

# DSFD: Dual Shot Face Detector

Trung-Nam Bui-Huynh\*  
huynhtrungnam2001@gmail.com  
Ho Chi Minh University Of Science  
Ho Chi Minh, Viet Nam

Xuan-Nam Cao  
Ho Chi Minh University Of Science  
Ho Chi Minh, Viet Nam  
cxnam@fit.hcmus.edu.vn

Kim-Khanh Chung\*  
ckkhanh19@clc.fitus.edu.vn  
Ho Chi Minh University Of Science  
Ho Chi Minh, Viet Nam

Hai-Quan Vu  
Ho Chi Minh University Of Science  
Ho Chi Minh, Viet Nam  
vhquan@vnuhcm.edu.vn

Minh-Nhut Pham  
Ho Chi Minh University Of Science  
Ho Chi Minh, Viet Nam  
nhut.phamminh@gmail.com

## ABSTRACT

Trong paper này, tác giả đã đưa ra một mô hình nhận diện khuôn mặt người với các cải thiện chính trên 3 khía cạnh cốt lõi bao gồm việc cải thiện chất lượng của bản đồ đặc trưng (Feature Enhance), thiết kế và áp dụng hàm mắt mát (Progressive Anchor Loss) để tận dụng hiệu quả hơn các đặc trưng và cuối cùng cải thiện độ khớp của neo (Improved Anchor Matching) bằng cách tích hợp chiến lược gán anchor vào trong việc tăng cường dữ liệu. Và cũng chính vì các sự cải thiện này đều có liên quan tới thiết kế hai luồng, nên mô hình đã được đặt tên là Dual Shot Face Detector (DFSD [6]). Mô hình được thực hiện trên các tập dữ liệu WIDER FACE và FDDB để chứng minh độ hiệu quả của mình.

## CCS CONCEPTS

• Computing methodologies → Machine learning

## KEYWORDS

Nhận diện, 2 mô đun, học đặc trưng, CNN, khuôn mặt

## ACM Reference Format:

Trung-Nam Bui-Huynh, Kim-Khanh Chung, Hai-Quan Vu, Xuan-Nam Cao, and Minh-Nhut Pham. 2024. DSFD: Dual Shot Face Detector. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 INTRODUCTION

Bài toán nhận diện khuôn mặt (Face detection) là một bài toán phổ biến đã được nghiên cứu từ rất lâu. Cho đến nay, bài toán này vẫn còn được quan tâm rộng rãi và có nhiều đóng góp to lớn trong việc tăng cường khả năng hiểu của máy tính, đặc biệt là trong dịch bệnh Covid-19. Về không gian, bài toán nhận diện khuôn mặt phát triển từ nhận diện trên

\*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

Conference acronym 'XX, April 7th, 2024, District 5, HCM

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00  
<https://doi.org/XXXXXXX.XXXXXXX>

ảnh hai chiều sang nhận diện trên chuỗi hình ảnh (video) ba chiều. Tuy nhiên, vẫn còn nhiều thách thức cần phải giải quyết trong bài toán này, đồng thời cũng nâng cao độ chính xác của các phương pháp nhận diện gương mặt hơn. Cụ thể và cơ bản nhất là nghiên cứu cải thiện nhận diện khuôn mặt trên hình ảnh để làm nền tảng cho những nghiên cứu sau này. Từ các thuật toán thủ công bằng tay như Adaboost cho đến các mạng lưới học sâu như mạng tích chập (convolutional neural network), tất cả đều cho thấy các kết quả và thành tựu ấn tượng, tuy nhiên các vấn đề về tỉ lệ đa súng, ánh sáng, ngoại hình,... vẫn là một thách thức cho mô hình đến ngày nay.

Những nghiên cứu trước có thể chia làm 2 loại chính: đầu tiên là mạng đề xuất khu vực (Region Proposal Network - RPS), đang được áp dụng trong các mô hình như Faster RCNN [10] và sử dụng nhận diện hai giai đoạn (two stage detection); thứ hai là nhận diện một giai đoạn (Single Stage Detector - SSD), phương pháp này trực tiếp dự đoán các bounding box và mức độ chính xác. Nhận diện khuôn mặt một giai đoạn đã thu hút được nhiều sự chú ý hơn gần đây bởi vì hiệu quả suy luận cao hơn và triển khai hệ thống đơn giản. Mặc dù là như vậy, những vấn đề vẫn còn tồn đọng trong các khía cạnh.

### 1.1 Học đặc trưng (feature learning)

Phân tích xuất đặc trưng rất quan trọng trong nhận diện khuôn mặt vì nó chứa thông tin để có được kết quả, tuy nhiên các phương pháp chỉ tổng hợp bản đồ đặc trưng giữa các tầng bằng đặc trưng giữa các lớp từ thấp tới cao và không quan tâm tới thông tin của lớp hiện tại cũng như quan hệ ngữ cảnh giữa các anchor bị loại bỏ.

### 1.2 Thiết kế hàm mắt mát (loss design)

Sự hội tụ của hàm mắt mát được dùng trong nhận diện vật thể bao gồm hàm mắt mát hồi quy cho vùng mặt và hàm mắt mát phân loại cho việc xác định liệu khuôn mặt đã được nhận diện hay chưa. Để giải quyết xa hơn các vấn đề về mắt cân bằng lớp, các hàm mắt mát như hàm mắt mát tiêu điểm (Focal loss) hoặc để sử dụng tất cả các đặc trưng gốc và cải thiện thì có hàm mắt mát thứ bậc (Hierarchical Loss). Tuy nhiên những hàm mắt mát này không quan tâm đến khả năng học một cách tiến bộ của các bản đồ đặc trưng trong cả các lớp và các phần của mô hình.



Hình 1: Mô hình ổn định với nhiều biến thể khác nhau về tỷ lệ, mờ, chiếu sáng, tư thế, che khuất, phản chiếu và trang điểm.

### 1.3 Khớp Neo (anchor matching)

Tập các anchor được xác định trước cho mỗi bản đồ đặc trưng được tạo ra bằng cách xếp các hộp với các tỉ lệ và khía cạnh khác nhau trên hình. Dã có vài cách nghiên cứu phân tích một chuỗi các tỉ lệ anchor hợp lý và chiến lược bù anchor để tăng các anchor dương (positive anchor). Tuy nhiên những chiến lược này bỏ qua lấy mẫu ngẫu nhiên trong tăng cường dữ liệu, và việc này vẫn đang gây ra sự mất cân bằng giữa anchor dương và âm.

Trong nghiên cứu ngày, tác giả đưa ra 3 kỹ thuật để giải quyết 3 vấn đề ở trên. Đầu tiên là giới thiệu mô đun cải thiện đặc trưng (Feature Enhance Module - FEM) để cải thiện sự phân biệt và ổn định của các đặc trưng. Thứ hai, tác giả thiết kế ra hàm mất mát cấp tiến cho các anchor (Progressive Anchor Loss - PAL), sử dụng kích thước của các progressive anchor cho không chỉ các lớp khác nhau mà còn là các phần khác nhau. Thứ ba là tác giả đưa ra cách tích hợp chiến lược phân vùng và tăng cường dữ liệu dựa trên anchor để có khớp tốt hơn giữa anchor và ground truth face (Improved Anchor Matching - IAM). Hình 1 cho thấy hiệu quả of DFSD trên đa dạng các điều kiện, đặc biệt là trên khuôn mặt rất nhỏ và khuôn mặt bị che nhiều. Tóm lại, đóng góp chính của nghiên cứu này bao gồm:

- Feature Enhance Module để tận dụng thông tin từ các lớp khác nhau và từ đó đạt được các đặc trưng phân biệt và ổn định
- Những giám sát phụ trợ giới thiệu trong các lớp trước thông qua một tập của các anchor nhỏ hơn để tận dụng tốt hơn các đặc trưng
- Một chiến lược cải thiện anchor matching để khớp các anchor và ground truth face nhiều nhất có thể để cung cấp khởi tạo tốt hơn của regressor
- Thí nghiệm toàn diện được tiến hành trên các tập dữ liệu nổi tiếng DFFB và WIDER FACE để trình bày tính ưu việt của mô hình DSFD của tác giả khi được so sánh với các phương pháp tốt khác

## 2 RELATED WORK

Để giải quyết những thách thức được nêu trên, những công trình nghiên cứu nổi bật về bài toán Nhận dạng khuôn mặt

có thể chia thành 3 nhóm dựa theo 3 khía cạnh chính còn tồn đọng:

Nhóm 1 tập trung vào các cải tiến trong quá trình học đặc trưng. Khi chưa có sự phát triển của phương pháp học sâu. Những công trình nghiên cứu từ rất sớm của bài toán Nhận dạng khuôn mặt tập trung vào các phương pháp trích xuất đặc trưng thủ công (hand-draft) như là đặc trưng Harr-like [12], thiết lập điểm kiểm soát (control point set) [1] hoặc biểu đồ định hướng cạnh (edge orientation histograms) [4]. Tuy phương pháp này cho ra kết quả khá tốt nhưng lại không hiệu quả đối với các trường hợp ảnh bị mờ, gương mặt bị che khuất hoặc góc mặt không trực diện. Với sự đổi mới và phát triển của phương pháp học sâu, trích xuất đặc trưng thủ công đã bị thay thế bởi mạng Nơ-ron (Convolutional Neural Networks - CNN). Ví dụ như là Face Attention Network [13], PyramidBox [11], FPN [5]. Tuy nhiên, các phương pháp này không xem xét thông tin lớp hiện tại. Khác với các phương pháp trên bỏ qua mối quan hệ ngữ cảnh giữa các neo, chúng tôi đề xuất một mô-đun nâng cao đặc trưng kết hợp các lớp Convolution giãn nở đa cấp bậc để nâng cao ngữ nghĩa của các đặc trưng.

Nhóm 2 tập trung vào việc sáng tạo ra hàm mất mát mới. Bình thường, đối tượng mất mát trong quá trình nhận dạng là tổng trọng số của mất mát trong quá trình phân lớp (ví dụ: mất mát softmax) và mất mát trong hộp hồi quy (ví dụ: mất L2). Wang và cộng sự. [14] thiết kế RepLoss để phát hiện người di bộ, giúp cải thiện hiệu suất trong các tình huống tắc nghẽn. FANet [16] tạo ra một kim tự tháp đặc trưng phân cấp độ và thể hiện sự mất mát có các cấp độ khác nhau trong kiến trúc của họ. Tuy nhiên, các neo được sử dụng trong FANet được giữ nguyên kích thước ở các giai đoạn khác nhau. Trong công việc này, chúng tôi chọn các kích thước neo khác nhau một cách thích ứng trong các giai đoạn khác nhau để tạo điều kiện thuận lợi cho các đặc trưng.

Nhóm 3 tập trung vào việc cải tiến quá trình khớp neo. Để làm cho mô hình mạnh mẽ hơn, hầu hết các phương pháp phát hiện [15, 17] đều thực hiện tăng cường dữ liệu bằng cách tiền xử lý, chẳng hạn như biến dạng màu sắc, lật ngang, cắt ngẫu nhiên và huấn luyện đa tỷ lệ. Tuy nhiên, các phương pháp này bỏ qua việc lấy mẫu ngẫu nhiên trong quá trình tăng cường dữ liệu, trong khi phương pháp của chúng tôi kết

hợp phép gán neo để cho việc khởi tạo dữ liệu tốt hơn trong việc so khớp neo.

### 3 DUAL SHOT FACE DETECTOR

Mô hình chính được sử dụng trong mô hình DSFD là VGG16, tác giả đã áp dụng các chỉnh sửa và thêm các cấu trúc phụ trợ nhằm cải thiện chất lượng của mô hình và từ đó đề xuất mạng Dual Shot Face Detector, mạng đề xuất cho thấy độ hiệu quả trên nhiều biến thể, và đặc biệt là trong điều kiện các khuôn mặt nhỏ và bị che khuất nhiều

#### 3.1 Tổng quan về VGG16

Về cơ bản, VGG16 là một mạng lưới tích chập (convolution neural network) được sử dụng để phân loại hình ảnh, điểm chính của mô hình là có 16 lớp, mỗi lớp sẽ giúp xử lý thông tin hình ảnh tăng dần và cải thiện độ chính xác cho dự đoán của mô hình. Mô hình có tính nhất quán khi tuân theo sự sắp xếp các lớp tích chập và max pool trong toàn bộ kiến trúc. Ở cuối cùng, mô hình có các lớp kết nối đầy (fully connected layers), và theo sau là hàm softmax cho kết quả đầu ra

#### 3.2 Luồng chính của DSFD

Bằng việc sử dụng VGG16 như là backbone mở rộng cho kiến trúc của mình, tác giả đã chỉnh sửa và từ đó tạo ra mô hình DSFD với 2 phần chính:

- Phần đầu tiên (first shot) dùng để tạo ra các bản đồ đặc trưng (feature map) như VGG16. Tuy nhiên tác giả chỉ lựa chọn một số lớp tích chập và kết nối đầy (conv3\_3, conv4\_3, conv5\_3, conv\_fc7, conv6\_2, conv7\_2) để tạo thành phần nhận diện đầu tiên cho mô hình (first shot detection layers), kết quả là có 6 feature map được tạo ra lần lượt là  $of_1, of_2, \dots, of_6$
- Phần thứ 2 (second shot) có tên gọi là mô đun cải thiện đặc trưng (Feature Enhance Module), dùng để cải thiện chất lượng của các feature map từ first shot của mô hình, các feature map được cải thiện có cùng kích thước với feature map gốc và được sử dụng theo kiểu Single Shot Detection để tạo thành các phần thứ 2 của mô hình (second shot detection layers), có tên lần lượt là  $ef_1, ef_2, \dots, ef_6$

#### 3.3 Feature Enhance Module

FEM được dùng để cải thiện các feature map ban đầu, hình 3 đại diện cho FEM, được lấy cảm hứng bởi FPN [7] và RFB [8], đầu tiên tác giả sử dụng bộ lọc  $1 \times 1$  để chuẩn hóa feature map hiện tại và của lớp trên, sau đó tác giả tăng chiều của feature map lớp trên lên để có kích thước bằng với feature map gốc. Tiếp đó tác giả thực hiện tích từng phần tử (element-wise product) giữa feature map hiện tại đã chuẩn hóa và feature map của lớp trên đã được tăng chiều dữ liệu. Cuối cùng tác giả chia kết quả có được thành 3 phần, được theo sau bởi 3 mạng con có số lớp tích chập giãn nở khác nhau, cuối cùng là quá trình nối tất cả lại để tạo thành feature map đã được cải thiện.

Để cải thiện một điểm (cell) ban đầu  $oc_{(i,j,l)}$ , FEM sử dụng thông tin từ các miền khác nhau bao gồm cell gốc của lớp trên  $oc_{(i,j,l)}$  và các cell xung quanh của lớp hiện tại:  $nc_{(i-e,j-e,l)}, nc_{(i,j-e,l)}, \dots, nc_{(i,j+e,l)}, nc_{(i+e,j+e,l)}$ . Cách cải thiện cell  $ec_{(i,j,l)}$  có thể được định nghĩa toán học như sau:

$$ec_{(i,j,l)} = f_{concat}(f_{dilation}(nc_{(i,j,l)})) \quad (1)$$

Có  $c_{i,j,l}$  là cell ở vị trí  $i, j$  trong feature map ở lớp thứ  $l$ ,  $f$  đại diện cho tập hợp các quá trình giãn nở tích chập, tích từng phần tử, tăng chiều hoặc phép nối.

#### 3.4 Progressive Anchor Loss

Khác với hàm mất mát truyền thống trong nhận diện, tác giả đã thiết kế kích thước anchor cải tiến cho không chỉ các lớp khác nhau, mà cho cả các phần (shot) khác nhau trong mô hình. Đặc trưng ở các lớp đầu thì thích hợp hơn cho các khuôn mặt nhỏ, tác giả đã gán kích thước nhỏ hơn cho các anchor ở first shot và sử dụng kích thước lớn hơn cho anchor ở second shot của mô hình. Hàm mất mát của dựa trên anchor cho nhiều tác vụ của second shot được định nghĩa như sau:

$$\begin{aligned} L_{SSL}(p_i, p_i^*, t_i, g_i, a_i) = & \frac{1}{N_{conf}} \sum_i L_{conf}(p_i, p_i^*) \\ & + \frac{\beta}{N_{loc}} \sum_i p_i^* L_{loc}(t_i, g_i, a_i) \end{aligned} \quad (2)$$

Có  $N_{conf}$  và  $N_{loc}$  chỉ ra số lượng anchor dương và âm, và số lượng anchor dương,  $L_{conf}$  là hàm mất mát softmax cho 2 phân lớp (mặt và nền), và  $L_{loc}$  là độ mượt của hàm mất mát  $L1$  giữa tham số hóa của box dự đoán  $t_i$  và box nhãn thực  $g_i$  khi sử dụng acnchor  $a_i$ . Khi  $p_i^* = 1$  ( $p_i^* = 0, 1$ ), anchor  $a_i$  là dương và mất mát cục bộ được kích hoạt.  $B$  là tham số để cân bằng hiệu quả của cả 2. So sánh với feature map đã cải thiện trong cùng lớp, feature map ban đầu có ít thông tin ngữ nghĩa hơn để phân lớp nhưng có thông tin cục bộ resolution cao để nhận dạng. Chính vì thế tác giả tin rằng feature map ban đầu có để nhận diện và phân loại các khuôn mặt nhỏ hơn. Hàm mất mát đa tác vụ cho first shot với một tập các anchor nhỏ hơn như sau:

$$\begin{aligned} L_{FSL}(p_i, p_i^*, t_i, g_i, sa_i) = & \frac{1}{N_{conf}} \sum_i L_{conf}(p_i, p_i^*) \\ & + \frac{\beta}{N_{loc}} \sum_i p_i^* L_{loc}(t_i, g_i, sa_i) \end{aligned} \quad (3)$$

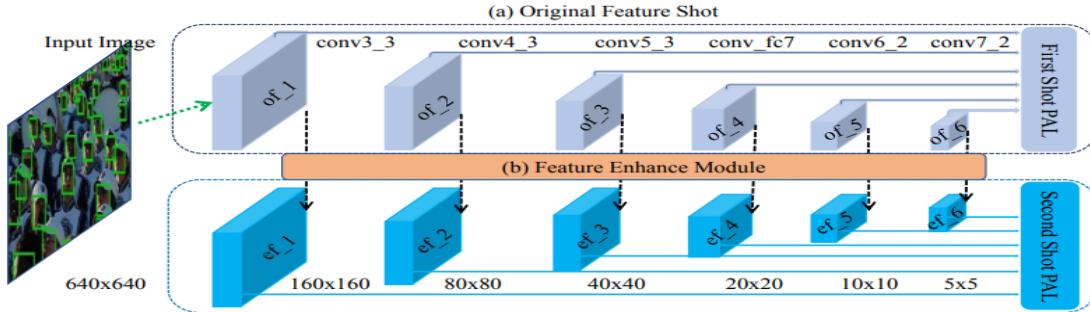
Có  $sa$  chỉ ra các neo nhỏ hơn trong các lớp phần đầu và cả 2 hàm mất mát có thể được tổng trọng số thành một hàm progressive anchor loss như sau:

$$L_{PAL} = L_{FSL}(sa) + \lambda L_{SSL}(a) \quad (4)$$

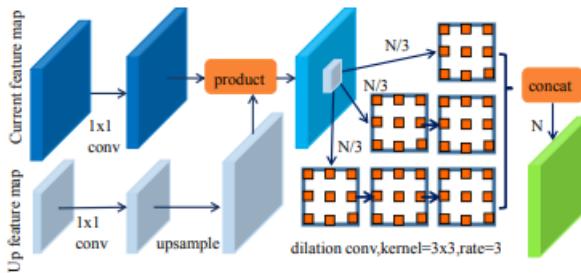
Chú ý rằng kích thước của neo trong first shot bằng một nửa của second shot, và lambda dùng để cân bằng giữa 2 hàm loss.

#### 3.5 Cải thiện so khớp neo

Phương pháp so khớp neo (anchor matching) hiện tại là 2 chiều giữa neo (anchor) và khuôn mặt chân thực, chính vì vậy mà thiết kế anchor và lấy mẫu khuôn mặt trong suốt quá trình tăng cường (augmentation) có tính cộng tác với nhau để khớp anchor và khuôn mặt nhất có thể, từ đó có được sự khôi



Hình 2: DSFD framework



Hình 3: Feature Enhance Module

tạo tốt hơn. Mục tiêu IAM của tác giả là giải quyết sự mâu thuẫn giữa tỉ lệ anchor rời rạc và và tỉ lệ khuôn mặt liên tục, nơi mà khuôn mặt được tăng cường bởi  $S_{input} * S_{face} / S_{anchor}$  ( $S$  là kích thước không gian) với xác suất 40% để tăng các positive anchor, ổn định hóa quá trình huấn luyện và từ đó cải thiện kết quả. Tác giả đặt tỉ lệ 1.5:1 dựa trên thống kê tỉ lệ khuôn mặt. Kích thước anchor cho feature map gốc được giảm một nửa so với feature map cải thiện. Thêm vào đó, với xác suất 2/5, tác giả sử dụng lấy mẫu dựa trên anchor như lấy mẫu thông tin anchor trong PyramidBox, nghĩa là lựa chọn ngẫu nhiên mặt trong hình, cắt ảnh con chứa mặt, và đặt tỉ lệ kích thước giữa ảnh con và ảnh được chọn là 640/rand (16, 32, 64, 128, 256, 512). Cho phần còn lại tỉ lệ 3/5, tác giả áp dụng tăng cường dữ liệu tương tự như SSD [9]. Để cải thiện tỉ lệ recall của khuôn mặt và chắc rằng khả năng phân loại anchor một cách đồng thời, tác giả đặt Intersection-over-Union (IoU) thresh hold 0.4 để gán anchor cho khuôn mặt có nhãn thực.

## 4 EXPERIMENTS

### 4.1 Chi tiết triển khai

Dầu tiên, chúng tôi trình bày chi tiết về việc triển khai mạng của mình. Các mạng backbone được khởi tạo bởi VGG/ResNet đã được huấn luyện trước trên ImageNet. Tất cả các tham số của lớp tích chập mới được thêm vào đều được khởi tạo bằng phương thức 'xavier'. Chúng tôi sử dụng:

- SGD: 0.9 momentum, giảm 0,0005 trọng lượng (weight)

- Batch size: 16
- Learning rate:
  - 40K bước đầu tiên:  $10^{-3}$
  - 10K bước tiếp theo:  $10^{-4}$
  - 10K bước còn lại:  $10^{-5}$

Trong quá trình suy luận, các đầu ra của lần bắn đầu tiên được bỏ qua và lần bắn thứ hai dự đoán top 5k các phát hiện có độ tin cậy cao nhất. Phương pháp "non-maximum suppression" được áp dụng với độ chồng chéo jaccard là 0.3 để tạo ra top 750 hộp giới hạn có độ tin cậy cao nhất cho mỗi hình ảnh. Dối với 4 tọa độ hộp giới hạn, chúng tôi làm tròn xuống tọa độ góc trên bên trái và làm tròn lên chiều rộng và chiều cao để mở rộng hộp giới hạn phát hiện. Mã nguồn chính thức đã được công bố tại: <https://github.com/Tencent/FaceDetection-DSFD>, và một phần nhỏ mã nguồn mô tả lại cơ bản DSFD được thể hiện tại: [https://colab.research.google.com/drive/1C1TbMVundr0\\_VB5GScMTCAZiYHSRHLFh?usp=sharing](https://colab.research.google.com/drive/1C1TbMVundr0_VB5GScMTCAZiYHSRHLFh?usp=sharing).

### 4.2 Phân tích phương pháp DSFD

Trong phần này, chúng tôi tiến hành các thử nghiệm và nghiên cứu cắt bớt phần trên tập dữ liệu WIDER FACE để đánh giá hiệu quả của một số đóng góp dựa trên 3 tiêu chí mà chúng tôi đã đề xuất: bao gồm mô-đun tăng cường đặc trưng, mắt mặt neo luỹ tiến và so khớp neo được cải tiến. Để so sánh một cách công bằng, chúng tôi sử dụng cùng cấu hình tham số cho tất cả các thử nghiệm, ngoại trừ các thay đổi cụ thể cho các thành phần. Tất cả các mô hình được huấn luyện trên tập dữ liệu huấn luyện WIDER FACE và được đánh giá trên tập dữ liệu xác thực. Để hiểu rõ hơn về DSFD, chúng tôi lựa chọn các mô hình cơ sở khác nhau để giảm bớt từng thành phần và xem chúng làm thế nào ảnh hưởng đến hiệu suất cuối cùng.

Mô hình	Dẽ	Trung bình	Khó
FSSD+VGG16	92.6%	90.2%	79.1%
FSSD+VGG16+FEM	93.0%	91.4%	84.6%

Bảng 1: Sự ảnh hưởng của Mô-đun tăng cường đặc trưng trên hiệu suất AP.)

- Mô-đun tăng cường đặc trưng: Kết quả Bảng 1 cho thấy mô-đun tăng cường đặc trưng của chúng tôi có thể cải thiện hiệu suất của FSSD dựa trên VGG16 từ từ 92.6%, 90.2%, 79.1% lên 93.0%, 91.4%, 84.6%.
- Mắt mát neo luỹ tiến: Kết quả Bảng 2 cho thấy mắt mát neo luỹ tiến của chúng tôi có thể cải thiện FSSD dựa trên Res50 bằng cách sử dụng FEM từ 95.0%, 94.1%, 88.0% đến 95.3%, 94.4%, 88.6%.
- Cải tiến so khớp neo: Kết quả Bảng 3 cho thấy rằng phương pháp cải tiến so khớp neo của chúng tôi có thể cải thiện FSSD dựa trên Res101 bằng cách sử dụng FEM từ 95.8%, 95.1%, 89.7% đến 96.1%, 95.2%, 90.0%. Cuối cùng, chúng tôi cũng có thể cải thiện DSFD của mình lên 96.6%, 95.7%, 90.4% với ResNet152 làm backbone.

Mô hình	Dẽ	Trung bình	Khó
FSSD+RES50	93.7%	92.2%	81.8%
FSSD+RES50+FEM	95.0%	94.1%	88.0%
FSSD+RES50+FEM+PAL	95.3%	94.4%	88.6%

Bảng 2: Sự ảnh hưởng của Mắt mát neo luỹ tiến trên hiệu suất AP.)

#### 4.3 So sánh với các phương pháp tiên tiến nhất cùng thời điểm

Chúng tôi đánh giá DSFD được đề xuất dựa trên hai điểm chuẩn nhận diện khuôn mặt phổ biến, bao gồm WIDER FACE [15] và Face Detection Data Set and Benchmark (FDDB) [3]. Mô hình của chúng tôi chỉ đào tạo trên tập huấn luyện WIDER FACE, sau đó được đánh giá trên cả hai điểm chuẩn mà không cần tinh chỉnh thêm. Chúng tôi cũng làm theo cách tương tự được sử dụng trong [13] để xây dựng kim tự tháp hình ảnh để thử nghiệm đa quy mô và sử dụng backbone mạnh hơn tương tự như [2].

Mô hình	Dẽ	Trung bình	Khó
FSSD+RES101	95.1%	93.6%	83.7%
FSSD+RES101+FEM	95.8%	95.1%	89.7%
FSSD+RES101+FEM+IAM	96.1%	95.2%	90.0%
FSSD+RES101+FEM+IAM+PAL	96.3%	95.4%	90.1%
FSSD+RES152+FEM+IAM+PAL	96.6%	95.7%	90.4%
FSSD+RES152+FEM+IAM+PAL+LargeBS	96.4%	95.7%	91.2%

Bảng 3: Sự ảnh hưởng của Cải tiến so khớp neo trên hiệu suất AP.)

Tập dữ liệu WIDER FACE: Từ Hình 4, DSFD của chúng tôi đạt được hiệu suất tốt nhất trong số tất cả các mô hình nhận diện khuôn mặt có kết quả nổi bật nhất ở thời điểm hiện tại dựa trên độ chính xác trung bình (AP) trên ba tập

hợp con: 96,6% (Dẽ), 95,7% (Trung bình) và 90,4% (Khó) trên bộ validation và 96,0% (Dẽ), 95,3% (Trung bình) và 90,0% (Khó) trên bộ testing.

Hình 5 cho thấy nhiều ví dụ hơn để chứng minh tác động của DSFD đối với việc xử lý các khuôn mặt ở những thách thức khác nhau. Trong đó, các bounding box màu xanh biểu thị độ tin cậy của máy dò là trên 0,8.

Tập dữ liệu FDDB: Sau khi thêm các chú thích bổ sung vào các khuôn mặt không được gắn nhãn [17], các kết quả tích cực giả trong mô hình của chúng tôi có thể giảm hơn nữa và vượt trội hơn tất cả các phương pháp khác.

## 5 CONCLUSIONS

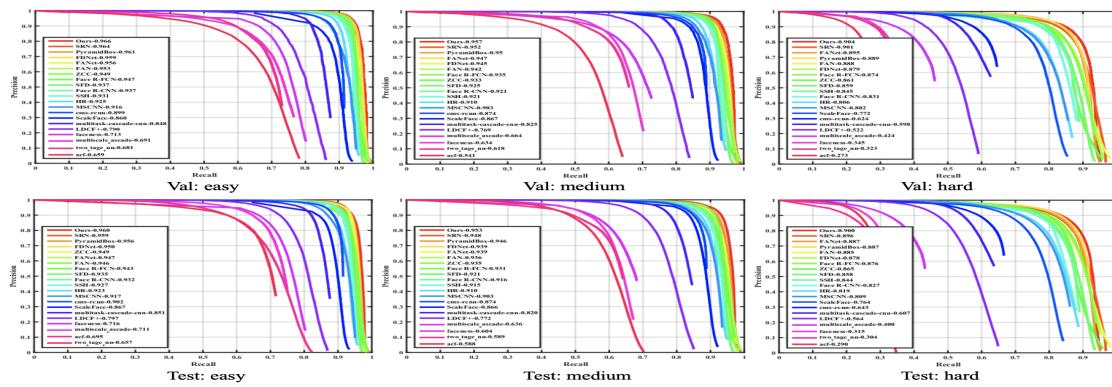
Trong nghiên cứu này, tác giả đã đưa ra mô hình DSFD với nhiều cải tiến. Đầu tiên là việc có thêm một mô-đun để cải thiện, tận dụng các feature map, từ đó nâng cao sự phân biệt và ổn định của mô hình. Thứ hai là sự áp dụng kích thước anchor nhỏ hơn cho mô-đun đầu của mô hình, giúp cho việc tận dụng các đặc trưng hiệu quả hơn. Cuối cùng là đưa ra cải thiện về cách khớp anchor và nhãn thực nhất có thể, từ đó làm cho sự khởi tạo tốt hơn. Tác giả thực hiện thí nghiệm trên các tập dữ liệu nổi tiếng, FDDB và WIDER FACE, để trình bày tính ưu việt của DSFD và so sánh với các mô hình nổi bật.

## ACKNOWLEDGMENTS

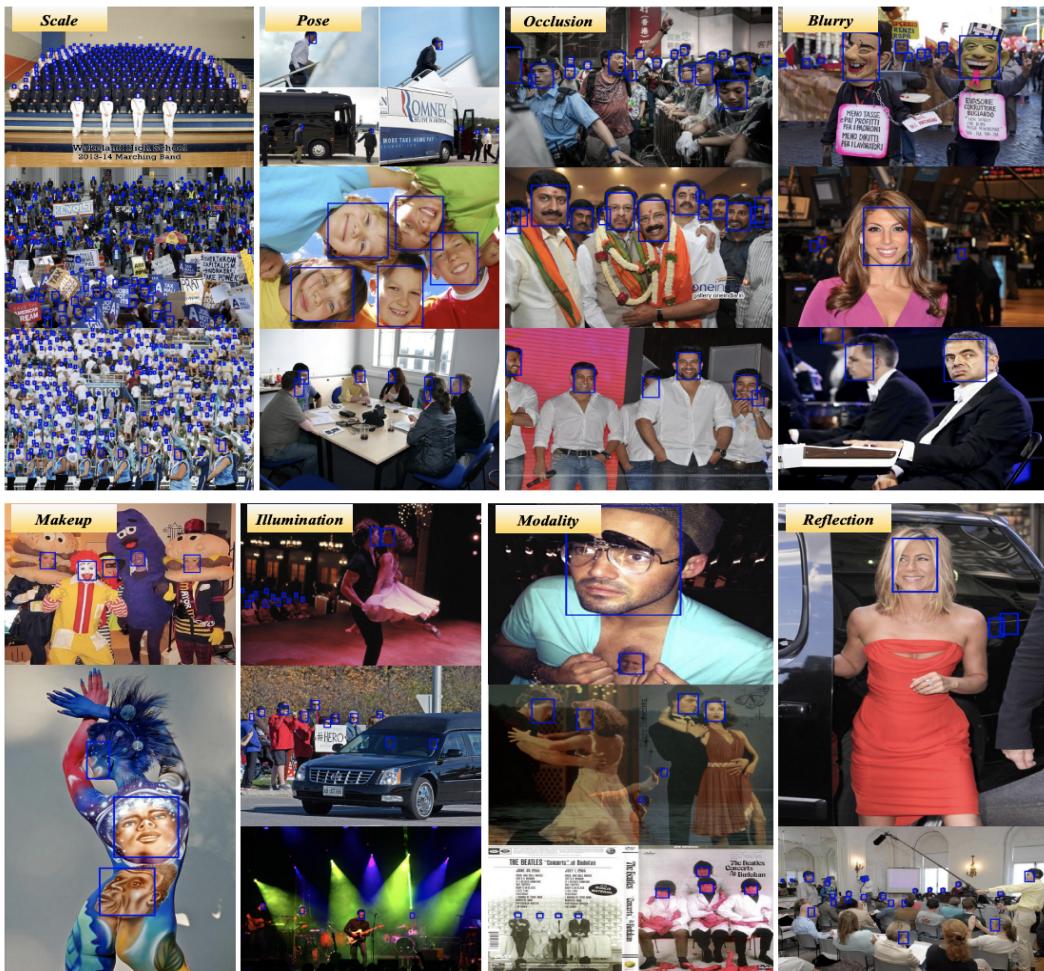
Không có acknowledgement.

## TÀI LIỆU

- [1] Yotam Abramson, Bruno Steux, and Hicham Ghorayeb. 2007. Yet Even Faster (YEF) real-time object detection. IJISTA 2 (01 2007), 102–112. <https://doi.org/10.1504/IJISTA.2007.012476>
- [2] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Li, and Xudong Zou. 2019. Selective Refinement Network for High Performance Face Detection.
- [3] Vudit Jain and Erik Learned-Miller. 2010. FDDB: A Benchmark for Face Detection in Unconstrained Settings. (12 2010).
- [4] K. Levi and Y. Weiss. 2004. Learning Object Detection from a Small Number of Examples: The Importance of Good Features. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2. IEEE Computer Society, Los Alamitos, CA, USA, 53–60. <https://doi.org/10.1109/CVPR.2004.145>
- [5] Jian Li, Jianjun Qian, and Jian Yang. 2017. Object detection via feature fusion based single network. In 2017 IEEE International Conference on Image Processing (ICIP). 3390–3394. <https://doi.org/10.1109/ICIP.2017.8296911>
- [6] Jian Li, Yabiao Wang, Chang'an Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. 2019. DSFD: dual shot face detector. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 5060–5069.
- [7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2117–2125.
- [8] Songtao Liu, Di Huang, et al. 2018. Receptive field block net for accurate and fast object detection. In Proceedings of the European conference on computer vision (ECCV). 385–400.
- [9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 21–37.



Hình 4: Các đường cong precision-recall trên tập hợp con validation và testing của WIDER FACE.



Hình 5: Minh họa DSFD của chúng tôi với nhiều thách thức lớn khác nhau như: tỷ lệ, tư thế, sự che khuất, mờ, trang điểm, chiếu sáng, thay đổi dáng và phản chiếu ánh sáng. Các bounding box màu xanh biểu thị độ tin cậy của máy dò trên 0,8.

- [10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
- [11] Xu Tang, Daniel K. Du, Zeqiang He, and Jingtuo Liu. 2018. PyramidBox: A Context-assisted Single Shot Face Detector. *CoRR abs/1803.07737* (2018). arXiv:1803.07737 <http://arxiv.org/abs/1803.07737>
- [12] Paul Viola and Michael Jones. 2004. Robust Real-Time Face Detection. *International Journal of Computer Vision* 57 (05 2004), 137–154. <https://doi.org/10.1023/B:VISI.0000013087.49260.fb>
- [13] Jianfeng Wang, Ye Yuan, and Gang Yu. 2017. Face Attention Network: An Effective Face Detector for the Occluded Faces. *CoRR abs/1711.07246* (2017). arXiv:1711.07246 <http://arxiv.org/abs/1711.07246>
- [14] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. 2018. Repulsion Loss: Detecting Pedestrians in a Crowd. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7774–7783. <https://doi.org/10.1109/CVPR.2018.00811>
- [15] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaou Tang. 2015. WIDER FACE: A Face Detection Benchmark. *CoRR abs/1511.06523* (2015). arXiv:1511.06523 <http://arxiv.org/abs/1511.06523>
- [16] Jialiang Zhang, Xiongwei Wu, Jianke Zhu, and Steven C. H. Hoi. 2017. Feature Agglomeration Networks for Single Stage Face Detection. *CoRR abs/1712.00721* (2017). arXiv:1712.00721 <http://arxiv.org/abs/1712.00721>
- [17] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. 2017. S<sup>3</sup>FD: Single Shot Scale-invariant Face Detector. *CoRR abs/1708.05237* (2017). arXiv:1708.05237 <http://arxiv.org/abs/1708.05237>

Received 20 February 2024; revised 12 March 2024; accepted 5 June 2024