




# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):  
(ví dụ: <https://www.youtube.com/watch?v=AWq7uw-36Ng>)
- Link slides (dạng .pdf đặt trên Github của nhóm):  
(ví dụ: <https://github.com/mynameuit/CS519.M11.KHCL/TenDeTai.pdf>)
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none"><li>• Họ và Tên: Thái Trần Khánh Nguyên</li><li>• MSSV: 1920188</li></ul> 	<ul style="list-style-type: none"><li>• Lớp: CS519.M11.KHCL</li><li>• Tự đánh giá (điểm tổng kết môn): 9.5/10</li><li>• Số buổi vắng: 0</li><li>• Số câu hỏi QT cá nhân: 10</li><li>• Số câu hỏi QT của cả nhóm: 4</li><li>• Link Github: <a href="https://github.com/ttknguyen/CS519.M11.KHCL">ttknguyen/CS519.M11.KHCL (github.com)</a></li><li>• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none"><li>○ Lên ý tưởng</li><li>○ Viết phần Tóm tắt, Nội dung và phương pháp</li><li>○ Thiết kế poster</li><li>○ Quay video YouTube</li></ul></li></ul>
<ul style="list-style-type: none"><li>• Họ và Tên: Nguyễn Khánh Như</li><li>• MSSV: 19520209</li></ul>	<ul style="list-style-type: none"><li>• Lớp: CS519.M11.KHCL</li><li>• Tự đánh giá (điểm tổng kết môn): 9.5/10</li><li>• Số buổi vắng: 0</li><li>• Số câu hỏi QT cá nhân: 10</li><li>• Số câu hỏi QT của cả nhóm: 4</li><li>• Link Github:</li></ul>

	<p><a href="https://github.com/ttknguyen/CS519.M11.KHCL">ttknguyen/CS519.M11.KHCL (github.com)</a></p> <ul style="list-style-type: none"> <li>• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: <ul style="list-style-type: none"> <li>○ Lên ý tưởng</li> <li>○ Viết phần Giới thiệu</li> <li>○ Làm Slide báo cáo</li> <li>○ Soạn kịch bản cho video Youtube</li> </ul> </li> </ul>
<ul style="list-style-type: none"> <li>• Họ và Tên: Lê Đoàn Thiện Nhân</li> <li>• MSSV: 19520197</li> </ul> 	<ul style="list-style-type: none"> <li>• Lớp: CS519.M11.KHCL</li> <li>• Tự đánh giá (điểm tổng kết môn): 9.5/10</li> <li>• Số buổi vắng: 0</li> <li>• Số câu hỏi QT cá nhân: 10</li> <li>• Số câu hỏi QT của cả nhóm: 4</li> <li>• Link Github: <p><a href="https://github.com/thiennhan2701/CS519.M11.KHCL">thiennhan2701/CS519.M11.KHCL (github.com)</a></p> </li> <li>• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: <ul style="list-style-type: none"> <li>○ Lên ý tưởng</li> <li>○ Viết phần Mục tiêu, kết quả mong đợi, tài liệu tham khảo.</li> <li>○ Thiết kế poster</li> <li>○ Làm hậu kỳ và xuất video Youtube</li> </ul> </li> </ul>

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

HƯỚNG TIẾP CẬN SỬ DỤNG THƯ VIỆN HỖ TRỢ CHO BÀI TOÁN NHẬN DẠNG CHỮ TRONG HÌNH ẢNH VÀ XÂY DỰNG BỘ DỮ LIỆU VỀ NHẬN DẠNG CHỮ TRONG HÌNH ẢNH CHO NGÔN NGỮ TIẾNG VIỆT

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

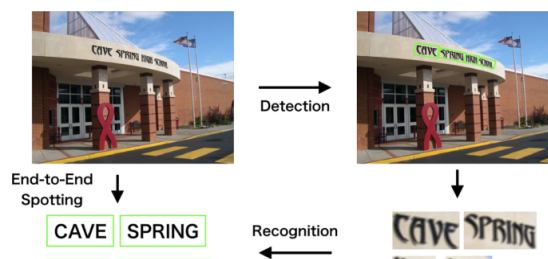
DICTIONARY-GUIDED SCENE TEXT RECOGNITION

## TÓM TẮT (Tối đa 400 từ)

Bài toán nhận dạng văn bản trong hình ảnh là một bài toán khó nhưng có nhiều ứng dụng thiết thực trong cuộc sống. Hiện nay, các phương pháp nhận dạng văn bản trong hình ảnh sử dụng từ điển để cải thiện hiệu suất nhận dạng. Tuy nhiên, cách tiếp cận còn đơn giản và tồn tại nhiều hạn chế vì họ chuyển kết quả đầu ra thành một từ xuất hiện trong từ điển dựa vào khoảng cách. Trong dự án này, chúng tôi nghiên cứu một cách tiếp cận khác để có thể tận dụng tối đa tiềm năng của bộ từ điển bằng cách sử dụng từ điển để tạo danh sách các kết quả có thể xảy ra. Bên cạnh đó, một bộ dữ liệu cho bài toán này trong ngữ cảnh Tiếng Việt sẽ được chúng tôi xây dựng cho cộng đồng.

## GIỚI THIỆU (Tối đa 1 trang A4)

Nhận dạng và phát hiện văn bản theo hình ảnh là một vấn đề nghiên cứu quan trọng và được nhiều chuyên gia quan tâm. Có rất nhiều ứng dụng đối với vấn đề này như lập bản đồ và bản địa hóa, điều hướng bằng robot, nâng cao khả năng tiếp cận cho người khiếm thị. Tuy nhiên trong thực tế, nhiều trường hợp văn bản trong tự nhiên không được rõ ràng do phong cách nghệ thuật, thời tiết hay những điều kiện ánh sáng bất lợi. Trong nhiều trường hợp, những điều bất lợi đó làm cho văn bản xuất



hiện trở nên mơ hồ và sẽ không thể giải quyết được nếu không có lập luận về ngôn ngữ của văn bản.

Một cách tiếp cận phổ biến để giải quyết vấn đề về sự mơ hồ của văn bản giúp cải thiện hiệu suất của mô hình hệ thống nhận dạng văn bản là đối với mỗi văn bản được phát hiện tiến hành tạo chuỗi ký tự có thể xảy ra nhất sau đó tìm trong từ điển từ giống với nó nhất dựa trên khoảng cách Levenshtein. Tuy nhiên, hướng tiếp cận này có ba vấn đề lớn. Đầu tiên, nhiều trường hợp văn bản là các từ nước ngoài hoặc được tạo ra không có trong từ điển, nên việc buộc đầu ra là một từ trong từ điển dẫn đến kết quả sai. Thứ hai, không có sự phản hồi trong quy trình suy luận khi huấn luyện mô hình. Thứ ba, khoảng cách dùng để xác định từ được xuất ra còn nhiều hạn chế trong một vài trường hợp như có cùng lúc nhiều từ có khoảng cách bằng nhau.

Trong dự án nghiên cứu này, hướng tiếp cận mới sẽ được chúng tôi nghiên cứu và thực hiện. Thay vì buộc đầu ra của mô hình là một từ trong từ điển, chúng tôi đặt giả thuyết rằng liệu có thể sử dụng từ điển để tạo ra danh sách các ứng cử viên phù hợp, sau đó sẽ được đưa trở lại mô-đun tính điểm để tìm ra đầu ra tương thích nhất với các đặc trưng đã được rút trích. Với việc cải thiện khả năng suy luận của mô hình chúng tôi hy vọng hướng tiếp cận này sẽ đạt được những thành tựu đáng kể khi so với các mô hình hiện tại. Ngoài ra, chúng tôi còn xây dựng một bộ dữ liệu về Scene Text Recognition dành cho Tiếng Việt có tên là VinText.

### **MỤC TIÊU** *(Viết trong vòng 3 mục tiêu)*

- Đề xuất cách tiếp cận mới, thay vì buộc mô hình phải dự đoán đầu ra là một từ thuộc từ điển thì tích hợp từ điển vào mô hình. Từ đó, tạo ra danh sách các từ ứng cử viên phù hợp, sau đó, đưa vào module tính điểm để tìm ra từ tương thích nhất với các đặc điểm ngoại hình.
- So sánh kết quả của mô hình Scene Text Recognition cơ bản và phiên bản cải tiến (tích hợp từ điển vào mô hình ban đầu) bằng cách sử dụng các bộ dataset TotalText, ICDAR2013, ICDAR2015 và VinText để đánh giá.
- Xây dựng bộ dataset cho bài toán Scene Text Recognition trong ngữ cảnh tiếng Việt.

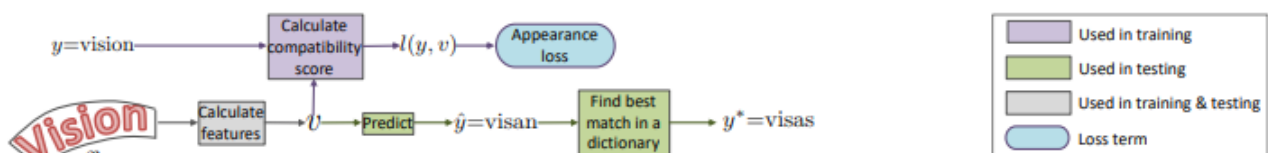
## NỘI DUNG VÀ PHƯƠNG PHÁP

### 1. Nội dung nghiên cứu:

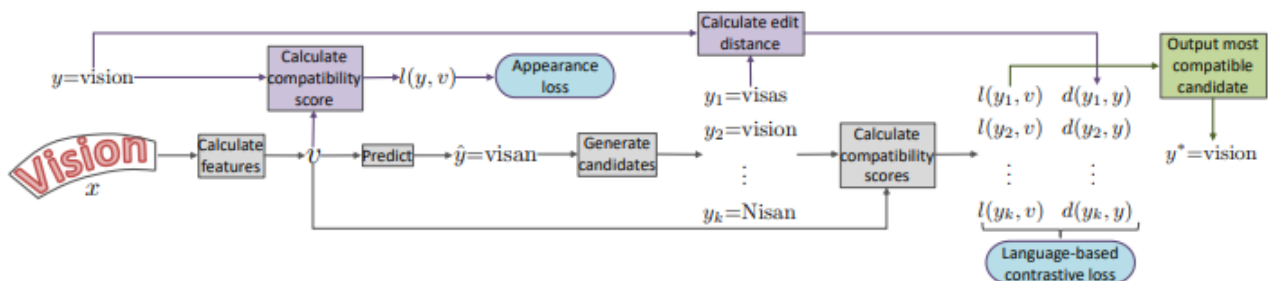
- Nghiên cứu về các mô hình detection tìm ra các mô hình tốt nhất để sử dụng trong quá trình nghiên cứu. Do dự án này sẽ tập trung phát triển vào phần recognition.
- Nghiên cứu về cách thức sinh danh sách ứng cử viên trong quá trình suy luận và huấn luyện
- Nghiên cứu về các khoảng cách để có thể tìm ra cách điều chỉnh khoảng cách thích hợp với hướng tiếp cận mới này.
- Thực hiện việc thu thập dữ liệu cho bộ dữ liệu VinText thông qua các cách như thu thập trên internet và chụp hình thực tế.

### 2. Phương pháp nghiên cứu:

- Đối với phần detection, các phương pháp phát hiện chữ trong hình ảnh tốt nhất hiện tại sẽ được nghiên cứu áp dụng (ABCNet [2], MaskTextSpotterV3[3]). Vì nghiên cứu này sẽ tập trung cải thiện phần recognition nên những phương pháp phát hiện tốt nhất được sử dụng sẽ góp phần làm tăng hiệu quả của mô hình tổng.



(a) The normal scene text recognition pipeline



(b) The proposed scene text recognition pipeline

- Nghiên cứu về việc sinh tập ứng cử viên (Candidate generation). Từ điển sẽ được nghiên cứu và được dùng trong việc tạo ra danh sách các từ ứng cử viên

trong cả quá trình suy luận (inference) và huấn luyện (training). Trong quá trình huấn luyện, chúng tôi đặt giả thuyết rằng danh sách cách từ ứng cử viên là  $k$  từ trong bộ từ điển có khoảng cách Levenshtein nhỏ nhất đến kết quả dự đoán ban đầu. Ví dụ: nếu kết quả dự đoán là visan và  $k = 10$  thì danh sách ứng viên sẽ là: visa, vise, vided, vises, visi, vising, vision, visit, visor, vista. Trong quá trình huấn luyện, chúng tôi nghiên cứu sao cho có thể sử dụng groundtruth ban đầu và kết quả dự đoán để tạo ra danh sách các từ ứng cử viên để có thể tạo ra  $k$  từ một cách chính xác hơn.

- Hàm mất mát sẽ được chúng tôi nghiên cứu và chỉnh sửa sao cho phù hợp với phương pháp mà chúng tôi đề xuất. Chúng tôi sẽ nghiên cứu và phát triển từ negative log likelihood và KL-divergence.
- Các phương pháp trước đây cũng sẽ được chúng tôi chạy thực nghiệm trong quá trình nghiên cứu để có thể so sánh với phương pháp mà chúng tôi đề xuất.
- Đối với phần thu thập dữ liệu, dữ liệu cho VinTex sẽ được thu thập bởi hai hình thức thông qua internet và chụp bởi các nhân viên thu thập dữ liệu. Bộ dữ liệu này sẽ lấy bối cảnh hằng ngày tại Việt Nam như các biển quảng cáo, biển báo giao thông, bảng thông báo, ...



- Dữ liệu sẽ được chia thành nhiều loại (như hình ...) và được gán nhãn theo từng từ có trong hình. Mỗi hình ảnh được chú thích bởi hai nhân viên chú thích độc lập. Nếu mối tương quan giữa các chú thích của họ nhỏ hơn 98%, chúng tôi

sẽ yêu cầu họ kiểm tra chéo lẫn nhau và giải quyết sự khác biệt. Bước cuối cùng, chúng tôi chọn chú thích từ một chú thích theo cách thủ công và kiểm tra trực quan chú thích để đảm bảo chú thích đó đáp ứng các yêu cầu chất lượng của chúng tôi.

## **KẾT QUẢ MONG ĐỢI**

- Hướng tiếp cận mới sẽ đạt được state-of-the-art khi so sánh (h-means) với các giải pháp trước đây (TextDragon, Boundary TextSpotter, CharNet, Mask TextSpotter(v2, v3), ABCNet<sup>pub</sup>, ABCNet, ABCNet+D) theo hai hướng: sử dụng Dictionary và không sử dụng Dictionary trên các bộ dữ liệu cho bài toán này như TotalText, ICDAR13, ICDAR15 và VinText
- Chứng minh rằng bộ dataset cho vấn đề Scene Text Recognition đến từ team Vin (VinText) là một bộ dataset có độ tin cậy cao và đáng để được đưa vào sử dụng rộng rãi.

## **TÀI LIỆU THAM KHẢO** (*Định dạng DBLP*)

- [1]. Chee Kheng Chng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017
- [2]. Y. Liu, Hao Chen, Chunhua Shen, Tong He, Lian-Wen Jin, and L. Wang: ABCNet: Real-time scene text spotting with adaptive bezier-curve network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020
- [3]. Minghui Liao, Guan Pang, J. Huang, Tal Hassner, and X. Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. ArXiv, 2020
- [4]. Wen-Yang Hu, Xiaocong Cai, J. Hou, Shuai Yi, and Zhiping Lin. Gtc: Guided training of ctc towards efficient and accurate scene text recognition. ArXiv, 2020.