

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

Nguyễn Đình Quân - 20521184, Nguyễn Hùng Phát - 22521074

December 25, 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

LỜI CẢM ƠN	i
Tóm tắt khóa luận	ii
Contents	iii
List of Figures	vi
List of Tables	viii
1 TỔNG QUAN	1
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
2 CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN	10
2.1 Tổng quan và ý nghĩa thực tiễn của bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong video đường phố Việt Nam	10
2.2 Các phương pháp tiếp cận	11
2.2.1 Phát hiện đối tượng	11
2.2.1.1 Cơ sở và hướng tiếp cận chung	11

2.2.1.2	Các nghiên cứu liên quan	11
2.2.2	Phát hiện và nhận dạng văn bản ngoại cảnh (Scene Text Detection and Recognition - STDR)	13
2.2.2.1	Phát hiện văn bản ngoại cảnh (Scene Text Detection - STD)	13
2.2.2.2	Nhận dạng văn bản (Scene Text Recognition - STR) . .	16
2.2.2.3	Nhận dạng văn bản ngoại cảnh đầu–cuối (End-to-End Scene Text Recognition)	17
3	PHƯƠNG PHÁP TIẾP CẬN	21
3.1	Tổng quan về phương pháp	21
3.2	Phát hiện biến hiệu	21
3.2.1	Phát hiện đối tượng	21
3.2.2	Phân đoạn đối tượng	24
3.3	Phát hiện và nhận dạng văn bản trên biến hiệu theo hướng tiếp cận hai giai đoạn (Two-Stage)	26
3.3.1	Phát hiện văn bản trên biến hiệu	26
3.3.2	Nhận dạng nội dung văn bản	31
3.4	Phát hiện và nhận dạng văn bản trên biến hiệu theo hướng tiếp cận một giai đoạn (One-Stage)	36
4	THỰC NGHIỆM VÀ ĐÁNH GIÁ	41
4.1	Tập dữ liệu	41
4.2	Thiết lập thực nghiệm	45
4.2.1	Phát hiện biến hiệu	45
4.2.2	Phát hiện và nhận dạng văn bản trên biến hiệu	46
5	KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	48
5.1	Kết luận	48
5.2	Hướng phát triển	48

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 [7]	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 [7]	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
2.1	Hình ảnh minh họa tổng quan quy trình phát hiện đối tượng. Ảnh đầu vào được xử lý qua mạng nơ-ron để trích xuất đặc trưng và dự đoán vị trí (hộp giới hạn) cùng phân loại các đối tượng (nhãn lớp) xuất hiện trong khung hình.	12
2.2	Hình ảnh minh họa quá trình nhận dạng văn bản ngoại cảnh đầu-cuối (End-to-End Scene Text Recognition)	18
3.1	Kiến trúc tổng quan của hệ thống phát hiện và nhận dạng văn bản trên biển hiệu	22
3.2	Kiến trúc tổng quan của YOLO [29]	23
3.3	Kiến trúc tổng quan của DETR [4]	23
3.4	Kiến trúc tổng quan của RT-DETR, được sử dụng trong RTDETRv2 [41]	24

3.5	Kiến trúc tổng quan của SegFormer [34]	26
3.6	Kiến trúc tổng quan của Mask2Former [6]	27
3.7	Kiến trúc tổng quan của PANet [?]	28
3.8	Kiến trúc tổng quan của DBNet++ [17]	28
3.9	Kiến trúc tổng quan của TextPMs [17]	29
3.10	Kiến trúc tổng quan của FAST [5]	30
3.11	Kiến trúc tổng quan của KPN [39]	31
3.12	Kiến trúc tổng quan của ViTSTR [1]	32
3.13	Kiến trúc tổng quan của PARSeq [3]	33
3.14	Kiến trúc tổng quan của CDistNet [42]	34
3.15	Kiến trúc tổng quan của SMTR [8]	35
3.16	Kiến trúc tổng quan của SVTRv2 [9]	36
3.17	Sơ đồ kiến trúc của TESTR [40]	37
3.18	Kiến trúc tổng quan của DeepSolo [36]	38
3.19	Pipeline tổng quan của UNITS [15]	39
3.20	Kiến trúc tổng quan của DNTextSpotter [27]	40
4.1	Phân bố số lượng hình ảnh trong ba tập con của SignboardText [7] . . .	42
4.2	Phân bố số lượng thể hiện văn bản (text instances) theo cấp độ nhãn (word-level và line-level) trong các tập con của SignboardText [7] . . .	43
4.3	Phân bố số lượng đối tượng biển hiệu theo từng tập con của SignboardText	44
4.4	Hình ảnh minh họa quá trình căn chỉnh biển hiệu (signboard alignment)	46

List of Tables

4.1 Thống kê tỷ lệ văn bản nằm trong vùng biển hiệu so với toàn bộ văn bản trong ảnh trên các tập con của SignboardText	44
--------------------------------------------------------------------------------------------------------------------------------------	----

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 đồng đúc. Theo khảo sát các nghiên cứu hiện có, cho đến nay mới chỉ có một nghiên cứu [28] 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 [7]

- **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.

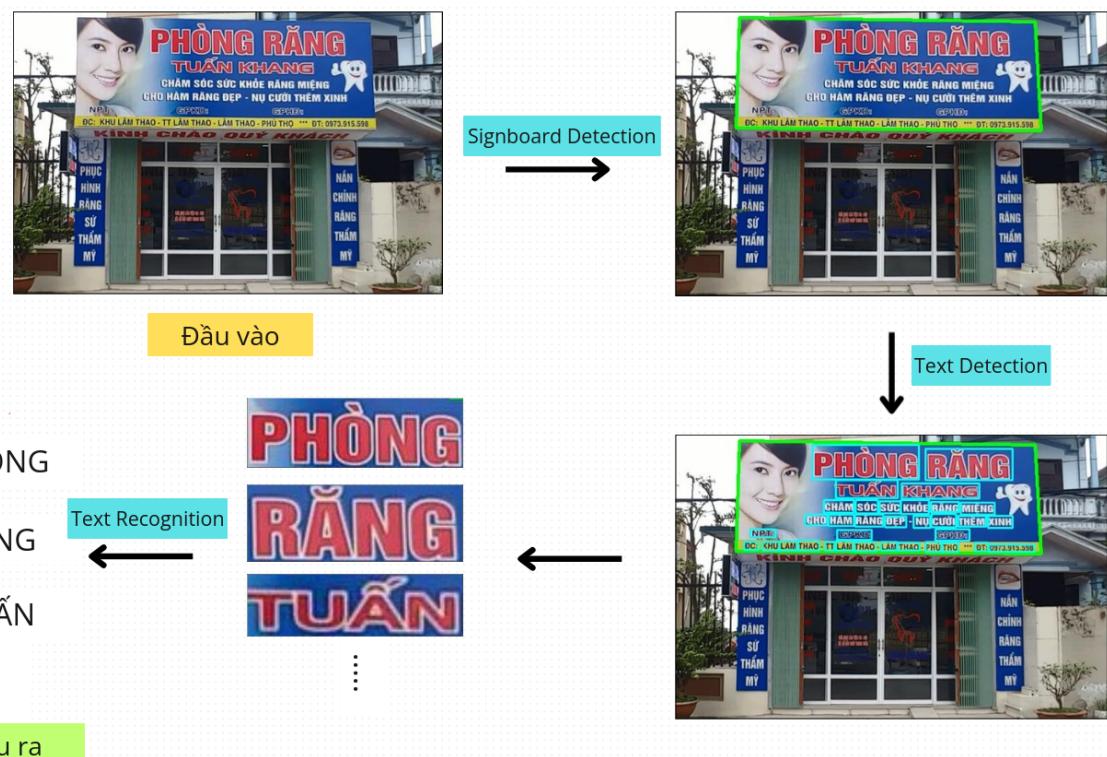


Hình 1.4: Hình ảnh minh họa sự đa dạng về hình dạng, kích thước, vật liệu và vị trí của biển hiệu, dựa trên bộ dữ liệu SignboardText [7]

- 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 [29], 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.

[4], RTDETR v2 [23], SegFormer [34], Mask2Former [6].

- **Phát hiện văn bản (Text Detection):** PANet [19], DBNet++ [17], TextPMs [37], FAST [5], KPN [39], các biến thể YOLO [29] với hộp giới hạn xoay.
- **Nhận dạng văn bản (Text Recognition):** ViTSTR [1], PARSeq [3], CDistNet [42], SMTR [8], SVTRv2 [9]
- **Text Spotting (End-to-End):** TESTR [40], DeepSolo [36], UNITS [15], DNTextSpotter [27]

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 [7] 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 [7], 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 CÁC NGHIÊN CỨU LIÊN QUAN

2.1 Tổng quan và ý nghĩa thực tiễn của bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong video đường phố Việt Nam

Trong môi trường giao thông đô thị tại Việt Nam, biển hiệu chứa đựng lượng lớn thông tin ngữ nghĩa cấp cao, phản ánh trực tiếp danh tính (tên cửa hàng, địa điểm) và loại hình kinh doanh/dịch vụ. Việc tự động trích xuất và hiểu các thông tin này từ luồng video không chỉ giúp giảm bớt việc gán nhãn dữ liệu thủ công mà còn mở ra nhiều ứng dụng thực tiễn hữu ích, có thể kể đến như:

- **Hệ thống dẫn đường thông minh:** Bổ sung thông tin các địa điểm thực tế (tên cửa hàng, địa điểm) từ biển hiệu vào hệ thống dẫn đường, giúp cải thiện độ chính xác của định vị và điều hướng.
- **Phân tích thông tin đô thị:** Tự động thống kê và phân loại các loại hình kinh doanh theo tuyến đường hoặc khu vực, phục vụ quy hoạch và nghiên cứu thị trường.
- **Cập nhật và làm giàu bản đồ số:** Tích hợp thông tin từ biển hiệu để tự động cập nhật cơ sở dữ liệu địa lý (GIS).

Trong bối cảnh này, để hiện thực hóa các ứng dụng trên, bài toán đặt ra nhiều thách thức kỹ thuật. Ngoài những khó khăn chung của nhận dạng văn bản trong cảnh (như đa dạng phông chữ, điều kiện ánh sáng), việc xử lý trong bối cảnh video hành trình tại Việt Nam còn phải đối mặt với: chất lượng hình ảnh thay đổi liên tục, góc quay và khoảng cách khác nhau, cùng sự xuất hiện của các biến hiệu với thiết kế đa dạng và ngôn ngữ phức tạp (kết hợp tiếng Việt và tiếng Anh). Những thách thức này nhấn mạnh tầm quan trọng và tính thực tiễn của việc nghiên cứu một giải pháp hiệu quả, phù hợp với đặc thù của bài toán.

2.2 Các phương pháp tiếp cận

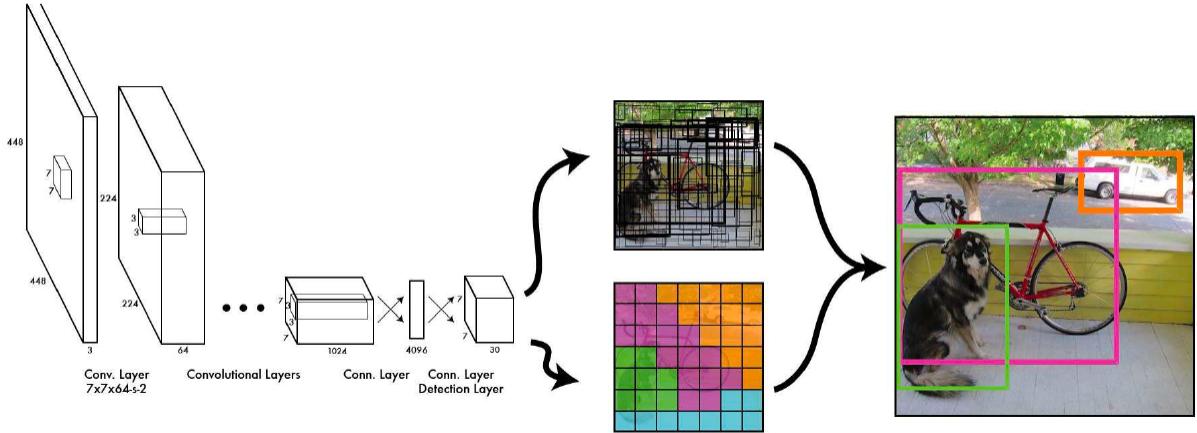
2.2.1 Phát hiện đối tượng

2.2.1.1 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. Hình [2.1](#) minh họa tổng quan quy trình phát hiện đối tượng.

2.2.1.2 Các nghiên cứu liên quan

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của học sâu (deep learning), bài toán phát hiện đối tượng đã đạt được những bước tiến vượt bậc cả về độ chính xác lẫn tốc độ, thúc đẩy sự phát triển mạnh mẽ của nhiều ứng dụng thực tế như giám sát thông minh, phân tích video, robot tự hành và xe tự lái. Sự gia tăng về độ phức tạp của dữ liệu ảnh, cùng yêu cầu ngày càng cao về độ chính xác và tốc độ xử lý, đã dẫn



Hình 2.1: Hình ảnh minh họa tổng quan quy trình phát hiện đối tượng. Ảnh đầu vào được xử lý qua mạng nơ-ron để trích xuất đặc trưng và dự đoán vị trí (hộp giới hạn) cùng phân loại các đối tượng (nhãn lớp) xuất hiện trong khung hình.

đến sự ra đời của nhiều hướng tiếp cận khác nhau cho bài toán này. Dựa trên khảo sát tổng quan của [45], các phương pháp phát hiện đối tượng tiên tiến hiện nay 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):** Các phương pháp hai giai đoạn tiếp cận bài toán phát hiện đối tượng bằng cách tách biệt quá trình đề xuất vùng chứa đối tượng (region proposal) và quá trình phân loại định vị chi tiết. Nhóm này tiêu biểu bởi các mô hình thuộc họ R-CNN, chẳng hạn như **R-CNN** [12], **Fast R-CNN** [11] và **Faster R-CNN** [30]. Trong đó, R-CNN sử dụng các thuật toán đề xuất vùng thủ công kết hợp với CNN để trích xuất đặc trưng, trong khi Fast R-CNN cải thiện hiệu quả bằng cách chia sẻ đặc trưng toàn ảnh. Faster R-CNN tiếp tục nâng cao hiệu suất bằng cách giới thiệu mạng Region Proposal Network (RPN), cho phép học tự động các vùng đề xuất.

Nhờ khả năng tách biệt rõ ràng giữa phát hiện và phân loại, các phương pháp hai giai đoạn thường đạt độ chính xác cao, đặc biệt trong các kịch bản phức tạp, nhưng đòi hỏi chi phí tính toán lớn và tốc độ xử lý chậm.

- **Các phương pháp một giai đoạn (One-Stage):** Khác với các phương pháp hai giai đoạn, các mô hình một giai đoạn thực hiện trực tiếp việc dự đoán nhãn lớp và

hộp giới hạn trong một bước duy nhất, không cần cơ chế đề xuất vùng riêng biệt. Với các đại diện nổi bật như **YOLO** [29], **SSD** [20] và **RetinaNet** [18]. YOLO tiếp cận phát hiện đối tượng như một bài toán hồi quy toàn cục, cho phép suy luận nhanh và phù hợp với các ứng dụng thời gian thực. SSD khai thác đặc trưng đa tỷ lệ nhằm cải thiện khả năng phát hiện các đối tượng có kích thước khác nhau. RetinaNet giải quyết vấn đề mất cân bằng giữa các lớp thông qua hàm mất mát Focal Loss, giúp nâng cao độ chính xác cho các đối tượng khó phát hiện.

Nhìn chung, các phương pháp một giai đoạn (One-Stage) đạt được sự cân bằng tốt giữa tốc độ và độ chính xác, nhưng đôi khi kém ổn định hơn trong các bối cảnh có mật độ đối tượng cao hoặc chồng lấn mạnh.

- **Các phương pháp dựa trên Transformer:** Gần đây, các phương pháp dựa trên Transformer đã tạo ra một bước chuyển quan trọng trong phát hiện đối tượng bằng cách xây dựng kiến trúc end-to-end, loại bỏ các thành phần được thiết kế thủ công như anchor boxes và thuật toán Non-Maximum Suppression (NMS). Diễn hình cho hướng đi này là mô hình **DETR** [4] trong đó bài toán phát hiện đối tượng được mô hình hóa như một bài toán gán tập (set prediction) thông qua cơ chế self-attention. Các biến thể sau đó của DETR tập trung vào cải thiện tốc độ hội tụ và hiệu suất suy luận, mở ra hướng nghiên cứu mới cho các mô hình phát hiện đối tượng. Bên cạnh đó, khả năng mô hình hóa quan hệ toàn cục và thiết kế kiến trúc end-to-end của Transformer cũng đã chứng minh hiệu quả trong nhiều bài toán thị giác máy tính liên quan, bao gồm cả phân đoạn ảnh.

2.2.2 Phát hiện và nhận dạng văn bản ngoại cảnh (Scene Text Detection and Recognition - STDR)

2.2.2.1 Phát hiện văn bản ngoại cảnh (Scene Text Detection - STD)

Cơ sở và hướng tiếp cận chung Phát hiện văn bản trong ảnh ngoại cảnh (Scene Text Detection) 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 mặt 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.

Các nghiên cứu liên quan

Dựa trên các nghiên cứu khảo sát và tổng quan gần đây [14, 26, 25, 13] về phát hiện văn bản trong ảnh ngoại cảnh, các phương pháp tiên tiến hiện nay có thể được phân thành ba nhóm chính: (i) dựa trên hồi quy (regression-based), (ii) dựa trên phân đoạn (segmentation-based) và (iii) 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. Liao và cộng sự đã đề xuất **TextBoxes** [16], với điều chỉnh hình dạng convolutional kernel và anchor của SSD để phù hợp với tỷ lệ co giãn đa dạng của văn bản cảnh, cải thiện khả năng phát hiện văn bản đa hướng. **EAST** [43] đề xuất một pipeline hiệu quả, dự đoán trực tiếp khung bao xoay (rotated box) hoặc tứ giác (quadrangle) từ đặc trưng hình ảnh, loại bỏ sự phụ thuộc vào các bước đề xuất vùng (region proposal) phức tạp. Để xử lý văn bản hình dạng bất kỳ, Liu và cộng sự đề xuất **ABCNet** [21], sử dụng đường cong Bezier (Bezier curve) làm biểu diễn tham số hóa linh hoạt cho đường biên văn bản, cho phép mô hình hóa chính xác các văn bản cong. **FCE-Net** [44] đề xuất một cách biểu diễn khác thông qua phép nhúng chuỗi Fourier (Fourier contour embedding), giúp biểu diễn hiệu quả và tái tạo các đường biên văn bản phi chuẩn từ đầu ra hồi quy.

Tuy nhiên, một hạn chế chung của hướng tiếp cận này là sự phụ thuộc vào các bước hậu xử lý (post-processing) phức tạp để phục hồi thể hiện văn bản từ đầu ra

hồi quy.

- **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. Long và cộng sự đề xuất **TextSnake** [22], mô hình hóa văn bản cong bằng một chuỗi các hình tròn linh hoạt (các "vảy rắn") dọc theo trục trung tâm. Baek và cộng sự **CRAFT** [2] dự đoán bản đồ "affinity" giữa các ký tự thông qua học chuyển giao từ dữ liệu tổng hợp, cung cấp hướng dẫn rõ ràng cho việc nhóm. **DRRG** [38] được đề xuất bởi Zhang và cộng sự, sử dụng Mạng Tích chập Đồ thị (Graph Convolutional Network - GCN) để mô hình hóa mối quan hệ cấu trúc giữa các thành phần văn bản và thực hiện việc nhóm một cách thông minh.

Mặc dù có thể biểu diễn chính xác các văn bản cong, hiệu quả cuối cùng của hướng tiếp cận này phụ thuộc nhiều vào các thuật toán nhóm (grouping) phức tạp, điều này có thể ảnh hưởng đến tốc độ xử lý và độ ổn định tổng thể.

- **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, phân loại mỗi pixel là văn bản hoặc nền, sau đó suy ra các vùng văn bản từ kết quả phân đoạn. **PANet** [19] được đề xuất để tổng hợp đặc trưng đa tỷ lệ và nhóm chính xác các pixel văn bản. Liao và cộng sự đề xuất **DBNet++** [17], tích hợp differentiable binarization và Adaptive Scale Fusion nhằm giảm thiểu bước hậu xử lý và cải thiện độ chính xác. **TextPMs** [37] sử dụng nhóm bản đồ xác suất và mô hình học lặp để phục hồi văn bản cong một cách chính xác. **FAST** [5] và **KPN** [39] tập trung vào việc xử lý hiệu quả các văn bản phi chuẩn và đa tỷ lệ.

Nhìn chung, phương pháp này đặc biệt hiệu quả trong việc xử lý các văn bản có hình dạng cong, hoặc nghiêng. Tuy nhiên, nó thường đòi hỏi chi phí tính toán cao hơn so với các phương pháp dựa trên hồi quy trực tiếp.

2.2.2.2 Nhận dạng văn bản (Scene Text Recognition - STR)

Cơ sở và hướng tiếp cận chung Nhận dạng văn bản trong ảnh ngoại cảnh (Scene Text Recognition) là một bài toán phức tạp trong lĩnh vực Thị giác Máy tính, liên quan đến việc đọc và xác định nội dung của văn bản xuất hiện trong các cảnh ảnh thực tế. Khác với các bài toán nhận dạng văn bản truyền thống trên tài liệu hoặc bảng biểu, văn bản ngoại cảnh thường xuất hiện trong môi trường không đồng nhất, chịu biến dạng, thay đổi lớn về ánh sáng, góc chụp, nền, phông chữ và hình thức trình bày. Mục tiêu là nhận diện chính xác nội dung của các ký tự và từ ngữ trong các vùng văn bản đã được khoanh vùng (cropped text instances), chuyển đổi từng hình ảnh văn bản riêng lẻ thành chuỗi ký tự tương ứng. Bài toán không chỉ đòi hỏi nhận diện ký tự riêng lẻ mà còn cần hiểu ngữ cảnh tổng thể của văn bản trong ảnh, bao gồm mối quan hệ giữa các ký tự, từ ngữ và bối cảnh, đặc biệt trong các điều kiện biến dạng, cong, nghiêng hoặc không chuẩn.

Các nghiên cứu liên quan Scene Text Recognition (STR) là một lĩnh vực nghiên cứu thu hút sự quan tâm mạnh mẽ trong cộng đồng thị giác máy tính. Trong các nghiên cứu khảo sát và tổng quan gần đây [14, 26, 25, 7], bài toán nhận dạng văn bản trong ảnh ngoại cảnh có thể được chia thành 2 loại chính dựa trên nguyên lý làm việc: (i) các phương pháp dựa trên phân đoạn ký tự (segmentation-based) và (ii) các phương pháp không dựa trên phân đoạn (segmentation-free).

- **Các phương pháp dựa trên phân đoạn ký tự (Segmentation-based):** Nhóm phương pháp này tiếp cận bài toán STR bằng cách dự đoán nhãn mức pixel cho từng ký tự hoặc thành phần ký tự, sau đó thực hiện nhận dạng dựa trên kết quả phân đoạn. **MaskTextSpotter** [24] được đề xuất bởi Lyu và cộng sự, sử dụng mạng phân đoạn ký tự kết hợp cơ chế spatial attention để nhận dạng văn bản hình dạng bất kỳ, khắc phục hạn chế do thiếu dữ liệu chú thích cấp ký tự. Để cải thiện độ chính xác trong bối cảnh phức tạp, Ye và cộng sự đề xuất **TextFuseNet** [35], một kiến trúc tích hợp thông tin đa cấp (character-, word-, và global-level), giúp phân đoạn ký tự mạnh mẽ hơn.

Tuy đạt hiệu quả cao trong việc biểu diễn văn bản phi chuẩn, các phương pháp trong nhóm này thường phụ thuộc vào chất lượng của bước phân đoạn ký tự, đồng

thời đòi hỏi quy trình hậu xử lý phức tạp, dẫn đến chi phí tính toán cao và khó mở rộng trong các kịch bản thực tế

- **Các phương pháp không dựa trên phân đoạn (Segmentation-free):** Nhóm phương pháp này tập trung vào việc trực tiếp ánh xạ toàn bộ vùng ảnh văn bản (word hoặc text line) thành chuỗi ký tự đầu ra, thông qua một kiến trúc encoder-decoder. Các phương pháp truyền thống trong nhóm này thường sử dụng mạng CNN để trích xuất đặc trưng thị giác, kết hợp với mô hình chuỗi như BiLSTM để nắm bắt quan hệ ngữ cảnh, và sử dụng CTC hoặc attention làm cơ chế dự đoán. Tiêu biểu là **CRNN [31]**, trong đó Shi và cộng sự kết hợp CNN, RNN và CTC loss để thực hiện nhận dạng chuỗi ký tự một cách hiệu quả.

Gần đây, các phương pháp tiên tiến dựa trên Transformer đã đạt được nhiều kết quả vượt trội. **ViTSTR [1]** được đề xuất nhằm áp dụng Vision Transformer trực tiếp cho bài toán STR, khai thác cơ chế self-attention để mô hình hóa quan hệ toàn cục trong chuỗi đặc trưng. **PARSeq [3]** tiếp tục mở rộng hướng tiếp cận này bằng cách kết hợp mô hình Transformer tự hồi quy với ngữ cảnh hai chiều, giúp cải thiện độ chính xác trong nhận dạng chuỗi dài và phức tạp. Bên cạnh đó, một số nghiên cứu gần đây như **CDistNet [42]**, **SMTR [8]** và **SVTRv2 [?]** tập trung vào việc thiết kế kiến trúc hiệu quả hơn cho STR, thông qua cải tiến cơ chế trích xuất đặc trưng, mô hình hóa chuỗi hoặc tối ưu hóa cấu trúc Transformer, nhằm cân bằng giữa độ chính xác và chi phí tính toán.

Nhìn chung, hướng tiếp cận này có kiến trúc tương đối gọn nhẹ và khả năng mở rộng tốt nhờ không phụ thuộc vào chú thích ký tự chi tiết. Tuy nhiên, việc không sử dụng phân đoạn tường minh khiến các phương pháp này gặp hạn chế khi xử lý văn bản có hình dạng phức tạp, cong hoặc biến dạng mạnh.

2.2.2.3 Nhận dạng văn bản ngoại cảnh đầu–cuối (End-to-End Scene Text Recognition)

Cơ sở và hướng tiếp cận chung Nhận dạng văn bản ngoại cảnh đầu–cuối (End-to-End Scene Text Recognition) hướng tới mục tiêu giải quyết đồng thời cả hai bài toán phát hiện văn bản (text detection) và nhận dạng văn bản (text recognition) trong một pipeline



Hình 2.2: Hình ảnh minh họa quá trình nhận dạng văn bản ngoại cảnh đầu-cuối (End-to-End Scene Text Recognition)

thống nhất, thay vì xử lý chúng như hai tác vụ tách biệt. Khác với các hệ thống truyền thống theo pipeline tuần tự, trong đó kết quả phát hiện văn bản được sử dụng làm đầu vào cho bước nhận dạng, các phương pháp end-to-end tìm cách tối ưu hóa toàn bộ quá trình từ ảnh đầu vào đến chuỗi ký tự đầu ra một cách thống nhất. Quy trình tổng quát của hướng tiếp cận này được minh họa trong [Hình 2.2](#)

Cách tiếp cận này cho phép mô hình học được mối quan hệ chặt chẽ giữa vị trí, hình dạng và nội dung của văn bản trong ảnh, từ đó giảm thiểu sự phụ thuộc vào các bước trung gian và hạn chế sai lệch lan truyền giữa các giai đoạn. Do đó, End-to-End Scene Text Recognition được xem là hướng tiếp cận hiệu quả cho các ứng dụng thực tế đòi hỏi độ chính xác cao và quy trình xử lý gọn nhẹ.

Các nghiên cứu liên quan Trong bối cảnh ngày càng nhiều ứng dụng thực tế yêu cầu trích xuất thông tin văn bản trực tiếp từ ảnh, chẳng hạn như dịch tự động, phân tích nội dung hình ảnh hay hỗ trợ người dùng trong môi trường thông minh, việc xử lý phát hiện và nhận dạng văn bản trong một pipeline thống nhất ngày càng trở nên cần thiết. Chính vì vậy, nhiều nghiên cứu gần đây tập trung vào bài toán nhận dạng văn bản ngoại cảnh đầu-cuối (End-to-End Scene Text Recognition), nhằm đồng thời giải quyết hai nhiệm vụ phát hiện và nhận dạng văn bản trong cùng một pipeline.

Theo các nghiên cứu khảo sát và tổng quan gần đây [14, 26, 25, 7], bài toán nhận dạng văn bản ngoại cảnh có thể được chia thành hai hướng tiếp cận chính: (i) các phương pháp hai giai đoạn (two-stage scene text spotters) và (ii) các phương pháp một giai đoạn (one-stage scene text spotters).

- **Các phương pháp hai giai đoạn (Two-Stage Scene Text Spotters):** Các phương pháp thuộc nhóm này tiếp cận bài toán text spotting bằng cách kết hợp một mô-đun phát hiện văn bản và một mô-đun nhận dạng văn bản riêng biệt trong một pipeline xử lý tuần tự. Tiêu biểu cho hướng tiếp cận này, **TextBoxes [16]** sử dụng bộ phát hiện dựa trên SSD và bộ nhận dạng CRNN, đặt nền móng cho kiến trúc hai giai đoạn trong bài toán text spotting. Tuy nhiên, việc tối ưu tách biệt hai mô-đun có thể gây lan truyền lỗi do thiếu sự phối hợp giữa phát hiện và nhận dạng, từ đó làm suy giảm độ chính xác tổng thể. Gần đây, **MaskTextSpotter [24]** sử dụng mô-đun Region-of-Interest (RoI) để trích xuất các vùng ứng viên và đưa vào nhánh Fast R-CNN nhằm sinh bản đồ phân đoạn ngữ nghĩa, cho phép xử lý hiệu quả văn bản có hình dạng bất kỳ. Một hướng tiếp cận khác được thể hiện trong **ABCNet [21]** với việc sử dụng BezierAlign, một phép biến đổi có tham số học được giúp chuyển đổi chính xác các vùng văn bản hình dạng bất kỳ (đặc biệt là văn bản cong) thành các đặc trưng đầu vào chuẩn cho bộ nhận dạng.

Mặc dù hướng tiếp cận hai giai đoạn (Two-Stage) mang lại hiệu quả cao nhờ khả năng kết hợp các mô-đun phát hiện và nhận dạng mạnh mẽ trong cùng một hệ thống, cấu trúc xử lý tuần tự và sự phụ thuộc vào các bước trung gian như đề xuất vùng (Region of Interest - RoI) khiến các phương pháp này gặp thách thức về hiệu suất và khả năng mở rộng, đặc biệt trong các ứng dụng yêu cầu xử lý nhanh và gọn.

- **Các phương pháp một giai đoạn (One-Stage Scene Text Spotters):** Nhằm giảm thiểu sự phụ thuộc vào các bước trung gian như đề xuất vùng (Region of Interest - RoI) và đơn giản hóa pipeline xử lý, các phương pháp một giai đoạn tích hợp trực tiếp phát hiện và nhận dạng văn bản vào một mạng duy nhất, cho phép dự đoán văn bản theo cách đầu-cuối (end-to-end). **PGNet [33]** dự đoán văn bản một cách trực tiếp thông qua việc học chuỗi các điểm trung tâm. Trong khi đó, **DeepSolo [36]**, lấy cảm hứng từ ABCNet, đề xuất cơ chế biểu diễn đường cong trung tâm Bezier đơn giản hơn kết hợp với công thức truy vấn mới, cho phép phân loại ký tự chỉ thông qua phép chiếu tuyến tính từ các đặc trưng truy vấn.

Bên cạnh đó, một số nghiên cứu gần đây như **TESTR** [40], **UNITS** [15] và **DNTextSpotter** [27] tập trung vào việc thiết kế kiến trúc thống nhất cho bài toán text spotting, thông qua khai thác Transformer, cơ chế truy vấn hoặc biểu diễn đặc trưng linh hoạt, nhằm cải thiện khả năng học đầu–cuối và giảm sự phụ thuộc vào các bước trung gian.

Như vậy, hướng tiếp cận một giai đoạn (One-Stage) giúp đơn giản hóa kiến trúc và giảm độ trễ suy luận nhờ loại bỏ các bước xử lý trung gian. Song, việc học đồng thời hai nhiệm vụ trong một kiến trúc duy nhất khiến mô hình dễ gặp khó khăn trong việc cân bằng giữa độ chính xác phát hiện và khả năng nhận dạng, đặc biệt trong các điều kiện dữ liệu phức tạp.

Chapter 3

PHƯƠNG PHÁP TIẾP CẬN

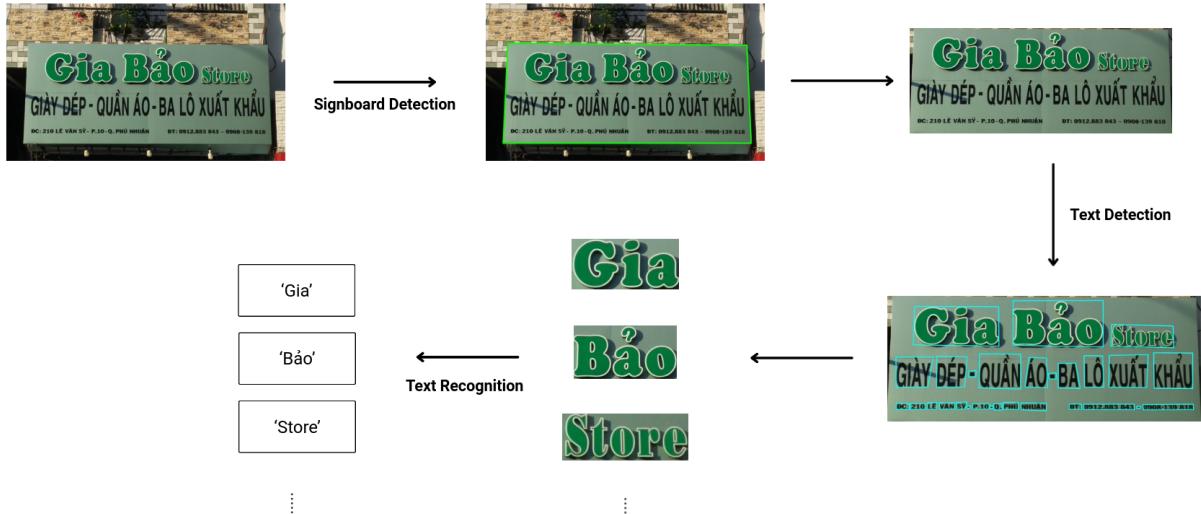
3.1 Tổng quan về phương pháp

Trong những năm gần đây, việc khai thác thông tin từ văn bản trên biển hiệu có ý nghĩa quan trọng trong việc xây dựng hệ sinh thái dữ liệu cho các ứng dụng dựa trên vị trí và quản lý đô thị. Xuất phát từ mục tiêu đó, khóa luận đề xuất một pipeline xử lý được thiết kế theo ba giai đoạn chính: (i) **Phát hiện biển hiệu**, (ii) **Phát hiện văn bản trên biển hiệu**, và (iii) **Nhận dạng nội dung văn bản**. Theo đó, pipeline được minh họa trong Hình 3.1, trong đó quá trình xử lý bắt đầu bằng việc phát hiện và trích xuất vùng biển hiệu từ ảnh đầu vào. Từ các vùng biển hiệu đã được cắt, hệ thống tiến hành phát hiện các vùng văn bản tương ứng, trước khi thực hiện nhận dạng nội dung văn bản từ các vùng đã được xác định. Dựa trên pipeline tổng thể này, khóa luận tập trung phân tích và lựa chọn một số phương pháp tiên tiến hiện nay cho từng giai đoạn xử lý. Việc lựa chọn này được thực hiện dựa trên các nghiên cứu khảo sát gần đây trong lĩnh vực. Nội dung chi tiết cho từng giai đoạn sẽ được trình bày lần lượt trong các mục sau.

3.2 Phát hiện biển hiệu

3.2.1 Phát hiện đối tượng

Trong giai đoạn phát hiện biển hiệu, khóa luận lựa chọn một số phương pháp tiêu biểu được đề cập trong các nghiên cứu gần đây [45] để tiến hành đánh giá thực nghiệm.

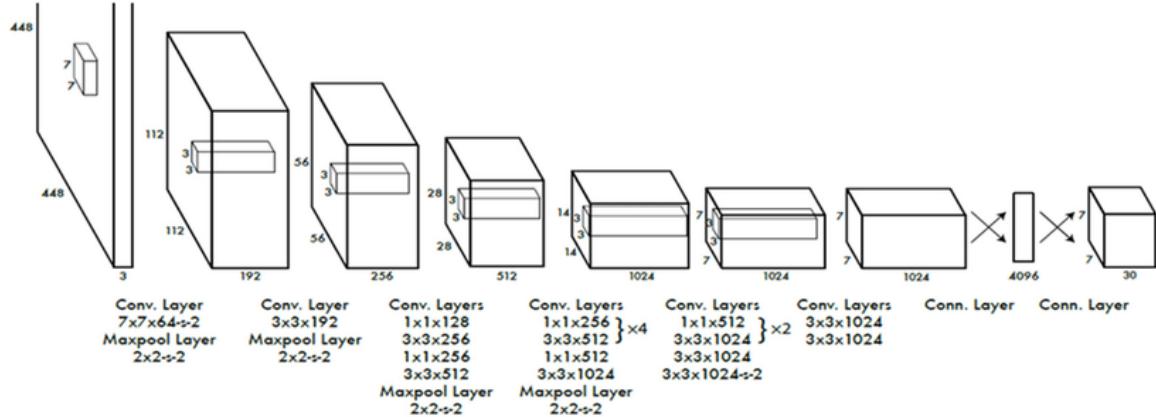


Hình 3.1: Kiến trúc tổng quan của hệ thống phát hiện và nhận dạng văn bản trên biển hiệu

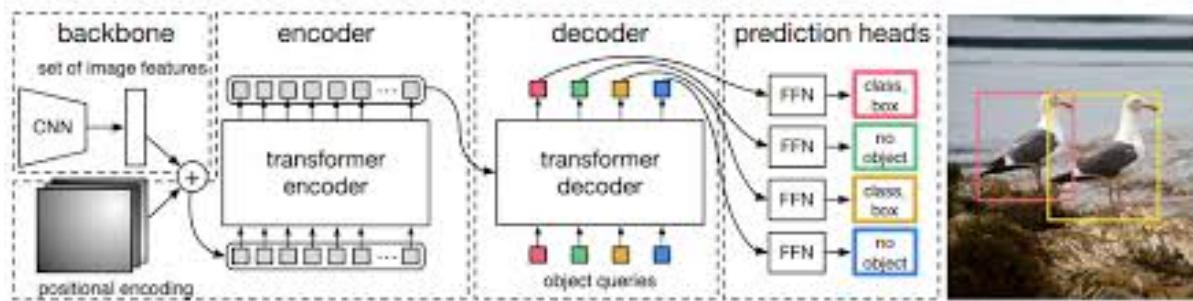
YOLO (You Only Look Once) YOLO [29] được đề xuất bởi Redmon và cộng sự, là đại diện tiêu biểu cho hướng tiếp cận một giai đoạn (one-stage) trong bài toán phát hiện đối tượng. Khác với các phương pháp hai giai đoạn, YOLO tiếp cận bài toán như một bài toán hồi quy toàn cục, trong đó mô hình dự đoán trực tiếp các hộp giới hạn (bounding boxes) cùng xác suất lớp (class probabilities) trên toàn bộ ảnh đầu vào. Kiến trúc tổng quát của YOLO được minh họa trong Hình 3.2, bao gồm ba thành phần chính: backbone dùng để trích xuất đặc trưng, neck nhằm kết hợp đặc trưng đa tỉ lệ, và detection head để thực hiện dự đoán hộp bao và nhãn phân loại.

Trải qua nhiều phiên bản phát triển, YOLO liên tục được cải tiến nhằm nâng cao độ chính xác trong khi vẫn duy trì hiệu quả tính toán. Do đó, khóa luận lựa chọn một số phiên bản YOLO gần đây, chẳng hạn như YOLOv8 và YOLOv11 để đưa vào thực nghiệm, qua đó đánh giá hiệu quả của phương pháp trong giai đoạn phát hiện biển hiệu. Bên cạnh đó, các biến thể YOLO hỗ trợ phát hiện hộp xoay (Oriented Bounding Box – OBB) cũng được đưa vào đánh giá, nhằm xử lý hiệu quả hơn các trường hợp biển hiệu có hướng nghiêng hoặc hình dạng không song song với trực ảnh.

DETR (DEtection TRansformer) DETR [4] được đề xuất bởi Carion và cộng sự, là mô hình phát hiện đối tượng đầu tiên hoàn toàn dựa trên kiến trúc Transformer. Khác với



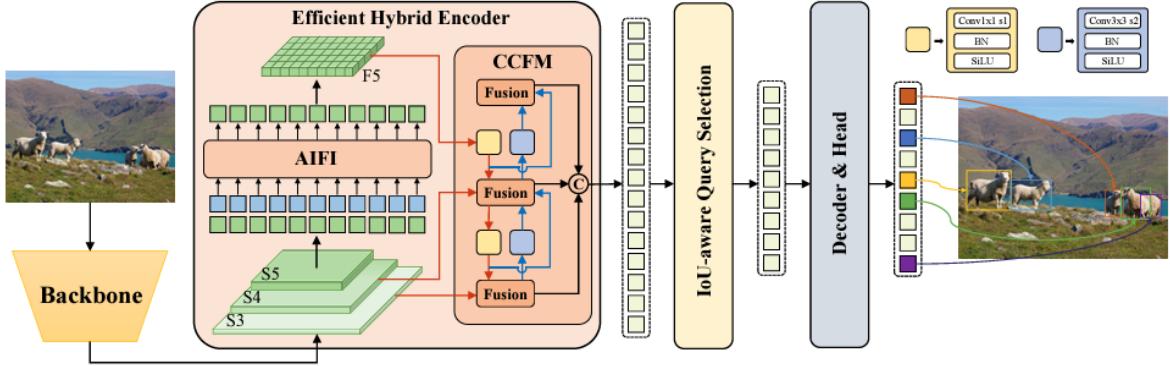
Hình 3.2: Kiến trúc tổng quan của YOLO [29]



Hình 3.3: Kiến trúc tổng quan của DETR [4]

các phương pháp dựa trên anchor truyền thống, DETR tiếp cận bài toán theo hướng dự đoán tập hợp (set prediction), trong đó mỗi đối tượng được ánh xạ trực tiếp thành một phần tử trong tập đầu ra thông qua cơ chế tự chú ý (self-attention). Cách tiếp cận này cho phép mô hình khai thác quan hệ ngữ cảnh toàn bộ ảnh và loại bỏ các bước hậu xử lý phức tạp như Non-Maximum Suppression (NMS). Hình 3.3 minh họa kiến trúc DETR, trong đó sử dụng CNN backbone để trích xuất đặc trưng, sau đó Transformer encoder và decoder phối hợp mô hình hóa ngữ cảnh toàn cục và sinh ra tập hợp các dự đoán cuối cùng thông qua các object queries học được.

Trong bối cảnh phát hiện biến hiệu, DETR được lựa chọn như một phương pháp đại diện cho hướng tiếp cận dựa trên Transformer nhằm đánh giá khả năng khai thác ngữ cảnh toàn cục. Đặc biệt, cơ chế dự đoán tập hợp của DETR giúp giảm thiểu sự phụ thuộc vào các giả định hình học cục bộ, từ đó phù hợp với các trường hợp biến hiệu có



Hình 3.4: Kiến trúc tổng quan của RT-DETR, được sử dụng trong RTDETRv2 [41]

bộ cục đa dạng.

RT-DETRv2 Dựa trên ý tưởng tiếp cận end-to-end của DETR cho bài toán phát hiện đối tượng, Zhao và cộng sự giới thiệu RT-DETRv2 [41] như một phiên bản cải tiến của RT-DETR, với mục tiêu tối ưu hóa hiệu suất thời gian thực trong khi vẫn duy trì độ chính xác cao. Mô hình này giữ nguyên ưu điểm loại bỏ các bước hậu xử lý như Non-Maximum Suppression (NMS), đồng thời được tăng cường bằng các cơ chế tối ưu nhằm cân bằng hiệu quả giữa tốc độ suy luận và chất lượng dự đoán. Kiến trúc của RTDETRv2, minh họa trong Hình 3.4, dựa trên thiết kế RT-DETR gốc, nổi bật với: (i) hybrid encoder kết hợp ưu điểm của CNN trong trích xuất đặc trưng hiệu quả và Transformer trong mô hình hóa ngữ cảnh toàn cục, (ii) cơ chế lựa chọn truy vấn thích ứng giúp giảm số lượng truy vấn không cần thiết, qua đó cải thiện tốc độ suy luận mà ít ảnh hưởng đến độ chính xác.

Trên cơ sở đó, RTDETRv2 được lựa chọn như một phương pháp đại diện cho hướng tiếp cận Transformer tối ưu hóa cho thời gian thực, đặc biệt phù hợp với các trường hợp yêu cầu tốc độ xử lý cao như phân tích video giao thông hoặc cảnh đường phố.

3.2.2 Phân đoạn đối tượng

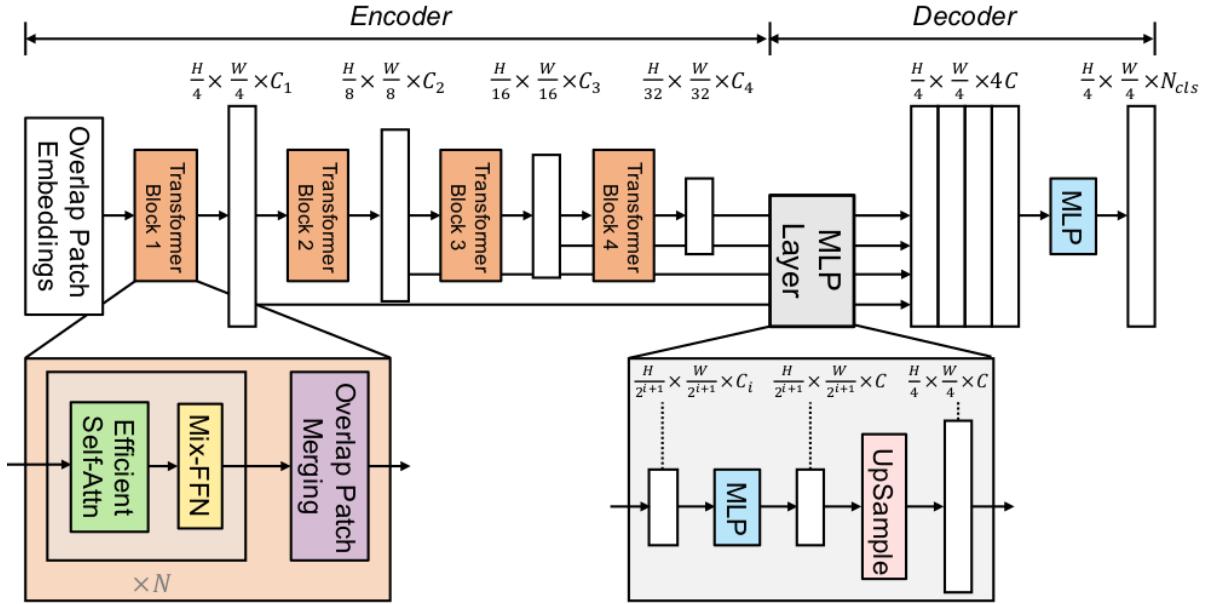
Bên cạnh các phương pháp phát hiện trực tiếp dựa trên bounding box, nhằm mở rộng góc nhìn đánh giá, khóa luận xem xét thêm một hướng tiếp cận gián tiếp thông qua bài

toán phân đoạn ngữ nghĩa (semantic segmentation). Theo hướng tiếp cận này, đối tượng được phân đoạn ở mức điểm ảnh, từ đó suy ra các vùng bao hình học phục vụ cho bài toán phát hiện.

Trong bối cảnh đó, theo nghiên cứu khảo sát gần đây [32], các kiến trúc dựa trên Transformer đã trở thành một hướng tiếp cận quan trọng và được quan tâm rộng rãi 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 tự chú ý (self-attention), các mô hình này cho thấy hiệu quả nổi bật trong việc xử lý các trường hợp phức tạp với sự đa dạng lớn về hình dạng và bối cảnh của đối tượng.

SegFormer SegFormer [34], được giới thiệu bởi Xie và cộng sự, là một kiến trúc phân đoạn ngữ nghĩa hiệu quả, kết hợp encoder Transformer phân cấp (hierarchical) và decoder MLP nhẹ, được minh họa trong Hình 3.5. Thiết kế này cho phép mô hình khai thác ngữ cảnh toàn cục ở nhiều tỷ lệ, đồng thời duy trì hiệu suất tính toán cao nhờ decoder đơn giản. Chính sự cân bằng giữa độ chính xác và tốc độ này khiến SegFormer trở thành một lựa chọn phù hợp để đánh giá hiệu quả của phân đoạn ngữ nghĩa trong việc phát hiện các biển hiệu, đặc biệt đối với các biển hiệu xuất hiện ở nhiều góc nghiêng khác nhau.

Mask2Former Mask2Former [6] được đề xuất bởi Cheng và cộng sự, đại diện cho một hướng tiếp cận phân đoạn thống nhất (unified framework) dựa trên Transformer, có khả năng xử lý linh hoạt 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). Kiến trúc của Mask2Former được trình bày chi tiết trong Hình 3.6, áp dụng cơ chế chú ý có mặt nạ (masked attention), trong đó mỗi truy vấn tập trung vào các vùng đặc trưng liên quan đến mặt nạ (mask) dự đoán, thay vì toàn bộ không gian ảnh. Cách tiếp cận này giúp mô hình cải thiện khả năng biểu diễn hình dạng và ranh giới chi tiết của các đối tượng, đặc biệt hiệu quả trong các trường hợp đối tượng chồng lấn hoặc có cấu trúc hình học phức tạp. Nhờ đó, Mask2Former được lựa chọn để đánh giá khả năng xử lý các biển hiệu trong những trường hợp bị chồng lấn.



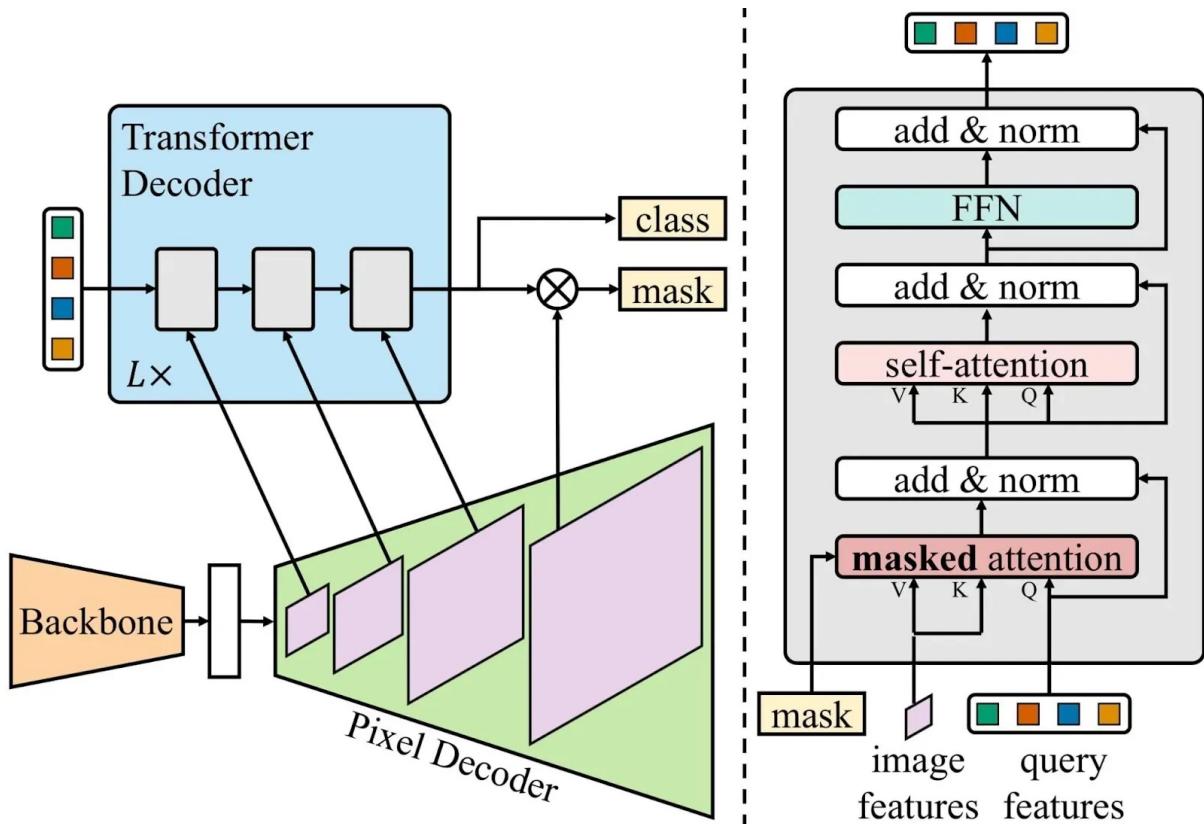
Hình 3.5: Kiến trúc tổng quan của SegFormer [34]

3.3 Phát hiện và nhận dạng văn bản trên biển hiệu theo hướng tiếp cận hai giai đoạn (Two-Stage)

3.3.1 Phát hiện văn bản trên biển hiệu

Trên cơ sở các vùng biển hiệu đã được xác định, hệ thống tiếp tục với nhiệm vụ phát hiện văn bản trên biển hiệu, được xem xét dưới góc độ của bài toán Scene Text Detection (STD). Để đánh giá hiệu quả, các phương pháp tiên tiến hiện nay được lựa chọn dựa trên các nghiên cứu khảo sát gần đây [14, 26, 25].

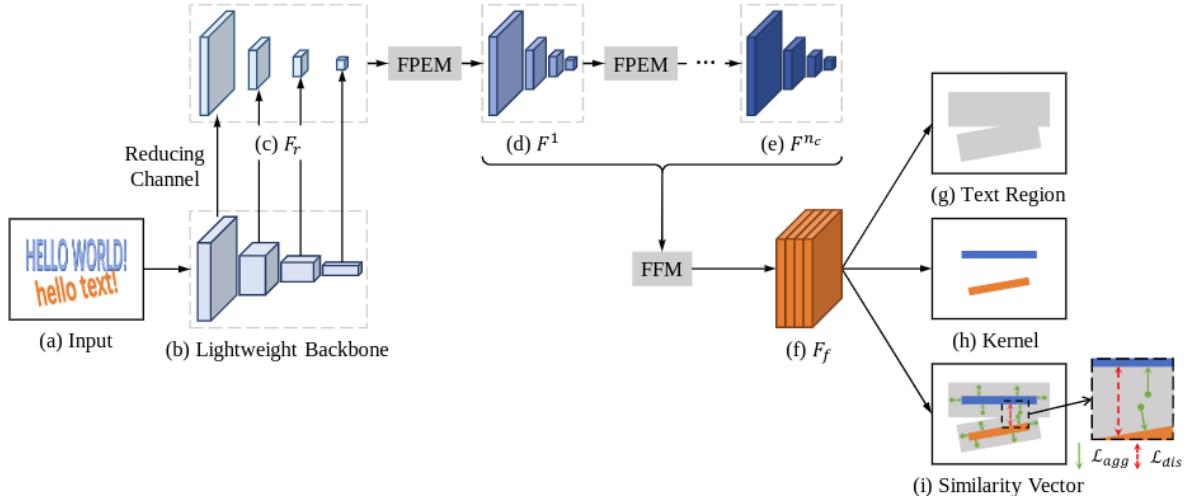
PANet PANet [?], được đề xuất bởi Liu và cộng sự, là một kiến trúc phát hiện văn bản hiệu quả dựa trên nguyên tắc phân đoạn. Mô hình bao gồm hai thành phần chính: Feature Pyramid Enhancement Module (FPEM) để tạo bản đồ đặc trưng đa tỷ lệ, và Feature Fusion Module (FFM) để tổng hợp các đặc trưng này. Kiến trúc của PANet được minh họa trong Hình 3.7. Nhờ khả năng tập hợp các pixel văn bản thành các thể hiện tương ứng trên bản đồ đặc trưng cuối cùng, PANet có thể phát hiện văn bản chính xác mà vẫn duy trì hiệu suất tính toán cao, phù hợp với bài toán phát hiện văn bản trên



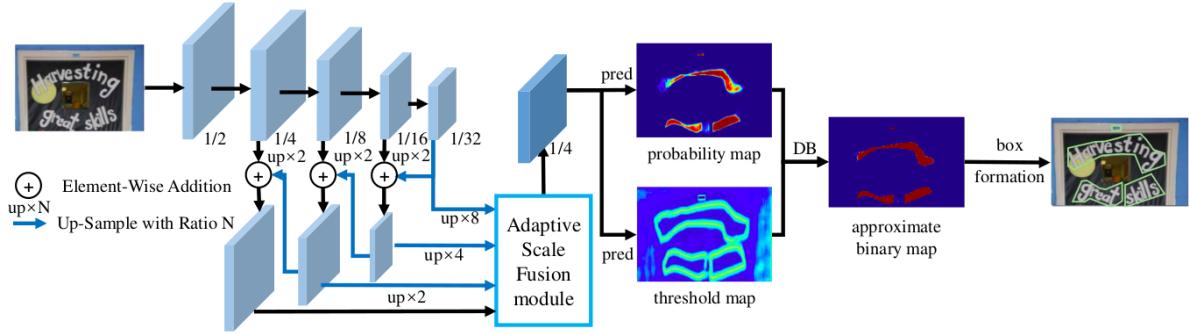
Hình 3.6: Kiến trúc tổng quan của Mask2Former [6]

biển hiệu với nhiều kích thước và hướng khác nhau.

DBNet++ DBNet++ [17], được đề xuất bởi Liao và cộng sự, là phiên bản cải tiến của DBNet, được thiết kế để phát hiện văn bản trong ảnh ngoại cảnh với độ chính xác cao và ổn định. Mô hình tích hợp cơ chế differentiable binarization (DB) trực tiếp vào mạng phân đoạn, giúp tạo mặt nạ (mask) văn bản chính xác và giảm đáng kể các bước hậu xử lý. Bên cạnh đó, module Adaptive Scale Fusion (ASF) được áp dụng để hợp nhất các đặc trưng đa tỷ lệ, nâng cao khả năng phát hiện các văn bản có kích thước khác nhau. Kiến trúc của DBNet++ được minh họa trong Hình 3.8. DBNet++ được lựa chọn nhờ khả năng xử lý hiệu quả các đường biên văn bản không rõ nét, đồng thời phát hiện chính xác cả các dòng chữ lớn (tiêu đề) và nhỏ (thông tin chi tiết) thường cùng xuất hiện trên một biển hiệu.



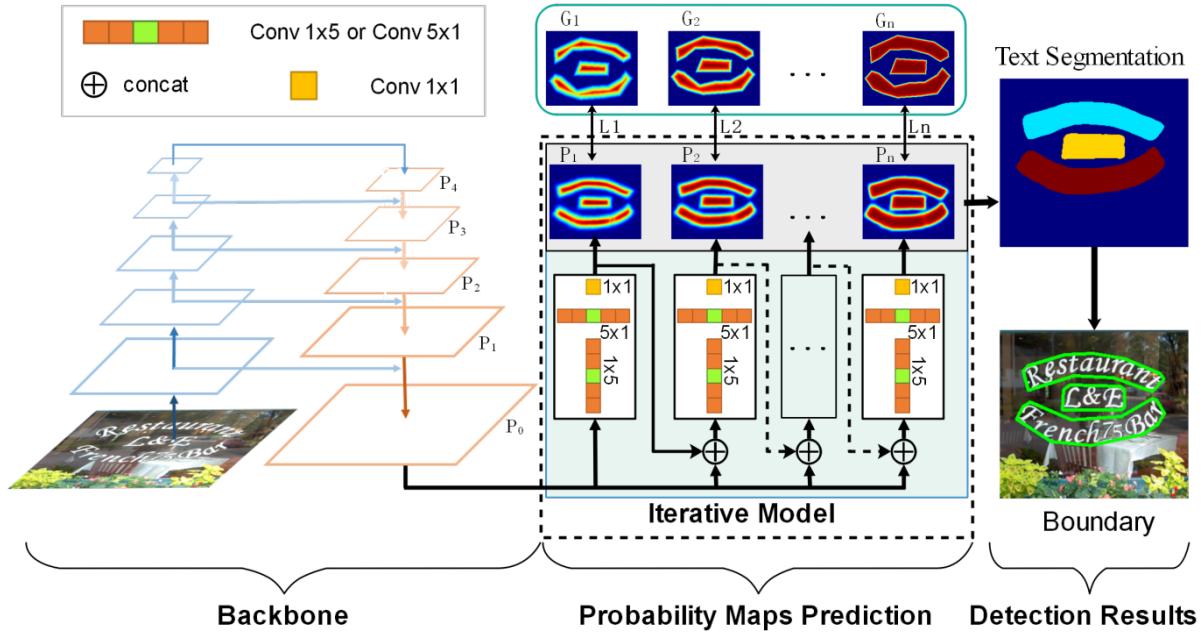
Hình 3.7: Kiến trúc tổng quan của PANet [?]



Hình 3.8: Kiến trúc tổng quan của DBNet++ [17]

TextPMs TextPMs [37] được đề xuất bởi Zhang và cộng sự, thay vì tạo trực tiếp mặt nạ (mask) nhị phân, TextPMs dự đoán một nhóm bản đồ xác suất (probability maps) bằng cách ánh xạ khoảng cách từ pixel đến đường biên đánh dấu (annotation boundary) thành giá trị xác suất, sử dụng các hàm Sigmoid Alpha (SAF). Sau khi dự đoán nhóm bản đồ xác suất, một mô hình học lặp (iterative model) được áp dụng để tổng hợp các bản đồ này, cung cấp thông tin đầy đủ cho việc tái tạo các thể hiện văn bản. Cuối cùng, thuật toán phát triển vùng (region growth) được sử dụng để gộp các bản đồ xác suất thành các đối tượng văn bản hoàn chỉnh. Quy trình dự đoán bản đồ xác suất và phát triển vùng của TextPMs được minh họa trong Hình 3.9

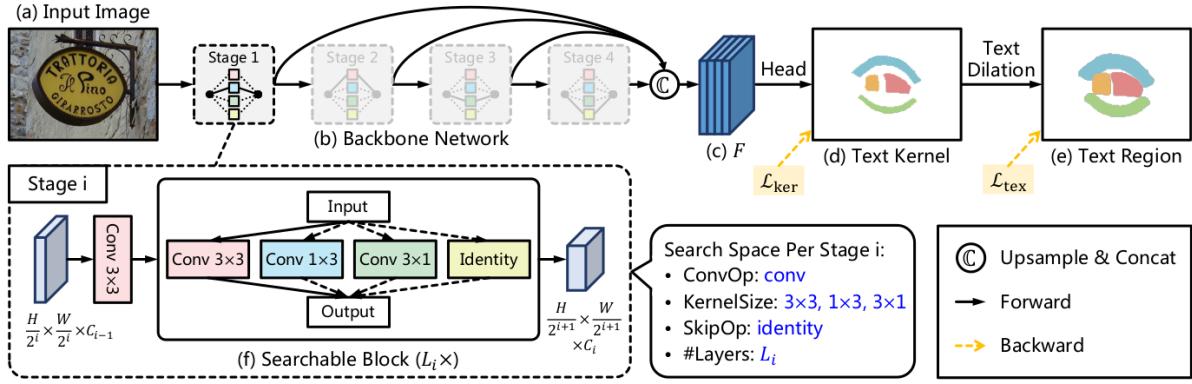
Việc lựa chọn TextPMs dựa trên khả năng phát hiện hiệu quả các văn bản với hình dạng bất thường (như cong hoặc nghiêng), kích thước khác nhau và hướng đa dạng,



Hình 3.9: Kiến trúc tổng quan của TextPMs [17]

đồng thời xử lý tốt các đường biên không rõ nét trên biển hiệu.

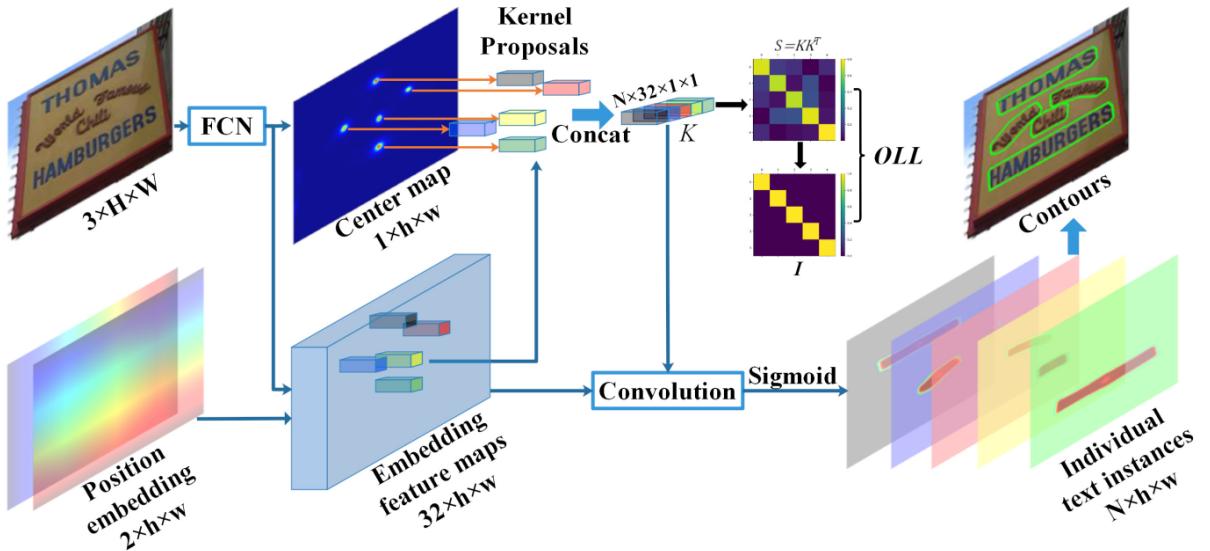
FAST Nhằm phát hiện các văn bản hình dạng tùy ý, đồng thời đảm bảo cả độ chính xác và tốc độ cao, Zhang và cộng sự đã giới thiệu FAST [5], tập trung vào việc đơn giản hóa mô hình và tối ưu hóa quá trình xử lý. Thay vì dựa vào các kiến trúc phức tạp và hậu xử lý nặng, FAST đề xuất một biểu diễn kernel tối giản (minimalist kernel) với đầu ra 1 kênh để mô hình hóa văn bản có hình dạng tùy ý, kết hợp với một quá trình hậu xử lý song song trên GPU nhằm ghép nhanh các dòng chữ với chi phí thời gian không đáng kể. Đồng thời, kiến trúc mạng của FAST được tối ưu hóa thông qua tìm kiếm kiến trúc mạng (neural architecture search) chuyên cho bài toán phát hiện văn bản, giúp trích xuất các đặc trưng mạnh mẽ và phù hợp hơn so với các mạng được thiết kế cho phân loại ảnh. Hình 3.10 cung cấp minh họa trực quan về kiến trúc tối ưu của FAST. Việc lựa chọn FAST dựa trên khả năng phát hiện hiệu quả các văn bản có hình dạng tùy ý, tối ưu cả về tốc độ lẫn độ chính xác, phù hợp với các biển hiệu xuất hiện ở nhiều kích thước, hình dạng và hướng khác nhau.



Hình 3.10: Kiến trúc tổng quan của FAST [5]

KPN Để giải quyết vấn đề tách các thể hiện văn bản liền kề trong hình ảnh ngoại cảnh, một thách thức thường gặp với các văn bản có hình dạng tùy ý. Zhang và cộng sự đề xuất KPN [39], sử dụng Kernel Proposal Network để dự đoán các bản đồ trung tâm Gaussian cho từng văn bản, từ đó trích xuất một tập hợp các kernel proposal động (dynamic convolution kernel) từ bản đồ đặc trưng embedding. Bên cạnh đó, để đảm bảo sự độc lập giữa các kernel, KPN áp dụng hàm mất mát học trực giao (orthogonal learning loss), kết hợp thông tin vị trí và thông tin ngữ nghĩa được mã hóa trong kernel. Các kernel này sau đó được áp dụng riêng rẽ lên bản đồ embedding nhằm tạo ra các bản đồ nhúng tương ứng với từng thể hiện văn bản, qua đó hỗ trợ phân tách rõ ràng các văn bản liền kề. Kiến trúc của KPN được minh họa trong Hình 3.11. Với các đặc điểm trên, KPN được lựa chọn nhờ khả năng phân tách chính xác các văn bản liền kề, đặc biệt phù hợp với các biến hiệu chứa nhiều dòng chữ gần nhau hoặc ký tự dày đặc.

YOLO (OBB) Bên cạnh các phương pháp tiên tiến cho phát hiện văn bản (Scene Text Detection - STD) đã được trình bày, khóa luận tiếp tục mở rộng đánh giá bằng cách áp dụng một số mô hình phát hiện đối tượng được giới thiệu ở Mục 3.2.1 (Phát hiện biến hiệu), cụ thể là phiên bản YOLOv8-OBB và YOLOv11-OBB. Do được huấn luyện ban đầu trên dữ liệu đối tượng tổng quát (general object), các mô hình này cần được fine-tune trên tập dữ liệu văn bản chuyên biệt. Việc đánh giá này nhằm xác định tính khả thi và hiệu quả của các kiến trúc phát hiện đối tượng khi chuyển giao (transfer) sang bài toán phát hiện văn bản, đặc biệt trong việc xử lý các dòng chữ nghiêng và có kích thước



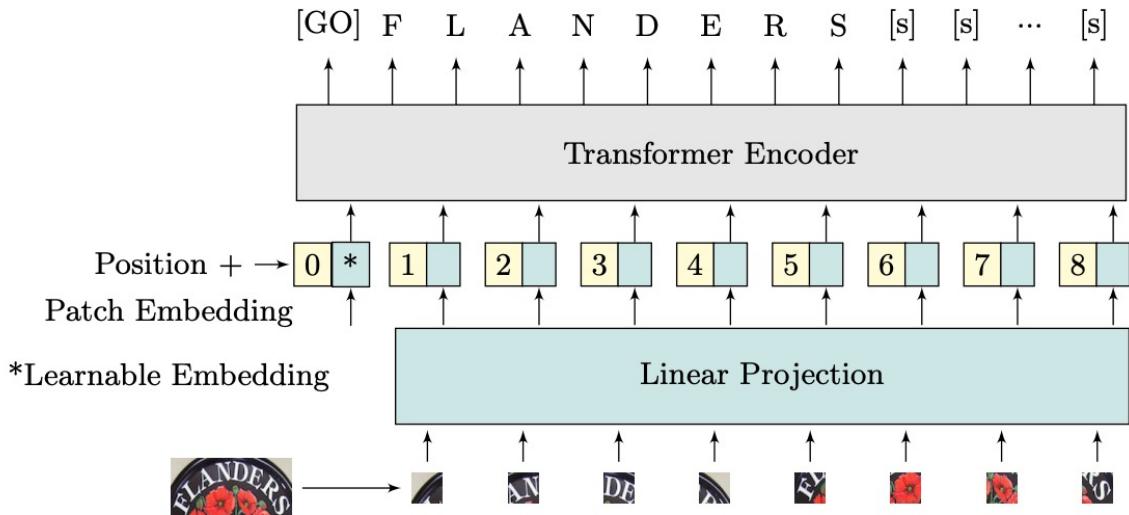
Hình 3.11: Kiến trúc tổng quan của KPN [39]

nhỏ trên biển hiệu.

3.3.2 Nhận dạng nội dung văn bản

Sau khi xác định các vùng chứa văn bản trên biển hiệu, hệ thống tiếp tục với giai đoạn nhận dạng nội dung văn bản. Giai đoạn này được tiếp cận như một bài toán Nhận dạng văn bản trong ảnh ngoại cảnh (Scene Text Recognition - STR), với mục tiêu chuyển đổi các vùng văn bản đã được phát hiện thành chuỗi ký tự tương ứng. Trên cùng cơ sở tiếp cận như giai đoạn phát hiện văn bản, khóa luận lựa chọn một số phương pháp STR tiên tiến hiện nay để tiến hành thực nghiệm và đánh giá, dựa trên các phân loại và hướng tiếp cận được tổng hợp từ các nghiên cứu khảo sát gần đây [14, 26, 25].

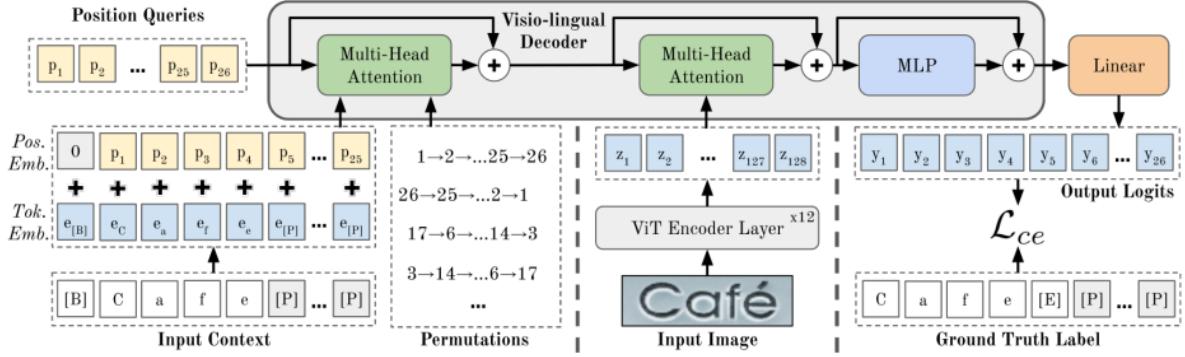
ViTSTR ViTSTR [1], được đề xuất bởi Atienza và cộng sự, đại diện cho một hướng tiếp cận đơn giản và hiệu quả khi áp dụng kiến trúc Vision Transformer cho bài toán nhận dạng văn bản trong ảnh ngoại cảnh. Mô hình sử dụng kiến trúc một giai đoạn, bao gồm 12 khối encoder Transformer giống nhau và không sử dụng decoder, như được minh họa trong Hình 3.12. Trong thiết kế này, việc dự đoán được thực hiện thông qua một lớp tuyến tính, ánh xạ trực tiếp các đặc trưng đã được mã hóa thành chuỗi ký tự đầu



Hình 3.12: Kiến trúc tổng quan của ViTSTR [1]

ra. Nhờ kiến trúc tối giản, ViTSTR đạt được hiệu quả tính toán cao, thể hiện qua tốc độ suy luận nhanh và số lượng tham số nhỏ, tạo ra một giải pháp nhẹ và nhanh phù hợp với giai đoạn nhận dạng văn bản trên biển hiệu. Ngoài ra, mô hình còn áp dụng các kỹ thuật tăng cường dữ liệu đa dạng nhằm cải thiện độ chính xác.

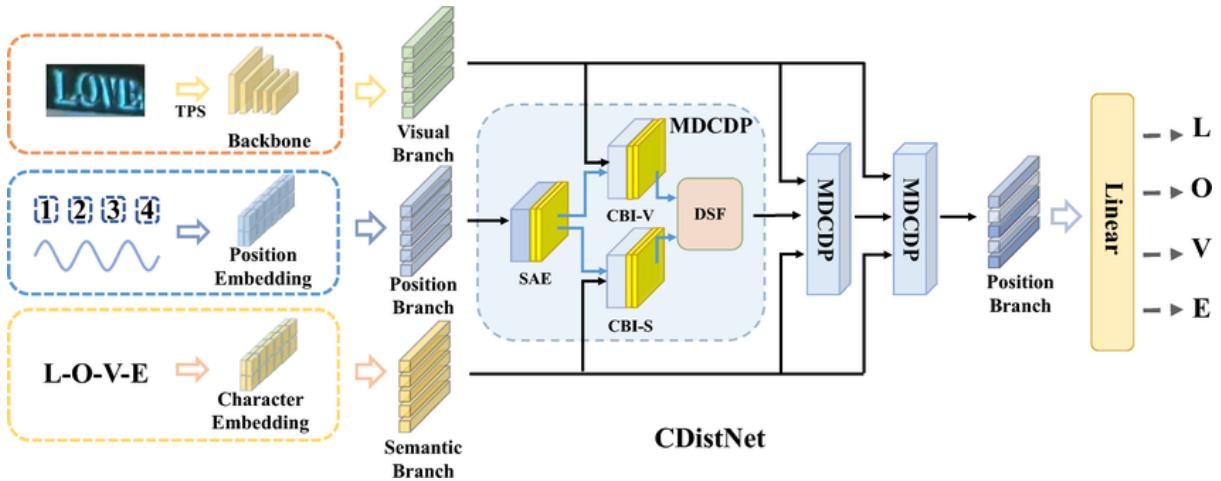
PARSeq Nhằm khắc phục những hạn chế của các mô hình ngôn ngữ tự hồi quy (autoregressive - AR) truyền thống, Bautista và cộng sự đã giới thiệu PARSeq [3]. Phương pháp này tận dụng kỹ thuật Permutation Language Modeling để huấn luyện một tập hợp các mô hình ngôn ngữ AR nội bộ có trọng số chung, qua đó cho phép kết hợp linh hoạt giữa suy luận không tự hồi quy mang tính độc lập ngữ cảnh (context-free non-AR) và suy luận tự hồi quy có xét đến ngữ cảnh chuỗi (context-aware AR). Trên cơ sở đó, PARSeq tích hợp cơ chế tinh chỉnh lặp (iterative refinement) dựa trên ngữ cảnh hai chiều nhằm khai thác hiệu quả thông tin ngữ cảnh mà không cần đến mô hình ngôn ngữ bên ngoài hay quy trình xử lý nhiều giai đoạn phức tạp. Nhờ vậy, mô hình thể hiện tính mạnh mẽ trước các văn bản có hướng và bố cục đa dạng, giúp nâng cao hiệu quả cho giai đoạn nhận dạng văn bản trên biển hiệu. Kiến trúc tổng quan của PARSeq được trình bày trong Hình 3.13.



Hình 3.13: Kiến trúc tổng quan của PARSeq [3]

CDistNet Để khắc phục hạn chế trong việc kết hợp thông tin thị giác và ngữ nghĩa vốn thường không được căn chỉnh chính xác, đặc biệt đối với các mẫu văn bản có bộ cục phức tạp hoặc biến dạng mạnh, Zheng và cộng sự đã đề xuất CDistNet [42]. Phương pháp này nhằm tăng cường mối liên kết chặt chẽ giữa hai miền đặc trưng, qua đó cải thiện khả năng căn chỉnh giữa đặc trưng và ký tự trong quá trình nhận dạng. CDistNet sử dụng một encoder gồm ba nhánh song song để trích xuất các nguồn thông tin bổ sung cho nhau, bao gồm đặc trưng thị giác từ ảnh đầu vào, đặc trưng ngữ nghĩa từ chuỗi ký tự, và embedding vị trí mô tả quan hệ không gian giữa các ký tự. Các đặc trưng này sau đó được đưa vào mô-đun Multi-Domain Character Distance Perception (MDCDP) tạo ra một embedding vị trí (positional embedding) để đồng thời truy vấn cả đặc trưng thị giác và ngữ nghĩa thông qua cơ chế chú ý chéo (cross-attention). Thông qua cơ chế này, CDistNet có khả năng trực tiếp mô hình hóa khoảng cách ký tự đa miền, bao gồm khoảng cách không gian, mối quan hệ ngữ nghĩa giữa các ký tự, cũng như sự liên kết giữa hai loại thông tin này. Cấu trúc encoder ba nhánh và mô-đun MDCDP của CDistNet được minh họa trong Hình 3.14.

Bằng cách xếp chồng nhiều mô-đun MDCDP, mô hình dần dần học được sự căn chỉnh chính xác hơn giữa vùng ảnh và ký tự, ngay cả trong các trường hợp nhận dạng khó. Nhờ đó, CDistNet hiệu quả giai đoạn nhận dạng văn bản trên biển hiệu với kiểu chữ biến dạng, xoay nghiêng hoặc bộ cục không chuẩn.

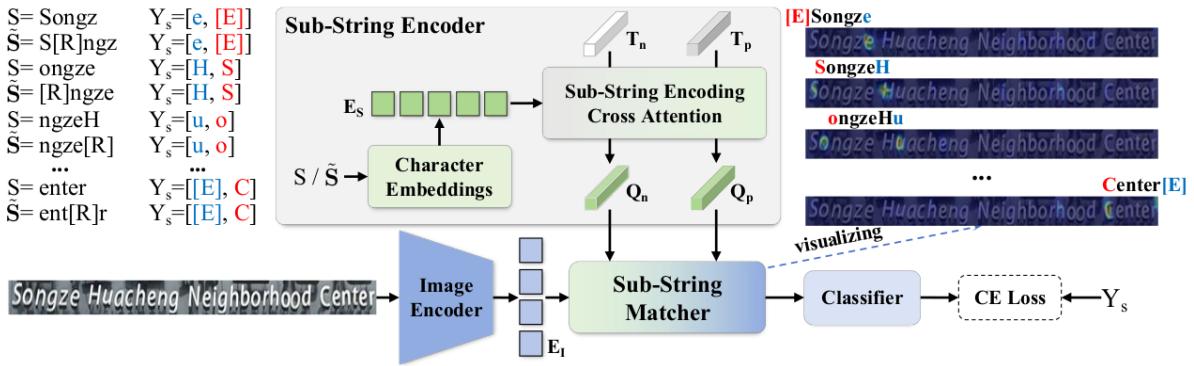


Hình 3.14: Kiến trúc tổng quan của CDistNet [42]

SMTR SMTR [8], được đề xuất bởi Du và cộng sự, áp dụng hướng tiếp cận dựa trên so khớp chuỗi con (sub-string matching) nhằm khắc phục hạn chế của các phương pháp truyền thống trong việc nhận dạng các chuỗi văn bản dài. Thay vì dự đoán toàn bộ chuỗi cùng lúc, SMTR thực hiện nhận dạng thông qua một quy trình lặp. Cụ thể, mô hình sử dụng hai mô-đun dựa trên cơ chế chú ý chéo (cross-attention), trong đó mô-đun thứ nhất mã hóa một chuỗi con gồm nhiều ký tự thành các truy vấn ngữ cảnh trước và sau, trong khi mô-đun thứ hai khai thác các truy vấn này để chú ý vào đặc trưng hình ảnh, đồng thời nhận dạng ký tự kế tiếp và ký tự liền trước của chuỗi con. Quá trình này được lặp lại nhiều lần, cho phép SMTR nhận dạng văn bản có độ dài tùy ý. Dựa trên cơ chế nhận dạng chuỗi con, SMTR có thể được huấn luyện trên các tập dữ liệu văn bản ngắn nhưng vẫn tổng quát tốt cho văn bản dài. Sơ đồ mô-đun và quy trình lặp của SMTR được trình bày trong Hình 3.15.

Ngoài ra, SMTR tích hợp chiến lược tăng cường suy luận (inference augmentation strategy) nhằm giảm thiểu sự nhầm lẫn giữa các chuỗi con tương tự. Nhờ đó, mô hình cải thiện đáng kể hiệu quả nhận dạng các chuỗi văn bản dài và phức tạp, đặc biệt phù hợp với các biến hiệu chứa nhiều từ hoặc các dòng chữ kéo dài.

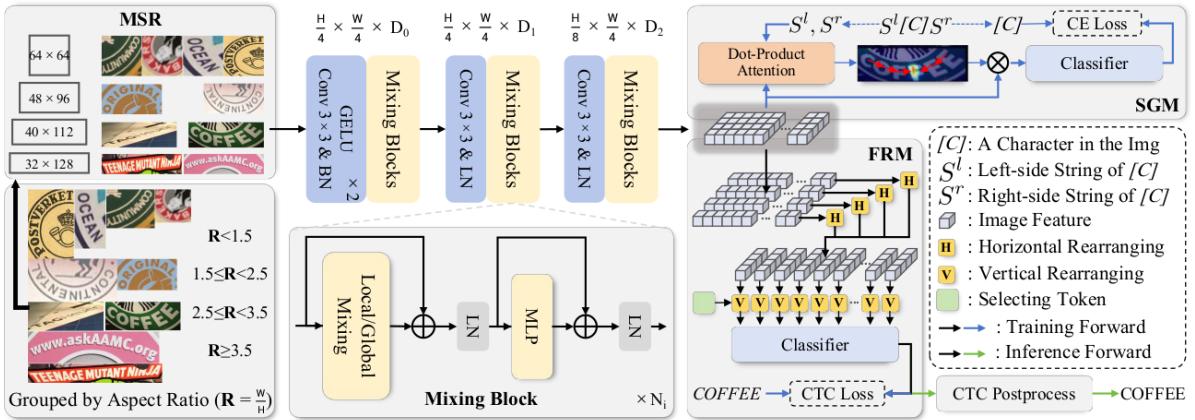
SVTRv2 Du và cộng sự đã giới thiệu SVTRv2 [9], là phiên bản mở rộng của SVTR, được đề xuất nhằm khắc phục những hạn chế về độ chính xác của các mô hình dựa trên



Hình 3.15: Kiến trúc tổng quan của SMTR [8]

Connectionist Temporal Classification (CTC) khi xử lý văn bản có hình dạng bất thường hoặc thiếu ngữ cảnh ngôn ngữ (linguistic missing), mặc dù các mô hình này vốn có ưu điểm về kiến trúc đơn giản và tốc độ suy luận nhanh so với các mô hình encoder-decoder. SVTRv2 áp dụng chiến lược đa kích thước (multi-size resizing) để điều chỉnh kích thước ảnh đầu vào phù hợp, tránh biến dạng nghiêm trọng, đồng thời giới thiệu mô-đun sắp xếp lại đặc trưng (feature rearrangement) để đảm bảo đặc trưng thị giác phù hợp với yêu cầu căn chỉnh của CTC. Bên cạnh đó, SVTRv2 tích hợp mô-đun định hướng ngữ nghĩa (semantic guidance module) nhằm đưa thông tin ngôn ngữ vào quá trình học đặc trưng thị giác, giúp mô hình tận dụng ngữ cảnh chuỗi để cải thiện độ chính xác. Đáng chú ý, mô-đun này chỉ được sử dụng trong giai đoạn huấn luyện và có thể loại bỏ hoàn toàn khi suy luận, do đó không làm tăng chi phí tính toán khi triển khai thực tế. Hình 3.16 minh họa kiến trúc tổng quan của SVTRv2.

Dựa trên sự kết hợp giữa hiệu quả suy luận của CTC và khả năng mô hình hóa các văn bản có hình dạng bất thường, cũng như khai thác ngữ cảnh ngôn ngữ, SVTRv2 đạt được sự cân bằng tốt giữa tốc độ và độ chính xác trong các trường hợp nhận dạng văn bản đa dạng như văn bản dài và văn bản có hình dạng phức tạp. Do đó, trong giai đoạn nhận dạng văn bản trên biển hiệu, SVTRv2 cho thấy tính phù hợp cao nhờ khả năng cân bằng giữa tốc độ suy luận và độ chính xác.

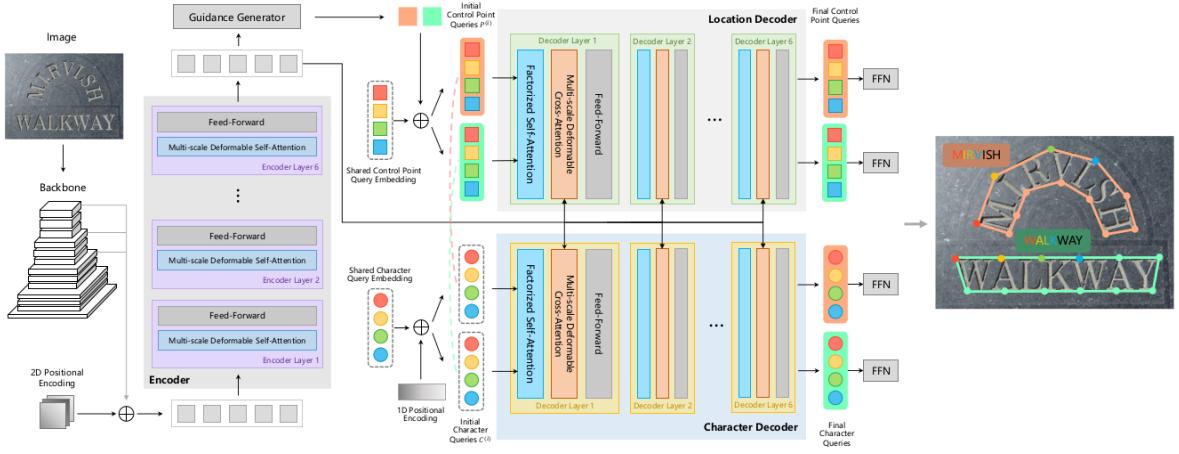


Hình 3.16: Kiến trúc tổng quan của SVTRv2 [9]

3.4 Phát hiện và nhận dạng văn bản trên biển hiệu theo hướng tiếp cận một giai đoạn (One-Stage)

Mặc dù hướng tiếp cận hai giai đoạn (two-stage), trong đó phát hiện và nhận dạng văn bản được thực hiện riêng biệt, đã cho thấy hiệu quả và tính linh hoạt cao trong bài toán nhận dạng văn bản trên biển hiệu, các phương pháp một giai đoạn (one-stage) ngày càng thu hút sự quan tâm nhờ khả năng suy luận trực tiếp từ ảnh đầu vào đến chuỗi ký tự đầu ra trong một mô hình thống nhất. Do đó, bên cạnh việc xây dựng và đánh giá pipeline hai giai đoạn, khóa luận tiến hành thực nghiệm so sánh giữa hai chiến lược tiếp cận: (i) hướng tiếp cận hai giai đoạn (two-stage), với hai mô-đun riêng biệt cho phát hiện và nhận dạng; và (ii) hướng tiếp cận một giai đoạn (one-stage), sử dụng các mô hình tiên tiến như TESTR, DeepSolo, UNITS, và DNTextSpotter.

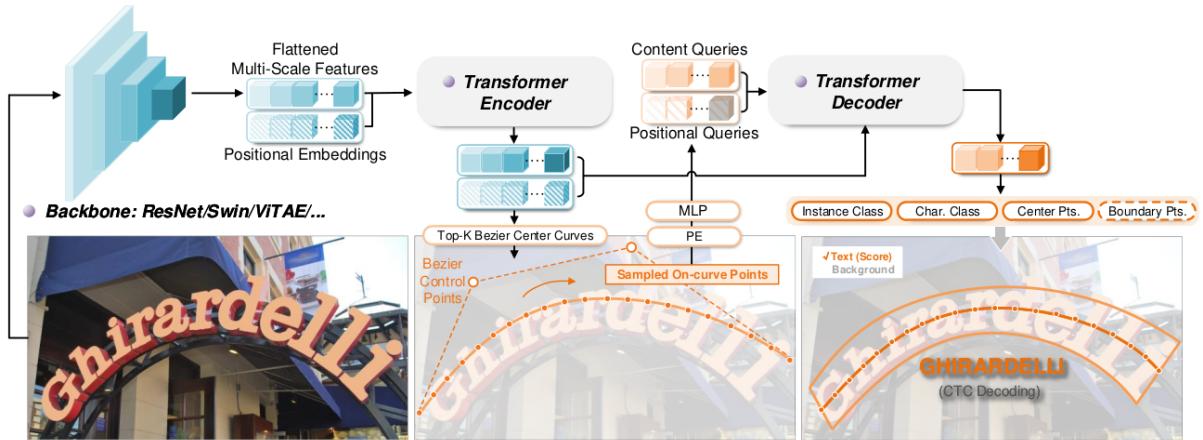
Mục tiêu của so sánh nhằm phân tích ưu nhược điểm của từng chiến lược trong bối cảnh nhận dạng văn bản trên biển hiệu, đặc biệt xét trên các khía cạnh như độ chính xác, tốc độ suy luận, mức độ phức tạp của hệ thống, đồng thời định hướng lựa chọn mô hình phù hợp để tinh chỉnh (fine-tuning) hiệu quả và tiết kiệm thời gian huấn luyện. Qua đó, khóa luận cung cấp cái nhìn tổng quan về khả năng ứng dụng thực tế của các phương pháp phát hiện và nhận dạng văn bản thông nhất (Text Spotting) hiện đại.



Hình 3.17: Sơ đồ kiến trúc của TESTR [40]

TESTR TESTR [40] được đề xuất bởi Zhang và cộng sự, nổi bật với việc áp dụng kiến trúc Transformer cho việc phát hiện và nhận dạng văn bản đầu-cuối (end-to-end) trong ảnh ngoại cảnh. Mô hình xây dựng dựa trên một bộ mã hóa (encoder) chung và hai bộ giải mã (decoder) song song, lần lượt đảm nhiệm việc hồi quy các điểm điều khiển của hộp chữ (text-box control point regression) và nhận dạng ký tự. Thiết kế này giúp TESTR loại bỏ hoàn toàn các thao tác trích xuất vùng quan tâm (RoI) và các quy trình hậu xử lý phức tạp dựa trên heuristic. Trên cơ sở đó, TESTR đặc biệt hiệu quả khi xử lý các văn bản uốn cong và có hình dạng bất kỳ nhờ biểu diễn linh hoạt bằng đường cong Bezier hoặc đa giác, thay vì chỉ sử dụng hộp giới hạn hình chữ nhật truyền thống. Bên cạnh đó, quy trình phát hiện đa giác có định hướng từ hộp giới hạn (box-to-polygon detection) được đề xuất nhằm nâng cao độ chính xác định vị. Kiến trúc tổng quan của TESTR được minh họa trong Hình 3.17.

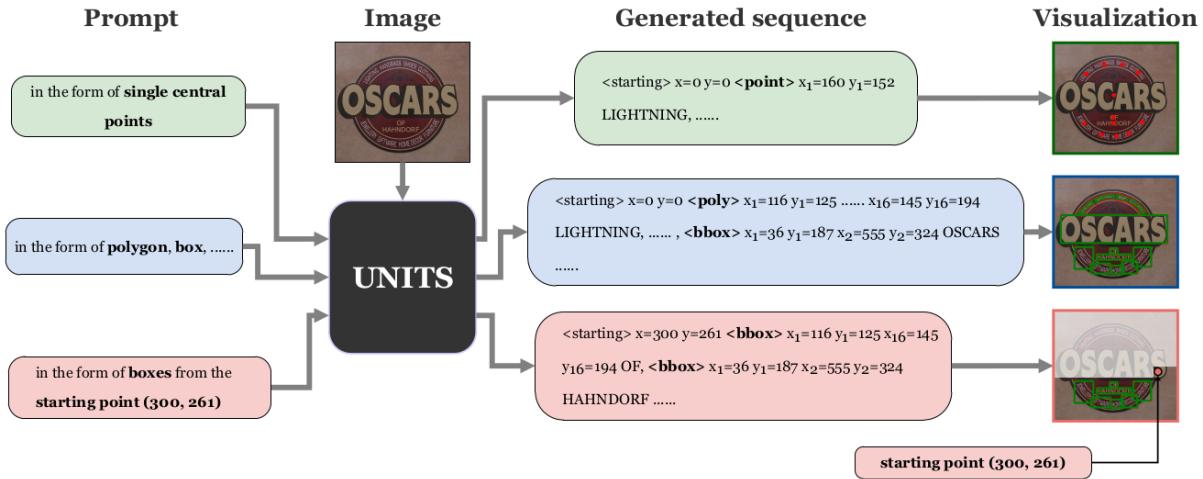
DeepSolo DeepSolo [36], được giới thiệu bởi Ye và cộng sự, nhằm giải quyết bài toán phát hiện và nhận dạng văn bản thông nhất (end-to-end) trong ảnh ngoại cảnh, nổi bật với khả năng xử lý đồng thời cả hai nhiệm vụ trong một mô hình duy nhất. Dựa trên nền tảng kiến trúc DETR, DeepSolo sử dụng một bộ giải mã (decoder) duy nhất với cơ chế Explicit Points Solo, cho phép mô hình học đồng thời để phát hiện và nhận dạng văn bản. Mỗi thực thể văn bản được biểu diễn dưới dạng chuỗi các điểm sắp xếp thứ tự, và được mô hình hóa thông qua các truy vấn điểm có thể học được (learnable explicit



Hình 3.18: Kiến trúc tổng quan của DeepSolo [36]

point queries). Sau khi đi qua decoder, các truy vấn này mã hóa thông tin ngữ nghĩa và vị trí của văn bản, từ đó có thể giải mã để xác định đường trung tâm (center line), biên giới (boundary), kiểu chữ (script) và độ tin cậy (confidence) thông qua các đầu ra dự đoán song song đơn giản. Nhờ những đặc điểm này, DeepSolo đạt hiệu quả cao cả về độ chính xác lẫn tốc độ huấn luyện trên các bộ dữ liệu chuẩn, đồng thời cung cấp giải pháp linh hoạt, phù hợp cho bài toán nhận dạng văn bản trên biển hiệu. Sơ đồ khối (block diagram) của DeepSolo được thể hiện trong Hình 3.18.

UNITS Nhằm khắc phục một số hạn chế về định dạng phát hiện và số lượng văn bản trong các mô hình tự hồi quy (auto-regressive) trước đó, Kil và cộng sự đã đề xuất UNITS (UNIfied Text Spotter) [15], nổi bật với khả năng thống nhất nhiều định dạng phát hiện, bao gồm tứ giác (quadrilateral) và đa giác (polygon), giúp mô hình xử lý văn bản có hình dạng bất kỳ. UNITS hoạt động theo cơ chế tạo chuỗi (sequence generation), trong đó thông tin của mỗi thể hiện văn bản (text instance) trong chuỗi đầu ra bao gồm token định dạng phát hiện, các token tọa độ cho việc định vị và chuỗi ký tự nhận dạng. Đặc biệt, kỹ thuật starting-point prompting được tích hợp, cho phép mô hình bắt đầu trích xuất văn bản từ một vị trí bất kỳ, từ đó có thể phát hiện nhiều thực thể văn bản vượt quá số lượng mà mô hình đã được huấn luyện. Pipeline hoạt động theo cơ chế tạo chuỗi (sequence generation) của UNITS được trình bày trong Hình 3.19. Nhờ khả năng mở rộng linh hoạt này, UNITS được lựa chọn là một giải pháp phù hợp cho giai đoạn phát

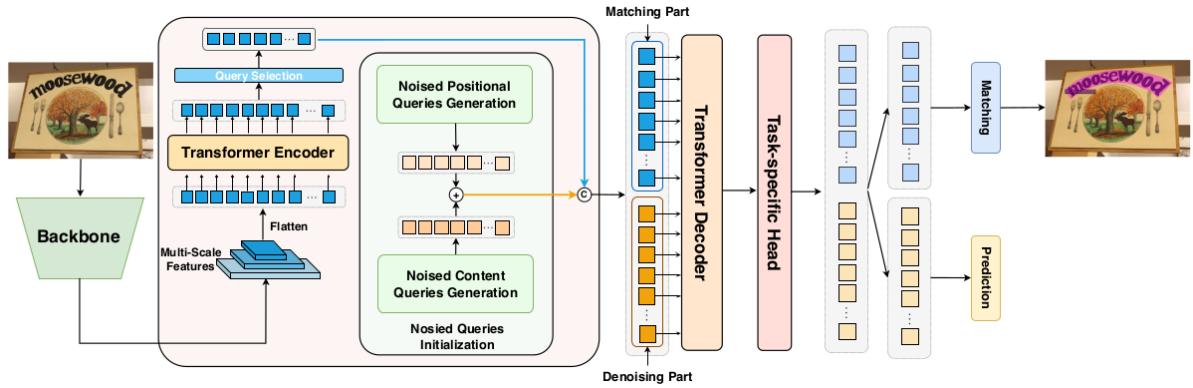


Hình 3.19: Pipeline tổng quan của UNITS [15]

hiện và nhận dạng văn bản trên biển hiệu, đặc biệt trong các trường hợp văn bản có hình dạng đa dạng và mật độ cao.

DNTTextSpotter DNTTextSpotter [27], được giới thiệu bởi Qiao và cộng sự, nhằm cải thiện độ ổn định và hiệu quả huấn luyện trong các phương pháp phát hiện và nhận dạng văn bản đầu-cuối (end-to-end text spotting) dựa trên kiến trúc Transformer. Các đặc trưng đa tỉ lệ (multi-scale features) được trích xuất từ backbone và bộ mã hóa (encoder), sau đó được đưa vào một bộ giải mã (decoder) với thiết kế hai nhánh đặc biệt: (i) gồm phần ghép cặp (matching part), sử dụng các truy vấn khởi tạo ngẫu nhiên và tính toán hàm mất mát thông qua thuật toán ghép cặp đồ thị hai phía (bipartite graph matching), và (ii) phần khử nhiễu (denoising part) căn chỉnh giữa vị trí và nội dung bằng các truy vấn vị trí nhiễu (noised positional queries) và truy vấn nội dung nhiễu (noised content queries). Trong đó, các truy vấn vị trí được tạo ra từ bốn điểm điều khiển Bezier của đường trung tâm, còn các truy vấn nội dung được khởi tạo thông qua phương pháp trượt ký tự có mặt nạ (masked character sliding), đồng thời một hàm mất mát bổ sung cho việc phân loại ký tự nền được tích hợp nhằm tăng cường khả năng nhận biết ngữ cảnh. Hình 3.20 minh họa kiến trúc DNTTextSpotter.

Dựa trên cơ chế khử nhiễu và khả năng căn chỉnh vị trí-nội dung hiệu quả, DNTText Spotter chọn cho bài toán phát hiện và nhận dạng văn bản trên biển hiệu với hình dạng



Hình 3.20: Kiến trúc tổng quan của DNTextSpotter [27]

đa dạng và phức tạp.

Chapter 4

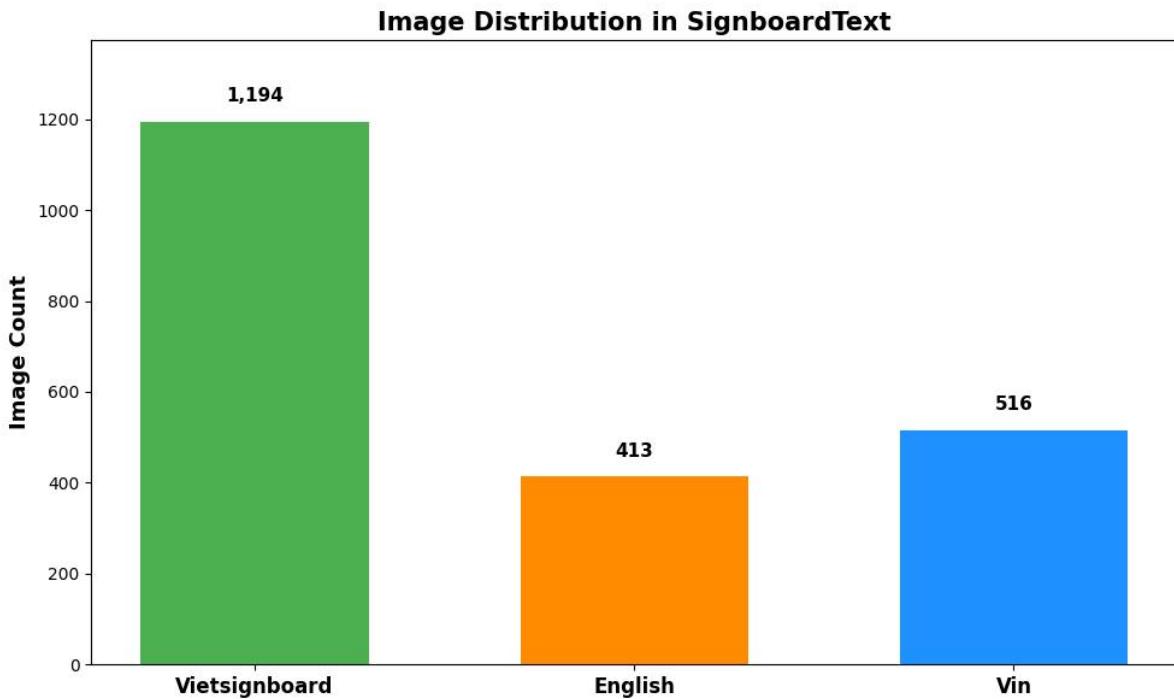
THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1 Tập dữ liệu

Để xây dựng và đánh giá hiệu quả cho bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong video đường phố Việt Nam, khóa luận sử dụng bộ dữ liệu **SignboardText** được giới thiệu bởi **Do và cộng sự [7]**. Bộ dữ liệu này cung cấp một tập dữ liệu chuyên biệt cho văn bản trên biển hiệu, với các thách thức đặc thù như văn bản đa ngôn ngữ (tiếng Anh và tiếng Việt), kiểu chữ nghệ thuật, đặc biệt sự xuất hiện của các dấu thanh (tone marks) trong tiếng Việt, một yếu tố có thể ảnh hưởng đáng kể đến độ chính xác của các phương pháp hiện tại vốn thường được huấn luyện trên các ngôn ngữ không dấu.

Dựa trên nghiên cứu nền tảng [7], khóa luận tiến hành phân tích và thống kê cấu trúc của bộ dữ liệu SignboardText, bao gồm ba tập con chính: Vietsignboard, English và Vin. Trong đó, tập Vietsignboard đóng vai trò là tập dữ liệu chính, với 1,327 ảnh được Do và cộng sự thu thập thủ công trên đường phố Việt Nam, trong khi các tập English và Vin được bổ sung với 413 và 516 ảnh, chọn lọc từ các bộ dữ liệu benchmark Total-Text, ICDAR2015 và VinText, nhằm tăng cường tính đa dạng cho dataset. Sự phân bố số lượng ảnh của ba tập con được minh họa trong [Hình 4.1](#).

Về định dạng gán nhãn (annotation), Vietsignboard cung cấp nhãn ở cả hai cấp độ: cấp độ từ (word-level) với 48.638 thể hiện (instances) và cấp độ dòng (line-level) với 10.950 thể hiện (instances). Trong khi đó, các tập English và Vin chỉ được gán nhãn ở cấp độ từ (word-level) với lần lượt 3,646 và 16,615 thể hiện (instances). Phân bố chi tiết

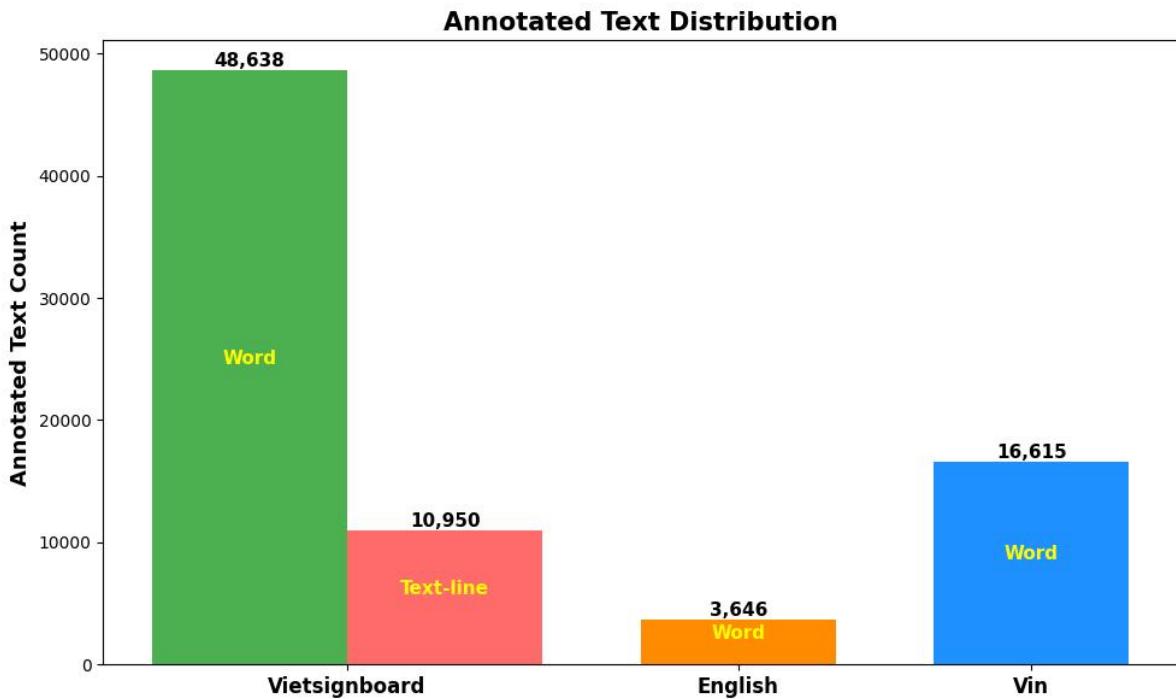


Hình 4.1: Phân bố số lượng hình ảnh trong ba tập con của SignboardText [7]

của các loại annotation này được trình bày trong Hình 4.2. Như vậy, trong bộ dữ liệu SignboardText, phần lớn văn bản được gán nhãn ở cấp độ từ (word). Đồng thời, việc kết hợp gán nhãn ở cả cấp độ từ (word) và dòng (line) trong tập Vietsignboard tạo điều kiện thuận lợi cho việc đánh giá linh hoạt hai nhiệm vụ phát hiện và nhận dạng văn bản.

Theo phân tích của Do và cộng sự [7], văn bản trong bộ dữ liệu SignboardText có hình dạng rất đa dạng, bao gồm các trường hợp văn bản nằm ngang (horizontal), văn bản có biên dạng tứ giác bất kỳ (arbitrary quadrilateral), cũng như văn bản đa hướng (multi-oriented). Đặc điểm này phản ánh sát với bối cảnh thực tế của biển hiệu ngoài trời, nơi các dòng chữ có thể được bố trí nghiêng, cong hoặc không song song với trục ảnh, qua đó đặt ra thách thức đáng kể cho các phương pháp phát hiện và nhận dạng văn bản tiên tiến hiện nay.

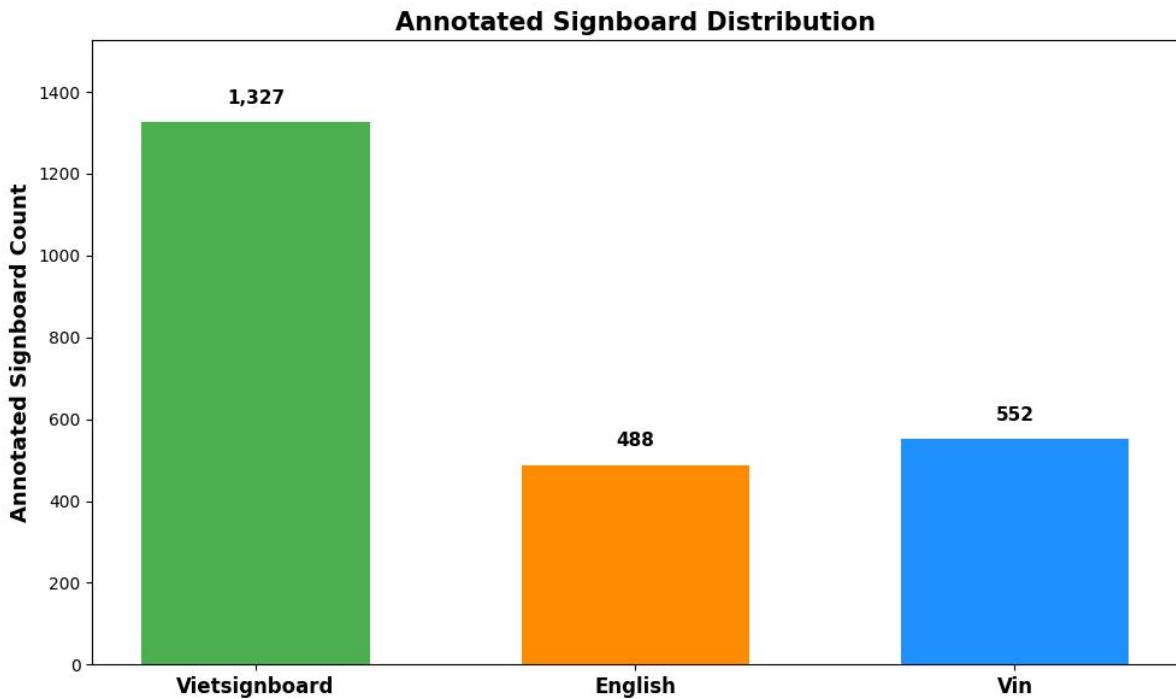
Mặc dù bộ dữ liệu SignboardText cung cấp hệ thống nhãn chi tiết cho văn bản ở nhiều cấp độ và hình dạng khác nhau, các nhãn (annotation) này vẫn chỉ tập trung vào các vùng văn bản, chưa bao quát toàn bộ đối tượng biển hiệu chứa văn bản. Trong khi đó, theo pipeline được đề xuất trong khóa luận, phát hiện biển hiệu đóng vai trò



Hình 4.2: Phân bố số lượng thể hiện văn bản (text instances) theo cấp độ nhãn (word-level và line-level) trong các tập con của SignboardText [7]

là bước tiền đề cho các giai đoạn phát hiện và nhận dạng văn bản phía sau. Do đó, nhằm hỗ trợ đánh giá giai đoạn phát hiện biển hiệu, khóa luận tiến hành mở rộng tập dữ liệu SignboardText bằng cách bổ sung lớp nhãn cho đối tượng biển hiệu. Cụ thể, toàn bộ 2.123 ảnh thuộc ba tập con Vietsignboard, English và Vin đã được gán nhãn thủ công, với sự hỗ trợ của công cụ PPOCRLLabel [10], để xác định các vùng chứa của biển hiệu trong ảnh. Trong đó, số lượng đối tượng biển hiệu được gán nhãn trong các tập Vietsignboard, English và Vin lần lượt là 1.327, 488 và 552. Phân bố số lượng các đối tượng này theo từng tập con được minh họa trong Hình 4.3.

Nhằm đánh giá mối quan hệ giữa các vùng văn bản và vùng biển hiệu trong tập dữ liệu SignboardText, khóa luận tiến hành thống kê tỷ lệ phần trăm văn bản nằm trong vùng biển hiệu so với toàn bộ văn bản xuất hiện trong ảnh, trên từng tập con của bộ dữ liệu. Dựa trên kết quả thống kê được trình bày trong Bảng 4.1, có thể nhận thấy rằng phần lớn văn bản trong tập dữ liệu SignboardText nằm bên trong các vùng biển hiệu đã được gán nhãn. Cụ thể, trung bình trên toàn bộ tập dữ liệu, khoảng 64,40% văn bản ở



Hình 4.3: Phân bố số lượng đối tượng biển hiệu theo từng tập con của SignboardText

Bảng 4.1: Thống kê tỷ lệ văn bản nằm trong vùng biển hiệu so với toàn bộ văn bản trong ảnh trên các tập con của SignboardText

	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

cấp độ từ (word-level) và 57,56% văn bản ở cấp độ dòng (line-level) thuộc về các đối tượng biển hiệu. Kết quả này cho thấy tập dữ liệu SignboardText phù hợp để đánh giá pipeline phát hiện và nhận dạng văn bản trên biển hiệu được sử dụng trong khóa luận.

Trong khuôn khổ khóa luận này, việc mở rộng tập dữ liệu chủ yếu tập trung vào bổ sung nhãn đối tượng biển hiệu (signboard annotation) cho tập dữ liệu ảnh tĩnh SignboardText hiện có, nhằm phục vụ trực tiếp cho quá trình huấn luyện và đánh giá mô hình. Bên cạnh đó, một tập dữ liệu video được thu thập trong môi trường đường phố Việt Nam và chỉ được sử dụng với mục đích minh họa cũng như kiểm tra khả năng tổng quát hóa của pipeline đề xuất trong bối cảnh thực tế.

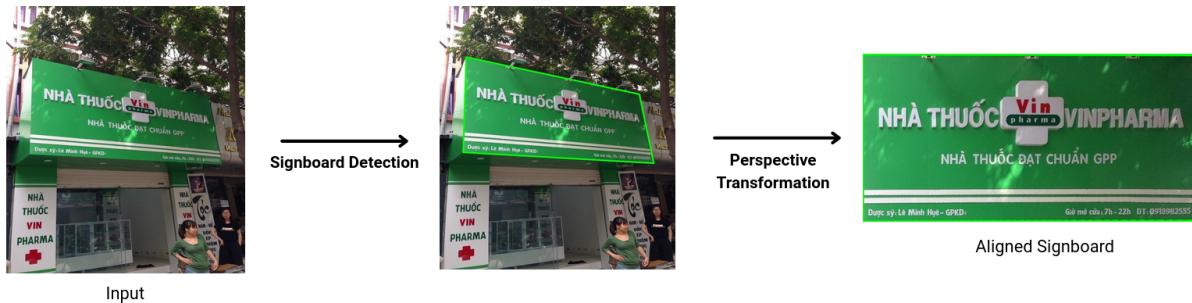
4.2 Thiết lập thực nghiệm

Trong khóa luận này, quá trình thực nghiệm được tổ chức theo hướng phân tách theo từng giai đoạn trong pipeline phát hiện và nhận dạng văn bản trên biển hiệu. Cách tổ chức này nhằm cho phép đánh giá độc lập hiệu quả của từng thành phần, đồng thời giảm chi phí huấn luyện và yêu cầu tài nguyên tính toán khi phải làm việc với nhiều mô hình khác nhau. Cụ thể, các thí nghiệm được thiết kế để lần lượt khảo sát từng giai đoạn chính, từ phát hiện biển hiệu, phát hiện và nhận dạng văn bản trên biển hiệu, cho đến việc xây dựng pipeline đầu-cuối (end-to-end). Kết quả thực nghiệm ở mỗi giai đoạn được sử dụng làm cơ sở để lựa chọn mô hình phù hợp, từ đó kết hợp và hình thành pipeline hoàn chỉnh cho bài toán đặt ra. Việc lựa chọn mô hình được thực hiện dựa trên sự cân bằng giữa độ chính xác, tốc độ xử lý và độ phức tạp mô hình, nhằm đảm bảo tính khả thi khi áp dụng pipeline trong bối cảnh xử lý video đường phố thực tế. Trên cơ sở này, các thiết lập thực nghiệm chi tiết cho từng giai đoạn sẽ được trình bày trong các mục tiếp theo.

4.2.1 Phát hiện biển hiệu

Trong pipeline phát hiện và nhận dạng văn bản trên biển hiệu, phát hiện biển hiệu đóng vai trò là bước khởi đầu, có ảnh hưởng trực tiếp đến hiệu quả của các giai đoạn xử lý phía sau. Do đặc thù về bối cảnh thu thập dữ liệu và sự khác biệt về miền dữ liệu so với các tập dữ liệu phát hiện đối tượng phổ biến, khóa luận tiến hành tinh chỉnh (fine-tuning) toàn bộ các mô hình khảo sát ở giai đoạn này nhằm đảm bảo khả năng thích ứng với môi trường đường phố Việt Nam.

Các mô hình phát hiện biển hiệu được phân nhóm dựa trên dạng biểu diễn đầu ra, bao gồm: (i) các phương pháp dự đoán vùng bao chữ nhật (rectangle bounding box), (ii) các phương pháp dự đoán vùng bao định hướng (oriented bounding box - OBB), và (iii) các phương pháp dựa trên phân đoạn đối tượng để xác định vùng biển hiệu dưới dạng đa giác (polygon). Việc phân nhóm này cho phép đánh giá một cách có hệ thống các đặc điểm và lợi thế của từng hướng tiếp cận trong bối cảnh bài toán đặt ra. Song song với



Hình 4.4: Hình ảnh minh họa quá trình căn chỉnh biển hiệu (signboard alignment)

đó, trong mỗi nhóm mô hình, các phương pháp được so sánh nhằm lựa chọn mô hình tốt nhất cho từng dạng đầu ra. Quá trình này tập trung đánh giá khả năng phát hiện trong điều kiện dữ liệu thực tế, đồng thời xem xét mức độ hiệu quả khi triển khai, làm cơ sở cho việc tích hợp các mô hình này vào pipeline và phục vụ các bước so sánh tổng thể ở các giai đoạn tiếp theo.

Bên cạnh đó, khóa luận tiến hành thực nghiệm bổ sung bước căn chỉnh biển hiệu (signboard alignment) đối với các mô hình có đầu ra là OBB hoặc polygon. Trong thiết lập này, vùng biển hiệu sau khi được phát hiện sẽ được biến đổi phối cảnh (perspective transformation) để đưa về dạng chuẩn, qua đó cho phép so sánh hiệu quả giữa trường hợp không căn chỉnh và có căn chỉnh biển hiệu. Thiết lập này nhằm đánh giá mức độ ảnh hưởng của bước căn chỉnh đối với chất lượng dữ liệu đầu vào cho các giai đoạn phát hiện và nhận dạng văn bản phía sau. Hình 4.4 minh họa ví dụ quá trình căn chỉnh biển hiệu, trong đó vùng biển hiệu được phát hiện với đầu ra dạng polygon được biến đổi phối cảnh để đưa về dạng hình chữ nhật chuẩn, phục vụ cho các bước phát hiện và nhận dạng văn bản tiếp theo.

4.2.2 Phát hiện và nhận dạng văn bản trên biển hiệu

Giai đoạn phát hiện và nhận dạng văn bản trên biển hiệu thực hiện hai nhiệm vụ chính: xác định vị trí các vùng văn bản và nhận dạng nội dung bên trong. Giai đoạn này được khảo sát theo hai hướng tiếp cận: hai giai đoạn (Two-Stage) và một giai đoạn (One-Stage). Các mô hình pretrained được sử dụng làm cơ sở đánh giá, từ đó lựa chọn hướng tiếp cận phù hợp nhằm tiết kiệm thời gian tinh chỉnh (fine-tune) mô hình. Bên

cạnh đó, ở giai đoạn này, khóa luận thực nghiệm với tất cả các văn bản xuất hiện trong ảnh, thay vì chỉ giới hạn ở văn bản trên biển hiệu trong tập SignboardText. Điều này cũng giúp mở rộng dữ liệu đánh giá, từ đó cải thiện độ tin cậy của kết quả thực nghiệm.

Hướng tiếp cận hai giai đoạn (Two-Stage)

Phát hiện văn bản

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

.....

References

- [1] Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In *International conference on document analysis and recognition*, pages 319–334. Springer, 2021. [vii](#), [7](#), [17](#), [31](#), [32](#)
- [2] Youngmin Baek, Bado Lee, Dongyo Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9365–9374, 2019. [15](#)
- [3] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pages 178–196. Springer, 2022. [vii](#), [7](#), [17](#), [32](#), [33](#)
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. [vi](#), [7](#), [13](#), [22](#), [23](#)
- [5] Zhe Chen, Jiahao Wang, Wenhui Wang, Guo Chen, Enze Xie, Ping Luo, and Tong Lu. Fast: Faster arbitrarily-shaped text detector with minimalist kernel representation. *arXiv preprint arXiv:2111.02394*, 2021. [vii](#), [7](#), [15](#), [29](#), [30](#)
- [6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. [vii](#), [7](#), [25](#), [27](#)

- [7] T. Do, T. Tran, T. Nguyen, D.-D. Le, and T. D. Ngo. Signboardtext: Text detection and recognition in in-the-wild signboard images. *IEEE Access*, 12:62942–62957, 2024. [vi](#), [vii](#), [4](#), [5](#), [7](#), [8](#), [16](#), [18](#), [41](#), [42](#), [43](#)
- [8] Yongkun Du, Zhineng Chen, Caiyan Jia, Xieping Gao, and Yu-Gang Jiang. Out of length text recognition with sub-string matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2798–2806, 2025. [vii](#), [7](#), [17](#), [34](#), [35](#)
- [9] Yongkun Du, Zhineng Chen, Hongtao Xie, Caiyan Jia, and Yu-Gang Jiang. Svtrv2: Ctc beats encoder-decoder models in scene text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20147–20156, 2025. [vii](#), [7](#), [34](#), [36](#)
- [10] Evezerest. PPOCRLLabel: Semi-automatic image annotation tool for ocr. GitHub repository, 2023. Available: <https://github.com/Evezerest/PPOCRLLabel>, accessed Jan. 10, 2025. [43](#)
- [11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. [12](#)
- [12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. [12](#)
- [13] Xu Han, Junyu Gao, Chuang Yang, Yuan Yuan, and Qi Wang. Spotlight text detector: Spotlight on candidate regions like a camera. *IEEE Transactions on Multimedia*, 2024. [14](#)

- [14] JianJun Kang, Mayire Ibrayim, and Askar Hamdulla. Overview of scene text detection and recognition. In *Proceedings of the 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 661–666, 2022. [14](#), [16](#), [18](#), [26](#), [31](#)
- [15] Taeho Kil, Seonghyeon Kim, Sukmin Seo, Yoonsik Kim, and Daehee Kim. Towards unified scene text spotting based on sequence generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15223–15232, 2023. [vii](#), [7](#), [20](#), [38](#), [39](#)
- [16] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. [14](#), [19](#)
- [17] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):919–931, 2022. [vii](#), [7](#), [15](#), [27](#), [28](#), [29](#)
- [18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. [13](#)
- [19] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. [7](#), [15](#)
- [20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. [13](#)
- [21] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network.

In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9809–9818, 2020. [14](#), [19](#)

- [22] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36, 2018. [15](#)
- [23] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024. [7](#)
- [24] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 67–83, 2018. [16](#), [19](#)
- [25] Fatemeh Naiemi, Vahid Ghods, and Hassan Khalesi. Scene text detection and recognition: A survey. *Multimedia Tools and Applications*, 81(1):20255–20290, 2022. [14](#), [16](#), [18](#), [26](#), [31](#)
- [26] Umapada Pal, Arnab Halder, Palaiahnakote Shivakumara, and Michael Blumenstein. A comprehensive review on text detection and recognition in scene images. *Artificial Intelligence and Applications*, 2(4):229–249, 2024. [14](#), [16](#), [18](#), [26](#), [31](#)
- [27] Qian Qiao, Yu Xie, Jun Gao, Tianxiang Wu, Shaoyao Huang, Jiaqing Fan, Ziqiang Cao, Zili Wang, and Yue Zhang. Dntextspotter: Arbitrary-shaped scene text spotting via improved denoising training. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10134–10143, 2024. [vii](#), [7](#), [20](#), [39](#), [40](#)
- [28] D. L. Quang, K. V. Sy, H. L. Viet, S. P. Bao, and H. B. Quang. Signboards detection from street-view image using convolutional neural network: A case study

in vietnam. In *Proceedings of the RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 394–397, Ho Chi Minh City, Vietnam, 2022. [2](#)

- [29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. [vi](#), [5](#), [7](#), [13](#), [22](#), [23](#)
- [30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. [12](#)
- [31] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016. [17](#)
- [32] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126:106669, 2023. [25](#)
- [33] Pengfei Wang, Chengquan Zhang, Fei Qi, Shanshan Liu, Xiaoqiang Zhang, Pengyuan Lyu, Junyu Han, Jingtuo Liu, Errui Ding, and Guangming Shi. Pgnet: Real-time arbitrarily-shaped text spotting with point gathering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2782–2790, 2021. [19](#)
- [34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. [vii](#), [7](#), [25](#), [26](#)

- [35] Jian Ye, Zhe Chen, Juhua Liu, and Bo Du. Textfusenet: Scene text detection with richer fused features. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 516–522, 2020. [16](#)
- [36] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. Deepsolo: Let transformer decoder with explicit points solo for text spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19348–19357, 2023. [vii, 7, 19, 37, 38](#)
- [37] Shi-Xue Zhang, Xiaobin Zhu, Lei Chen, Jie-Bo Hou, and Xu-Cheng Yin. Arbitrary shape text detection via segmentation with probability maps. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):2736–2750, 2022. [7, 15, 28](#)
- [38] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9699–9708, 2020. [15](#)
- [39] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chun Yang, and Xu-Cheng Yin. Kernel proposal network for arbitrary shape text detection. *IEEE transactions on neural networks and learning systems*, 34(11):8731–8742, 2022. [vii, 7, 15, 30, 31](#)
- [40] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9519–9528, 2022. [vii, 7, 20, 37](#)
- [41] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. [vi, 24](#)
- [42] Tianlun Zheng, Zhineng Chen, Shancheng Fang, Hongtao Xie, and Yu-Gang Jiang. Cdistnet: Perceiving multi-domain character distance for robust text recognition.

- [43] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. 14
- [44] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3123–3131, 2021. 14
- [45] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 12, 21