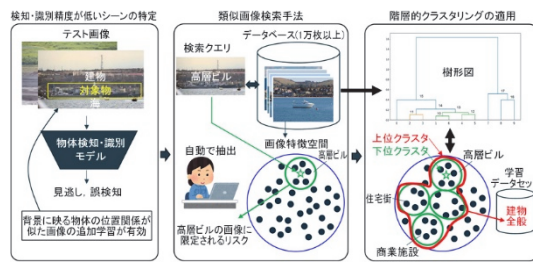


自律移動体の物体検知モデルの ロバスト性向上のための画像検索技術

Image Retrieval Techniques for Enhancing the Robustness of Object Detection Models
in Autonomous Vehicles



小林 周^{*1}
Amane Kobayashi

山野 優里子^{*2}
Yuriko Yamano

木内 祐介^{*3}
Yusuke Kinouchi

杉本 喜一^{*4}
Kiichi Sugimoto

洋上監視や沿岸警備を目的とした無人機システムなどではカメラ画像から警備の対象となる目標物や障害物を検知・識別する画像認識機能が不可欠である。高い画像認識性能を実現するためには検知・識別精度が低下するシーンの画像を学習させる必要があるが、学習用画像の取得作業及び特定シーンの画像の探索には多大な労力を要する。そこで、検知・識別精度が低い画像を入力として、ネットに大量に公開されている画像データベースから自動で特定シーンの画像を検索する手法を開発することで、当該作業の工数削減を実現した。当開発技術は、深層学習を用いた画像認識機能を有する製品に広く適用する予定である。

1. はじめに

深層学習を用いた画像認識技術の発展は目覚ましく、画像上の人物や船舶など特定の物体を検知・識別する手法の社会実装が盛んである⁽¹⁾。その一例として、三菱重工業株式会社(以下、当社)では、洋上監視や沿岸警備に活用する無人機システムにおいて、無人航空機(Uncrewed Aerial Vehicle 以下、UAV)に搭載したカメラから不審船を検知する機能に物体検知・識別手法を適用している。本手法では、事前に船舶などの検知・識別対象物の画像を学習しており、運用を想定する洋上・沿岸部の様々な場所を背景とした学習用画像を大量に用意する必要がある。学習した物体検知・識別モデルを用いてテストし、対象物を見逃す、あるいは、対象物でないものを誤って検知するといった精度が低い画像に対しては、当該シーンの画像の追加学習を実施する。精度が低いシーンがなくなるまで、追加学習を繰り返すことで、性能を改善していく必要がある。

追加学習用の画像は、検知・識別精度の低いシーンの画像と背景に映る建物や海といった物体の位置関係が同じ画像が望ましい。このような画像の取得作業には、試験計画の立案から撮影許可申請、機材の調達などを要し、追加学習のたびに取得作業を繰り返すのはコストが高い。そこで、当社ではインターネット上に大量に公開されている画像データベースを活用する手法を開発し、適用してきた⁽²⁾。しかしながら、1万枚を超える大規模な画像データベースより、学習に資する特定シーンの画像を探索する作業にもまた多大な労力を要しており、開発効率を向上させる上での課題となっている。また、追加学習用画像としてピックアップした画像が例えば“高層ビル”のような特殊な建物のみが背景に映る画像にバリエーションが限定されてしまうと、追加学習によって得られる検知・識別モデルのロバスト性の悪化のリスクがある。

このような課題に対する打ち手として、大量の画像群から階層的クラスティングを用いてシーンごとに画像を分類し、それらの中から所望の特定シーンの画像を、類似画像検索を用いて自動

*1 総合研究所 制御システム研究部 博士(理学)

*2 総合研究所 制御システム研究部

*3 総合研究所 制御システム研究部 主席研究員 技術士(機械部門)

*4 総合研究所 制御システム研究部 主幹研究員 技術士(情報工学部門)

でピックアップする手法を開発した。階層的クラスタリングは，“高層ビル”が映った画像，“高層ビル”や“商業施設”，“住宅街”など建物全般が映った画像，建物のみならず人工構造物全般の画像，といった具合に分類するシーンの粒度を調整することが可能である。これに類似画像検索を適用し，所望の特徴を持った画像をクエリ（ユーザが検索エンジンに入力する画像）として，似た特徴を持った特定シーンの画像をピックアップする。開発技術では，階層的クラスタリングのシーンの粒度をパラメトリックに調整することで，ユーザが求めるバリエーション豊かな画像群を学習対象とすることが可能となった。

2. 階層的クラスタリングを用いた特定シーン画像検索技術

本章では，図 1 に示すように，学習済みモデルのテスト時に，見逃し・誤検知といった検知・識別精度が低いシーンが特定された状況を起点として，大規模データベースから特定シーンの画像の選定に至るまでに有用な類似画像検索と階層的クラスタリング，それらを組み合わせた当社開発の画像検索技術について説明する。既存技術として，まず類似画像検索手法について述べ，次に背景に映る物体の位置関係に注目した類似画像検索となるように施す前処理について言及する。最後に，これらと階層的クラスタリングを組み合わせた特定シーン画像検索技術を説明する。

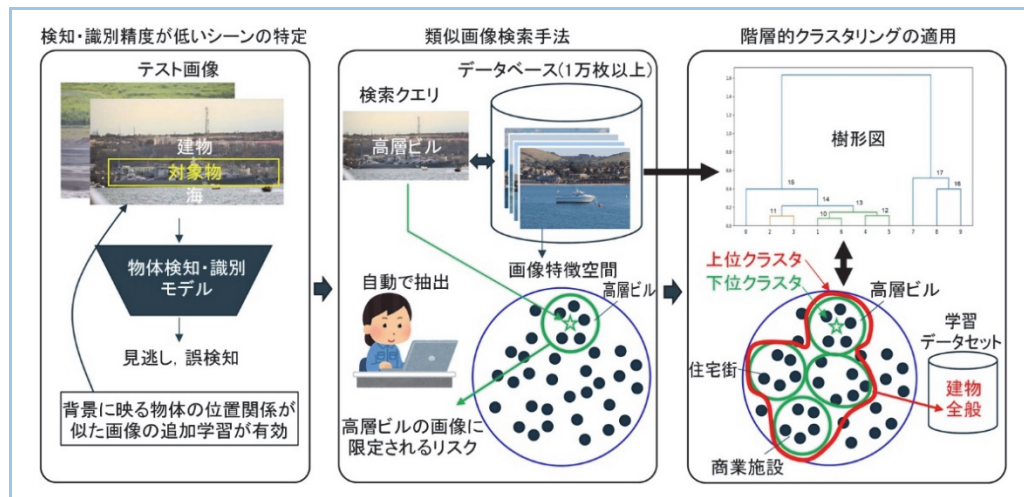


図1 開発手法の概要

(1) 類似画像検索手法

まず，比較したい画像全てを，ニューラルネットワークを有する特徴抽出器に入力し，特徴量ベクトルを出力する。2 枚の画像の特徴の類似度は，以下の式で特徴量ベクトル同士のコサイン類似度により算出する。

$$\cos(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j / \|\mathbf{x}_i\| \|\mathbf{x}_j\|$$

ここで， $\mathbf{x}_i$ はクエリ画像の特徴量ベクトル， $\mathbf{x}_j$ は検索対象のデータベースに含まれる画像の特徴量ベクトルである。データベースに含まれる全ての画像に対してコサイン類似度を算出し，高いスコアとなるものがクエリ画像に近い特徴を持った画像であると判定することができる。ただし，単純に類似度の高い画像を学習候補としてピックアップすると，過学習のリスクがある。そのため，クエリ画像に対して近い画像でもある程度のバリエーションを網羅した画像をピックアップする工夫が必要となる。

(2) 背景に注目した前処理

類似画像検索では，画像の全ての領域に対して特徴抽出器を適用するが，検知・識別の対象となる船舶などが映っている場合，それらの種別や姿勢・向きの特徴の影響が大きく，背景に注目した類似画像検索ができない。そこで，図 2 に示すように対象物の周辺の領域をトリミングし，対象物の領域を黒塗りする前処理を施行する。こうすることで，例えば上が建物，下が海という背景の物体位置関係のみに注目した画像を作成することができる。前処理をした画像に

対して、特徴抽出器を用いて特徴量ベクトルを出力する。特徴抽出器は多くの画像特徴をとらえることができるように学習されたニューラルネットワークであり、IT 企業各社が公開しているものを利用することができる。本報では、背景に映る物体の位置関係の特徴としてとらえたいという目的があるので、画像上の位置関係を小区分に分けてその並びを特徴量化することが可能な Vision Transformer⁽³⁾を特徴抽出器として採用した。トリミング処理及び特徴抽出処理はクエリ画像のみならず、検索対象のデータベースに含まれる画像全てに対して適用する。



図2 背景に注目した前処理

(3) 特定シーン画像検索技術

類似画像検索処理を実行する前に、データベースの画像に対して、階層的クラスタリングを適用し、樹形図で表される階層構造を形成する。樹形図は以下の①～④の手順にて作成する。

- ① データベース内の全ての画像の特徴ベクトルの距離を総当たりで計測
- ② ベクトル間距離が近い画像のペアをクラスタリングし、小規模なクラスタを形成
- ③ クラスタ間距離を評価し、クラスタ同士を合算
- ④ ③を繰り返し、大規模なクラスタを形成

このとき、②、③の特徴ベクトル $\mathbf{u}, \mathbf{v}$ の距離計測及びクラスタ間距離計測は、python ライブラリを利用すれば種々のアルゴリズムを適用できる⁽⁴⁾が、本報では以下のユークリッド距離 r_k 及びクラスタ間距離 R を評価する average linkage 採用した。

$$r_k = \|\mathbf{u} - \mathbf{v}\|_2, R = \frac{1}{N} \sum_{k=1}^N r_k$$

ここで、 N は2つのクラスタ内に含まれる画像ペアの組合せ総数を示す。図3は小規模なデータベースに対して階層的クラスタリングを適用した結果得られる樹形図の例である。縦軸は、画像 No.0～9 の距離 r_k ならびにクラスタ No.10～17 間の距離 R を示している。下位から上位に遷移するとクラスタ規模が大きくなり、含まれる画像のバリエーションが拡大する。所望の特定のシーンのバリエーションを調整する方法を次に示す。

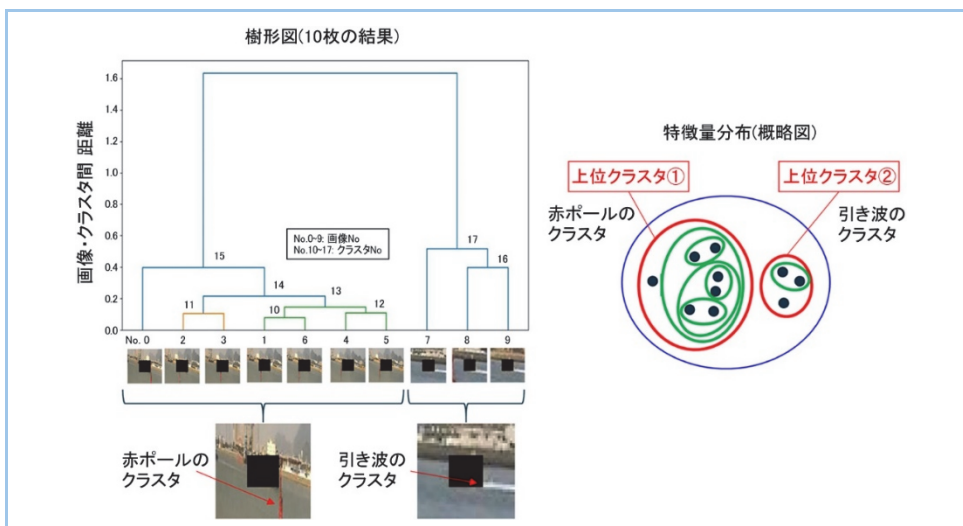


図3 階層的クラスタリングの適用例

図 4 に特定シーンの画像が含まれるクラスタの判定フローを示す。まず、(1)と同様に類似画像検索を実行し、データベースに含まれる全ての画像とクエリ画像のコサイン類似度を評価し、最も類似度の高い画像を決定する。次に、クエリ画像の属する 2 枚以上のクラスタで樹形図の最下位のクラスタに含まれる画像のコサイン類似度の分布の標準偏差 σ を算出する。分布の最頻値をピーク位置として、次に合算される同じレベルのクラスタとのピーク間距離 d を計測する。ピーク間距離 d が以下の条件を満足するとき、合算して上位レベルのクラスタで同じ処理を実行する。

$$d \leq n\sigma, \quad n: \text{整数}$$

上記の処理を繰り返し実行すると、いずれは上記式の条件を満足せず、合算できるクラスタがない階層レベルに達することになる。このレベルに至るまでに合算してきたクラスタを特定シーンのクラスタとして採用する。パラメータ n の大きさを調整することで、特定シーンのクラスタ内の画像のバリエーションが調整可能となる。

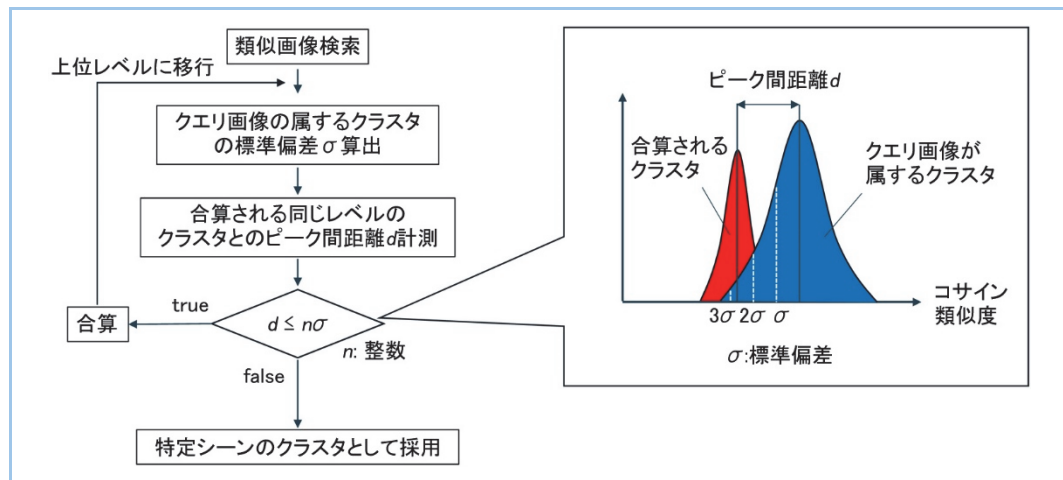


図4 特定シーンの画像クラスタ判定フロー

3. 公開データベースを用いた評価

本章では、2 章で説明した特定シーン画像検索技術を、公開データベース SeaShips⁽⁵⁾を 7080 枚、Mcships⁽⁶⁾を 3000 枚、合計 10080 枚を検索対象として評価した。階層的クラスタリングを適用した結果、図 5 に示す樹形図が得られ、画像特徴ごとにクラスタを整理することができたが、これを目視で分析することは現実的ではないことがわかる。このデータベースに対して、上が建物、下が海となる特定のシーンの画像クラスタを判定することを問題設定とし、図 6 に示すクエリ画像 3 枚を設定した。これらは同じ建物に分類されるが、それぞれ高層ビル、住宅街、商業施設が上に映った画像である。

3 枚それぞれ独立に類似画像検索を実行し、最も類似度の高い画像が属するクラスタをスタートとして上位のクラスタの合算を繰り返し、その都度クラスタ内の画像のコサイン類似度の分布の幅 3σ とピーク間距離を算出した結果を図 6 に示す。コサイン類似度の分布の幅は合算を繰り返すごとに広がる傾向にあり、クラスタ内に含まれる画像のバリエーションが増大する様子がわかる。ピーク間距離は合算の度にピーク位置が変化するため、下位から上位に遷移する過程で決まった傾向はみられないが、No.1964 のクラスタに対して合算処理を実行するタイミングで 3σ を超える大きな差分となり、合算後のクラスタ内にはクエリ画像と遠い画像群が混在することを検知することができる。以上の分析は 2 章(3)で述べたロジックに適合するものであり、No.1964 が特定シーンの画像クラスタとして判定することができた。

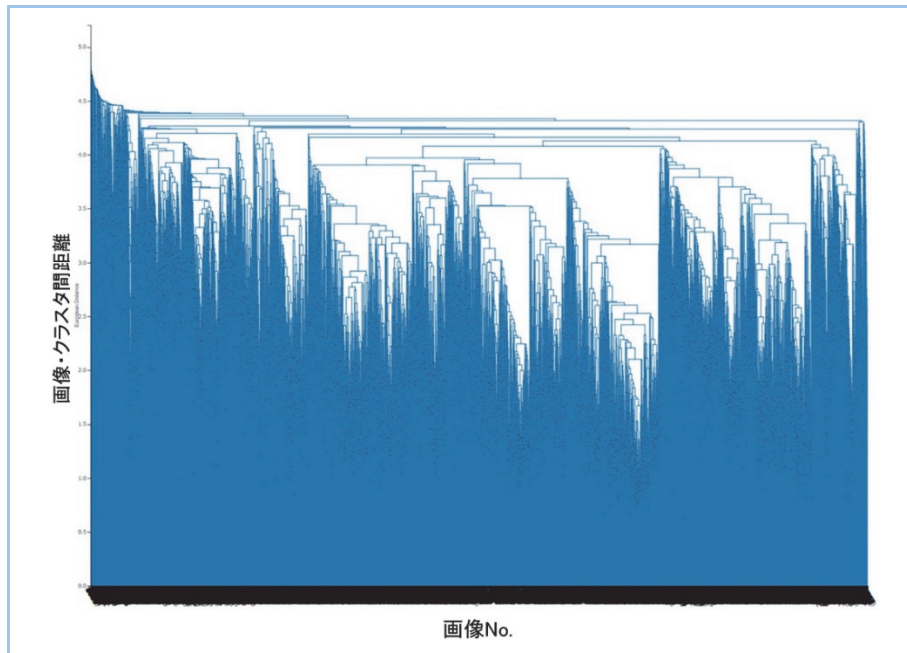


図5 階層的クラスタリングを用いて可視化した公開データベースの樹形図

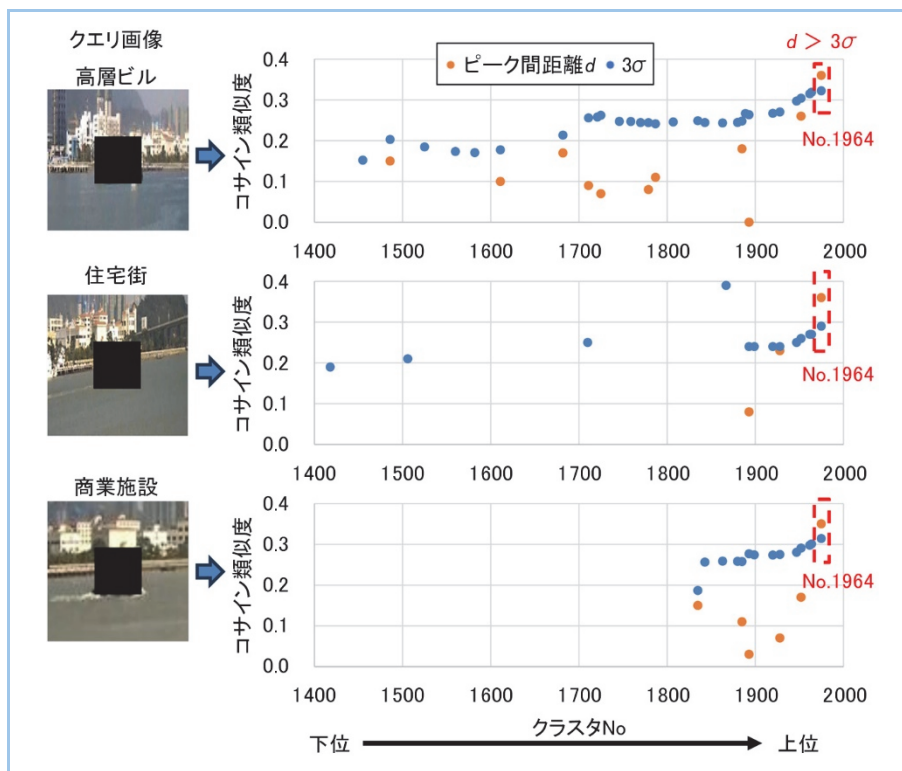


図6 クエリ画像(左)と特定シーン画像クラスタ判定結果(右)

No.1964 のクラスタに含まれる画像は図 7 に示すように所望の物体位置関係を示す画像の他に、誤って抽出された画像が含まれていた。また、No.1964 に含まれない画像の中に所望の画像があり、見落としとなったものも存在した。ただし、建物ではないにせよ人工構造物が含まれているものを誤抽出したり、小さく映っている建物を見逃したりと重篤な検索エラーは生じていないといえる。検索結果の精度を評価するため、見落とし率(当該クラスタに含まれない見落としした画像の枚数 / データベースに含まれる所望の画像の総数)と誤抽出率(所望の画像ではない画像枚数 / 当該クラスタに含まれる画像の総数)を算出した結果、見落とし率 5.7%, 誤抽出率 2.3%となった。大規模なデータベースから所望の画像の大部分を抽出し、誤抽出の割合も極めて低い結果となった。

上記の結果より、本報の検索技術の特性について考察した。まず、図 5 の樹形図からわかるよ

うに、1 万枚を超える大規模データベースの樹形図を目視で分析するためには、本手法における特定シーンのクラスターの自動判定ロジックは不可欠であった。加えて、クエリ画像は 3 枚とも同じクラスターを特定シーンとして判定できたことから、高層ビルや商業施設など建物の種別に依存することなく、建物全般の画像を抽出することが可能な技術といえる。従来の類似画像検索手法であれば、高層ビルのクエリでは、高層ビルが映っている画像のみを抽出する。そのため、他の建物を抽出するためには、所望の種別の建物の画像すべてをクエリとして手元に用意する必要があったのに対して、本開発技術であれば建物の画像どれか 1 枚用意すれば十分である。次に判定クラスターに含まれる画像バリエーションについて、ピーク間距離の閾値(図 4)を広めに設定することで、建物のみならず人工構造物全般をとらえることや、狭く設定することで高層ビルのみに限定するなど調整できる。即ち、ピーク間距離の閾値を追加学習の候補とする画像のバリエーションを調整するパラメータとして運用することが可能である。

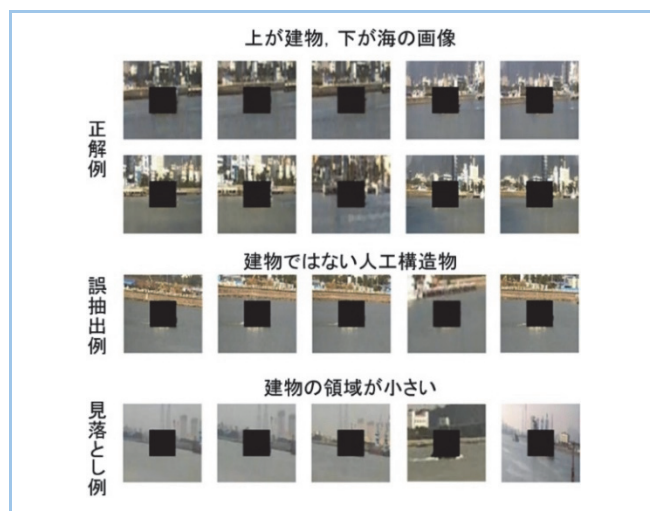


図7 特定シーンと判定された画像と誤抽出・見落とし例

4. まとめ

本報では、物体検知・識別モデルの精度が低いシーンの画像を大規模データベースより自動で検索し、追加学習データとして抽出する方法を開発した。本手法により、画像取得や当該シーンの探索に要する作業工数の削減ならびに開発効率の向上に寄与する。また、検索クエリを高層ビルの画像とした場合でも高層ビルのみではなく、高層ビルを含む建物全般のバリエーションを満足する画像クラスターを抽出することが可能であり、追加学習における過学習の防止や様々な条件でロバストに機能する質の高い物体検知・識別モデルの構築が期待できる。今後は、様々なアプリケーションに展開し、課題点を洗い出すことで、画像認識技術のさらなる高度化を目指す。

参考文献

- (1) L. Jiao et al., A Survey of Deep Learning-Based Object Detection, IEEE Access, vol. 7, pp. 128837-128868, 2019.
- (2) 小林ほか, 深層学習を用いた画像認識における多段階学習による少数の実画像で構築可能な物体検出技術の開発, 三菱重工技報 Vol.61 No.1 (2024) p.1~7
- (3) Dosovitskiy et al., AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE, arXiv:2010.11929v2, 2021.
- (4) The SciPy community, Hierarchical clustering, <https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html>
- (5) Shao et al., SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection, Transactions on Multimedia, vol. 20, no. 10, pp. 2593-2604, 2018.
- (6) Zheng and Zhang, "Mcships: A Large-Scale Ship Dataset For Detection And Fine-Grained Categorization In The Wild," ICME, pp. 1-6, 2020.