



大規模画像言語モデルを用いた物体検出・分類による橋梁用ケーブル点検に関する研究

富谷, 竜一 ; ハスコエト トリスタン エワン, マリ ; 百田, 陽平 ; 峰地, 慎一 ; 堀井, 智紀 ; 高島, 遼一 ; 滝口, 哲也

(Citation)

神戸大学都市安全研究センター研究報告, 29:107-113

(Issue Date)

2025-03

(Resource Type)

departmental bulletin paper

(Version)

Version of Record

(JaLCOI)

<https://doi.org/10.24546/0100496657>

(URL)

<https://hdl.handle.net/20.500.14094/0100496657>



大規模画像言語モデルを用いた物体検出・分類による 橋梁用ケーブル点検に関する研究

Bridge cable inspection using object detection and
classification with vision-language models

富谷 竜一¹⁾
Ryuichi Tomiya
Tristan Hascoet²⁾
百田 陽平³⁾
Yohei Momota
峰地 慎一³⁾
Shinichi Mineji
堀井 智紀³⁾
Tomonori Horii
高島 遼一⁴⁾
Ryoichi Takashima
滝口 哲也⁵⁾
Tetsuya Takiguchi

概要: 橋梁の定期点検は人による直接目視点検が基本であり、多大な労力を要する。また、橋梁においてケーブルは特殊材であり、点検には高い熟練度が必要となる。この課題を解決するために、画像認識技術を活用した省力化と安定化が考えられるが、学習データ量の不足が懸念点となる。近年提案されている、大量の一般データを事前学習した大規模画像言語モデルは、少量データによる追加学習で特定タスクへ適応できるため、学習データ量の不足を補うことができる。本稿では、橋梁用ケーブル点検の省力化と安定化を目的として、大規模画像言語モデルを用いた物体検出・画像分類を活用した異常検出手法を提案し、その有効性を示した。

キーワード: 橋梁用ケーブル、大規模画像言語モデル (Vision-Language モデル)、物体検出、画像分類

1. 序論

橋梁の定期点検は、従来から人による直接目視点検が基本であり、多大な労力が必要となる。特に、橋梁においてケーブルは特殊材であり、点検には高い熟練度や高度な専門知識が必要となる。近年、この課題を解決するために、画像認識技術を活用した省力化と安定化が進んでいるが、学習データ量の不足が懸念点となっている。近年提案されている大量の一般データを用いて事前学習した大規模画像言語モデル (Vision-Language モデル) は、少量データによる追加学習で特定タスクへ適応できるため、学習データ量の不足を補うことができると考えられる。本稿では、橋梁用ケーブル点検の省力化と安定化を目的として、画像認識技術だけではなく、大規模画像言語モデルを橋梁用ケーブルの点検に活用する方法について検討する。

2. 関連研究

近年、橋梁の維持管理において、画像認識技術を用いた異常検出の研究が発展している。これらの技術は、従来の目視点検に比べ、省力化と安定化の観点から重要視されている。Flotzinger^{ら 1)}は、橋梁異常検出のためのベンチマークとして、dacl10k データセットを提供した。このデータセットは、多様な橋梁構造の損傷や異常の画像を含んでおり、異常検出モデルの開発や性能評価に活用されている。Esteghamati^{ら 2)}は、セグメンテーション技術を用いて、鉄製の橋梁画像の腐食領域の推定を実現した。この技術により、腐食領域と深刻度の推定が可能になり、構造物の劣化状況を把握できる。これらの研究において共通して指摘されている課題はデータ量不足で、さらなる性能向上のためには、腐食の深刻度や照明条件などが異なる多様な画像を多く取得することが重要であるとされている。

また、画像認識技術の分野において、大規模画像言語モデル(Vision-Language モデル)と呼ばれる、画像認識技術と自然言語処理を組み合わせた研究が発展しており、大規模な一般データで学習された事前学習済モデルが公開されている。これらのモデルは、画像特徴量と言語特徴量を合わせて大規模な学習を行うことで、未学習のカテゴリの画像を理解するゼロショット能力をもつため、任意のテキストでリクエストを設定できる。また、特定タスクで利用する際、それらのデータでファインチューニングを行うことで、さらなる精度向上が期待される。本稿では、ゼロショット能力を持ち、少量データで特定タスクに適応可能である大規模画像言語モデルの特性を、橋梁用ケーブル点検に活用する。

3. 橋梁用ケーブル異常検出フロー



図1 大規模画像言語モデル(Vision-Language モデル)を用いた橋梁用ケーブル異常検出フロー

本研究で提案する異常検出フローを図1に示す。まず、検出モデル GLIP³⁾を用いて部品を検出する。次に、セグメンテーションモデル SAM2⁴⁾を用いてセグメンテーションマスクを作成し、部品抽出画像を取得する。最後に、部品抽出画像に対して、クラス分類モデル CLIP⁵⁾を用いて異常クラスを分類し、部品ごとの異常状態を取得する。

GLIP は大規模画像言語モデルで、カテゴリを自由に設定できる物体検出モデルである。2,700 万もの大量の画像とテキストのペア（一般データセットやインターネット上から収集したデータ）で学習されており、画像と任意のテキストを入力すると、テキストで指定した物体が検出可能である。本稿では事前学習済モデルをベースとして、橋梁用ケーブル部品検出タスクに適応させている。SAM2 は大規模画像セグメンテーションモデルで、バウンディングボックスをプロンプトとして入力すると、その領域内の物体に対してセグメンテーションが可能である。画像と、GLIP で獲得したボックスを入力して、検出した部品のセグメンテーションマスクを作成する。CLIP は大規模画像言語モデルで、カテゴリを自由に設定できる画像分類モデルである。4 億もの大量の画像とテキストのペア（インターネット上から収集したデータ）で学習されており、画像とテキストの類似度が計算可能である。部品抽出画像と、テキストで指定した異常クラス候補それぞれの類似度を計算し、一番高いクラスを分類結果とする。本稿では事前学習済モデルをベースとして、異常クラス分類タスクに適応させている。本稿では、橋梁ケーブルの画像を撮影して、部品と異常クラスのアノテーションを実施し、実験を行った。対象となる部品はストランドロープ・ケーブルバンド・ケーブルバンドボルトであり、対象となる異常クラスは赤錆全面・赤錆 50%・赤錆点在・赤錆なし(白錆)である。部品検出、異常クラス分類それぞれについて、画像のみを用いた画像認識モデルと比較を行った。

(1) GLIP を用いた橋梁ケーブル部品検出

GLIP を用いてストランドロープ・ケーブルバンド・ケーブルバンドボルトを検出する実験を行った。“wire strand, cable clamp, cable-clamp bolt”というテキストを入力して物体を検出させる。追加学習データの画像数を変化させて、GLIP 事前学習済モデルのファインチューニングの効果を調査した。比較対象として、一

般データ(COCO データセット)事前学習済画像認識モデル YOLOv5x⁶⁾を用いて、ファインチューニングを行った。30 枚の画像をテストデータ(バウンディングボックスの数はストランドロープ 63 個、ケーブルバンド 17 個、ケーブルバンドボルト 53 個)として、精度は $\text{IoU} \geq 0.5$ を正解条件としたときの mean Average Precision (mAP)を用いて評価した。

図 2 から、GLIP は、YOLOv5x と比較して、少数データでのファインチューニングで高い精度のモデル学習が可能であることがわかる。これは、言語特徴量と同時に適応できる大規模画像言語モデルの利点を活用できたからであると考えられる。GLIP で最も良い精度となったモデルの AP は、ストランドロープで 76.20%、ケーブルバンドで 87.08%、ケーブルバンドボルトで 69.54%となった。それぞれ検出スコアの閾値を下げると Recall は 90%を超えており、ほぼ漏れなく部品を検出することが可能である。図 3 は GLIP の検出例で、左は正解例、右はストランドロープ 1 本に対して 2 つ検出がある誤検出の例である。図 4 は GLIP と YOLOv5x (いずれも最も良い精度となったモデルを使用) の検出例の比較で、GLIP は正しく部品を検出できているが、YOLOv5x はケーブルバンドボルトに対して 3 つ検出漏れを起こしている。以上のことから、大規模画像言語モデルを用いることで、少ない学習データ数で高い精度のモデルを構築できることを示した。

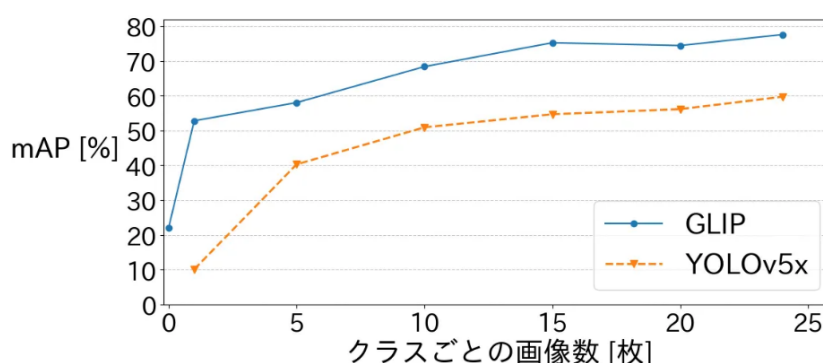


図 2 各検出モデルの学習における学習データ数と精度 (mAP) の関係

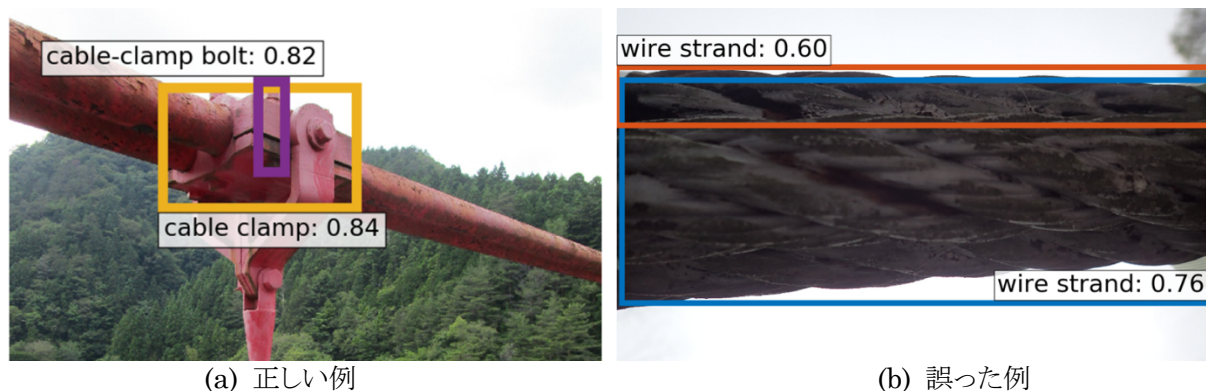


図 3 GLIP による部品検出例



図 4 GLIP と YOLOv5x の部品検出例の比較。YOLOv5x は GLIP が正しく検出できているケーブルバンドボルトについて検出漏れを起こしている。

(2) SAM2 を用いたマスク生成

部品ごとの異常状態を判別するために、それぞれの部品についてピクセル単位でセグメンテーションマスクを生成する。特に橋梁用ケーブルの場合は、多くの部品が密集していることが多く、画像認識技術を用いた部品ごとの異常状態の把握には、マスク生成が不可欠であると考えられる。大規模事前学習済セグメンテーションモデル SAM2 は、バウンディングボックスをプロンプトとして入力することで、バウンディングボックスで囲んだ対象のマスクを生成可能である。

図 5 に、GLIP が出力した橋梁用ケーブルの部品(ストランドロープ)の検出結果であるバウンディングボックスを、SAM2 に入力したときのマスク生成例を示す。バウンディングボックスで囲まれた部品ごとのマスクを生成できていることが分かる。上から 2 本目のロープは、左側の日光が強く当たっている部分が同一物体としてみなされず、一部マスクが生成できていない部分はあるものの、ほぼ部品全体を覆うようなマスクを生成できているため、部品ごとの異常状態の把握が可能になることが考えられる。また、SAM2 は事前学習済モデルであり、本タスクに特化していないが、この例のように検出結果を活用することで、部品ごとのマスク生成に有用であると考えられる。さらに、より高精度なセグメンテーションを可能にするためには、橋梁用ケーブル画像において部品ごとにマスクのアノテーションを付与し、SAM2 をファインチューニングすることが必要であると考えられる。本稿では、事前学習済モデルをそのまま用いて(3)の実験を行っている。



(a) 入力: GLIP によるストランドロープの検出結果 (b) 出力: バウンディングボックスごとのマスク

図 5 SAM2 による部品セグメンテーション結果。GLIP による検出結果であるバウンディングボックスをプロンプトとして入力することにより、部品ごとのマスクを生成できている。

(3) CLIP を用いた異常クラス分類

SAM2 を用いて部品抽出に成功したストランドロープの画像に対して、赤錆全面・赤錆 50%・赤錆点在・赤錆なし(白錆)に分類する実験を行った。“cable with full red rust”、“cable with half red rust”、“cable with pointed red rust”、“cable with no red rust”というテキストと画像の類似度を計算する。比較対象として、一般データ(ImageNet)事前学習済画像分類モデル VGG16⁷⁾を用いて、ファインチューニングを行った。赤錆全面 48 枚、赤錆 50% 53 枚、赤錆点在 67 枚、白錆 54 枚を 5 分割したものをテストデータとして、平均の分類精度(Accuracy)を用いて評価した。

図 6 から、どちらも学習データ数が増加するにつれて精度が向上しているが、CLIP を用いることで、少数データでのファインチューニングで高い精度のモデル学習が可能であることがわかる。これは、言語特徴量と同時に適応できる大規模画像言語モデルの利点を活用できたからであると考えられる。CLIP で最も良い精度となったモデルは、4 クラス分類において 67.25%の正解率であった。図 7(a)は分類例である。左は正しく異常状態を分類できている。右は不正解例で、正しくは「赤錆 50%」であるが、「赤錆なし」と分類されている。追加学習データはデータ数が少なく、類似した画像で「赤錆 50%」とラベル付与されたデータがなく、図 7(b)のように、形や角度が類似した「赤錆なし」とラベル付与されたデータが含まれるため、誤ったと考えられる。誤分類を防ぐため、データ拡張など、追加学習データ増強も重要である。

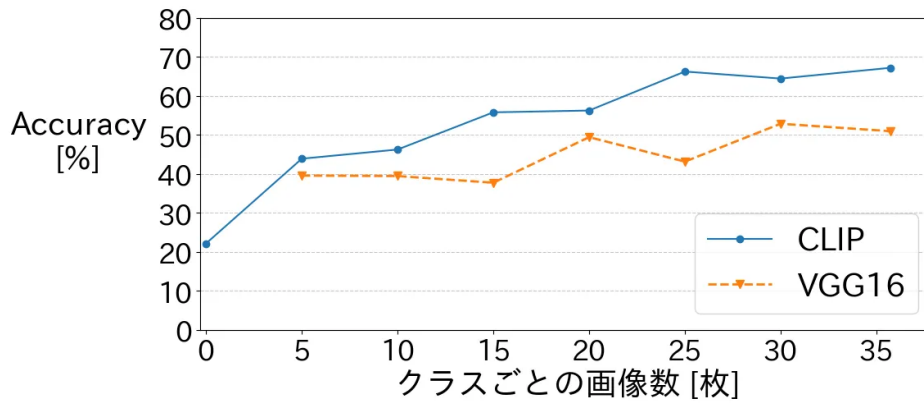


図 6 各分類モデルの学習における学習データ数と精度 (Accuracy) の関係



図 7 ファインチューニングを行った CLIP を用いた異常クラス分類の例とそのファインチューニングに用いた学習データの例

図 8 に、SAM2 を用いて生成したマスク画像を学習・推論に使用する手法(提案手法)と SAM2 を用いずに検出したボックスを切り取った画像を学習・推論に使用する手法の検出精度の比較を示している。SAM2 を用いた方が精度の上昇率が高いことが分かる。これは、マスク画像を用いて画像分類を行った方が、部品の上に注目することができるため、よりよい異常状態の把握が可能になることを示している。図 9 は、SAM2 によるマスク生成を行うことによって正しく異常クラスを分類できている例であり、部品(ストランドロープ)のみに注目することによって、錆の状態を正しく把握できたと考えられる。以上より、SAM2 を用いると、部品ごとの異常状態を取得できる上に、セグメンテーションを用いた方が異常クラス分類の精度が高くなることから、SAM2 を用いることは有用であると分かった。

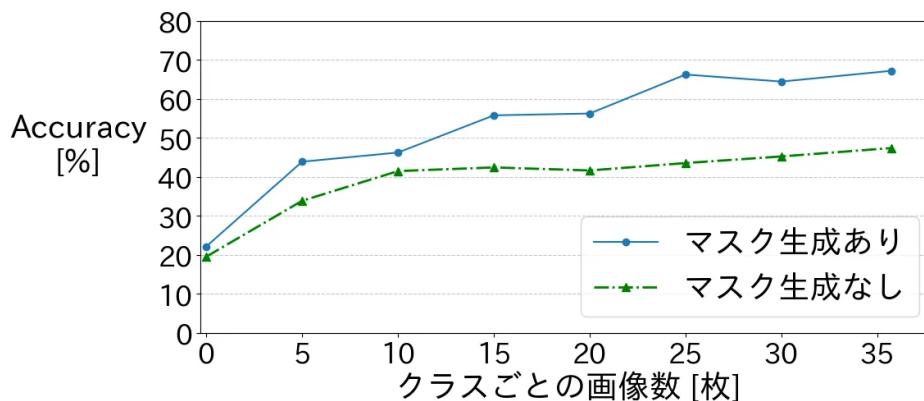
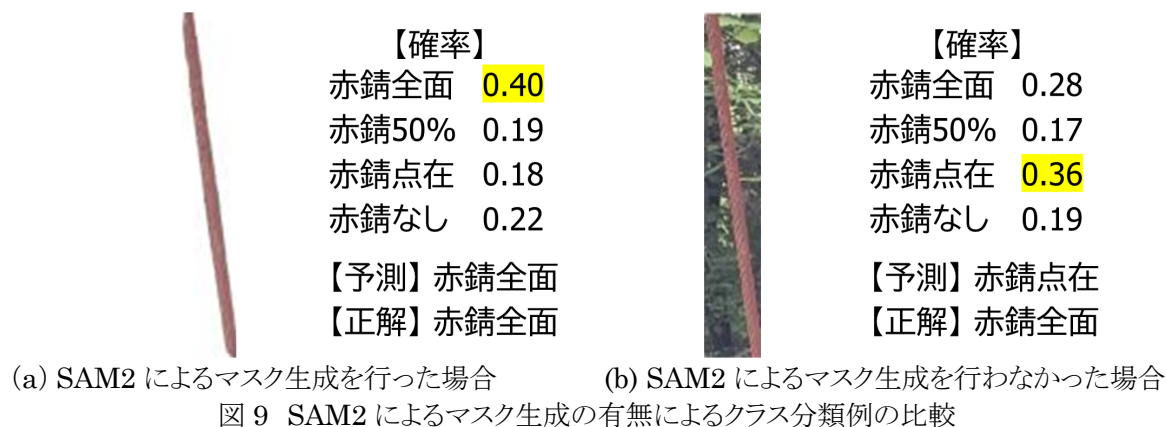


図 8 SAM2 によるマスク生成を行った場合、行わなかった場合の「CLIP 学習における学習データ数と精度の関係」の比較



4. おわりに

本稿では、橋梁用ケーブルの異常検出において、大規模画像言語モデルを用いた手法を提案し、大規模画像言語モデルを活用することで、少量データによる異常検出モデル開発が可能であることを示した。また、大規模画像セグメンテーションモデルである SAM2 を用いることにより、よりよい部品ごとの異常状態の把握が可能になることを示した。実際の橋梁用ケーブル点検に本稿で提案したフローを用いるための今後の課題として、少量データによるデータ拡張、SAM2 による部品抽出精度の改善が考えられる。

参考文献

- 1) Flotzinger, J., Rösch, P. J., Benz, C., Ahmad, M., Cankaya, M., Mayer, H., Rodehorst, V., Oswald, N., Thomas, B., Dacl-Challenge: Semantic Segmentation During Visual Bridge Inspections, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 716-725, 2024.
- 2) Esteghamati, M. Z., Ahmed, M. H., Young, S., Halling, M., A multi-task deep learning-based framework for structural details and corrosion detection in steel bridges, *Bridge Maintenance, Safety, Management, Digitalization and Sustainability*, pp. 3530-3537, 2024.
- 3) Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J. N., Chang, K. W., Gao, J., Grounded language-image pre-training, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965-10975, 2022.
- 4) Ravi, N., Gabeur, V., Hu, Y. T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. J., Carion, N., Wu, C. Y., Girshick, R., Dollár, P., Feichtenhofer, C., SAM 2: Segment Anything in Images and Videos, *arXiv preprint arXiv:2408.00714*, 2024.
- 5) Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., Learning Transferable Visual Models From Natural Language Supervision, *Proceedings of the 38th International Conference on Machine Learning*, pp.8748-8763, 2021.
- 6) Jocher, G., Ultralytics YOLOv5, <https://github.com/ultralytics/yolov5>, 2020. (2025/3/14 閲覧)
- 7) Simonyan, K., Zisserman, A., Very deep convolutional networks for large-scale image recognition, *3rd International Conference on Learning Representations (ICLR 2015)*, pp.1-14, 2015.

筆者: 1) 富谷 竜一、システム情報学研究科システム情報学専攻、学生; 2) Tristan Hascoet、理化学研究所 計算科学研究センター; 3) 百田 陽平、峰地 慎一、堀井 智紀、神鋼鋼線工業株式会社; 4) 高島 遼一、都市安全研究センター、准教授; 5) 滝口 哲也、都市安全研究センター、教授

Bridge cable inspection using object detection and classification with vision-language models

Ryuichi Tomiya
Tristan Hascoet
Yohei Momota
Shinichi Mineji
Tomonori Horii
Ryoichi Takashima
Tetsuya Takiguchi

Abstract

Conventional bridge inspections primarily rely on direct visual inspection by human inspectors, which requires significant costs. Moreover, cables in bridges are made of specialized materials, necessitating a high level of expertise for inspection. To address these challenges, optimizing and stabilizing the inspection process through computer vision has been considered, however, a shortage of training data remains a concern. Recently proposed large-scale Vision-Language models, which are pre-trained on massive amounts of general data, can adapt to specific tasks with additional training on a small dataset, thus addressing the lack of training data. In this paper, we propose an anomaly detection method for bridge cable inspection that utilizes object detection and image classification based on large-scale Vision-Language models, demonstrating its effectiveness in improving the efficiency and reliability of the inspection process.

©2025 Research Center for Urban Safety and Security, Kobe University, All rights reserved.