A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis

Tóm tắt báo tuần 3-4

Ngọc Thuận - IPSAL LAB

January 2023

Tóm tắt nội dung

Bài báo này đưa ra một *Dataset* (*DeepFish*) gần với thực tế nhất với nhiều đặc điểm như: kích thước lớn, chất lượng hình ảnh tốt, được thu thập khách quan, phục vụ đa chức năng với mục tiêu phát triển các thuật toán *Computer Vision* và *DeepLearning*... Ứng dụng ra thực tế cho nghiên cứu, nuôi trồng và phát triển ngành thủy sản.

1 Mục đích chính

Có thể tóm tắt mục đích chính của bài báo này như sau:

1. Đưa ra một bộ dữ liệu chuẩn mô tả được độ phức tạp và đa dạng của môi trường tự nhiên dưới nước so với các *datasets* trước đó!
2. Cung cấp các nhãn bổ sung (*additional labels*) để phân tích toàn diện hơn về các môi trường dưới nước.
3. Cho thấy tầm quan trọng của việc tiền huấn luyện (*pretrained*) để đạt được các kết quả tốt.
4. Cung cấp các kết quả để đánh giá các phương pháp mới.

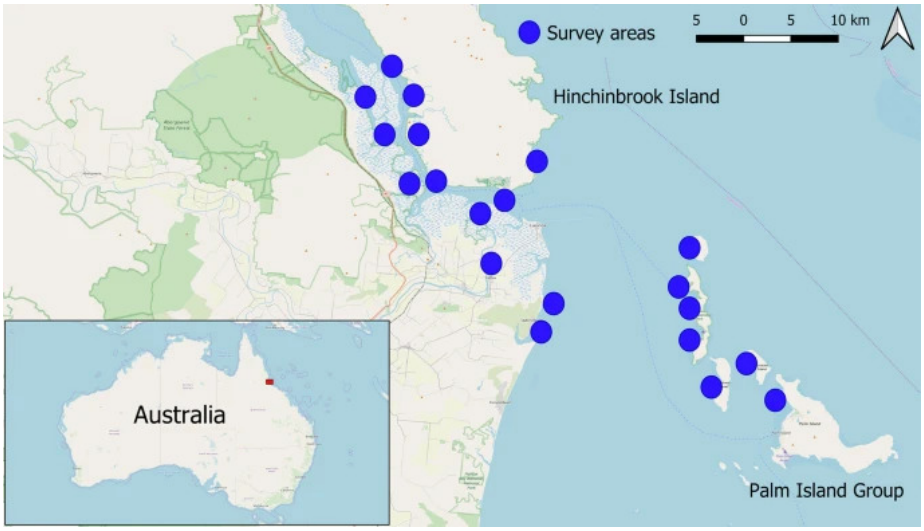
2 Cụ thể

Ta sẽ lướt qua những nét nổi bật của bộ dữ liệu này.

2.1 Dataset

- Bắt đầu từ công trình của Bradley và các đồng nghiệp (40 nghìn ảnh).
- Mục đích ban đầu không để phục vụ cho các tác vụ Học máy (*Machine Learning*). Nhưng đã được chỉnh sửa thành một cơ sở dữ liệu toàn diện hơn cho lĩnh vực này. Với mong muốn sẽ kích hoạt được các thuật toán mới!

2.2 Data collection



Hình 1: Locations where the DeepFish images were acquired. Most of DeepFish has been acquired from the Hinchinbrook/Palm Islands region in North Eastern Australia, the rest from Western Australia (not shown in the map). (The map was created using QGIS version 3.8, which is available at https://qgis.org).

- Được thu thập từ 20 môi trường sống khác nhau từ các vùng viên biển xa xôi của vùng nhiệt đới Australia. Hình 1
- Quá trình thu thập khách quan khi *camera* được gắn vào mạn tàu sau đó được hạ xuống và giữ khoảng cách $100m$ ghi lại các quần xã tự nhiên dưới nước. Điều này khách quan và hiệu quả hơn trong nghiên cứu và đánh giá so với các phương pháp thu thập dữ liệu khác (ví dụ như lặn, ...)

2.3 Additional annotations

Ban đầu các nhãn gốc chỉ phù hợp cho các tác vụ phân loại (*classification*) giữa tiền cảnh (*foreground*) và hậu cảnh (*background*), có cá và không có cá (bất kể số lượng). Và việc này không cho phép phân tích loại tiết hơn về môi trường sống. Các chú thích bổ sung ra đời để khắc phục hạn chế này:

- Chú thích cấp điểm (*Point-level annotations*) mục tiêu là giúp mô hình có thể học để đếm cá. Và ứng dụng nó trong việc tự động giám sát quần thể cá (*automatically monitor fish population*), tránh nguy cơ đánh bắt quá mức (*overfishing*)
- Chú thích cho mỗi *Pixel* (*Per-pixel annotations*) mục tiêu là huấn luyện và đánh giá mô hình để phân đoạn cá trên các hình ảnh. Việc này sẽ hữu

ích trong việc ước lượng kích thước, hình dáng, và trọng lượng của cá, và sẽ hữu ích trong việc đánh cá thương mại (*commercial trawling*)

Để ý rằng riêng môi trường *thảm tảo thưa* (*Sparse algal bed*) không có cả hai loại chú thích trên. Lí do là không thể phân biệt rõ ràng giữa cá nhỏ và đá, cũng như một số yếu tố khác.

2.4 Dataset splits

Chia tập dữ liệu gốc thành các tập con cho các tác vụ của Thị giác máy tính: *FishClf - Classification*, *FishLoc - Counting and Locaization*, *FishSeg - Segmentation*
Mỗi tập dữ liệu con có chia hình ảnh đã được chú thích thành các tập: *training*, *validation* và *test*. Điểm khác là dữ liệu được chia dựa trên môi trường và số lượng quần thể của cá. Trong đó số luôn đảm bảo số ảnh hậu cảnh và tiền cảnh là như nhau, và 50% cho việc huấn luyện, 30% cho việc test và 20% cho việc xác thực.
Kết quả cho ta:
(*training, validation and test*)

- 19,3883, 7,953, 11,930 cho *FishClf*
- 1,600, 640, 960 cho *FishLoc*
- 310, 124, 186 cho *FishSeg*

2.5 Comparison to other datasets

Hình 2: A comparison of fish datasets. (a) QUT1, (b) Fish4Knowledge8, (c) Rockfish35, and (d) our proposed dataset DeepFish. (a–c) Datasets are acquired from constrained environments, whereas DeepFish has more realistic and challenging environments.

Đặc điểm dữ liệu

- QUT: Đã được hậu chỉnh (*post-processed*) (nền trắng, chỉ cá!)
- Rockfish and Fish4Knowledge: Có môi trường đấy, nhưng cá lúc nào cũng ở trọng tâm ảnh

Còn *DeepFish*:

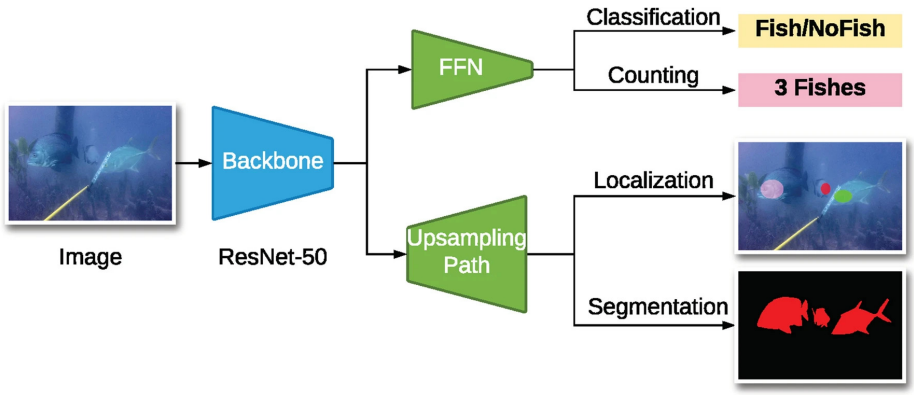
1. Bao quát về môi trường sống dưới nước của cá
2. Đa dạng nhiều môi trường sống với các đặc điểm dưới nước khác nhau
3. Ảnh được thu thập tại chỗ, khách quan, trực tiếp không qua chỉnh sửa

Tác vụ

QUT (*Clf*), Rockfish and Fish4Knowledge (*Detection*). Riêng *DeepFish* một mình 4 tác vụ: Clf, Cnt, Loc, Seg.

Nhìn chung, bộ dữ liệu *DeepFish* vượt xa các bộ dữ liệu trước đó về kích thước, mức độ phong phú của chú thích cũng như độ phức tạp và biến đổi của ảnh.

2.6 Methods and experiments



Hình 3: Deep learning methods. The architecture used for the four computer vision tasks of classification, counting, localization, and segmentation consists of two components. The first component is the ResNet-50 backbone which is used to extract features from the input image. The second component is either a feed-forward network that outputs a scalar value for the input image or an upsampling path that outputs a value for each pixel in the image.

Học chuyển tiếp (*Tranfer learning*). Sử dụng *ResNet-50*, *pre-training* bởi *ImageNet*. Khởi tạo ngẫu nhiên *trọng số* (*weights*) cho *ResNet-50* theo phương pháp của Xavier (*Xavier's method*).

2.7 Classification results

Phân loại: *Foreground* (tiền cảnh - có cá) và *Background* (hậu ảnh - không cá).
Accuracy:

$$ACC = \frac{TP + TN}{N}$$

với TP: *True Positives*, TN: *True negatives*, N: the totall number of images.
Hiểu cơ bản là số lượng phân loại đúng chia tổng số lượng ảnh.

1. Sử dụng *Feed-forward network* (FFN) để tính xác suất.
2. 3 layers và 2 class output layer
3. Loss function: *Cross entropy*
4. Optimizer: *Adam*

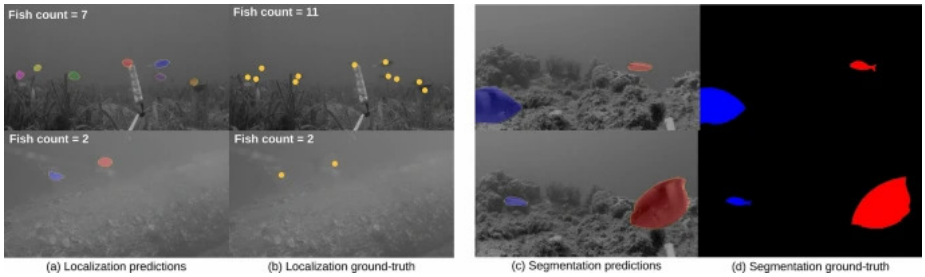
...

2.8 Couting results

Sử dụng *mean absolute error* để đánh giá mức độ hiệu quả của mô hình:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{C}_i - C_i|$$

trong đó C_i là số lượng cá có trong ảnh, $\hat{C}_i$ là số lượng cá đoán được từ mô hình.
Mô hình đếm khá giống mô hình phân loại, tuy nhiên khác biệt là với mô hình đếm chỉ có một nút đầu ra duy nhất!



Hình 4: Qualitative results on counting, localization, and segmentation. (a) Prediction results of the model trained with the LCFCN loss21. (b) Annotations that represent the (x, y) coordinates of each fish within the images. (c) Prediction results of the model trained with the focal loss25. (d) Annotations that represent the full segmentation masks of the corresponding fish.

2.9 Localization results

Tác vụ này xác định các vị trí có cá trong ảnh, và việc này khó hơn nhiều so với phân loại và đếm bởi đôi khi các con cá còn có thể chồng chéo lên nhau!
Cũng giống với *Couting task* tác vụ này cũng sử dụng dữ liệu trên *FishLoc dataset*, tuy nhiên *MAE* không thể nói lên được độ hiệu quả của mô hình, bởi đôi khi mô hình chọn sai đối tượng nhưng vẫn thỏa mãn. Vì vậy để đánh giá độ chính xác cho tác vụ định vị, ta sử dụng *Grid Average Mean Absolute Error (GAME)*

$$GAME = \sum_{i=1}^{4} GAME(L), GAME(L) = \frac{1}{N}\left(\sum_{l=1}^{4^L} |D_i^l - \hat{D}_i^l|\right)$$

trong đó D_i^l là số lượng chú thích cấp điểm trong vùng l, $\hat{D}_i^l$ là dự đoán của mô hình. *GAME(L)* chia bức ảnh thành một lưới của 4^L vùng không chồng lên nhau và sau đó tính tổng điểm *MAE* trên các vùng này.
Tác vụ xác định vị trí chứa *ResNet-50* và *unsampling path*. Và output là xác suất mỗi điểm để rơi vào lớp cá.
Hàm mất mát: *LCFCN*, *LCFCN* được huấn luyện sử dụng 4 chức năng khách quan: *image-level loss, point-level loss, split-level loss* và *false positive loss*.

- Image-level loss: xác định tất cả những *pixels* là nền của ảnh
- Point-level loss: xác định các tâm cá
- Split-level loss: dự đoán sao cho một vùng không có quá một chú thích cấp điểm
- False possitive loss: ngăn chặt mô hình dự đoán các đốm màu cho các khu vực không có chú thích cấp điểm.

2.10 Segmentation results

Ta đánh giá mô hình sử dụng chỉ số chuẩn *Jaccard*, là số được định nghĩa là tổng số *pixels* được dự đoán đúng, chia cho tổng số *pixels* được gắn với nhãn với lớp đó. Thường được biết là IoU (*intersection-over-union*), được cho bởi $\frac{TP}{TP+FP+FN}$
Trong tác vụ này, thay vì sử dụng *cross entropy*, ta sử dụng *focal loss function*, thứ sẽ thích hợp hơn khi mà số lượng pixel hậu cảnh nhiều hơn nhiều cho với pixel tiền cảnh. Còn lại thì sẽ giống với tác vụ xác định vị trí.

2.11 Pretrained Model

[ht!] Mô hình tiền huấn luyện thực sự cho kết quả tốt hơn trong tất cả các tác vụ trên.

	Classification	Counting	Localization		Segmentation
	Accuracy	MAE	MAE	GAME	mIoU
Random weights	0.65	1.30	1.22	1.30	0.49
Pretrained weights	0.99	0.38	0.21	1.22	0.93

Hình 5: Classification results were evaluated on the FishClf dataset, counting and localization on the FishLoc dataset, and segmentation on the FishSeg dataset.

3 Tái bút

Đây là lần đầu tôi tóm tắt một bài báo, kiến thức cũng nửa có nửa không so với những kiến thức yêu cầu của bài báo. Cho nên trong quá trình đọc, dịch, hiểu và tóm tắt. Nếu có sai sót nào, xin phép được bỏ qua ạ! Tôi sẽ cố gắng hơn trong các bài sau!

4 Bài báo gốc

Ta có thể truy cập vào bài báo gốc theo đường link sau: `https://www.nature.com/articles/s41598-020-71639-x`