

# Text Detection and Recognition on Signboards in Vietnamese Street-View Videos

Nguyễn Đình Quân - 20521184, Nguyễn Hùng Phát - 22521074

December 22, 2025

# LỜI CẢM ƠN

Trước tiên, em xin gửi lời cảm ơn chân thành và sâu sắc đến Ban Giám hiệu nhà trường và Khoa Khoa học Máy tính đã tạo điều kiện học tập và nghiên cứu thuận lợi trong suốt thời gian em theo học tại Trường Đại học Công nghệ Thông tin.

Em xin bày tỏ lòng biết ơn đặc biệt đến Thầy Đỗ Văn Tiến, đã trực tiếp giảng dạy và tận tình hướng dẫn em trong quá trình thực hiện đề tài khóa luận. Những định hướng, chỉ dẫn rõ ràng cùng sự hỗ trợ quý báu từ thầy đã là tiền đề quan trọng giúp em hoàn thành tốt công việc nghiên cứu và viết báo cáo đúng tiến độ. Em cũng xin cảm ơn thầy vì đã cung cấp tài liệu, giải đáp thắc mắc và luôn tạo môi trường học tập tích cực, hiệu quả.

Trong suốt quá trình thực hiện đề tài, em đã có cơ hội vận dụng những kiến thức nền tảng đã được học, đồng thời tích cực học hỏi, tìm tòi thêm các kiến thức mới. Đây là một trải nghiệm quý báu giúp em trưởng thành hơn trong tư duy và kỹ năng làm việc nghiên cứu.

Mặc dù đã nỗ lực hoàn thành đề tài với tinh thần nghiêm túc và cầu thị, nhưng do hạn chế về thời gian và kinh nghiệm, khóa luận không thể tránh khỏi những thiếu sót. Em rất mong nhận được sự thông cảm, góp ý chân thành từ các thầy cô để em có thể tiếp tục hoàn thiện và phát triển trong tương lai.

Em xin chân thành cảm ơn!

# TÓM TẮT KHÓA LUẬN

aaaaa.....

# Contents

<b>LỜI CẢM ƠN</b>	<b>i</b>
<b>Tóm tắt khóa luận</b>	<b>ii</b>
<b>Contents</b>	<b>iii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>vii</b>
<b>1 TỔNG QUAN</b>	<b>1</b>
1.1 Đặt vấn đề . . . . .	1
1.2 Mục tiêu và phạm vi . . . . .	7
1.2.1 Mục tiêu . . . . .	7
1.2.2 Phạm vi . . . . .	8
1.3 Đóng góp của khóa luận . . . . .	8
1.4 Cấu trúc khóa luận . . . . .	9
<b>2 CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP TIẾP CẬN</b>	<b>10</b>
2.1 Phát hiện đối tượng . . . . .	10
2.1.1 Cơ sở và hướng tiếp cận chung . . . . .	10
2.1.2 Phương pháp tiếp cận . . . . .	11
2.2 Phát hiện và nhận dạng văn bản ngoại cảnh (Scene Text Detection and Recognition) . . . . .	13

2.2.1	Phát hiện văn bản ngoại cảnh (Scene Text Detection) . . . . .	13
2.2.1.1	Cơ sở và hướng tiếp cận chung . . . . .	13
2.2.1.2	Phương pháp tiếp cận . . . . .	14
2.2.2	Nhận dạng văn bản (Text Recognition) . . . . .	15
2.2.2.1	Cơ sở và hướng tiếp cận chung . . . . .	15
2.2.2.2	Phương pháp tiếp cận . . . . .	15
2.2.3	End-to-End (End-to-End Text Recognition) . . . . .	15
2.2.3.1	Cơ sở và hướng tiếp cận chung . . . . .	15
2.2.3.2	Phương pháp tiếp cận . . . . .	15
<b>3</b>	<b>PHƯƠNG PHÁP</b>	<b>16</b>
3.1	Hệ thống phát hiện và nhận dạng chữ trên biển hiệu . . . . .	16
3.1.1	Bài toán và mục tiêu . . . . .	16
3.1.2	Kiến trúc tổng thể của pipeline . . . . .	17
3.1.3	Quy ước biểu diễn hình học (BBox/OBB/Mask) . . . . .	17
3.1.4	Chiến lược “phát hiện biển hiệu trước” . . . . .	18
3.2	Mô-đun phát hiện biển hiệu (Signboard Detection/Segmentation) . . . . .	18
3.2.1	Phát hiện biển hiệu bằng object detection . . . . .	18
3.2.1.1	Lý do ưu tiên OBB cho biển hiệu . . . . .	18
3.2.1.2	Thiết lập fine-tune . . . . .	19
3.2.2	Phân đoạn biển hiệu bằng segmentation . . . . .	19
3.2.2.1	Tạo patch biển hiệu từ mask . . . . .	20
3.3	Chuẩn hóa hình học (Align/Rectify) . . . . .	20
3.3.1	Động cơ của bước Align . . . . .	20
3.3.2	Align dựa trên OBB/4 điểm . . . . .	20
3.4	Mô-đun phát hiện văn bản (Text Detection) . . . . .	20
3.4.1	Bài toán phát hiện văn bản trong patch biển hiệu . . . . .	20
3.4.2	Lựa chọn hướng tiếp cận: segmentation-based vs OBB-based . .	21
3.4.3	Quy trình fine-tune text detector . . . . .	21
3.4.4	Chỉ số đánh giá cho text detection . . . . .	21

3.5	Mô-đun nhận dạng văn bản (Text Recognition) . . . . .	22
3.5.1	Bài toán nhận dạng chuỗi ký tự . . . . .	22
3.5.2	Tiền xử lý patch chữ . . . . .	22
3.5.3	Mô hình nhận dạng và lý do lựa chọn . . . . .	22
3.5.4	Chỉ số đánh giá cho text recognition . . . . .	22
3.6	Hậu xử lý và chuẩn hoá tiếng Việt . . . . .	23
3.6.1	Chuẩn hoá Unicode và lọc nhiễu ký tự . . . . .	23
3.6.2	Gộp kết quả theo dòng/khối (tuỳ chọn) . . . . .	23
3.7	Tổng hợp cấu hình thực nghiệm trong pipeline . . . . .	23
3.7.1	Nguyên tắc chọn mô hình cho pipeline cuối . . . . .	23
3.7.2	Các biến thể pipeline được so sánh . . . . .	24
3.8	Tóm tắt chương . . . . .	24

<b>4</b>	<b>THỰC NGHIỆM VÀ ĐÁNH GIÁ</b>	<b>27</b>
4.1	Dữ liệu . . . . .	27
4.1.1	Nguồn dữ liệu . . . . .	27
4.1.2	Thống kê bộ dữ liệu dùng trong khóa luận . . . . .	27
4.1.3	Đặc trưng hình học và kích thước mẫu . . . . .	28
4.1.4	Tỷ lệ văn bản thuộc vùng biển hiệu . . . . .	28
4.1.5	Chia tập Train/Validation/Test . . . . .	28
4.2	Tiền xử lý . . . . .	29
4.2.1	Chuẩn hóa nhãn và định dạng . . . . .	29
4.2.2	Sinh dữ liệu cho Text Recognition . . . . .	29
4.2.3	Tiền xử lý tăng tính tổng quát . . . . .	29
4.3	Mô hình thực nghiệm . . . . .	30
4.3.1	Bài toán phát hiện biển hiệu . . . . .	30
4.3.2	Bài toán phát hiện văn bản . . . . .	30
4.3.3	Bài toán nhận dạng văn bản . . . . .	30
4.3.4	Bài toán Text Spotting (end-to-end và two-stage) . . . . .	30
4.4	Độ đo đánh giá . . . . .	30

4.4.1	Signboard detection (BBox/OBB) . . . . .	30
4.4.2	Signboard segmentation . . . . .	31
4.4.3	Text detection . . . . .	31
4.4.4	Text recognition . . . . .	31
4.4.5	Text spotting . . . . .	31
4.5	Kết quả thực nghiệm . . . . .	32
4.5.1	Kết quả phát hiện biển hiệu (BBox/OBB) . . . . .	32
4.5.2	Kết quả phân đoạn biển hiệu (Segmentation) . . . . .	32
4.5.3	Kết quả text detection (mô hình tiền huấn luyện) . . . . .	32
4.5.4	Kết quả text detection sau fine-tune . . . . .	33
4.5.5	Kết quả text recognition (mô hình tiền huấn luyện) . . . . .	33
4.5.6	Kết quả text recognition sau fine-tune . . . . .	33
4.5.7	Kết quả text spotting . . . . .	33
4.5.8	Dánh giá pipeline: Text detection trên vùng biển hiệu . . . . .	33
4.5.9	Dánh giá pipeline: Text Recognition trên vùng biển hiệu . . . . .	34
<b>5</b>	<b>KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN</b>	<b>36</b>
5.1	Kết luận . . . . .	36
5.2	Hướng phát triển . . . . .	36

# List of Figures

1.1	Văn bản trong ảnh ngoại cảnh . . . . .	2
1.2	Văn bản trên biển hiệu . . . . .	3
1.3	Hình ảnh minh họa sự đa dạng về phông chữ, kích thước, chữ nghệ thuật, và đa ngôn ngữ [?] . . . . .	4
1.4	Hình ảnh minh họa sự đa dạng về hình dạng, kích thước, vật liệu và vị trí của biển hiệu, dựa trên bộ dữ liệu SignboardText [?] . . . . .	5
1.5	Hình ảnh minh họa đầu vào và đầu ra của bài toán. Bài toán nhận đầu vào là ảnh hoặc khung hình video và trả về danh sách các biển hiệu, trong đó mỗi biển hiệu được xác định bằng vùng chứa (bounding region) và nội dung văn bản tương ứng đã được nhận dạng. . . . .	6
3.1	Kiến trúc tổng thể hệ thống đọc chữ trên biển hiệu: Signboard Detection/Segmentation → (Align/Rectify) → Text Detection → Text Recognition → Post-processing.	25
3.2	Minh họa bước Align/Rectify dựa trên 4 điểm (OBB) để giảm nghiêng và phối cảnh trước khi phát hiện/nhận dạng chữ. . . . .	26

# List of Tables

4.1	Thống kê kích thước (px) trên SignboardText dùng trong khóa luận. . . . .	28
4.2	Tỷ lệ văn bản nằm trong vùng biển hiệu (%). . . . .	28
4.3	Chia tập dữ liệu theo nhiệm vụ. . . . .	29
4.4	Kết quả signboard detection với <b>rectangle bounding box</b> . . . . .	32
4.5	Kết quả signboard detection với <b>oriented bounding box (OBB)</b> . . . . .	32
4.6	Kết quả signboard segmentation. . . . .	33
4.7	So sánh các mô hình text detection tiền huấn luyện trên SignboardText. .	33
4.8	Kết quả text detection sau fine-tune trên SignboardText. . . . .	34
4.9	So sánh các mô hình text recognition tiền huấn luyện trên SignboardText.	34
4.10	Kết quả text recognition sau fine-tune (Exact-match và Norm-match). .	35
4.11	Kết quả text spotting (Hmean %) với one-stage và two-stage. . . . .	35
4.12	Kết quả Text Detection trên vùng biển hiệu với các biến thể pipeline. .	35
4.13	Kết quả Text Recognition trên vùng biển hiệu (pipeline). . . . .	35

# Chapter 1

## TỔNG QUAN

### 1.1 Đặt vấn đề

Phát hiện và nhận dạng văn bản trong ảnh ngoại cảnh (*Scene Text Detection and Recognition – STDR*) là một bài toán quan trọng trong thị giác máy tính, thu hút nhiều sự quan tâm nhờ tính ứng dụng rộng rãi như dịch tự động, hỗ trợ dẫn đường, hay phân tích biển báo giao thông. Với đầu vào là ảnh tĩnh hoặc các khung hình video, bài toán STDR hướng tới việc xác định vị trí xuất hiện và nội dung của văn bản (Hình 1.1).

Trong số các loại văn bản ngoại cảnh, **văn bản trên biển hiệu** (Hình 1.2) có ý nghĩa đặc biệt do thường chứa các thông tin quan trọng như *tên địa điểm*, *cơ sở kinh doanh* hoặc *loại hình dịch vụ*. Chính vì vậy, bài toán **phát hiện và nhận dạng văn bản trên biển hiệu** (*Text Detection and Recognition on Signboard*) trở thành một nhánh nghiên cứu quan trọng của STDR, với nhiều tiềm năng ứng dụng trong hệ thống dẫn đường thông minh, phân tích thông tin đô thị, và bổ sung thông tin ngữ nghĩa cho bản đồ số.

Tuy nhiên, bài toán phát hiện và nhận dạng văn bản trên biển hiệu đặt ra nhiều thách thức. Thách thức đầu tiên xuất phát từ đặc điểm của văn bản, như sự đa dạng về phông chữ, kích thước, hướng, bô cục; văn bản có thể bị nghiêng, cong, chồng chép hoặc hòa lẫn vào nền phức tạp, cùng với các phong cách thiết kế nghệ thuật và yếu tố đa ngôn ngữ (Hình 1.3). Đặc biệt đối với tiếng Việt, khó khăn còn gia tăng do hệ thống dấu thanh (sắc, huyền, hỏi, ngã, nặng) và các ký tự đặc biệt (ô, ê, ă, â, ơ, ư), làm tăng đáng kể tập ký tự cần nhận dạng và dễ gây nhầm lẫn giữa các chữ có hình dáng tương tự (ví dụ giữa



Hình 1.1: Văn bản trong ảnh ngoại cảnh

*a, â, ă, á).*

Thách thức thứ hai bắt nguồn từ đặc điểm của biển hiệu và bối cảnh môi trường xung quanh, biển hiệu đa dạng về hình dạng, kích thước, vật liệu và thường xuất hiện ở các vị trí phức tạp trong ảnh (Hình ??), chẳng hạn như bị che khuất một phần, chịu ảnh hưởng của phản xạ ánh sáng, hoặc nằm trong các bối cảnh đong đúc. Theo khảo sát các nghiên cứu hiện có, cho đến nay mới chỉ có một nghiên cứu [?] tập trung vào phát hiện biển hiệu trên đường phố Việt Nam, trong khi hướng tiếp cận kết hợp cả phát hiện đối tượng biển hiệu lẫn nhận dạng nội dung văn bản trên đó vẫn còn rất ít được khai thác.

Hơn nữa, khi mở rộng phạm vi từ ảnh tĩnh sang **video hành trình**, bài toán còn phải đổi mới với những thách thức đặc thù như hiện tượng mờ do chuyển động, chất lượng hình ảnh bị giới hạn bởi camera hành trình, cùng với sự biến đổi liên tục về điều kiện ánh sáng và góc quay. Những yếu tố này khiến nhiệm vụ phát hiện và nhận dạng văn



Hình 1.2: Văn bản trên biển hiệu

bản trong video trở nên phức tạp hơn nhiều so với trên ảnh đơn lẻ.

Từ những thách thức nêu trên, bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong bối cảnh **video hành trình** có thể được định nghĩa một cách cụ thể như sau (hình ảnh minh họa trực quan tại Hình 1.5):

- **Đầu vào (Input):** Các hình ảnh hoặc khung hình thực tế được trích xuất từ video camera hành trình trên đường phố Việt Nam, chứa các cảnh có biển hiệu trong nhiều điều kiện khác nhau, bao gồm ban ngày/ban đêm, trời nắng/mưa và các góc nhìn đa dạng.
- **Đầu ra (Output):** Đối với mỗi hình ảnh (hoặc khung hình video) đầu vào, bài toán cần trả về hai thông tin chính:



Hình 1.3: Hình ảnh minh họa sự đa dạng về phông chữ, kích thước, chữ nghệ thuật, và đa ngôn ngữ [?]

- **Vị trí của biển hiệu:** Danh sách các vùng (bounding regions) xác định vùng chứa biển hiệu trong ảnh.
- **Thông tin văn bản trên từng biển hiệu:** Ứng với mỗi biển hiệu, cung cấp vị trí và nội dung văn bản đã được nhận dạng trên biển hiệu đó.

*(Kết quả đầu ra có thể được trực quan hóa trực tiếp trên ảnh đầu vào hoặc tích hợp để xử lý liên tục cho luồng video.)*

Trước những thách thức thực tế và dựa trên các kết quả nghiên cứu trước đây cho thấy rằng hướng nghiên cứu kết hợp (phát hiện biển hiệu và nhận dạng văn bản) vẫn còn ít được khai thác, khóa luận này đặt ra mục tiêu phát triển một **pipeline end-to-end** cho bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong video được quay bởi camera hành trình trên đường phố. Pipeline hướng tới việc:

- Xác định vùng chứa biển hiệu (signboard detection) và vùng chứa văn bản bên trong mỗi biển hiệu (text detection) trong từng khung hình video.



Hình 1.4: Hình ảnh minh họa sự đa dạng về hình dạng, kích thước, vật liệu và vị trí của biển hiệu, dựa trên bộ dữ liệu SignboardText [?]

- Trích xuất và chuyển đổi nội dung văn bản từ các vùng văn bản đã phát hiện thành dạng văn bản có thể đọc được, hỗ trợ hai ngôn ngữ chính là tiếng Việt và tiếng Anh, hướng tới việc cung cấp thông tin đầu ra có ích cho các tác vụ truy xuất hoặc khai thác thông tin trong tương lai.

Để đạt được các mục tiêu trên, khóa luận sẽ tiến hành khảo sát, thực nghiệm so sánh và lựa chọn các phương pháp tiên tiến nay cho từng tác vụ con, đồng thời so sánh hai hướng tiếp cận chính cho bài toán text spotting. Các phương pháp cụ thể được xem xét bao gồm:

- **Phát hiện biển hiệu (Signboard Detection):** Các biến thể YOLO [?], DETR [?],



Hình 1.5: Hình ảnh minh họa đầu vào và đầu ra của bài toán. Bài toán nhận đầu vào là ảnh hoặc khung hình video và trả về danh sách các biển hiệu, trong đó mỗi biển hiệu được xác định bằng vùng chứa (bounding region) và nội dung văn bản tương ứng đã được nhận dạng.

RTDETR v2 [?], SegFormer [?], Mask2Former [?].

- **Phát hiện văn bản (Text Detection):** PANet [?], DBNet++ [?], TextPMs [?], FAST [?], KPN [?].
- **Nhận dạng văn bản (Text Recognition):** ViTSTR [?], PARSeq [?], CDistNet [?], SMTR [?], SVTRv2 [?]
- **Text Spotting (End-to-End):** TESTR [?], DeepSolo [?], UNITS [?], DNTextSpotter [?]

Trên cơ sở kết quả đánh giá và so sánh từ thực nghiệm cho từng tác vụ con, một pipeline end-to-end sẽ được xây dựng bằng cách lựa chọn phương pháp tối ưu cho mỗi tác vụ và xác định kiến trúc hiệu quả nhất cho giai đoạn xử lý văn bản thông qua so sánh hướng tiếp cận two-stage (tích hợp các phương pháp phát hiện và nhận dạng văn bản đã chọn) với các mô hình end-to-end tiên tiến.

## 1.2 Mục tiêu và phạm vi

### 1.2.1 Mục tiêu

Trong khóa luận này, sinh viên đề ra các mục tiêu như sau:

- Mở rộng và chuẩn bị tập dữ liệu ảnh tĩnh SignboardText [?] bằng cách bổ sung nhãn đối tượng biển hiệu (*signboard*), nhằm hỗ trợ đánh giá bài toán phát hiện biển hiệu.
- Thực nghiệm, so sánh và đánh giá một số phương pháp tiên tiến nay cho từng tác vụ con (phát hiện biển hiệu, phát hiện văn bản, nhận dạng văn bản) trên tập dữ liệu được chuẩn bị, từ đó rút ra ưu điểm, nhược điểm của từng phương pháp.
- Xây dựng một pipeline end-to-end cho bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong video hành trình tại Việt Nam.

### 1.2.2 Phạm vi

Phạm vi của khóa luận được giới hạn nhằm đảm bảo tính tập trung và khả thi, bao gồm các công việc sau:

- Mở rộng tập dữ liệu tập trung vào việc bổ sung nhãn đối tượng biển hiệu (signboard annotation) cho tập dữ liệu ảnh tĩnh SignboardText hiện có. Dữ liệu video được thu thập chỉ nhằm mục đích minh họa và kiểm tra tính tổng quát của mô hình, với điều kiện chính là ban ngày. Các tình huống phức tạp (ban đêm, thời tiết xấu) không nằm trong phạm vi xem xét.
- Khảo sát và thực nghiệm được giới hạn trong một tập hợp các phương pháp tiên tiến cho các hướng tiếp cận phổ biến và hiệu quả hiện nay. Việc so sánh không bao quát toàn bộ các phương pháp trong lĩnh vực, mà tập trung vào những phương pháp phù hợp và khả thi với dữ liệu và mục tiêu của khóa luận.
- Pipeline end-to-end tập trung vào bài toán phát hiện và nhận dạng văn bản trên biển hiệu và hướng tới việc cung cấp thông tin đầu ra vị trí và nội dung văn bản, làm cơ sở cho các tác vụ truy xuất thông tin trong tương lai.

## 1.3 Đóng góp của khóa luận

Các đóng góp chính của khóa luận bao gồm:

- **Mở rộng bộ dữ liệu:** Bổ sung nhãn đối tượng biển hiệu (signboard bounding box) cho tập dữ liệu ảnh tĩnh SignboardText [?], hỗ trợ thực nghiệm và đánh giá cho bài toán phát hiện biển hiệu. Đồng thời, thu thập một tập dữ liệu video hành trình thực tế để phục vụ minh họa và kiểm tra tính tổng quát.
- **Thực nghiệm và đánh giá:** Tiến hành cài đặt, thực nghiệm và so sánh một số phương pháp tiên tiến cho ba tác vụ thành phần: phát hiện biển hiệu, phát hiện văn bản và nhận dạng văn bản. Kết quả đánh giá đi kèm phân tích ưu/nhược điểm cụ thể trong bối cảnh dữ liệu tiếng Việt và cảnh quan đường phố.

- **Phát triển pipeline end-to-end:** Trên cơ sở kết quả thực nghiệm, phát triển một pipeline cho bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong video hành trình trên đường phố Việt Nam. Pipeline hướng tới việc cung cấp đầu ra vị trí và nội dung văn bản, làm cơ sở cho các tác vụ truy xuất thông tin trong tương lai.

## 1.4 Cấu trúc khóa luận

Nội dung khóa luận được tổ chức như sau:

**Chương 1:** Tổng quan bài toán, bối cảnh, động lực, mục tiêu, phạm vi và đóng góp.

**Chương 2:** Cơ sở lý thuyết và các nghiên cứu liên quan đến phát hiện biển hiệu, phát hiện/nhận dạng văn bản và các kỹ thuật xử lý video.

**Chương 3:** Các phương pháp và pipeline đề xuất cho bài toán phát hiện và nhận dạng văn bản biển hiệu trong video, bao gồm mô tả kiến trúc hệ thống và mô-đun xử lý.

**Chương 4:** Thực nghiệm và đánh giá trên tập dữ liệu SignboardText mở rộng và dữ liệu video hành trình; phân tích kết quả và thảo luận.

**Chương 5:** Xây dựng ứng dụng minh họa và mô tả các chức năng khai thác thông tin văn bản biển hiệu.

**Chương 6:** Kết luận và hướng phát triển trong tương lai.

# Chapter 2

## CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP TIẾP CẬN

### 2.1 Phát hiện đối tượng

#### 2.1.1 Cơ sở và hướng tiếp cận chung

Phát hiện đối tượng (Object Detection) là một bài toán trong lĩnh vực Thị giác Máy tính (Computer Vision), đóng vai trò trung tâm trong nhiều ứng dụng thực tiễn như giám sát an ninh, lái xe tự động, và tương tác người máy. Khác với nhiệm vụ phân loại ảnh truyền thống vốn chỉ xác định loại đối tượng xuất hiện trong toàn bộ ảnh, phát hiện đối tượng yêu cầu mô hình không chỉ nhận diện đúng loại đối tượng mà còn xác định chính xác vị trí của chúng thông qua các hộp giới hạn (bounding boxes). Thách thức của bài toán này nằm ở việc phải xử lý đồng thời nhiều đối tượng với sự đa dạng lớn về kích thước, tư thế, góc nhìn, điều kiện ánh sáng và mức độ chồng lấn giữa các đối tượng.

Dựa trên tổng quan của [?], các phương pháp phát hiện đối tượng hiện đại có thể được phân loại thành ba hướng tiếp cận chính xét theo kiến trúc và quy trình xử lý:

- **Các phương pháp hai giai đoạn (Two-stage)** hoạt động dựa trên nguyên tắc tách biệt quá trình đề xuất vùng (region proposal) và phân loại. Nhóm này tiêu biểu bởi các mô hình thuộc họ R-CNN, chẳng hạn như R-CNN [?], Fast R-CNN [?] và Faster R-CNN [?]. Các phương pháp này thường đạt độ chính xác cao nhưng đòi hỏi chi phí tính toán lớn và tốc độ xử lý chậm.

- **Các phương pháp một giai đoạn (One-stage)** thực hiện trực tiếp việc dự đoán lớp và vị trí mà không có bước đề xuất vùng riêng biệt, với các đại diện nổi bật như YOLO [?], SSD [?] và RetinaNet [?]. Cách tiếp cận này giúp cân bằng tốt hơn giữa tốc độ và độ chính xác, phù hợp với các ứng dụng thời gian thực.
- **Các phương pháp dựa trên Transformer** gần đây tạo ra bước đột phá với kiến trúc end-to-end, loại bỏ sự phụ thuộc vào các thành phần được thiết kế thủ công (hand-crafted) như anchor và thuật toán Non-Maximum Suppression (NMS). Điển hình cho hướng đi này là mô hình DETR [?] và các biến thể tối ưu hóa tốc độ của nó.

### 2.1.2 Phương pháp tiếp cận

Trong bối cảnh của khóa luận này, bài toán phát hiện biển hiệu đòi hỏi sự cân bằng giữa tốc độ xử lý, độ chính xác và khả năng xử lý các đối tượng có hướng (oriented objects). Vì vậy, khóa luận tập trung lựa chọn và đánh giá một số phương pháp tiên tiến hiện nay, tiêu biểu cho các hướng tiếp cận khác nhau, dựa trên mức độ phù hợp với các yêu cầu của bài toán.

- **YOLO (You Only Look Once) [?]:** YOLO là đại diện tiêu biểu cho hướng tiếp cận một giai đoạn (one-stage). Kiến trúc của YOLO dựa trên việc chia ảnh đầu vào thành một lưới (grid), mỗi ô lưới chịu trách nhiệm dự đoán đồng thời các bounding box và xác suất lớp. Cách tiếp cận trực tiếp này mang lại tốc độ suy luận rất cao, phù hợp cho các ứng dụng thời gian thực. Đặc biệt, biến thể YOLO-OBB (Oriented Bounding Box) mở rộng khả năng phát hiện vật thể xoay, là một lựa chọn rất phù hợp cho bài toán phát hiện biển hiệu.
- **DETR (DEtection TRansformer) [?]:** DETR là mô hình phát hiện đối tượng đầu tiên hoàn toàn dựa trên kiến trúc Transformer. Mô hình sử dụng một tập hợp cố định các "truy vấn" (object queries) để tương tác với đặc trưng hình ảnh và dự đoán trực tiếp một tập các bounding box, nhờ đó loại bỏ hoàn toàn nhu cầu về các thành phần thủ công như anchor boxes và NMS.

- **RTDETRv2 [?]:** RT-DETRv2 là phiên bản cải tiến từ RT-DETR, được đề xuất với mục tiêu tối ưu hóa hiệu suất thời gian thực (real-time performance) trong khi vẫn duy trì độ chính xác cao. Mô hình này giữ nguyên ưu điểm end-to-end của DETR, loại bỏ sự phụ thuộc vào NMS, và được tối ưu hóa thông qua các cơ chế như hybrid encoder cùng cơ chế lựa chọn truy vấn (query selection) nhằm cân bằng giữa độ chính xác và hiệu quả tính toán.

Bên cạnh các phương pháp phát hiện trực tiếp dựa trên bounding box, để mở rộng góc nhìn đánh giá, khóa luận xem xét một hướng tiếp cận gián tiếp thông qua bài toán phân đoạn ngữ nghĩa (semantic segmentation). Theo hướng tiếp cận này, đối tượng trước hết được phân đoạn ở mức điểm ảnh, từ đó suy ra vùng bao hình học của đối tượng, phục vụ cho bài toán phát hiện. Trong bối cảnh đó, theo tổng quan của [?], các kiến trúc dựa trên Transformer đã trở thành một hướng tiếp cận được quan tâm rộng rãi và ngày càng quan trọng trong bài toán phân đoạn ảnh, đặc biệt là phân đoạn ngữ nghĩa. Nhờ khả năng mô hình hóa ngữ cảnh toàn cục thông qua cơ chế self-attention, các mô hình này cho thấy hiệu quả nổi bật trong việc xử lý các kịch bản có cấu trúc phức tạp và sự đa dạng lớn về hình dạng đối tượng.

- **SegFormer [?]:** SegFormer là một kiến trúc phân đoạn ngữ nghĩa hiệu quả dựa trên Transformer, kết hợp giữa encoder Transformer phân cấp và decoder MLP nhẹ. Thiết kế này cho phép mô hình khai thác ngữ cảnh toàn cục thông qua self-attention, đồng thời duy trì hiệu quả tính toán cao nhờ cấu trúc decoder đơn giản. Nhờ đó, SegFormer đạt được sự cân bằng tốt giữa độ chính xác và tốc độ suy luận, phù hợp với các ứng dụng phân đoạn ngữ nghĩa trong bối cảnh thực tế.
- **Mask2Former [?]:** Mask2Former là một kiến trúc phân đoạn dựa trên Transformer theo hướng tiếp cận thống nhất (unified framework), có khả năng xử lý nhiều bài toán phân đoạn khác nhau như phân đoạn ngữ nghĩa (semantic segmentation), phân đoạn theo thể hiện (instance segmentation) và phân đoạn toàn cảnh (panoptic segmentation). Mask2Former áp dụng cơ chế masked attention, cho phép mô hình tập trung vào các vùng đối tượng tiềm năng thông qua các mask dự đoán, từ đó

cải thiện khả năng biểu diễn hình dạng và biên đối tượng. Cách tiếp cận này giúp Mask2Former thể hiện hiệu quả trên các trường hợp có đối tượng chồng lấn hoặc hình dạng phức tạp.

## 2.2 Phát hiện và nhận dạng văn bản ngoại cảnh (Scene Text Detection and Recognition)

### 2.2.1 Phát hiện văn bản ngoại cảnh (Scene Text Detection)

#### 2.2.1.1 Cơ sở và hướng tiếp cận chung

Phát hiện văn bản (Text Detection) trong ảnh ngoại cảnh hướng tới mục tiêu xác định và khoanh vùng các khu vực chứa văn bản. Khác với các tác vụ phát hiện đối tượng truyền thống, phát hiện văn bản trong ảnh ngoại cảnh phải đổi mới với nhiều thách thức do sự đa dạng về hình dạng, kích thước, hướng và bố cục của văn bản, cũng như các trường hợp văn bản bị nghiêng, cong, chồng chéo hoặc mờ. Do đó, bài toán này đòi hỏi kết hợp các kỹ thuật phát hiện đối tượng với các phương pháp chuyên biệt cho văn bản nhằm xác định chính xác và hiệu quả các vùng chứa văn bản.

Dựa trên các nghiên cứu khảo sát gần đây [?, ?, ?] về phát hiện văn bản trong ảnh ngoại cảnh, các phương pháp tiên tiến hiện nay có thể được phân thành ba hướng tiếp cận chính: dựa trên hồi quy (regression-based), dựa trên phân đoạn (segmentation-based) và dựa trên thành phần liên thông (connected component-based).

- **Các phương pháp dựa trên hồi quy (Regression-based):** Hướng tiếp cận này giải quyết bài toán phát hiện văn bản tương tự như phát hiện đối tượng, bằng cách trực tiếp dự đoán tọa độ các vùng văn bản dưới dạng hộp chữ nhật hoặc đa giác. Nhờ kiến trúc tối ưu cho việc dự đoán tọa độ, các phương pháp này thường có tốc độ suy luận nhanh, phù hợp với các ứng dụng thời gian thực. Tuy nhiên, một hạn chế chung của hướng tiếp cận này là thường yêu cầu các bước hậu xử lý (post-processing) phức tạp, đồng thời gặp khó khăn khi xử lý văn bản cong hoặc có hình dạng phức tạp. Một số phương pháp tiên tiến theo hướng này bao gồm TextBoxes

[?], EAST [?], FCE-Net [?], ABCNet [?].

- **Các phương pháp dựa trên thành phần liên thông (Connected Component-based):** Các phương pháp này tập trung vào việc phát hiện và nhóm các thành phần ảnh có đặc trưng tương đồng (như màu sắc, kết cấu hoặc cường độ) để hình thành các vùng văn bản hoàn chỉnh. Hướng tiếp cận này đạt hiệu quả trong các trường hợp đơn giản nhưng thường kém hiệu quả khi gặp nền phức tạp, văn bản cong hoặc có hình dạng phức tạp. Đồng thời, việc nhóm các thành phần riêng lẻ cũng đòi hỏi các bước hậu xử lý phức tạp để tái cấu trúc văn bản. Một số phương pháp tiêu biểu trong nhóm này gồm TextSnake [?], DRRG [?].
- **Các phương pháp dựa trên phân đoạn (Segmentation-based):** Nhóm phương pháp này xem phát hiện văn bản như một bài toán phân đoạn mức pixel, trong đó mỗi pixel được phân loại là văn bản hoặc nền. Từ kết quả phân đoạn, các vùng văn bản được suy ra và phục hồi thông qua các bước hậu xử lý. Cách tiếp cận này cho phép xử lý hiệu quả các văn bản có hình dạng phi chuẩn, như văn bản cong hoặc nghiêng, nhưng thường tồn kém tính toán hơn. Một số phương pháp tiên tiến hiện nay với hướng tiếp cận này bao gồm PANet [?], DBNet++ [?], TextPMs [?], FAST [?] và KPN [?].

### 2.2.1.2 Phương pháp tiếp cận

Trong bối cảnh đầy thách thức của phát hiện văn bản trên biển hiệu thực tế, khóa luận lựa chọn một số phương pháp tiên tiến hiện nay nhằm đánh giá khả năng xử lý các văn bản có hình dạng đa dạng. Những phương pháp này có khả năng duy trì độ chính xác đồng thời giảm thiểu các bước hậu xử lý phức tạp. Việc đánh giá các mô hình không chỉ giúp so sánh hiệu quả giữa các giải pháp hiện nay, mà còn cung cấp cái nhìn tổng quan về những hướng tiếp cận khả thi cho bài toán.

- **PANet [?]:** PANet là một kiến trúc phân đoạn hiệu quả với hai thành phần chính gồm Feature Pyramid Enhancement Module (FPEM), chịu trách nhiệm tạo bản đồ đặc trưng đa tỷ lệ, và Feature Fusion Module (FFM), thực hiện tổng hợp các

đặc trưng này. Bằng cách áp dụng phương pháp pixel aggregation trên bản đồ đặc trưng cuối cùng, PANet có thể nhóm chính xác các pixel văn bản vào các thể hiện tương ứng, đạt hiệu quả cao nhờ quy trình phân đoạn có chi phí tính toán thấp.

- **DBNet [?]:** DBNet++ là phiên bản cải tiến của DBNet, tích hợp cơ chế differentiable binarization (DB) trực tiếp vào mạng phân đoạn để tạo mask văn bản chính xác và ổn định, giảm đáng kể các bước hậu xử lý. Đồng thời, Adaptive Scale Fusion (ASF) được áp dụng để hợp nhất các đặc trưng đa tỷ lệ, cải thiện khả năng xử lý các văn bản có kích thước khác nhau.
- **TextPMs [?]:**
- **FAST [?]:**
- **KPN [?]:**

## 2.2.2 Nhận dạng văn bản (Text Recognition)

### 2.2.2.1 Cơ sở và hướng tiếp cận chung

### 2.2.2.2 Phương pháp tiếp cận

## 2.2.3 End-to-End (End-to-End Text Recognition)

### 2.2.3.1 Cơ sở và hướng tiếp cận chung

### 2.2.3.2 Phương pháp tiếp cận

# Chapter 3

## PHƯƠNG PHÁP

### 3.1 Hệ thống phát hiện và nhận dạng chữ trên biển hiệu

#### 3.1.1 Bài toán và mục tiêu

Mục tiêu của khóa luận là xây dựng một hệ thống tự động “đọc chữ trên biển hiệu” trong ảnh/video đường phố. Dữ liệu đầu vào có thể là ảnh tĩnh hoặc chuỗi khung hình video; đầu ra là (i) vị trí biển hiệu, (ii) vị trí vùng chữ trên biển hiệu, và (iii) nội dung văn bản được nhận dạng.

Do đặc trưng dữ liệu street-view chứa nhiều nhiễu nền và văn bản đa dạng về hình dạng, hướng, kích thước, nhóm đề xuất một pipeline dạng module hoá gồm ba thành phần chính:

- **Phát hiện biển hiệu (Signboard Detection/Segmentation):** khoanh vùng biển hiệu để giảm không gian tìm kiếm và loại nhiễu nền.
- **Phát hiện văn bản (Text Detection):** phát hiện vùng chữ bên trong biển hiệu (ưu tiên OBB/đa giác khi chữ nghiêng).
- **Nhận dạng văn bản (Text Recognition):** nhận dạng chuỗi ký tự từ vùng chữ đã crop/rectify.

### 3.1.2 Kiến trúc tổng thể của pipeline

Hình 3.1 minh họa luồng xử lý tổng thể. Pipeline được thiết kế theo hướng “tách nhiệm vụ” để linh hoạt thay thế mô-đun (detector/recognizer) tùy điều kiện tài nguyên hoặc yêu cầu tốc độ.

Với mỗi khung hình/ảnh đầu vào  $I$ , hệ thống thực hiện:

1. Dự đoán vùng biển hiệu  $\mathcal{S} = \{s^{(i)}\}$  bằng mô hình detection (bbox/OBB) hoặc segmentation (mask).
2. Chuẩn hoá hình học biển hiệu (tuỳ chọn) để giảm méo phôi cảnh, thu được ảnh biển hiệu đã crop:  $I_{sb}^{(i)}$ .
3. Chạy Text Detector trên  $I_{sb}^{(i)}$  để lấy tập vùng chữ  $\mathcal{B} = \{b^{(j)}\}$ .
4. Với mỗi  $b^{(j)}$ , thực hiện crop & rectify để tạo patch chữ  $\hat{I}^{(j)}$ .
5. Chạy Text Recognizer để dự đoán chuỗi ký tự  $s^{(j)}$  và độ tin cậy  $c^{(j)}$ .
6. Hậu xử lý (chuẩn hoá Unicode, lọc theo confidence, gộp theo dòng/khối nếu cần).

### 3.1.3 Quy ước biểu diễn hình học (BBox/OBB/Mask)

Trong khóa luận, để mô tả vùng quan tâm (biển hiệu hoặc chữ), nhóm sử dụng một trong ba dạng:

- **Rectangle bounding box (BBox):**  $b = (x_{min}, y_{min}, x_{max}, y_{max})$ .
- **Oriented bounding box (OBB):** hộp quay, biểu diễn bởi (i) 4 đỉnh hoặc (ii)  $(x_c, y_c, w, h, \theta)$ .
- **Segmentation mask:** mặt nạ nhị phân  $m(x, y) \in \{0, 1\}$ .

OBB/mask đặc biệt hữu ích khi biển hiệu hoặc chữ bị nghiêng/biến dạng do phôi cảnh.

### 3.1.4 Chiến lược “phát hiện biển hiệu trước”

Thay vì phát hiện chữ trực tiếp trên toàn ảnh street-view (nhiều nhiễu), nhóm lựa chọn chiến lược:

$$I \xrightarrow{\text{Signboard module}} \{I_{sb}^{(i)}\} \xrightarrow{\text{Text detection/recognition}} \text{outputs}$$

Lợi ích chính:

- **Giảm nhiễu nền:** hạn chế text detector bị đánh lạc hướng bởi biển số xe, poster nhỏ, vật thể nền.
- **Giảm chi phí tính toán:** text detector/recognizer chạy trên patch nhỏ thay vì ảnh full-HD.
- **Tăng độ ổn định:** vùng biển hiệu thường chứa text liên quan, giúp mô hình tập trung đúng ngữ cảnh.

## 3.2 Mô-đun phát hiện biển hiệu (Signboard Detection/Segmentation)

### 3.2.1 Phát hiện biển hiệu bằng object detection

Nhóm thực nghiệm fine-tune các mô hình object detection hiện đại trên SignboardText để dự đoán vị trí biển hiệu dưới dạng hộp chữ nhật (BBox) hoặc hộp quay (OBB). Về mặt phương pháp, object detector thực hiện ánh xạ:

$$f_{\text{det}}(I) \rightarrow \{(b^{(i)}, p^{(i)})\}$$

trong đó  $b^{(i)}$  là bbox/obb,  $p^{(i)}$  là confidence.

#### 3.2.1.1 Lý do ưu tiên OBB cho biển hiệu

Biển hiệu trong street-view thường bị nghiêng theo phôi cảnh (góc quay camera) hoặc đặt lệch. Do đó, OBB giúp:

- bám sát hình dạng biển hiệu hơn BBox,
- giảm vùng nền bị crop dư thừa,
- cải thiện bước align/rectify và tăng chất lượng patch đưa vào text detector.

### **3.2.1.2 Thiết lập fine-tune**

Trong thực nghiệm, các mô hình detection được fine-tune theo cùng một giao thức để đảm bảo so sánh công bằng:

- Huấn luyện trong 50 epochs.
- Áp dụng augmentation cơ bản cho bài toán street-view (phóng to/thu nhỏ, xoay nhẹ, thay đổi sáng/tương phản, motion blur mức nhẹ nếu cần).
- Tiêu chí chọn checkpoint dựa trên mAP (đặc biệt là AP50 và AP50–95).

Lưu ý: các siêu tham số chi tiết (batch size, lr, scheduler) có thể thay đổi tùy GPU; khóa luận tập trung mô tả pipeline và so sánh theo cùng một chuẩn đánh giá.

### **3.2.2 Phân đoạn biển hiệu bằng segmentation**

Đối với các biển hiệu có hình dạng không đều hoặc bị che khuất một phần, bounding box có thể bao quá nhiều nền, làm giảm chất lượng text detection. Vì vậy nhóm bổ sung hướng segmentation để dự đoán mask biển hiệu:

$$f_{\text{seg}}(I) \rightarrow m(x, y)$$

Sau đó, mask được dùng để:

- crop theo vùng mask (tight crop),
- hoặc giữ nguyên kích thước ảnh nhưng xóa nền ngoài mask (background suppression).

### 3.2.2.1 Tạo patch biển hiệu từ mask

Từ mask  $m$ , nhóm lấy hộp bao nhỏ nhất (min bounding rectangle) hoặc OBB của vùng mask để crop ra patch biển hiệu  $I_{sb}$ . Trong một số trường hợp, việc giữ mask để “triệt nền” giúp text detector ổn định hơn khi nền phía sau biển hiệu quá phức tạp.

## 3.3 Chuẩn hóa hình học (Align/Rectify)

### 3.3.1 Động cơ của bước Align

Bước Align nhằm giảm tác động của phôi cảnh và nghiêng xoay, giúp:

- text detector phát hiện ổn định hơn (đặc biệt với chữ thẳng hàng),
- recognizer giảm lỗi do kí tự bị kéo dãn/biến dạng.

### 3.3.2 Align dựa trên OBB/4 điểm

Với OBB biểu diễn bởi 4 đỉnh  $\{(x_k, y_k)\}_{k=1}^4$ , nhóm thực hiện phép biến đổi phôi cảnh (homography) để đưa biển hiệu (hoặc vùng chữ) về mặt phẳng chuẩn. Khi đó:

$$\hat{I} = \text{WarpPerspective}(I, \mathbf{H})$$

với  $\mathbf{H}$  là ma trận homography  $3 \times 3$  ước lượng từ cặp điểm nguồn–đích.

## 3.4 Mô-đun phát hiện văn bản (Text Detection)

### 3.4.1 Bài toán phát hiện văn bản trong patch biển hiệu

Với patch biển hiệu  $I_{sb}$ , text detector dự đoán tập vùng chữ:

$$f_{\text{textdet}}(I_{sb}) \rightarrow \mathcal{B} = \{(b^{(j)}, p^{(j)})\}_{j=1}^N$$

Trong đó  $b^{(j)}$  có thể là bbox/obb/polygon tùy mô hình;  $p^{(j)}$  là độ tin cậy.

### 3.4.2 Lựa chọn hướng tiếp cận: segmentation-based vs OBB-based

Nhóm xét hai hướng chính phù hợp dữ liệu signboard:

- **Segmentation-based (ví dụ: DBNet++, TextPMs):** linh hoạt với chữ cong, chữ dày/nhỏ, nhưng tách instance gần nhau có thể khó.
- **OBB-based (ví dụ: YOLOv8-obb, YOLOv11-obb):** dự đoán hộp quay trực tiếp, phù hợp chữ nghiêng và cho tốc độ cao.

### 3.4.3 Quy trình fine-tune text detector

Sau khi đánh giá các mô hình tiền huấn luyện, nhóm chọn các mô hình có kết quả tốt để fine-tune trên SignboardText. Quy trình huấn luyện gồm:

- Chuẩn hoá dữ liệu nhãn về định dạng của mô hình (bbox/obb hoặc polygon).
- Augmentation tập trung vào hiện tượng street-view: blur nhẹ, thay đổi ánh sáng, scale, rotate nhỏ.
- Hậu xử lý:
  - NMS (hoặc NMS xoay cho OBB) để loại bỏ dự đoán trùng.
  - Lọc theo confidence threshold để giảm false positives.

### 3.4.4 Chỉ số đánh giá cho text detection

Khó khăn sử dụng Precision/Recall/Hmean:

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad H = \frac{2PR}{P+R}$$

Tiêu chí khớp (matching) giữa dự đoán và ground-truth dựa trên IoU (hoặc IoU cho OBB/polygon tùy dạng nhãn).

## 3.5 Mô-đun nhận dạng văn bản (Text Recognition)

### 3.5.1 Bài toán nhận dạng chuỗi ký tự

Với mỗi vùng chữ  $b^{(j)}$ , ta crop/rectify được patch chữ  $\hat{I}^{(j)}$ . Text recognizer dự đoán chuỗi:

$$f_{\text{textrec}}(\hat{I}^{(j)}) \rightarrow (\mathbf{s}^{(j)}, c^{(j)})$$

trong đó  $\mathbf{s}^{(j)}$  là chuỗi ký tự,  $c^{(j)}$  là độ tin cậy (hoặc xác suất trung bình theo token).

### 3.5.2 Tiền xử lý patch chữ

Để tăng độ ổn định cho recognizer, nhóm áp dụng các bước chuẩn hoá đầu vào:

- Resize về kích thước chuẩn theo yêu cầu mô hình.
- Giữ tỉ lệ (aspect ratio) khi có thể; dùng padding để tránh méo chữ.
- (Tùy chọn) tăng tương phản cục bộ nhẹ cho chữ mờ/thiếu sáng.

### 3.5.3 Mô hình nhận dạng và lý do lựa chọn

Các mô hình recognizer hiện đại (đặc biệt transformer-based) phù hợp văn bản tự nhiên do khả năng học phụ thuộc dài và chống biến dạng tốt. Trong khóa luận, nhóm đánh giá các mô hình tiền huấn luyện và chọn mô hình có hiệu quả tốt để fine-tune trên SignboardText (đặc biệt trên Vietsignboard và Vin).

### 3.5.4 Chỉ số đánh giá cho text recognition

Khó khăn sử dụng:

- **Exact-match accuracy:** dự đoán đúng hoàn toàn chuỗi ký tự.
- **Normalized-match accuracy:** so khớp sau khi chuẩn hoá (ví dụ: chuẩn hoá Unicode/NFC, bỏ khoảng trắng dư, chuẩn hoá dấu câu, hoặc chuẩn hoá chữ hoa-thường theo quy ước).

## 3.6 Hậu xử lý và chuẩn hoá tiếng Việt

### 3.6.1 Chuẩn hoá Unicode và lọc nhiễu ký tự

Văn bản tiếng Việt thường phát sinh lỗi về mã Unicode (tổ hợp dấu) hoặc ký tự nhiễu do nền phức tạp. Do đó, nhóm thực hiện:

- Chuẩn hoá Unicode về một chuẩn thống nhất (ví dụ NFC).
- Loại bỏ ký tự không hợp lệ theo tập ký tự mục tiêu (alphabet) của dữ liệu.
- Lọc dự đoán theo confidence; ưu tiên kết quả có độ tin cậy cao hơn trong các trường hợp trùng lắp vùng chữ.

### 3.6.2 Gộp kết quả theo dòng/khối (tùy chọn)

Trong trường hợp text detector trả về nhiều box theo từng từ/ký tự, có thể gộp theo dòng dựa trên:

- độ gần theo trực ngang,
- độ chênh lệch góc quay,
- và khoảng cách giữa các hộp.

Bước này giúp tạo câu/nhãn biến hiệu hoàn chỉnh hơn (tùy yêu cầu đầu ra).

## 3.7 Tổng hợp cấu hình thực nghiệm trong pipeline

### 3.7.1 Nguyên tắc chọn mô hình cho pipeline cuối

Từ đánh giá mô-đun (signboard det/seg, text det, text rec), nhóm chọn cấu hình kết hợp dựa trên:

- **Độ chính xác:** ưu tiên Hmean cao cho text detection và accuracy cao cho recognition.
- **Tốc độ:** cân bằng FPS để phù hợp xử lý video.
- **Tính ổn định:** ưu tiên cấu hình ít nhạy với nền phức tạp và góc nhìn.

### 3.7.2 Các biến thể pipeline được so sánh

Khóa luận so sánh một số biến thể theo cấu trúc:

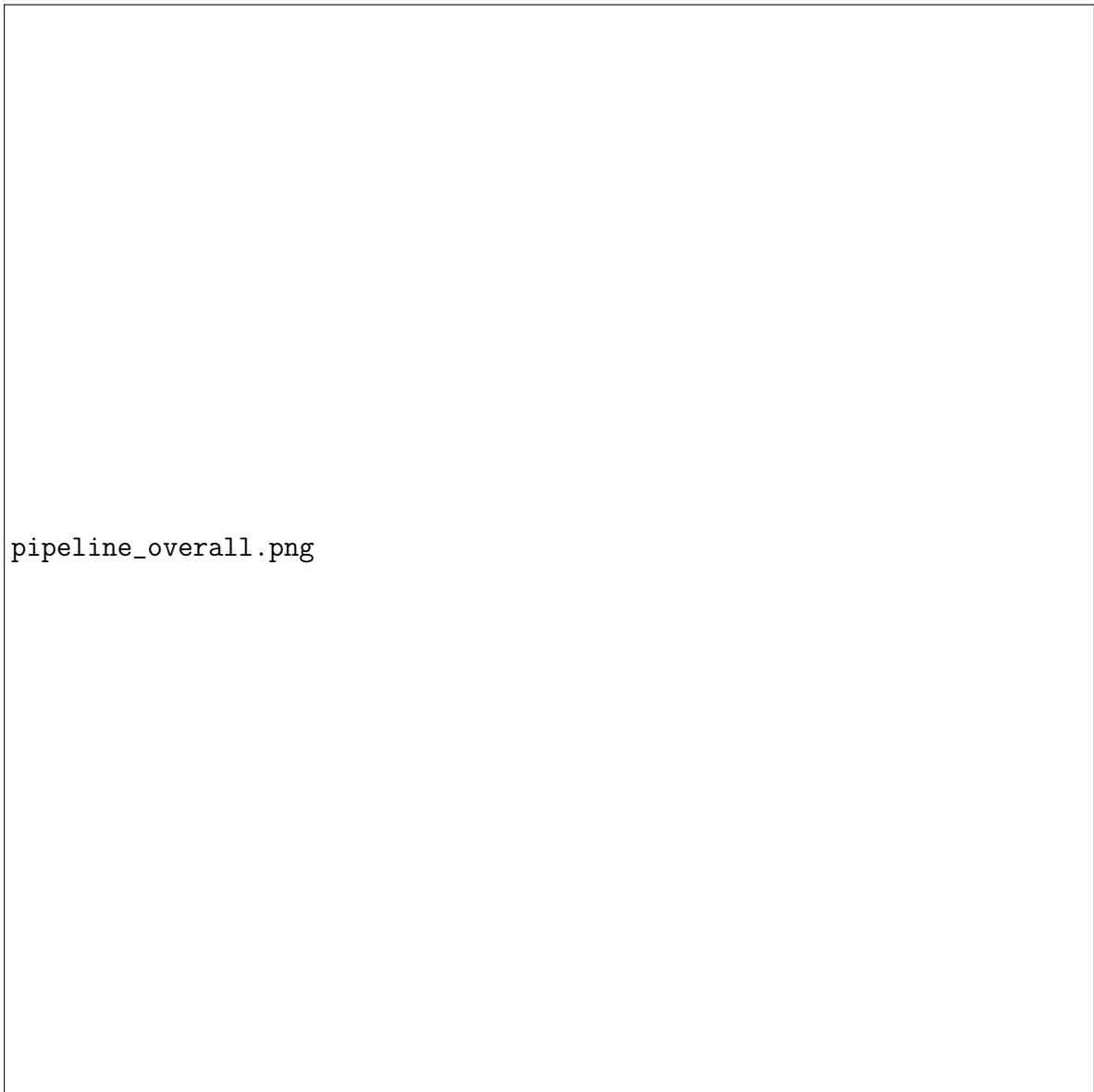
(Signboard module) + (Text detector) + (Text recognizer) + (Align)

Trong đó:

- Signboard module gồm: object detection (BBox/OBB) hoặc segmentation.
- Text detector gồm: segmentation-based hoặc OBB-based.
- Text recognizer: mô hình transformer-based fine-tune.
- Align: bật/tắt theo cấu hình để đánh giá tác động của chuẩn hoá hình học.

## 3.8 Tóm tắt chương

Chương này đã trình bày phương pháp xây dựng hệ thống đọc chữ trên biển hiệu theo hướng pipeline module hóa. Hệ thống gồm ba thành phần chính: phát hiện/phan đoạn biển hiệu để giới hạn vùng quan tâm, phát hiện văn bản trong patch biển hiệu (ưu tiên OBB/segmentation tùy bối cảnh), và nhận dạng chuỗi ký tự bằng mô hình recognizer hiện đại. Ngoài ra, chương cũng mô tả bước Align/Rectify để giảm méo hình học, cùng các bước hậu xử lý và chuẩn hoá tiếng Việt nhằm tăng độ ổn định đầu ra. Các cấu hình pipeline khác nhau sẽ được đánh giá chi tiết bằng thực nghiệm trong chương tiếp theo.



pipeline\_overall.png

Hình 3.1: Kiến trúc tổng thể hệ thống đọc chữ trên biển hiệu: Signboard Detection/Segmentation → (Align/Rectify) → Text Detection → Text Recognition → Post-processing.



align\_rectify.png

Hình 3.2: Minh họa bước Align/Rectify dựa trên 4 điểm (OBB) để giảm nghiêng và phôi cảnh trước khi phát hiện/nhận dạng chữ.

# Chapter 4

## THỰC NGHIỆM VÀ ĐÁNH GIÁ

### 4.1 Dữ liệu

#### 4.1.1 Nguồn dữ liệu

Trong khóa luận này, nhóm sử dụng bộ dữ liệu **SignboardText** được giới thiệu trong bài báo “*SignboardText: Text Detection and Recognition in In-the-Wild Signboard Images*” (IEEE), với các tác giả Tien Do, Thuyen Tran, Thua Nguyen, Duy-Dinh Le và Thanh Duc Ngo. [?]

Bộ dữ liệu tập trung vào văn bản trên biển hiệu trong điều kiện tự nhiên (in-the-wild), phù hợp với bối cảnh street-view/video hành trình do có nhiều biến thiên về phông chữ, kích thước, hướng, bố cục, đa ngôn ngữ và nền phức tạp.

#### 4.1.2 Thống kê dữ liệu

Theo thống kê trên tập dữ liệu mà nhóm sử dụng (các biểu đồ Dataset Statistic), SignboardText được chia thành ba nhóm:

- **Vietsignboard**: 1,194 ảnh.
- **English**: 413 ảnh.
- **Vin**: 516 ảnh.

Số lượng văn bản được gán nhãn (annotated text):

Bảng 4.1: Thống kê kích thước (px) trên SignboardText dùng trong khóa luận.

Loại	Image		Word		Line		Signboard	
	W	H	W	H	W	H	W	H
Min	190	86	16	7	21	8	20	28
Max	4608	4160	3556	1125	3952	1032	4536	4057
Mean	1015.99	710.75	128.03	60.75	255.33	52.68	706.33	330.32

Bảng 4.2: Tỷ lệ văn bản nằm trong vùng biển hiệu (%).

Nhóm	Vietsignboard		English		Vin		Avg	
	word	line	word	line	word	line	word	line
Text proportion (%)	64.50	57.56	72.68	–	56.03	–	64.40	57.56

- **Vietsignboard:** 48,638 word và 10,950 text-line.
- **English:** 3,646 word.
- **Vin:** 16,615 word.

Số lượng biển hiệu được gán nhãn (annotated signboard):

- **Vietsignboard:** 1,327 biển hiệu.
- **English:** 488 biển hiệu.
- **Vin:** 552 biển hiệu.

### 4.1.3 Đặc trưng hình học và kích thước mẫu

Bộ dữ liệu hỗ trợ nhiều dạng hình học văn bản (*horizontal, arbitrary quadrilateral, multi-oriented*) và mang tính đa ngôn ngữ (ML). Thống kê kích thước ảnh/văn bản/biển hiệu (px) được tổng hợp ở Bảng ??.

### 4.1.4 Tỷ lệ văn bản nằm trong vùng biển hiệu

Bảng ?? thể hiện tỷ lệ phần trăm văn bản thuộc vùng biển hiệu so với toàn bộ văn bản xuất hiện trong ảnh.

Bảng 4.3: Chia tập dữ liệu theo nhiệm vụ.

Loại dữ liệu	Train	Validation	Test	Task
Images	1357	340	426	Signboard/Text Detection
Word Images	44238	10661	13988	Text Recognition

#### 4.1.5 Chia tập Train/Validation/Test

Để phục vụ huấn luyện và đánh giá, nhóm sử dụng cách chia dữ liệu như Bảng ??.

## 4.2 Tiết xuỷ lý

### 4.2.1 Chuẩn hóa nhãn và định dạng dữ liệu

Nhóm thực hiện chuẩn hóa nhãn để phục vụ nhiều họ mô hình khác nhau:

- **Signboard detection:** chuyển nhãn biển hiệu về BBox hoặc OBB tùy mô hình.
- **Signboard segmentation:** biểu diễn biển hiệu bằng mask nhị phân.
- **Text detection:** chuẩn hóa nhãn vùng chữ về polygon/quad hoặc OBB/bbox tùy hướng tiếp cận.

### 4.2.2 Tạo dữ liệu cho Text Recognition

Từ nhãn *word-level*, nhóm cắt (crop) vùng chữ để tạo tập **Word Images**. Với các trường hợp văn bản bị nghiêng, nhóm áp dụng bước **rectify/align** (warp theo 4 điểm) trước khi resize về kích thước đầu vào của recognizer.

### 4.2.3 Tăng cường dữ liệu (Augmentation)

Trong huấn luyện/fine-tune, nhóm sử dụng augmentation mức cơ bản phù hợp street-view:

- thay đổi sáng/tương phản,
- scale và rotate nhỏ,

- blur nhẹ (tùy chọn) để mô phỏng rung/mờ.

## 4.3 Mô hình thực nghiệm

### 4.3.1 Phát hiện biển hiệu (Signboard Detection)

Nhóm đánh giá hai hướng:

- **Object detection (BBox/OBB):** DETR, YOLOv8, RT-DETRv2, YOLOv11; và các biến thể **YOLOv8-obb**, **YOLOv11-obb**.
- **Segmentation (mask):** SegFormer và Mask2Former.

### 4.3.2 Phát hiện văn bản (Text Detection)

Nhóm so sánh các mô hình text detection: PANet, DBNet++, TextPMs, FAST, KPN. Sau đó, chọn các mô hình có kết quả tốt để fine-tune trên SignboardText.

### 4.3.3 Nhận dạng văn bản (Text Recognition)

Nhóm đánh giá các recognizer: ViTSTR, PARSeq, CDistNet, SMTR, SVTRv2; và thực hiện fine-tune cho các mô hình tốt nhất trên dữ liệu Word Images.

### 4.3.4 Text spotting (End-to-end / Two-stage)

Nhóm so sánh:

- **End-to-end one-stage:** TESTR, DeepSolo, UNITS, DNTextSpotter.
- **Two-stage:** kết hợp (Text Detector) + (Text Recognizer), gồm TextPMs+SVTRv2, DBNet++ + SVTRv2, FAST + SVTRv2.

## 4.4 Độ đo đánh giá

### 4.4.1 Signboard detection

Nhóm sử dụng các chỉ số chuẩn của object detection:

- **AP<sub>50</sub>** (IoU = 0.5),
- **AP<sub>50:95</sub>** (trung bình IoU từ 0.5 đến 0.95),
- **FPS** để đánh giá tốc độ suy luận.

### 4.4.2 Signboard segmentation

Với segmentation, nhóm sử dụng:

- **mIoU** và **mAccuracy**,
- **FPS**.

### 4.4.3 Text detection

Khóa luận sử dụng Precision/Recall/Hmean:

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad H = \frac{2PR}{P+R}.$$

Tiêu chí khớp giữa dự đoán và ground-truth dựa trên IoU (với bbox/obb) hoặc IoU theo polygon/quad (tùy dạng nhãn).

### 4.4.4 Text recognition

Khóa luận sử dụng:

- **Exact-match accuracy**: đúng hoàn toàn chuỗi ký tự,
- **Normalized-match accuracy**: so khớp sau chuẩn hóa (Unicode/NFC, khoảng trắng, dấu câu theo quy ước).

Bảng 4.4: Kết quả signboard detection với rectangle bounding box.

Model	Year	Params(M)	Epochs	FPS	AP <sub>50</sub>	AP <sub>50:95</sub>
DETR	2020	41.50	50	42.82	89.30	68.95
YOLOv8	2023	3.01	50	133.01	86.85	74.20
RT-DETRv2	2024	42.73	50	26.44	90.71	81.22
YOLOv11	2024	2.59	50	96.06	88.00	73.47

Bảng 4.5: Kết quả signboard detection với oriented bounding box (OBB).

Model	Year	Params(M)	Epochs	FPS	AP <sub>50</sub>	AP <sub>50:95</sub>
YOLOv8-obb	2023	3.01	50	133.01	93.04	79.08
YOLOv11-obb	2024	2.59	50	96.06	93.20	80.78

#### 4.4.5 Text spotting

Nhóm sử dụng **Hmean** cho kết quả end-to-end (phát hiện + nhận dạng) và báo cáo thêm **FPS**.

### 4.5 Kết quả thực nghiệm

#### 4.5.1 Kết quả phát hiện biển hiệu (BBox/OBB)

**Nhận xét:** RT-DETRv2 đạt AP<sub>50:95</sub> cao nhất ở thiết lập BBox, trong khi YOLOv8 đạt tốc độ suy luận cao nhất. Ở thiết lập OBB, YOLOv11-obb nhỉnh hơn về AP<sub>50:95</sub>, phù hợp với biển hiệu bị nghiêng/phối cảnh.

Bảng 4.6: Kết quả signboard segmentation.

<b>Model</b>	<b>Year</b>	<b>Params(M)</b>	<b>Epochs</b>	<b>FPS</b>	<b>mIoU(%)</b>	<b>mAccuracy(%)</b>
SegFormer	2021	3.8	50	97.5	89.03	93.92
Mask2Former	2022	47.4	50	15.9	90.48	94.44

Bảng 4.7: So sánh các mô hình text detection tiền huấn luyện trên SignboardText.

<b>Model</b>	<b>Year</b>	<b>Params(M)</b>	<b>Vietsignboard</b>			<b>English</b>			<b>Vin</b>		
			P	R	H	P	R	H	P	R	H
PANet	2020	12.25	81.00	82.25	81.62	61.00	72.56	66.28	81.71	75.82	78.66
DBNet++	2022	26.43	89.86	80.31	84.82	73.38	60.95	65.59	91.52	74.18	81.94
TextPMs	2022	36.43	90.27	84.85	87.48	78.82	77.58	78.20	93.41	81.32	86.95
FAST	2023	10.58	83.98	86.32	85.13	64.34	80.25	71.42	84.45	79.09	81.69
KPN	2023	58.24	81.19	81.85	81.52	63.49	85.90	73.01	83.49	78.37	80.85

#### 4.5.2 Kết quả phân đoạn biển hiệu (Segmentation)

#### 4.5.3 Kết quả text detection (mô hình tiền huấn luyện)

#### 4.5.4 Kết quả text detection sau fine-tune

#### 4.5.5 Kết quả text recognition (mô hình tiền huấn luyện)

#### 4.5.6 Kết quả text recognition sau fine-tune

#### 4.5.7 Kết quả text spotting

#### 4.5.8 Đánh giá pipeline trên vùng biển hiệu

Nhóm so sánh các biến thể pipeline theo cấu trúc:

(Signboard module) + (Text detector) + (Text recognizer) + (Align).

Kết quả Text Detection trên vùng biển hiệu được tổng hợp ở Bảng ??.

Kết quả Text Recognition trong pipeline được tổng hợp ở Bảng ??.

Bảng 4.8: Kết quả text detection sau fine-tune trên SignboardText.

Model	Vietsignboard			English			Vin		
	P	R	H	P	R	H	P	R	H
DBNet++	90.40	81.94	85.96	84.42	61.75	71.33	92.22	76.34	83.53
TextPMs	90.36	85.29	87.75	83.80	82.39	83.09	92.98	84.46	88.51
YOLOv8-obb	91.14	83.94	87.39	83.77	81.42	82.58	93.34	77.57	84.73
YOLOv11-obb	91.49	82.97	87.02	84.16	78.70	81.34	93.10	76.84	84.19

Bảng 4.9: So sánh các mô hình text recognition tiền huấn luyện trên SignboardText.

Model	Year	Params(M)	Vietsignboard		English		Vin		Speed (ms)
			Word	Line	Word	Line	Word	Line	
ViTSTR	2021	85.48	56.70	29.01	48.71	–	50.82	–	9.80
PARSEq	2022	23.83	80.76	59.08	78.28	–	72.96	–	12.71
CDistNet	2024	65.46	63.25	42.43	68.13	–	56.39	–	120.33
SMTTR	2025	15.82	79.58	64.54	76.77	–	70.64	–	23.47
SVTRv2	2025	21.02	80.82	65.64	78.33	–	72.34	–	18.81

## 4.6 Tóm tắt chương

Chương này đã mô tả bộ dữ liệu SignboardText và các thống kê quan trọng, quy trình tiền xử lý, các mô hình được lựa chọn để đánh giá cho từng mô-đun (phát hiện/phân đoạn biển hiệu, phát hiện văn bản, nhận dạng văn bản, và text spotting), cùng các độ đo đánh giá tương ứng. Kết quả thực nghiệm cho thấy các mô hình OBB phù hợp hơn cho biển hiệu/văn bản bị nghiêng; đồng thời các cấu hình two-stage và pipeline “phát hiện biển hiệu trước” mang lại hiệu quả tổng thể tốt hơn trong bối cảnh dữ liệu street-view.

Bảng 4.10: Kết quả text recognition sau fine-tune (Exact-match và Norm-match).

Model	Vietsignboard		English		Vin	
	Exact	Norm	Exact	Norm	Exact	Norm
PARSeq	79.19	80.26	60.68	62.11	76.95	78.67
SMTR	77.40	78.47	59.64	61.07	75.06	76.58
SVTRv2	76.39	77.96	57.55	58.59	76.46	78.10

Bảng 4.11: Kết quả text spotting (Hmean %) với one-stage và two-stage.

Model	Year	Params(M)	Viet		English		Vin		FPS
			Word	Line	Word	Line	Word	Line	
TESTR	2022	49.48	48.77	7.43	60.43	—	50.11	—	11.11
DeepSolo	2023	42.59	47.61	8.02	70.71	—	53.32	—	12.21
UNITS	2023	101.00	63.19	9.75	88.00	—	64.88	—	1.28
DNTTextSpotter	2025	42.73	49.17	9.20	73.20	—	53.57	—	12.40
TextPMs+SVTRv2	—	57.45	66.48	9.79	68.59	—	68.50	—	5.52
DBNet++ +SVTRv2	—	47.45	66.58	8.22	59.21	—	67.32	—	—
FAST +SVTRv2	—	31.60	64.22	9.06	56.90	—	62.18	—	—

Bảng 4.12: Kết quả Text Detection trên vùng biển hiệu với các biến thể pipeline.

Signboard	Text Det	Text Rec	Vietsignboard			English			P
			P	R	H	P	R	H	
RTDETRv2	YOLOv8-obb	PARSeq	91.79	87.59	89.64	80.13	72.16	75.94	92.41
YOLOv11-obb	YOLOv8-obb	PARSeq	89.34	86.76	88.03	73.47	67.54	70.38	89.39
SegFormer	YOLOv8-obb	PARSeq	83.50	89.05	86.19	67.04	79.95	72.93	82.53
YOLOv11-obb + Align	YOLOv8-obb	PARSeq	89.78	86.79	88.26	73.48	67.41	70.32	89.25
SegFormer + Align	YOLOv8-obb	PARSeq	85.42	87.27	86.34	68.09	73.31	70.60	84.35

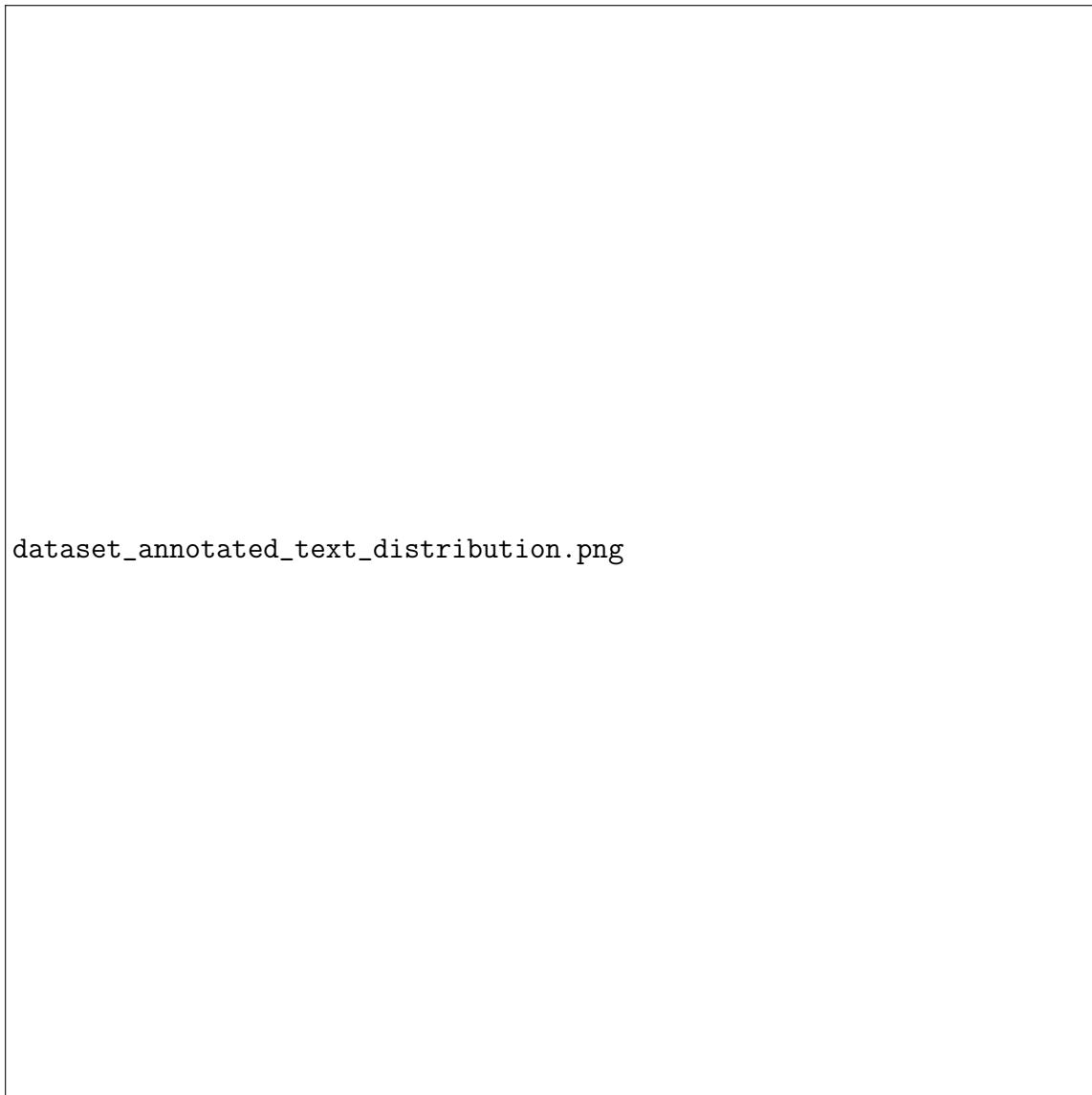
Bảng 4.13: Kết quả Text Recognition trên vùng biển hiệu (pipeline).

Signboard	Text Det	Text Rec	Vietsignboard		English		Vin	
			Exact	Norm	Exact	Norm	Exact	Norm
RTDETRv2	YOLOv8-obb	PARSeq	71.36	72.32	48.68	49.70	70.35	72.23
YOLOv11-obb	YOLOv8-obb	PARSeq	70.23	71.19	41.65	42.66	69.00	70.69
SegFormer	YOLOv8-obb	PARSeq	67.91	68.89	45.63	46.29	66.68	68.25
YOLOv11-obb + Align	YOLOv8-obb	PARSeq	70.46	71.43	41.69	42.70	68.32	70.02
SegFormer + Align	YOLOv8-obb	PARSeq	68.16	69.16	42.35	43.08	67.34	68.72



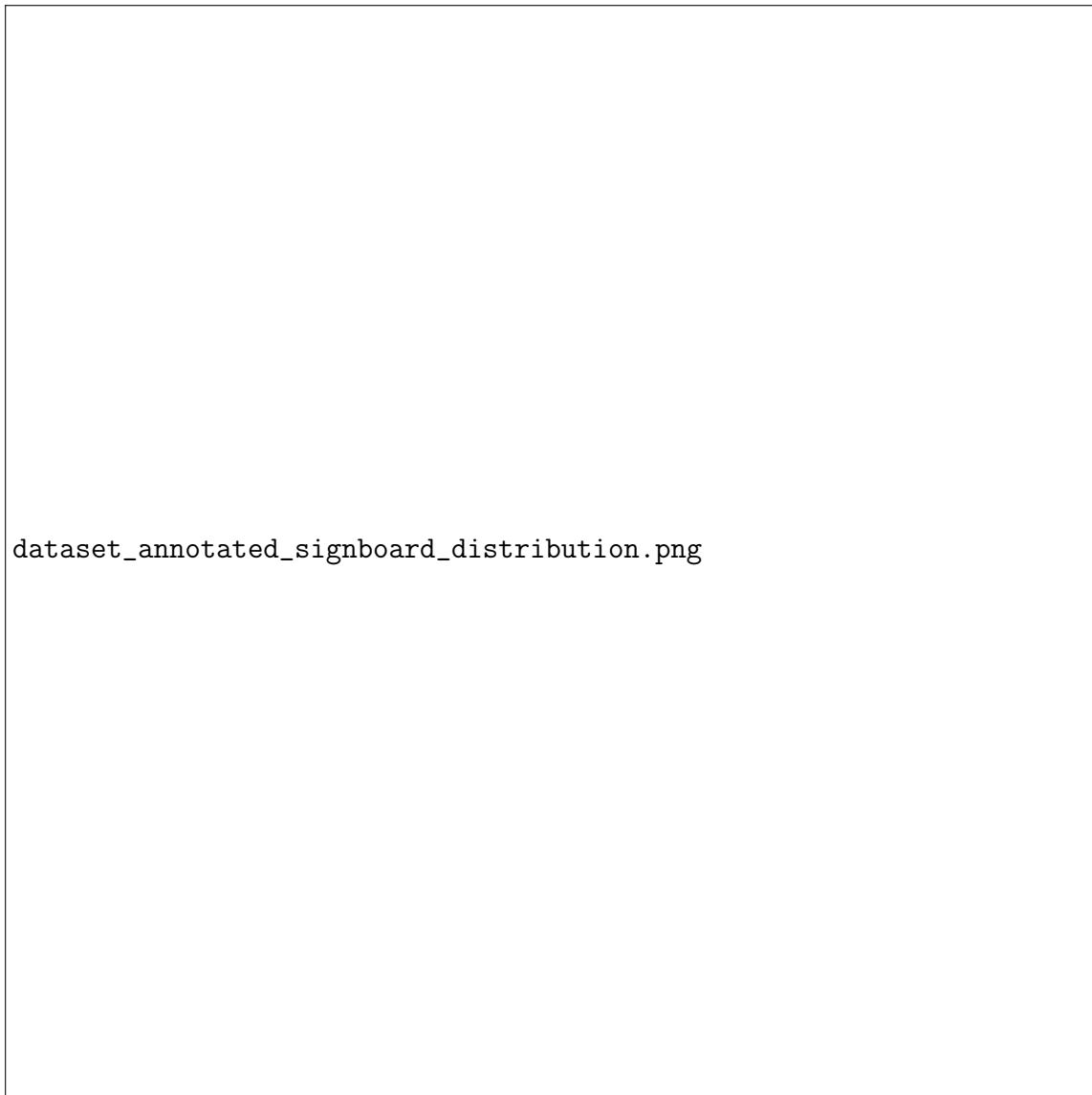
dataset\_image\_distribution.png

Hình 4.1: Phân bố số lượng ảnh theo nhóm trong SignboardText.



dataset.annotated\_text\_distribution.png

Hình 4.2: Phân bố số lượng văn bản được gán nhãn (word/line) theo nhóm trong SignboardText.



dataset\_annotated\_signboard\_distribution.png

Hình 4.3: Phân bố số lượng biển hiệu được gán nhãn theo nhóm trong SignboardText.



Hình 4.4: Ví dụ định tính: phát hiện/nhận dạng văn bản trước và sau khi căn chỉnh biến hiệu (Align).

# **Chapter 5**

## **KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

### **5.1 Kết luận**

.....:

- aaaa.

### **5.2 Hướng phát triển**

Để khắc phục những hạn chế trên và nâng cao hơn nữa tính hiệu quả, tính khả dụng và tính mở rộng của hệ thống, các hướng phát triển trong tương lai được đề xuất như sau:

**Tối ưu hóa khả năng mở rộng dữ liệu:**

- aaa
- aaa

**Tăng cường khả năng tương tác và thích ứng với người dùng:**

- Thiết kế giao diện người dùng aaaaaaaaaaaaaaa

**Tích hợp truy vấn hình thức thoại (Spoken Query Integration):** Phát triển hệ thống

.....