

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

| | |
|--|------------|
| LỜI CẢM ƠN | i |
| Tóm tắt khóa luận | ii |
| Contents | iii |
| List of Figures | iv |
| List of Tables | v |
| 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 Cơ sở lý thuyết | 10 |
| 2.1.1 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 |
| 2.1.2 Các phương pháp tiếp cận | 11 |
| 2.1.2.1 Phát hiện đối tượng | 11 |

| | | |
|-------------------|---|-----------|
| 2.1.2.2 | Cơ sở và hướng tiếp cận chung | 11 |
| 2.1.2.3 | Các nghiên cứu liên quan | 11 |
| 2.2 | Phát hiện và nhận dạng văn bản ngoại cảnh (Scene Text Detection and Recognition - STDR) | 13 |
| 2.2.1 | Phát hiện văn bản ngoại cảnh (Scene Text Detection - STD) | 13 |
| 2.2.1.1 | Cơ sở và hướng tiếp cận chung | 13 |
| 2.2.1.2 | Các nghiên cứu liên quan | 14 |
| 2.2.2 | Nhận dạng văn bản (Scene Text Recognition - STR) | 16 |
| 2.2.2.1 | Cơ sở và hướng tiếp cận chung | 16 |
| 2.2.2.2 | Các nghiên cứu liên quan | 16 |
| 2.2.3 | Nhận dạng văn bản ngoại cảnh đầu–cuối (End-to-End Scene Text Recognition) | 18 |
| 2.2.3.1 | Cơ sở và hướng tiếp cận chung | 18 |
| 2.2.3.2 | Các nghiên cứu liên quan | 18 |
| 3 | PHƯƠNG PHÁP | 21 |
| 3.1 | Hệ thống phát hiện và nhận dạng chữ trên biển hiệu | 21 |
| 4 | THỰC NGHIỆM VÀ ĐÁNH GIÁ | 22 |
| 4.1 | Dữ liệu | 22 |
| 4.2 | Tiền xử lý | 22 |
| 4.3 | Tập câu truy vấn đánh giá | 22 |
| 4.4 | Độ đo đánh giá | 22 |
| 4.5 | Kết quả thực nghiệm | 22 |
| 5 | KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN | 23 |
| 5.1 | Kết luận | 23 |
| 5.2 | Hướng phát triển | 23 |
| References | | 25 |

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 |

List of Tables

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 [26] 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 [27], 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 [22], SegFormer [31], Mask2Former [6].

- **Phát hiện văn bản (Text Detection):** PANet [18], DBNet++ [16], TextPMs [34], FAST [5], KPN [36].
- **Nhận dạng văn bản (Text Recognition):** ViTSTR [1], PARSeq [3], CDistNet [38], SMTR [8], SVTRv2 [9]
- **Text Spotting (End-to-End):** TESTR [37], DeepSolo [33], UNITS [14], DNTextSpotter [25]

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 Cơ sở lý thuyết

2.1.1 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 đổi mới 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.

2.1.2 Các phương pháp tiếp cận

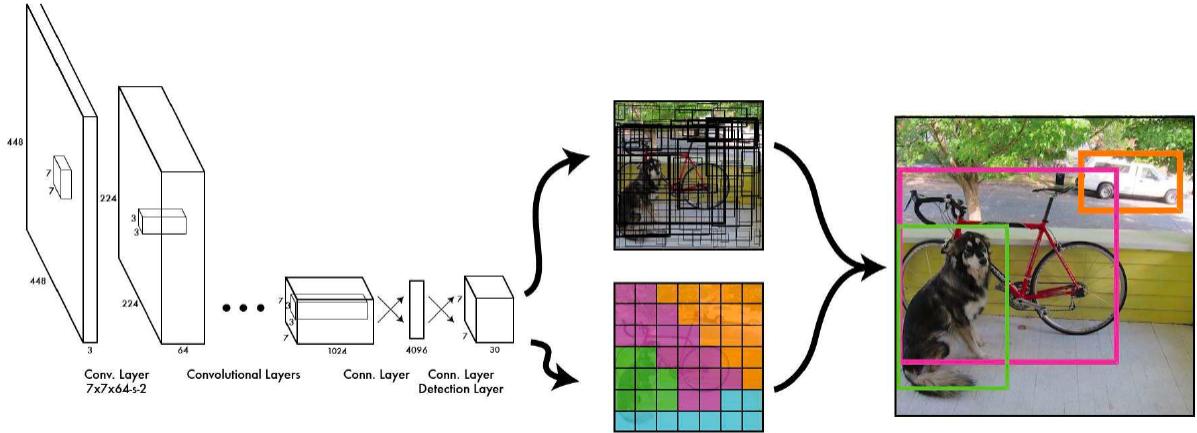
2.1.2.1 Phát hiện đối tượng

2.1.2.2 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.

2.1.2.3 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 đến sự ra đời của nhiều hướng tiếp cận khác nhau cho bài toán này. Dựa trên khảo sát



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.

tổng quan của [41], 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 [11]**, **Fast R-CNN [10]** và **Faster R-CNN [28]**. 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** [27], **SSD** [19] và **RetinaNet** [17]. 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 Phát hiện và nhận dạng văn bản ngoại cảnh (Scene Text Detection and Recognition - STDR)

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

2.2.1.1 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 đối mặt với nhiều thách thức do sự đa dạng về hình dạng, kích thước, hướng và bộ 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.

2.2.1.2 Các nghiên cứu liên quan

Dựa trên các nghiên cứu khảo sát và tổng quan gần đây [13, 24, 12] 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** [15], 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** [39] 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** [20], 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** [40] 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** [21], mô hình hóa văn bản cong bằng một chuỗi các hình tròn linh hoạt (các "vảy rắn") dọc theo trực 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** [35] 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** [18] 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++** [16], tích hợp differentiable binarization và Adaptive Scale Fusion nhằm giảm thiểu bước hậu xử lý và cải thiện độ chính xác. **TextPMs** [34] 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** [36] 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 Nhận dạng văn bản (Scene Text Recognition - STR)

2.2.2.1 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.

2.2.2.2 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 [13, 24, 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** [23] 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** [32],

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 [29]**, 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 [38]**, **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.



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

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

2.2.3.1 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 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.

2.2.3.2 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 [13, 24, 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** [15] 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** [23] 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** [20] 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 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** [30] dự đoán văn bản một cách trực tiếp thông qua việc học chuỗi các điểm trung tâm. Trong khi đó, **DeepSolo** [33], 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** [37], **UNITS** [14] và **DNTTextSpotter** [25] 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

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

Chapter 4

THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1 Dữ liệu

4.2 Tiền xử lý

4.3 Tập câu truy vấn đánh giá

4.4 Độ đo đánh giá

4.5 Kết quả thực nghiệm

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. [7](#), [17](#)
- [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. [7](#), [17](#)
- [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. [5](#), [13](#)
- [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. [7](#), [15](#)
- [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. [7](#)

- [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. [4](#), [5](#), [7](#), [8](#), [16](#), [19](#)
- [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. [7](#), [17](#)
- [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. [7](#)
- [10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. [12](#)
- [11] 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](#)
- [12] 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](#)
- [13] 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](#), [19](#)
- [14] 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. [7](#), [20](#)

- [15] 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](#)
- [16] 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. [7](#), [15](#)
- [17] 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](#)
- [18] 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](#)
- [19] 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](#)
- [20] 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](#)
- [21] 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](#)
- [22] 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](#)

- [23] 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](#)
- [24] 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](#), [19](#)
- [25] 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. [7](#), [20](#)
- [26] 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](#)
- [27] 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. [5](#), [13](#)
- [28] 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](#)
- [29] 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](#)

- [30] 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. [20](#)
- [31] 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. [7](#)
- [32] 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](#)
- [33] 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. [7, 20](#)
- [34] 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](#)
- [35] 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](#)
- [36] 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. [7, 15](#)

- [37] 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. [7](#), [20](#)
- [38] Tianlun Zheng, Zhineng Chen, Shancheng Fang, Hongtao Xie, and Yu-Gang Jiang. Cdinstnet: Perceiving multi-domain character distance for robust text recognition. *International Journal of Computer Vision*, 132(2):300–318, 2024. [7](#), [17](#)
- [39] 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](#)
- [40] 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](#)
- [41] 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](#)