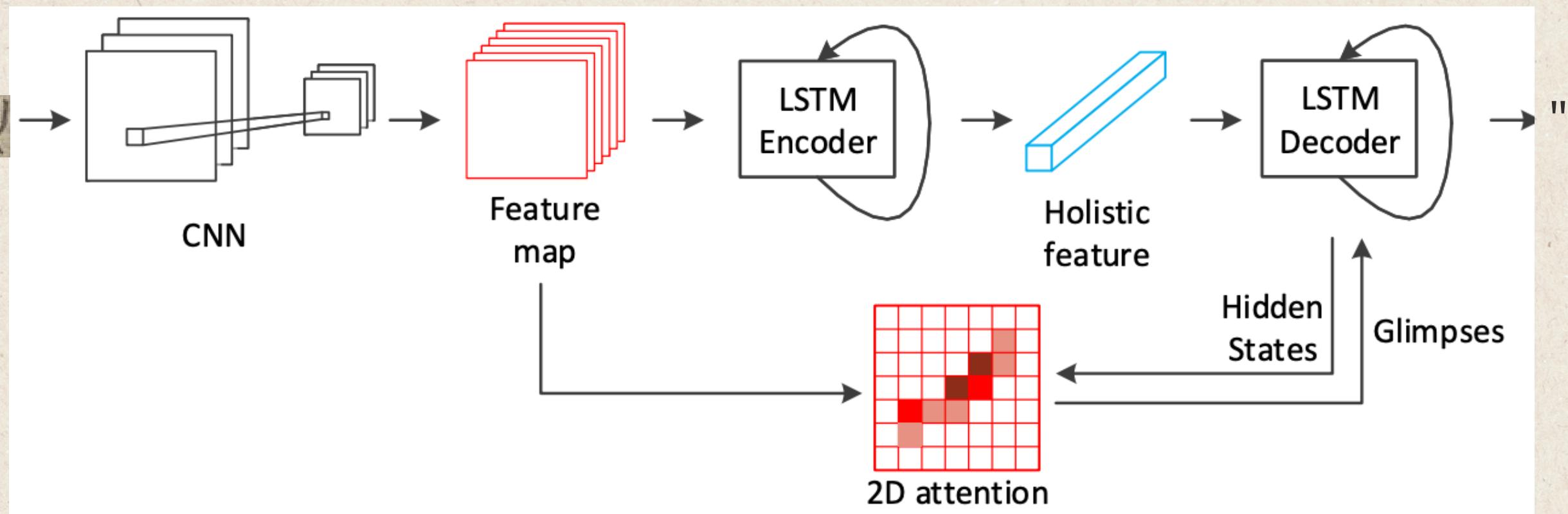
Phương pháp Segmentation-based với DBNet

Cài đặt thử nghiệm trên Paddle OCR



# Text Recognition

Phương pháp Recognition với SAR



# Text Recognition

Phương pháp Recognition với SAR(SHOW, ATEND AND READ)

Cài đặt thử nghiệm trên Paddle CCR

Thiết lập	Thông số
Epoch	50
lr	scheduler(0.01, 0.001, 0.0001)
resize	[3.48,48,160]
width downsample ratio	0.25



# Kết quả thực nghiệm

Phương pháp đánh giá text detection

- Dùng độ đo precision, recall và f1 score để đánh giá. Với phép so sánh giữa nhãn và dự đoán dựa trên IoU (Intersection over Union)

Phương pháp đánh giá recognition

- Dùng Sequence acc và norm edit distance.
- Với norm edit distance dựa trên Levenshtein distance (khoảng cách chỉnh sửa), đo số lượng thao tác tối thiểu (thêm, xóa, thay thế ký tự) để chuyển từ chuỗi dự đoán sang chuỗi thực tế (ground truth) chia cho độ dài chuỗi dài nhất giữa pred và gt.



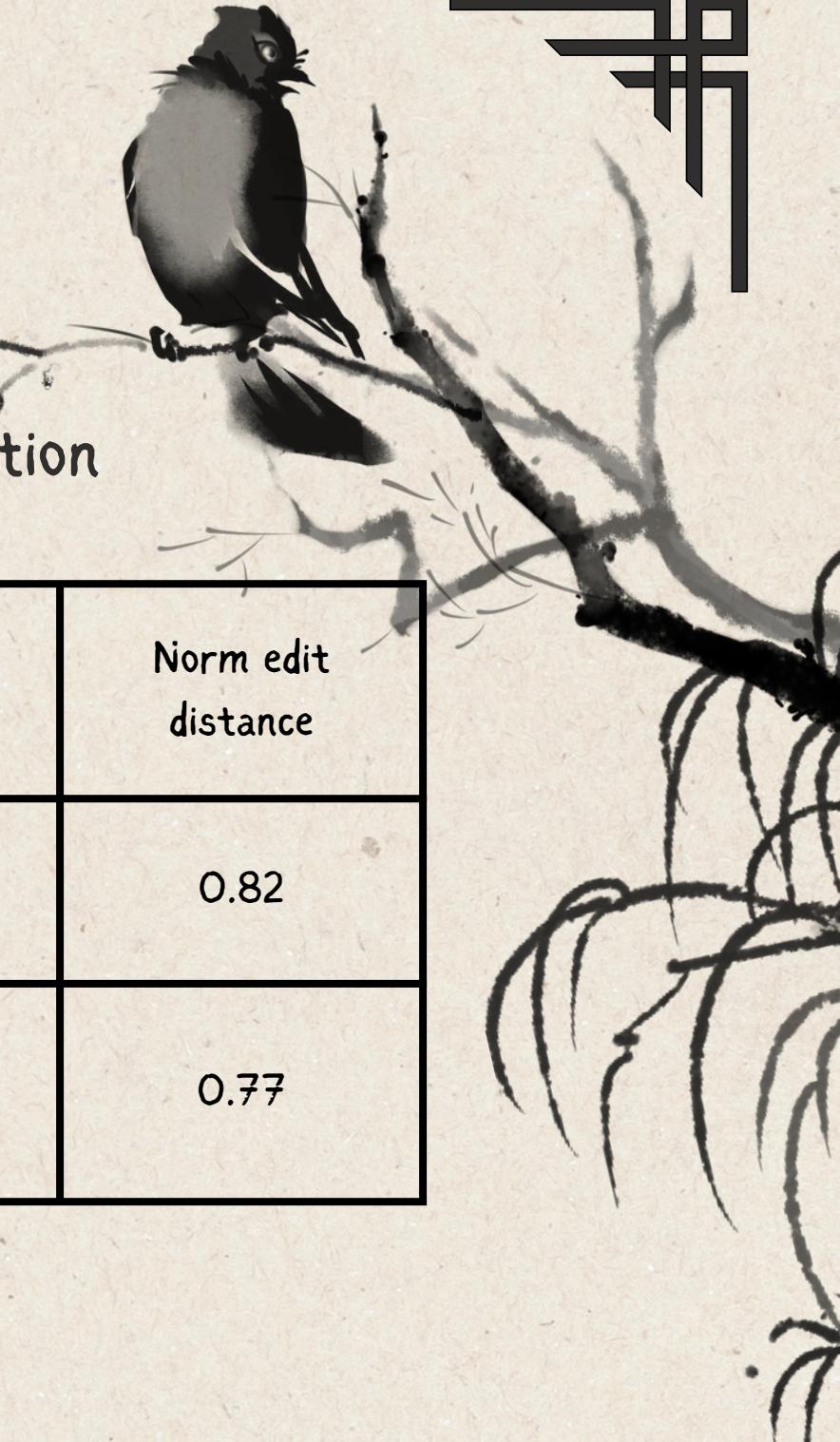
# Kết quả thực nghiệm

Kết quả cho text detection

	Mô hình	Precision	Recall	F1
Thơ	DBNET	0.99	1	0.99
Văn xuôi	DBNET	0.95	0.97	0.96

Kết quả cho text recognition

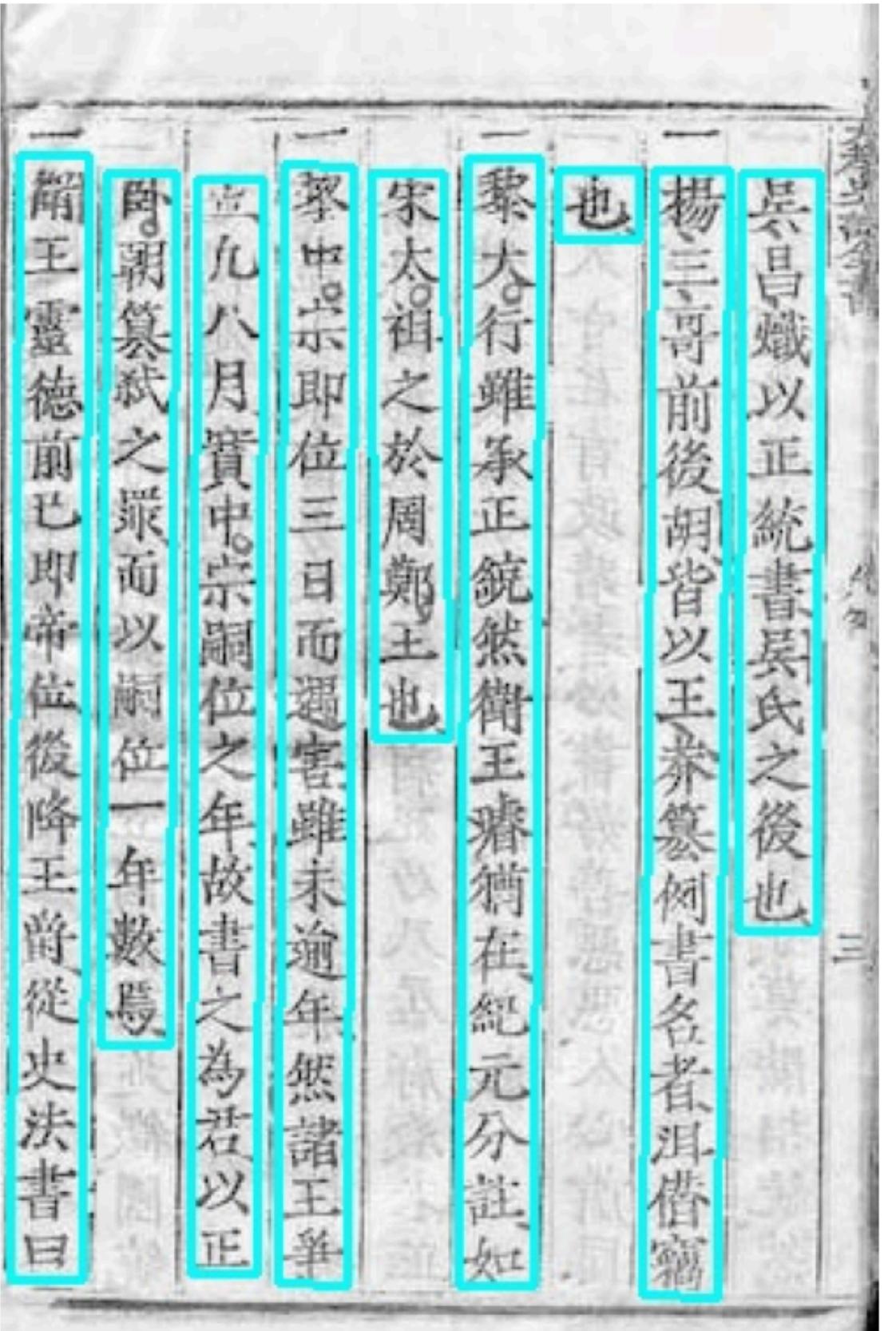
	Mô hình	Sequence Acc	Norm edit distance
Thơ	SAR	0.47	0.82
Văn xuôi	SAR	0.18	0.77



# Demo pipeline end-to-end

Case 0: PRED='吳昌熾以正統書吳是之後也楊三哥前後胡皆以王莽篡例書名者沮僭竊也黎六行雖承正統然衛王璿猶在紀元分註宋太祖之於周鄭王也眾中宗即位三日而遇害雖未逾年然諸王爭立凡八月實中宗嗣位之年故書之爲君以正臥朝篡弑之罪而以嗣位一年數焉衛王靈德前已即帝位後降王爵從史法書曰'

Detection Result - Case 0



# Kết Luận

- Đề tài “OCR cho chữ Nôm trong tài liệu tiếng Việt” không chỉ góp phần vào việc số hóa và bảo tồn di sản văn hóa Việt Nam mà còn mở ra cơ hội nghiên cứu sâu hơn về ứng dụng trí tuệ nhân tạo trong xử lý ngôn ngữ tự nhiên và thị giác máy tính.





Thank you