RESEARCH ARTICLE

YOLOF-SE: CẢI TIẾN MÔ HÌNH YOLOF BẰNG CƠ CHẾ CHÚ Ý NHẰM NÂNG CAO HIỆU SUẤT PHÁT HIỆN ĐỐI TƯỢNG

Tôn Quang Toại[1*], Lưu Gia Khang[1,2]

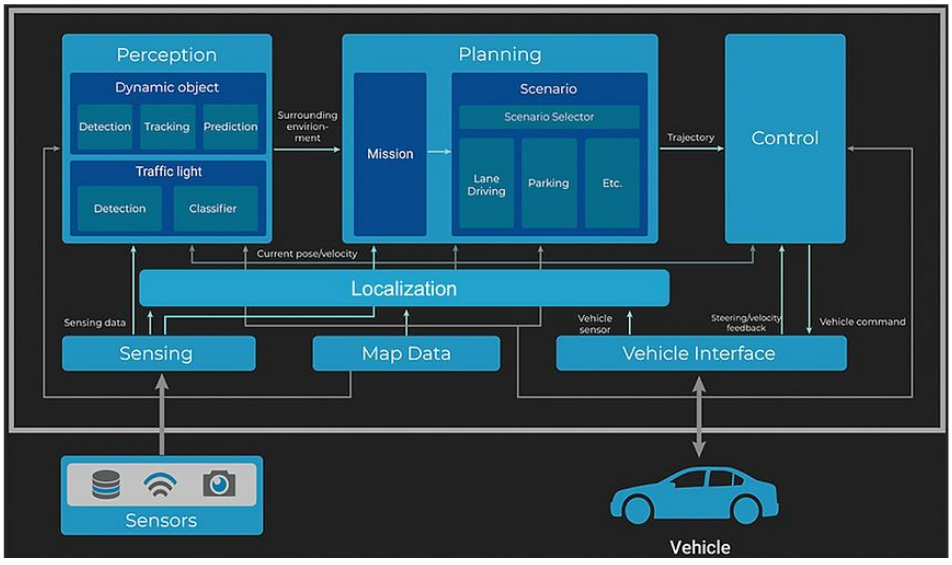
[1] Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ -Tin học TP.HCM
[2] Công ty trách nhiệm hữu hạn Emage Development
toaitq@huflit.edu.vn, khanglg@huflit.edu.vn

TÓM TẮT—Phát hiện đối tượng (object detection) là một trong những vấn đề trọng tâm của thị giác máy tính, đặc biệt trong các lĩnh vực như robot và xe tự hành, nơi việc nhận thức và hiểu biết chính xác môi trường xung quanh là điều kiện tiên quyết để vận hành an toàn và hiệu quả. Theo hướng tiếp cận truyền thống, các mô hình phát hiện đối tượng thường đòi hỏi chi phí tính toán đáng kể và thời gian huấn luyện kéo dài, gây khó khăn cho việc triển khai thực tế. Để giải quyết những thách thức này, một phiên bản cải tiến của mô hình YOLOF (You Only Look One-level Feature), gọi là YOLOF-SE, được đề xuất với mục tiêu nâng cao hiệu quả và độ chính xác trong phát hiện đối tượng. Phương pháp đề xuất tăng cường phần xương sống (backbone) bằng cơ chế chú ý (attention) nhằm tinh chỉnh biểu diễn đặc trưng, giúp mô hình tập trung tốt hơn vào các khu vực quan trọng trong ảnh. Cách tiếp cận này giữ được sự đơn giản và tốc độ vốn có của YOLOF, đồng thời cải thiện hiệu suất nhận diện. Mô hình YOLOF-SE được đánh giá trên hai tập dữ liệu tiêu chuẩn: MS COCO và BDD100K. Trên MS COCO, mô hình đạt AP là 40.4%, nằm trong nhóm hiệu suất hàng đầu đối với các phương pháp phát hiện một giai đoạn. Trên BDD100K, tập dữ liệu thiết kế cho kịch bản lái xe tự động, mô hình đạt AP là 33.2%, cho thấy khả năng thích ứng tốt trong môi trường phức tạp và đa dạng. Kết quả này chứng minh rằng phương pháp đề xuất có thể cân bằng giữa độ chính xác và hiệu quả, và là lựa chọn phù hợp cho các ứng dụng thực tế, nơi tài nguyên tính toán bị hạn chế.

Từ khóa—Attention Mechanism, Object Detection, YOLOF, MS COCO, BDD100K.

I. GIỚI THIỆU

Trong kiến trúc tổng thể của hệ thống xe tự hành hoặc robot, chuỗi xử lý chính bao gồm bốn thành phần trụ cột: Localization (định vị), Perception (nhận thức), Planning (lập kế hoạch) và Control (điều khiển) (hình 1). Localization sử dụng các cảm biến như GNSS, IMU và LiDAR để xác định chính xác vị trí cũng như hướng di chuyển của phương tiện. Perception phân tích dữ liệu từ cảm biến hình ảnh hoặc đám mây điểm để phát hiện, nhận dạng và theo dõi các đối tượng xung quanh. Tiếp đó, Planning xây dựng lộ trình tối ưu và tính toán quỹ đạo di chuyển an toàn, trong khi Control chuyển đổi quỹ đạo này thành các tín hiệu điều khiển cụ thể để vận hành phương tiện. Trong chuỗi xử lý này, Perception giữ vai trò trung tâm, bởi chất lượng nhận thức môi trường sẽ quyết định hiệu quả của các bước tiếp theo như lập kế hoạch và điều khiển. Do đó, nâng cao độ chính xác và hiệu quả của hệ thống phát hiện đối tượng trong phân hệ Perception là yếu tố then chốt, đảm bảo tính an toàn và khả năng vận hành tin cậy của toàn bộ hệ thống xe tự hành và robot.



Hình 1. *Sơ đồ luồng dữ liệu và các thành phần chức năng chính trong hệ thống vận hành xe tự hành thời gian thực [1]*

[*] Coressponding Author

Phát hiện đối tượng 2D là một trong những nhiệm vụ nền tảng của phân hệ Perception, đóng vai trò nhận diện các thực thể quan trọng như xe cộ, người đi bộ hay biển báo giao thông từ dữ liệu đầu vào. Đây là bước khởi đầu thiết yếu, ảnh hưởng trực tiếp đến độ chính xác của toàn bộ chuỗi xử lý, từ lập kế hoạch đến điều khiển. Bài toán này đặt ra yêu cầu phải cân bằng giữa ba yếu tố: độ chính xác, tốc độ suy luận và khả năng triển khai trên phần cứng giới hạn, là điều kiện phổ biến trong các hệ thống robot và xe tự hành thực tế.

Các mô hình phát hiện đối tượng hai giai đoạn như Faster R-CNN [2] đã chứng minh hiệu quả vượt trội nhờ cơ chế Region Proposal Network (RPN) kết hợp với nhánh phân loại và hồi quy, đạt độ chính xác cao trên nhiều tập dữ liệu chuẩn như MS COCO [3] và PASCAL VOC [4], [5], [6]. Tuy nhiên, hạn chế về tốc độ suy luận và chi phí tính toán khiến chúng khó áp dụng trong môi trường thời gian thực. Ngược lại, các mô hình một giai đoạn như SSD (Single Shot MultiBox Detector) [7] và YOLO (You Only Look Once) [8], [9], [10] mang lại cải tiến đáng kể về tốc độ. SSD tận dụng nhiều tầng đặc trưng từ backbone để phát hiện đối tượng ở nhiều tỉ lệ, giúp cải thiện hiệu quả trên đối tượng nhỏ. Trong khi YOLO đơn giản hóa toàn bộ quy trình thành một bài toán hồi quy duy nhất, đạt tốc độ vượt trội nhưng hạn chế với đối tượng chồng lấn hoặc kích thước nhỏ.

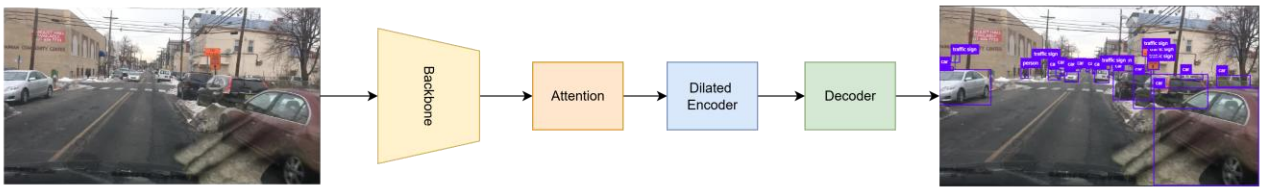
Để tăng khả năng phát hiện đa tỉ lệ, các mô hình hiện đại thường tích hợp FPN, kết hợp đặc trưng từ tầng sâu và tầng nông nhằm cải thiện độ chính xác trên nhiều kích thước vật thể. Tuy nhiên, FPN lại làm tăng đáng kể số tham số và chi phí tính toán, ảnh hưởng tiêu cực đến tốc độ suy luận khi triển khai trên thiết bị nhúng hoặc GPU tầm trung.Để giải quyết, YOLOF (You Only Look One-level Feature) [11], [12], [13] được đề xuất như một biến thể đơn giản hóa, thay thế FPN bằng cách khai thác duy nhất tầng C5 của backbone ResNet [14] và mở rộng trường nhìn nhờ phép tích chập giãn cách (dilated convolution) [15]. Thiết kế này giúp giữ tốc độ cao và hiệu quả trên đối tượng trung bình đến lớn, nhưng vẫn hạn chế trong phát hiện đối tượng nhỏ, thường phổ biến trong giao thông đô thị như người đi bộ, xe máy hay biển báo.

Một thách thức khác đến từ kiến trúc backbone truyền thống như ResNet hoặc RegNet [16], các backbone này xử lý đồng đều giữa các kênh đặc trưng mà không xét đến tầm quan trọng khác nhau của từng kênh. Điều này làm giảm khả năng tập trung vào các tín hiệu quan trọng trong bối cảnh phức tạp như ánh sáng yếu hoặc vật thể bị che khuất. Để khắc phục, nghiên cứu này đề xuất một phiên bản cải tiến từ YOLOF, gọi là YOLOF-SE, trong đó backbone được tích hợp cơ chế chú ý kênh Squeeze-and-Excitation (SE) [17]. Cơ chế SE cho phép tái phân bổ trọng số giữa các kênh dựa trên ngữ cảnh toàn cục, từ đó nâng cao khả năng biểu diễn đặc trưng mà không làm gia tăng đáng kể chi phí tính toán.

Kiến trúc đề xuất vẫn giữ thiết kế đơn tầng với chập giãn cách nhằm duy trì tốc độ và độ phân giải không gian, đồng thời cải thiện khả năng biểu diễn nhờ khối SE. Các đóng góp chính của nghiên cứu bao gồm: (i) phát triển một mô hình YOLOF-SE để phát hiện đối tượng 2D nhẹ và hiệu quả dựa trên YOLOF, loại bỏ hoàn toàn FPN để giảm độ phức tạp; (ii) tích hợp cơ chế chú ý SE nhằm tăng cường khả năng học đặc trưng; và (iii) tiến hành đánh giá thực nghiệm trên tập dữ liệu MS COCO và BDD100K [18] để kiểm chứng hiệu quả trong môi trường giao thông thực tế có nhiều biến thiên về thời tiết, ánh sáng và mật độ giao thông.

Phần còn lại của bài báo được tổ chức như sau: phần II trình bày chi tiết kiến trúc của mô hình; phần III trình bày kết quả thực nghiệm trên tập dữ liệu MS COCO và BDD100K; phần IV trình bày một số thảo luận và đề xuất các hướng nghiên cứu tiếp theo; phần V trình bày lời cảm ơn; cuối cùng, phần VI trình bày một số tài liệu tham khảo.

II. PHƯƠNG PHÁP ĐỀ XUẤT



Hình 2. *Kiến trúc tổng thể của mô hình YOLOF-SE, bao gồm bốn thành phần chính: Backbone, Attention, Dilated Encoder và Decoder*

Mô hình YOLOF được thiết kế với cấu trúc đơn tầng nhằm đơn giản hóa quá trình suy luận và tăng tốc độ xử lý. Thay vì áp dụng FPN như các phương pháp phát hiện đa tỉ lệ, YOLOF chỉ khai thác một tầng đặc trưng sâu (tầng thứ 5, ký hiệu là C5) và sử dụng các khối chập giãn cách trong encoder để mở rộng trường nhìn (receptive field). Nhờ vậy, mô hình đạt hiệu quả tốt trong phát hiện các đối tượng trung bình và lớn, đồng thời duy trì tốc độ suy luận cao. Tuy nhiên, việc bỏ qua các tầng đặc trưng nông khiến YOLOF gặp hạn chế trong việc xử lý các đối tượng nhỏ, chiếm tỷ lệ lớn trong môi trường giao thông thực tế.

Bên cạnh đó, một thách thức khác đến từ khả năng biểu diễn đặc trưng của backbone. Các kiến trúc phổ biến như ResNet thường xử lý đồng đều giữa các kênh, không xem xét mức độ quan trọng khác nhau, dẫn đến việc mô hình có thể bỏ qua các tín hiệu then chốt trong những tình huống phức tạp như ánh sáng yếu, thời tiết bất lợi hoặc vật thể bị che khuất. Để khắc phục hạn chế này, nghiên cứu này tích hợp cơ chế chú ý ngay sau backbone (Hình 2). Cơ chế này cho phép tái phân bổ trọng số giữa các kênh dựa trên ngữ cảnh toàn cục, giúp mô hình tập trung mạnh hơn vào những tín hiệu quan trọng mà không làm tăng đáng kể chi phí tính toán. Như minh họa trong Hình 2, kiến trúc tổng thể của mô hình gồm ba thành phần chính: (i) backbone trích xuất đặc trưng ban đầu, (ii) encoder với chập giãn cách và cơ chế chú ý để khuếch đại ngữ cảnh, và (iii) detection head thực hiện dự đoán vị trí và phân loại đối tượng.

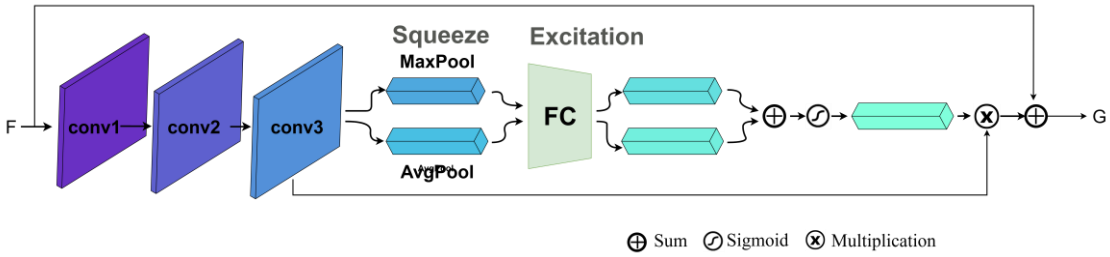
Thiết kế này vừa kế thừa ưu điểm đơn giản, tốc độ cao của YOLOF, vừa nâng cao khả năng biểu diễn đặc trưng, từ đó cải thiện hiệu quả phát hiện đối tượng trong các kịch bản giao thông đô thị nhiều biến động.

A. BACKBONE

Trong kiến trúc YOLOF-SE đề xuất, chúng tôi sử dụng ResNet đã được huấn luyện trước trên tập dữ liệu ImageNet làm backbone để trích xuất đặc trưng từ ảnh đầu vào. Với một ảnh $I \in \mathbb{R}^{H\times W\times 3}$, backbone tạo ra một bản đồ đặc trưng ở tầng C5, được ký hiệu là $F \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C}$, trong đó $C = 2048$ là số kênh đặc trưng đầu ra, và hệ số giảm mẫu là 32. Các đặc trưng này mang ngữ nghĩa ở mức cao, đóng vai trò nền tảng cho quá trình suy luận của encoder và decoder (còn gọi là detection head) ở các tầng tiếp theo. Với thiết kế này, backbone không chỉ trích xuất thông tin thị giác cơ bản mà còn biểu diễn cấu trúc phân cấp của ảnh, từ các cạnh và hình khối đơn giản ở tầng nông cho tới những khái niệm ngữ nghĩa phức tạp hơn ở tầng sâu. Điều này cho phép mô hình có được một không gian biểu diễn giàu thông tin, hỗ trợ trực tiếp cho nhiệm vụ phát hiện đối tượng trong môi trường giao thông thực tế với nhiều biến thiên về hình dạng, kích thước và ngữ cảnh.

B. ATTENTION

Mặc dù backbone cung cấp các đặc trưng ngữ nghĩa mạnh, một hạn chế lớn của ResNet là xử lý đồng đều giữa các kênh mà không xét đến mức độ quan trọng khác nhau. Điều này khiến mô hình dễ bỏ qua các tín hiệu then chốt, đặc biệt trong bối cảnh giao thông phức tạp. Để khắc phục, mô hình tích hợp thêm khối chú ý kênh SE (Squeeze-and-Excitation) ngay sau backbone (Hình 3).

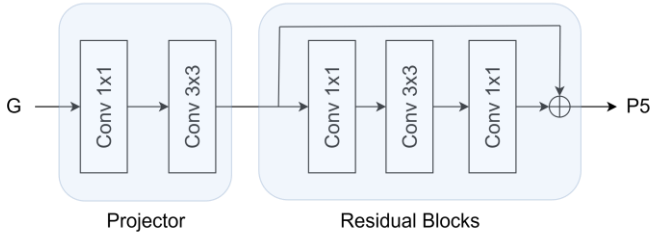


Hình 3. *Cấu trúc khối Attention theo cơ chế Squeeze-and-Excitation (SE), bao gồm giai đoạn squeeze (trích xuất thông tin kênh toàn cục) và excitation (tái phân bổ trọng số kênh)*

Cụ thể, sau khi backbone sinh ra bản đồ đặc trưng $F \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C}$ với $C = 2048$, bản đồ đặc trưng F đi qua ba tầng tích chập tuần tự (conv1–conv3) để thu được biểu diễn trung gian $U \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C}$. Biểu diễn này có cùng kích thước không gian và số kênh với F, nhưng đã được làm giàu thêm thông tin ngữ nghĩa, tạo tiền đề cho giai đoạn hiệu chỉnh kênh. Từ U, bước squeeze nén thông tin không gian bằng cách kết hợp Global Average Pooling và Global Max Pooling, tạo ra vector ngữ cảnh $z \in \mathbb{R}^C$ với $z_c = \frac{1}{H'.W'}\sum_{i,j} U_c(i,j) + max_{i,j}U_c(i,j)$, trong đó $H' = \frac{H}{32}, W' = \frac{W}{32}$. Ở bước excitation, vector z được đưa qua hai tầng fully connected cùng ReLU và sigmoid để sinh ra trọng số kênh $s = \sigma\big(W_2\delta(W_1 z)\big)$, trong đó δ là ReLU và σ là sigmoid. Trọng số này được nhân theo từng kênh với U để tạo ra U', sau đó kết hợp với tín hiệu gốc thông qua skip connection, thu được đầu ra $G = F + U'$. Thiết kế này vừa bảo toàn thông tin ban đầu, vừa nhấn mạnh vào các kênh quan trọng, giúp mô hình tập trung hơn vào tín hiệu hữu ích và loại bỏ nhiễu.

C. DILATED ENCODER

Đầu ra từ khối Attention, ký hiệu là G, tiếp tục được đưa vào Dilated Encoder (Hình 4). Thành phần này được thiết kế nhằm mở rộng trường nhìn của đặc trưng mà không làm giảm độ phân giải không gian. Dilated Encoder gồm hai phần: Projector và Residual Blocks.



Hình 4. *Cấu trúc khối Dilated Encoder, trong đó đặc trưng đầu vào G lần lượt đi qua Projector và Residual Blocks, sau đó kết hợp với kết nối tắt để tạo ra bản đồ đặc trưng P5*

Trong projector, đặc trưng G trước hết được đưa qua một phép tích chập 1×1 để giảm số chiều kênh từ C xuống C_p (thường là 512), biểu diễn bằng công thức $G' = W_{1\times1} * G$. Tiếp đó, một phép tích chập 3×3 được áp dụng lên G' nhằm tái tổ chức không gian đặc trưng mà vẫn duy trì kích thước $\frac{H}{32} \times \frac{W}{32}$.

Đầu ra G' sau projector được đưa vào residual block, bao gồm ba phép tích chập liên tiếp: 1×1 để giảm số chiều kênh, 3×3 giãn cách với hệ số d để mở rộng receptive field, và 1×1 để khôi phục lại số chiều ban đầu. Kết quả của chuỗi tích chập này sau đó được cộng với đầu vào ban đầu thông qua một skip connection, nhờ đó mô hình vừa bảo toàn thông tin gốc vừa duy trì sự ổn định trong lan truyền gradient. Toàn bộ quá trình có thể viết gọn thành $P5 = W_{1\times1}^{(2)} * \delta\left(W_{3\times3,d} * \delta\left(W_{1\times1}^{(1)} * G'\right)\right) + G'$, trong đó δ là ReLU. Kết quả $P5$ là bản đồ đặc trưng có receptive field rộng hơn, chứa nhiều ngữ cảnh hơn nhưng vẫn duy trì độ phân giải không gian. Đây là bước trung gian quan trọng, giúp mô hình đạt sự cân bằng giữa hiệu quả tính toán và khả năng phát hiện đối tượng ở nhiều kích thước khác nhau.

D. DECODER

Đặc trưng đầu vào của decoder là bản đồ đặc trưng $P5$, thu được từ Dilated Encoder. Như đã phân tích ở trên, $P5$ có kích thước $N \times 512 \times \frac{H}{32} \times \frac{W}{32}$ trong đó N là batch size, 512 là số kênh đặc trưng sau projector, còn $\frac{H}{32} \times \frac{W}{32}$ lần lượt là chiều cao và chiều rộng của bản đồ đặc trưng sau khi đã được giảm mẫu từ ảnh gốc. Bản đồ $P5$ chứa các thông tin ngữ cảnh phong phú và đóng vai trò trung gian để tạo ra dự đoán cuối cùng.

Từ đặc trưng $P5$, decoder chia thành hai nhánh song song: classification head và regression head. Trong classification head, đặc trưng $P5$ đi qua hai tầng tích chập liên tiếp để sinh ra đầu ra có kích thước $O_{cls} \in \mathbb{R}^{N\times(K.A)\times\frac{H}{32}\times\frac{W}{32}}$, trong đó K là số lớp và A là số anchor trên mỗi vị trí không gian. Điều này có nghĩa là tại mỗi vị trí (i, j), classification head xuất ra $K \cdot A$ xác suất, tương ứng với các lớp khác nhau trên từng anchor.

Ngược lại, regression head được thiết kế phức tạp hơn để đảm bảo độ chính xác cao khi dự đoán hộp giới hạn. Đặc trưng $P5$ đi qua bốn tầng tích chập, cho ra đầu ra $O_{reg} \in \mathbb{R}^{N\times(4.A)\times\frac{H}{32}\times\frac{W}{32}}$. Mỗi anchor tại một vị trí (i, j) dự đoán một vector 4 chiều (x, y, w, h), trong đó (x, y) là toạ độ tâm hộp giới hạn (bounding box) và (w, h) là chiều rộng, chiều cao hộp giới hạn.

Bên cạnh hai đầu ra chính, mô hình còn học một bản đồ implicit objectness từ regression head. Bản đồ này có kích thước $O_{obj} \in \mathbb{R}^{N\times A\times\frac{H}{32}\times\frac{W}{32}}$, trong đó mỗi phần tử $o_{n,a,i,j}$ phản ánh mức độ tin cậy của anchor thứ a tại vị trí (i, j). Thay vì yêu cầu huấn luyện trực tiếp, objectness được học một cách ẩn, đóng vai trò như một trọng số hiệu chỉnh. Khi tính toán xác suất phân loại cuối cùng, mô hình nhân xác suất phân loại gốc với trọng số này, tức là $\hat{p}_c = p_c . o$, trong đó p_c là xác suất đối tượng thuộc lớp c và o là objectness của anchor tương ứng.

Cách thiết kế này có hai lợi ích rõ rệt. Thứ nhất, classification head được giữ nhẹ hơn (chỉ hai tầng tích chập), giúp tăng tốc độ suy luận, trong khi regression head được tăng cường độ sâu (bốn tầng tích chập) để đảm bảo độ chính xác định vị. Thứ hai, việc kết hợp implicit objectness giúp cải thiện tính ổn định, giảm thiểu dự đoán sai ở các vùng nền và nâng cao chất lượng phát hiện đối tượng trong môi trường giao thông phức tạp.

III. THỰC NGHIỆM

A. TẬP DỮ LIỆU

Để bảo đảm quá trình đánh giá mô hình mang tính toàn diện và phản ánh được cả khía cạnh tổng quát hóa lẫn khả năng ứng dụng trong thực tế, hai tập dữ liệu MS COCO 2017 và BDD100K được lựa chọn cho thực nghiệm. Trong khi MS COCO là tập dữ liệu chuẩn mực toàn cầu để đo lường hiệu quả của các mô hình phát hiện đối tượng trong nhiều tình huống thị giác máy tính phổ biến, BDD100K lại được thiết kế chuyên biệt cho bối cảnh giao thông, phản ánh chính xác những điều kiện mà hệ thống xe tự hành thường gặp phải. Việc kết hợp cả hai tập dữ liệu cho phép

đánh giá đồng thời khả năng biểu diễn đặc trưng ở phạm vi rộng và năng lực thích ứng trong môi trường đặc thù (Bảng 1).

Bảng 1. *Bảng 3 dưới đây minh họa sự khác biệt quan trọng giữa hai tập dữ liệu.*

Tập dữ liệu	Số lượng ảnh	Nhiều thành phố	Nhiều điều kiện thời tiết	Nhiều thời điểm	Nhiều loại cảnh
MS COCO	143.575	Không	Không	Không	Không
BDD100K	120.000	Có	Có	Có	Có

MS COCO (Microsoft Common Objects in Context) bao gồm 143.575 hình ảnh với khoảng 1,5 triệu nhãn đối tượng thuộc 80 lớp khác nhau. Tập dữ liệu này đa dạng về kích thước, góc nhìn, mức độ che khuất và bối cảnh nền của các đối tượng, qua đó tạo ra thách thức cho mô hình trong việc xử lý tình huống phức tạp và đối tượng chồng lấn. Tuy nhiên, MS COCO không bao hàm những yếu tố đặc thù của môi trường giao thông như biến thiên thời tiết, thay đổi theo thời gian trong ngày hay sự khác biệt cảnh quan giữa các thành phố. Do đó, MS COCO chủ yếu phù hợp để đánh giá khả năng tổng quát hóa và hiệu quả phát hiện đối tượng trong các kịch bản thị giác máy tính tổng quát.

BDD100K (Berkeley DeepDrive 100K) là tập dữ liệu quy mô lớn trong lĩnh vực xe tự hành, bao gồm 120.000 hình ảnh gắn nhãn đối tượng. Điểm mạnh của BDD100K là tính đa dạng như: dữ liệu được thu thập từ nhiều thành phố, trong nhiều điều kiện thời tiết (nắng, mưa, sương mù), ở các thời điểm khác nhau (ban ngày, ban đêm), và với nhiều loại cảnh giao thông (đường cao tốc, khu dân cư, nút giao đông đúc,...). Nhờ đặc điểm này, BDD100K phản ánh chính xác các biến thiên môi trường phức tạp của bối cảnh giao thông thực tế, qua đó trở thành chuẩn đánh giá quan trọng cho các mô hình phát hiện đối tượng phục vụ hệ thống xe tự hành.

B. THIẾT LẬP THỰC NGHIỆM

Các thí nghiệm được thực hiện trên máy trạm trang bị GPU NVIDIA RTX A5000 (24 GB bộ nhớ), CPU đa luồng 32 lõi và RAM 128 GB. Mô hình YOLOF-SE được cài đặt bằng PyTorch, kết hợp CUDA/cuDNN để khai thác khả năng tính toán song song. Trong toàn bộ quá trình huấn luyện, chế độ Automatic Mixed Precision (AMP) được kích hoạt nhằm giảm chi phí bộ nhớ và tăng tốc độ xử lý.

Ảnh đầu vào được chuẩn hóa theo thống kê ImageNet. Các phép tăng cường dữ liệu gồm lật ngang ngẫu nhiên (xác suất 0.5) và padding để phù hợp với stride 32 của backbone. Đối với tập dữ liệu BDD100K, bổ sung thêm biến đổi màu nhẹ và làm mờ Gaussian nhằm tăng khả năng bền vững trong điều kiện ánh sáng yếu và môi trường ban đêm.

Mô hình YOLOF-SE sử dụng một mức đặc trưng duy nhất $P5 \in \mathbb{R}^{N\times 512\times \frac{H}{32}\times \frac{W}{32}}$ làm đầu vào cho decoder. Tại mỗi vị trí trong bản đồ đặc trưng, 9 anchor được tạo từ ba tỉ lệ khung $\{0.5,1.0,2.0\}$ và ba thang tỉ lệ $\{1,2^{1/3},2^{2/3}\}$. Chiến lược gán nhãn dựa trên IoU động: anchor dương khi có IoU cao với hộp thật, âm khi IoU thấp ($\leq$ 0.4), và bỏ qua các trường hợp trung gian.

Trong quá trình huấn luyện, nhánh phân loại sử dụng Focal Loss để xử lý mất cân bằng mẫu, còn nhánh hồi quy sử dụng Distance-IoU (DIoU) Loss để tối ưu hoá vị trí hộp giới hạn. Hàm mất mát tổng thể được kết hợp bằng cách cộng có trọng số hai thành phần này.

C. KẾT QUẢ THỰC NGHIỆM

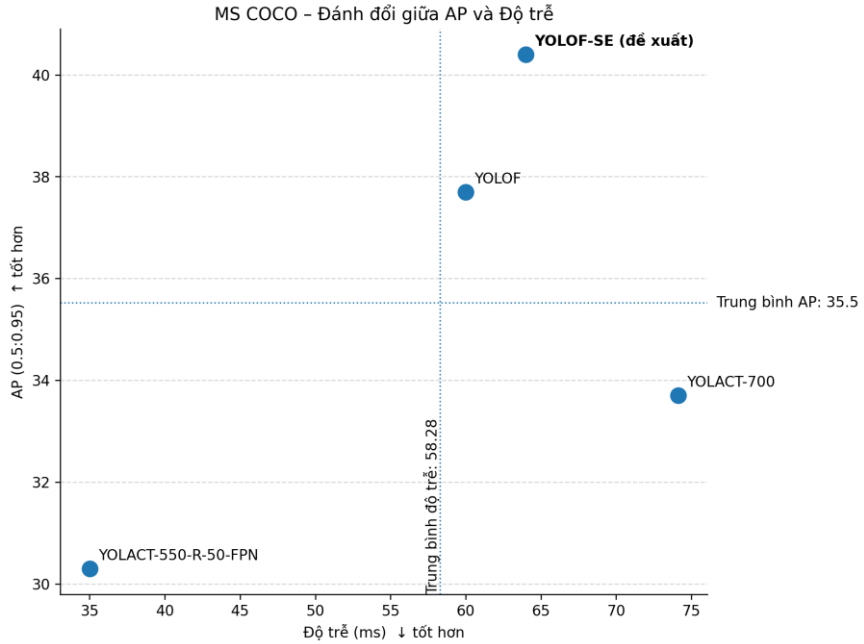
1. KẾT QUẢ TRÊN TẬP DỮ LIỆU MS COCO

Bảng 2 trình bày kết quả đánh giá trên tập dữ liệu MS COCO. Các phương pháp YOLACT-700 và YOLACT-550-R-50-FPN cho thấy giới hạn về độ chính xác, lần lượt đạt 33.7% AP và 30.3% AP. Trong đó, YOLACT-550-R-50-FPN có thời gian suy luận ngắn nhất (35.01 ms) nhưng độ chính xác thấp nhất, còn YOLACT-700 đạt AP cao hơn một chút (33.7%) song lại có độ trễ lớn nhất (74.13 ms). YOLOF gốc cải thiện đáng kể hiệu năng, đạt 37.7% AP ở tốc độ 60 ms/ảnh nhờ tận dụng đặc trưng one-level và tích chập giãn cách để mở rộng trường nhìn. Đáng chú ý, mô hình đề xuất YOLOF-SE đạt 40.4% AP với 64 ms/ảnh, cao hơn YOLOF gốc 2.7% AP trong khi độ trễ chỉ tăng nhẹ.

Bảng 2. *So sánh các mô hình trên tập dữ liệu MS COCO*

Phương pháp	AP[0.5:0.95] (%)	Thời gian (ms)
YOLACT-700	33.7	74.13
YOLACT-550-R-50-FPN	30.3	35.01
YOLOF	37.7	60.00
YOLOF-SE (Mô hình đề xuất)	40.4	64.00

Hình 5 minh họa trực quan sự đánh đổi giữa độ chính xác và tốc độ suy luận trên MS COCO. Kết quả cho thấy YOLOF-SE nổi bật hơn hẳn, đạt AP cao nhất trong nhóm so sánh mà độ trễ vẫn ở mức hợp lý. Trong khi các phương pháp YOLACT thể hiện sự mất cân đối rõ rệt giữa độ chính xác và tốc độ, YOLOF-SE duy trì được sự cân bằng hiệu quả. Sự cải thiện này xuất phát từ việc tích hợp cơ chế chú ý kênh SE vào backbone, giúp mô hình tập trung vào đặc trưng quan trọng, đồng thời tận dụng hiệu quả encoder giãn cách để mở rộng khả năng biểu diễn mà không làm gia tăng đáng kể chi phí tính toán.



Hình 5. *So sánh đánh đổi giữa độ chính xác (AP) và độ trễ trên tập dữ liệu MS COCO. Mô hình YOLOF-SE (đề xuất) đạt AP cao nhất với độ trễ hợp lý, vượt trội hơn so với các phương pháp đối chứng*

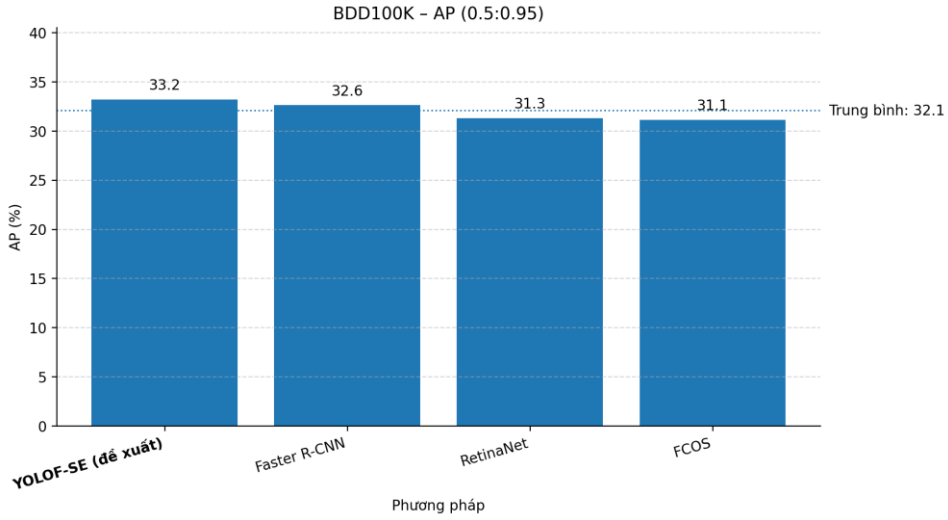
2. KẾT QUẢ TRÊN TẬP DỮ LIỆU BDD100K

Bảng 3 trình bày kết quả đánh giá của các mô hình trên tập dữ liệu BDD100K, phản ánh khả năng thích ứng của mô hình trong môi trường giao thông thực tế. Các phương pháp đối chứng như FCOS [19], RetinaMask [20] và Faster R-CNN, được trích dẫn từ Visual Intelligence and Systems Group at ETH Zürich [21] đạt lần lượt 31.1% AP, 31.3% AP và 32.6% AP. Trong đó, Faster R-CNN thể hiện mức cải thiện rõ rệt so với hai phương pháp còn lại. Tuy nhiên, mô hình đề xuất YOLOF-SE đạt 33.2% AP, cao hơn tất cả các phương pháp so sánh, chứng tỏ hiệu quả của việc kết hợp cơ chế chú ý kênh SE và tích chập giãn cách trong encoder.

Bảng 3. *So sánh các mô hình trên tập dữ liệu BDD100K*

Phương pháp	AP[0.5:0.95] (%)
FCOS	31.1
RetinaNet	31.3
Faster R-CNN	32.6
YOLOF-SE (Mô hình đề xuất)	33.2

Hình 6 minh họa trực quan sự khác biệt về độ chính xác trên tập dữ liệu BDD100K. Có thể thấy YOLOF-SE vượt trội hơn so với FCOS, RetinaNet và Faster R-CNN, đạt mức AP cao nhất trong nhóm. Kết quả này cho thấy mô hình có khả năng xử lý tốt những tình huống phức tạp trong giao thông thực tế, chẳng hạn như phát hiện người đi bộ trong điều kiện ánh sáng yếu, phương tiện dưới trời mưa hoặc các biển báo bị che khuất. Điều này khẳng định tiềm năng ứng dụng của YOLOF-SE trong bối cảnh xe tự hành và các hệ thống thị giác máy tính ngoài phòng thí nghiệm.



Hình 6. *So sánh độ chính xác (AP) trên tập dữ liệu BDD100K. YOLOF-SE (đề xuất) đạt 33.2% AP, cao hơn so với Faster R-CNN, RetinaNet và FCOS, cho thấy hiệu năng vượt trội trong bối cảnh giao thông thực tế*

Tóm lại, các kết quả trên hai tập dữ liệu MS COCO và BDD100K cho thấy YOLOF-SE đạt hiệu năng vượt trội cả về khả năng tổng quát hóa lẫn khả năng thích ứng trong môi trường giao thông thực tế. Trên MS COCO, mô hình duy trì sự cân bằng hiệu quả giữa độ chính xác và độ trễ suy luận, trong khi trên BDD100K, YOLOF-SE thể hiện ưu thế rõ rệt về độ chính xác so với các phương pháp đối chứng. Những kết quả này khẳng định tính đúng đắn của thiết kế mô hình

IV. THẢO LUẬN

Kết quả thực nghiệm trên MS COCO và BDD100K đã chứng minh tính hiệu quả của mô hình YOLOF-SE trong việc duy trì sự cân bằng giữa độ chính xác và tốc độ suy luận. So với YOLOF gốc, mô hình YOLOF-SE đạt được mức tăng đáng kể về độ chính xác nhờ việc tích hợp cơ chế chú ý kênh và encoder giãn cách. Trên MS COCO, độ chính xác tăng 2.7 % AP mà không làm gia tăng đáng kể chi phí tính toán. Trên BDD100K, YOLOF-SE đạt 33.2% AP, cao hơn tất cả các phương pháp đối chứng, cho thấy khả năng thích ứng tốt trong các tình huống giao thông đa dạng và phức tạp.

Tuy nhiên, vẫn còn tồn tại một số hạn chế. Thứ nhất, việc chỉ sử dụng một mức đặc trưng duy nhất khiến mô hình vẫn gặp khó khăn khi phát hiện các đối tượng rất nhỏ trong môi trường phức tạp, chẳng hạn như biển báo ở khoảng cách xa hoặc người đi bộ bị che khuất một phần. Thứ hai, mô hình chưa khai thác triệt để các chiến lược tăng cường dữ liệu và huấn luyện đa thang (multi-scale training), có thể cải thiện hiệu năng trên những kịch bản biến thiên mạnh về kích thước và góc nhìn. Cuối cùng, mặc dù cơ chế chú ý kênh đã nâng cao khả năng biểu diễn, nhưng đây vẫn là một dạng cơ chế chú ý tương đối đơn giản; khả năng kết hợp với các cơ chế chú ý không gian hoặc chú ý xuyên tầng vẫn chưa được kiểm chứng.

Trong tương lai, có thể phát triển nghiên cứu theo ba hướng chính. Một là kết hợp cơ chế tổng hợp đặc trưng đa tỉ lệ nhẹ thay cho FPN truyền thống, nhằm cải thiện khả năng phát hiện đối tượng nhỏ nhưng vẫn duy trì tốc độ. Hai là mở rộng sang các dạng cơ chế chú ý tiên tiến hơn, chẳng hạn như self-attention hoặc transformer-based attention, để tăng cường khả năng học đặc trưng ngữ cảnh phức tạp. Ba là tiến hành các thử nghiệm triển khai trực tiếp trên nền tảng embedded GPU hoặc xe tự hành thực tế, nhằm đánh giá tính ổn định của mô hình trong môi trường hoạt động ngoài phòng thí nghiệm.

Nhìn chung, các kết quả thu được khẳng định tiềm năng của mô hình đề xuất như một giải pháp phát hiện đối tượng hiệu quả cho hệ thống xe tự hành, đồng thời mở ra nhiều hướng nghiên cứu hứa hẹn để tiếp tục nâng cao hiệu năng và khả năng ứng dụng thực tiễn.

V. LỜI CẢM ƠN

Nghiên cứu được tài trợ bởi Trường Đại học Ngoại ngữ - Tin học Thành phố Hồ chí Minh trong khuôn khổ Đề tài mã số H2024-07.

VI. TÀI LIỆU THAM KHẢO

[1] Autoware, “Autoware - the world’s leading open-source software project for autonomous driving.” Accessed: Sep. 10, 2025. [Online]. Available: https://github.com/pixmoving-moveit/Autoware

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[3] M. and B. S. and H. J. and P. P. and R. D. and D. P. and Z. C. L. Lin Tsung-Yi and Maire, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, T. and S. B. and T. T. Fleet David and Pajdla, Ed., Cham: Springer International Publishing, 2014, pp. 740–755.

[4] M. Everingham, L. Van~Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *Int J Comput Vis*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[5] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *Int J Comput Vis*, vol. 111, pp. 98–136, 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:207252270

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," 2012.

[7] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016. doi: 10.1007/978-3-319-46448-0_2.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[9] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors Latency (ms) MS COCO mAP (%)," 2025, doi: 10.0.

[10] M. Lei *et al.*, "YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception," Sep. 2025, [Online]. Available: http://arxiv.org/abs/2506.17733

[11] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You Only Look One-level Feature," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13034–13043, 2021.

[12] H. Wei, Q. Zhang, Y. Qin, X. Li, and Y. Qian, "YOLOF-F: you only look one-level feature fusion for traffic sign detection," *Vis Comput*, vol. 40, no. 2, pp. 747–760, Feb. 2024, doi: 10.1007/s00371-023-02813-1.

[13] Q. Wang, Y. Qian, Y. Hu, C. Wang, X. Ye, and H. Wang, "M2YOLOF: Based on effective receptive fields and multiple-in-single-out encoder for object detection," *Expert Syst Appl*, vol. 213, 2023, doi: 10.1016/j.eswa.2022.118928.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[15] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *CoRR*, vol. abs/1511.07122, 2015.

[16] J. Xu, Y. Pan, X. Pan, S. Hoi, Z. Yi, and Z. Xu, "RegNet: Self-Regulated Network for Image Classification," *IEEE Trans Neural Netw Learn Syst*, vol. 34, no. 11, pp. 9562–9567, Nov. 2023, doi: 10.1109/TNNLS.2022.3158966.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[18] F. Yu *et al.*, "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 2633–2642. doi: 10.1109/CVPR42600.2020.00271.

[19] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9626–9635, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:91184137

[20] C.-Y. Fu, M. Shvets, and A. C. Berg, "RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free," *ArXiv*, vol. abs/1901.03353, 2019.

[21] ETH VIS Group, "BDD100K Model Zoo." Accessed: Sep. 20, 2025. [Online]. Available: https://github.com/SysCV

ENHANCING YOLOF WITH ATTENTION MECHANISM FOR IMPROVED OBJECT DETECTION PERFORMANCE

Ton Quang Toai, Luu Gia Khang

ABSTRACT— Object detection is one of the central problems in computer vision, particularly in fields such as robotics and autonomous driving, where accurate perception and understanding of the surrounding environment is a prerequisite for safe and efficient operation. In traditional approaches, object detection models often require significant computational costs and lengthy training times, which hinder practical deployment. To address these challenges, an improved version of the YOLOF (You Only Look One-level Feature) model is proposed, focusing on enhancing efficiency and accuracy in object detection. The proposed method strengthens the backbone with an attention mechanism to refine feature representations, enabling the model to better focus on important regions in the image. This approach preserves the inherent simplicity and speed of YOLOF while improving detection performance. The model is evaluated on two benchmark datasets: MS COCO and BDD100K. On MS COCO, the model achieves an AP of 40.4%, ranking among the top-performing one-stage detection methods. On BDD100K, a dataset designed for autonomous driving scenarios, the model achieves an AP of 33.2%, demonstrating strong adaptability in complex and diverse environments. These results confirm that the proposed method strikes a balance between accuracy and efficiency, making it a suitable choice for real-world applications where computational resources are limited.

Keywords—Attention Mechanism, Object Detection, YOLOF, MS COCO, BDD100K

Tôn Quang Toại tốt nghiệp thạc sĩ Công nghệ thông tin năm 2006 tại trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM. Hiện ông là Phó trưởng khoa, Khoa Công nghệ thông tin của Trường đại học Ngoại ngữ - Tin học TP.HCM (HUFLIT). Lĩnh vực nghiên cứu quan tâm của ông liên quan Object tracking, LiDAR và Deep learning.

Lưu Gia Khang tốt nghiệp cử nhân Công nghệ thông tin tại trường Đại học Ngoại ngữ - Tin học TP.HCM. Hiện ông đang là giảng viên thỉnh giảng tại Khoa Công nghệ thông tin của Trường đại học Ngoại ngữ - Tin học TP.HCM (HUFLIT) và đồng thời là kỹ sư Trí tuệ nhân tạo tại công ty TNHH Emage Development. Lĩnh vực nghiên cứu của ông liên quan đến Computer vision, Deep learning và Robot.