

ドローンによる河川巡視の自動化を見据えた、 河川敷に点在する投棄物の自動検出手法に関する検討

A STUDY ON AUTOMATED WASTE DETECTION METHODS ON RIVERBEDS: TOWARDS THE AUTOMATION OF RIVER INSPECTION VIA UAV

小島 崇¹・横山 翔一²・吉田 圭介³

Takashi KOJIMA, Shouichi YOKOYAMA and Keisuke YOSHIDA

¹株式会社東京建設コンサルタント 環境防災研究所 (〒170-0004 東京都豊島区北大塚1-15-6)

²学生会員 岡山大学大学院 学術研究院環境生命自然科学学域 (〒700-8530 岡山市北区津島中3-1-1)

³正会員 岡山大学准教授 学術研究院環境生命自然科学学域 (〒700-8530 岡山市北区津島中3-1-1)

本研究は、UAVによる河川巡視の高度化を目的に、教師データの作成負担を解消する新たな物体検出手法を提案している。従来の物体検出モデルは高速・高精度だが、未知のゴミへの対応や教師データの拡充に課題があった。そこで、Autoencoderを用いた教師なしモデルで「背景と異なる物=ゴミ」を抽出し、事前学習済みのVLMでその種類を判別する手法を構築した。高度10mの画像による検証の結果、AutoencoderはYOLOが苦手とする微細なゴミを効果的に検出可能であること、VLM(Qwen3-v1)は高精度で様々なゴミを検出可能であることを確認した。しかし、高度30mや60mではVLMは著しく検出精度が低下した。一方でAutoencoderは30mや60mでも高精度を維持していた。VLM単体での検出は計算時間の長さや位置情報の誤差に課題があるが、本手法のようにAutoencoderで解析対象を絞り込み、VLMの復元率で補正を行うことで、効率かつ精度の高い検出が可能となった。この手法はゴミ検出に留まらず、災害時の人命救助や未経験の河川異状検知など、多様な河川管理業務への応用が期待される。

Key Words: UAV, River patrol, Autoencoder, Vision language model, AI

1. はじめに

河川管理において、モニタリング技術やAI技術の進歩・普及に伴い、DXの促進が様々な試行されている。特に河川巡視では、ドローンを用いた巡視の自動化に向けて、レベル3.5飛行の試行とマニュアルの整備が進められている¹⁾²⁾。ドローンの活用では、河川流速の推定³⁾や3次元地形計測から粒度分布の推定⁴⁾といった適用性に関する研究もおこなわれている。一方で、ドローンの画像に対してAIを適用する試みでは、河川の不法投棄物の検出⁵⁾⁶⁾や地被分類⁷⁾⁸⁾などが報告されている。

これまでの多くの適用事例では、Object DetectionやSemantic SegmentationといったAIは教師ありモデルが用いられており、新たに適用を行う際には、学習用データの蓄積とアノテーション作業が必要となっていた。これに対して、岡野ら⁹⁾や英ら¹⁰⁾はマルチモーダルモデル(画像を入力し、自然言語の回答を得るモデル: Vision

Language Model以降ではVLM)を用いる事で、教師あり学習モデルを用いながら、パブリックデータによる学習済みモデルを適用することで、新規の学習や転移学習を行うことなく、様々な河川投棄物に対して検出を可能としている。この様にVLMを用いる事で、画像内の物体を説明(分類)可能であることは報告されている。しかし、Object Detectionと同様の計数と位置の把握については十分な検証が行われていない。

Object Detectionを用いた河川ゴミの検出について研究が進められ、生成AI(diffusion model)を用いることで、疑似的に生成された教師データを学習させて汎用性の高いモデルの開発についても同様に進められている⁹⁾。

本稿では、実際の河川敷に様々なゴミを仮設し、ドローンを用いて複数の高度から撮影した画像に対して、生成AIにより作成された教師データを学習したObject Detectionを適用し課題・特徴を整理するとともに、VLMを用いた検出を適用し、位置・数の推論が可能かを検証する。また、教師なしモデルとしてAutoencoderを用いた

表-1 河川敷に配置したゴミの種類

ゴミの種類	缶	段ボール	ゴミ袋	ペットボトル	発砲スチロール	その他
プロンプト上の表現	Can	Cardboard	Plastic bag	Plastic bottle	Styrofoam	Othrs
長辺の寸法	10～15cm	60cm	70～90cm	15～20cm	25～40cm	40～60cm

図-1 階段画像の高度による見え方の違い
(上:10m, 中:30m, 下:60m)

方法についても適用し、教師なしモデルの可能性を示す。最後に、AutoencoderとVLMを組み合わせることで、精度向上について提案する。

2. 検証対象画像

本検討では、実際の河川敷を対象に検証を行う。しかし、様々なゴミが投棄されている場所を探し撮影する事は困難であることから、河川敷に表-1に示すゴミをランダムに配置した物をUAVで撮影した画像を用いる。

(1) 撮影条件と検証画像

画像は好天の日中の正午ごろの影が短くなる時間帯を選んで撮影を行った(2025年8月22日14時～15時ごろ)。対象フィールドは河川敷の「草地」と堤防に設置された階段と裸地を含む「階段」の2か所を対象とした。UAVには、DJI Mavic 2 Proを用い、カメラは1インチCMOS, FOV:77°, 静止画解像度:5,472×3,648ピクセルを用いた。対象地点を中心に同一高度で複数地点撮影しオルソ化した画像を解析に用いている。高度によって撮影範囲が広がることから、検証範囲を定め、画像内の検証範囲を切り出し適用する。高度30m, 60mは先行研究⁹⁾の推論データと同じ画像を用いている。

検証画像は表-2に示す6枚として、人が目視にて確認したゴミを内容別に整理した結果を正解として、以降の検討に用いる。高度別にゴミの見え方がどの様になるかを図-1に階段、図-2に草地の画像を示す。高度60mでの画像は、ゴミの輪郭がぼやけて不鮮明な画像となっている。特にペットボトルや缶などの小さなゴミは形状を

表-2 検証対象画像とゴミの一覧

撮影場所	高度 (m)	画像横 (pixel)	画像縦 (pixel)	ゴミの総数 (個)	缶 (個)	段ボール (個)	ゴミ袋 (個)	ペットボトル (個)	発砲スチロール (個)	その他 (個)
草地	10	13,376	10,688	80	18	5	23	26	7	1
	30	3,932	3,322							
	60	2,064	1,593							
階段	10	9,424	12,464	88	20	5	27	27	7	2
	30	2,992	4,688							
	60	1,894	2,807							

図-2 草地画像の高度による見え方の違い
(上:10m, 中:30m, 下:60m)

確認する事が難しい。草地の評価範囲は横: 28.9m, 縦: 38.2m, 階段は横: 41.0m, 縦: 32.8mとなっており、高度10mから撮影した場合、長辺が23cmのペットボトルは75ピクセル程度となる。したがって、高度60mでは、短辺が数ピクセル程度になることから、缶やペットボトルは30m程度の高度から撮影する必要があると考えられる。

(2) 高解像度のファイルの取り扱い

検証対象の画像は高度10mでは約10,000×10,000ピクセルと非常に大きな画像となっている。実際にドローンによる巡視を考えた場合、複数垂直画像からオルソ画像を作成し、一連の大きな画像ファイルとなる事が想定される。一方でAIモデルに対して巨大な画像ファイルを読み込ませ、一括で処理する事はメモリ容量の制約から難しいと考えられる。本検討では、1,024×1,024のタイル状に分割し、オーバーラップ率20%とすることで、境界付近のゴミが分割されるのを防ぐとともに、重複部分については、NMS(Non-Maximum Suppression)を用いて検出結果の最適化を行う。

3. 河川の投棄物検出手法

本稿では大きく3つのAIモデルを用いる。(1)生成AIを用いて作成された河川の投棄物(ゴミ)の教師データAIGC⁹⁾を学習させたObject Detectionモデル、(2)汎用VLM、(3)ゴミの無い河川敷画像を学習させたAutoencoder(教師なしモデル)。また、Autoencoderではゴミの内容を分類する事ができないため、VLMを用いてゴミの分類及び誤検出のチェックを行う。

表-3 Object Detectionによる検出結果

モデル名	サイズ	草地									階段								
		10m			30m			60m			10m			30m			60m		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Yolov8	n	0.210	0.731	0.326	0.750	0.570	0.648	0.795	0.403	0.535	0.313	0.733	0.439	0.814	0.402	0.539	0.870	0.241	0.377
	x	0.299	0.667	0.413	0.831	0.747	0.787	0.783	0.468	0.585	0.243	0.570	0.340	0.870	0.460	0.602	0.833	0.241	0.374
Yolo11	n	0.290	0.833	0.431	0.798	0.848	0.822	0.810	0.610	0.696	0.339	0.756	0.468	0.786	0.506	0.615	0.923	0.289	0.440
	x	0.340	0.680	0.453	0.806	0.684	0.740	0.793	0.546	0.646	0.355	0.709	0.473	0.800	0.506	0.620	0.946	0.422	0.583
Yolo26	n	0.311	0.590	0.407	0.750	0.342	0.470	0.667	0.026	0.050	0.366	0.733	0.488	0.968	0.345	0.509	0.966	0.337	0.500
	x	0.278	0.692	0.397	0.831	0.684	0.750	0.800	0.416	0.547	0.127	0.256	0.170	0.882	0.172	0.289	0.882	0.181	0.300

(1) Object Detection

Object Detectionモデルとして、Ultralytics社のYoloを用いる。Yoloには複数のバージョンとそれぞれパラメータ数の異なるサイズ(n,s,m,l,x)があり、本稿では Yolo v8, 11, 26の3モデルについて、2サイズn, xを適用しバージョン、サイズによる精度の違いを6パターン比較検証する。

(2) VLM

著者らは既往の研究⁹⁾においてLLaVA v1.5¹⁰⁾を用いてゴミの分類を行っているが、新しい高精度なモデルが様々提案されていることから、本稿ではLLaVA v1.6, Qwen3-vl¹¹⁾, Ministral-3¹²⁾, Gemma3¹³⁾について適用し、それぞれのモデルの特性と河川のゴミ検出に適したモデルの選定を行う。ここで、VLMについてもパラメータ数の異なるモデルがそれぞれ公開されているが、本研究ではローカル環境 (GPU 1台) で実行可能なモデルの中で評価を行う事とした。

(3) Autoencoder

Autoencoderを用いた異常 (ゴミ) の検出では、画像内の復元誤差の分布から求めるが、MSE等の画像全体の復元率に基づくLoss関数を用いた学習では、微小な異常や緻密な形状の違いによる異常を検出する事が難しい¹⁴⁾。河川のゴミについて考えた場合、缶やペットボトルといったゴミはUAVの飛行高度・カメラの画角・解像度によるが、十分な解像度で撮影できない可能性が高い。本研究では高度10mと非常に低い高度から撮影をしており、形状を鮮明にとらえていると考えられるが、高度30mや60mの画像では、草本や周辺の状態によって形状を判読する事が難しいものも存在しているため、Paulら¹⁴⁾の提案するSSIMを損失関数に加え、形状・コントラストの復元精度を高める学習を行った。厳密には、SSIMをマルチスケールに拡張したMS-SSIM¹⁵⁾とMSEを1:1の割合で損失を評価している。また、学習データには検証対象画像のゴミの無いエリアを切り取り学習させている。本来は、ゴミの無い異なる場所で撮影した画像を用いる事が望ましいが、本稿ではAutoencoderによるゴミの検出可能性を確認する意図で試行しているため学習データと検証データの類似性が高いデータを用いており、Autoencoderによる推論精度が高めに評価される可能性がある。異なる地点での画像を用いた検証は今後実施し、汎化性の確認等を進める予定である。

Autoencoderの入出力画像のサイズは16×16, 32×32,

64×64の3つサイズについて検証を行った。AutoencoderのネットワークにはUnet¹⁶⁾を用い、Encoderの出力サイズが4×4(16次元)になるように、それぞれの階層を設定している。

4. AIモデルの検出結果

本章では、表-3で示したAIモデルそれぞれのゴミ検出結果を示す。ここで、ゴミのクラス(ペットボトルや段ボールなど)については評価を行わず、一律ゴミ(Waste)として扱う。

(1) Object Detectionによる検出結果

Object Detectionによるゴミ検出精度を表-4に示す。10mの高精細な画像が最も精度が高くなると期待されるが、表-3から高度30mのF1スコアが全体的に高い結果となった。この原因として、推論時のConfidence(信頼度)の閾値を先行研究⁹⁾と同じ0.1としているため、10mの正解の検出数は30mや60mより高く、検出漏れが少ない結果である一方で誤検出も多くなり、結果精度が低下している。Confidenceの閾値を最適化することで、10mの誤検出が減り精度が向上すると考えられるが、共通の学習データ(AIGC)に基づいて決定された閾値を用いると、誤検出が増える問題が生じる。汎化性の高い教師データを作成するとともに、適用時のConfidenceの閾値の決定方法については、引き続き検討が必要と考えられた。

モデルやサイズによる精度の違いは高さや場所による違いに比べ小さく、適用する画像の影響が最も大きい事がわかった。また、場所による違いでは同じ高度で比較すると、草地の検出精度が高い結果となった。これは、階段の縁石や目地等の人工的な構造や色彩の分布が教師データに含まれていないため、検出精度が低下したと考えられる。

(2) VLMによる検出結果

同様にVLMについてもタイル単位に下記の図-3のプロンプトを用いてゴミに該当するオブジェクトを指定し、ゴミの位置を取得しObject Detectionと同様の精度評価を行った。結果を表-4に示す。この結果から、高度30m, 60では検出が極めて困難な事がわかる。また、VLMのモデルによる精度の違いも顕著でQwen3-vlは他のモデルに

表-4 VLMによる検出結果

モデル名	サイズ	草地									階段								
		10m			30m			60m			10m			30m			60m		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Gemini3	4b	0.025	0.064	0.036	0.167	0.152	0.159	0.000	0.000	0.000	0.033	0.070	0.045	0.068	0.058	0.062	0.000	0.000	0.000
	12b	0.038	0.231	0.065	0.044	0.013	0.020	0.333	0.013	0.025	0.043	0.186	0.070	0.000	0.000	0.000	0.000	0.000	0.000
	27b	0.188	0.192	0.190	0.154	0.051	0.076	0.250	0.013	0.025	0.060	0.035	0.044	0.111	0.012	0.021	0.200	0.012	0.023
LLaVA	7b	0.004	0.013	0.006	0.095	0.025	0.040	0.000	0.000	0.000	0.117	0.244	0.159	0.040	0.023	0.029	0.000	0.000	0.000
	13b	0.011	0.039	0.017	0.035	0.038	0.037	0.000	0.000	0.000	0.006	0.012	0.008	0.000	0.000	0.000	0.008	0.012	0.010
	34b	0.137	0.410	0.206	0.029	0.038	0.033	0.000	0.000	0.000	0.005	0.012	0.007	0.000	0.000	0.000	0.011	0.024	0.015
Ministral3	3b	0.116	0.103	0.109	0.222	0.127	0.161	0.333	0.026	0.048	0.077	0.035	0.048	0.200	0.081	0.115	0.143	0.012	0.022
	8b	0.047	0.103	0.064	0.097	0.127	0.110	0.114	0.052	0.071	0.053	0.081	0.064	0.155	0.149	0.152	0.053	0.024	0.033
	14b	0.359	0.244	0.290	0.344	0.139	0.198	0.667	0.026	0.050	0.222	0.093	0.131	0.539	0.081	0.140	0.000	0.000	0.000
Qwen3-vl	2b	0.781	0.321	0.455	0.500	0.013	0.025	0.000	0.000	0.000	0.880	0.256	0.396	0.000	0.000	0.000	0.000	0.000	0.000
	4b	0.833	0.385	0.526	0.333	0.013	0.024	0.000	0.000	0.000	0.769	0.349	0.480	0.667	0.023	0.044	0.000	0.000	0.000
	8b	0.882	0.385	0.536	0.333	0.025	0.047	0.000	0.000	0.000	0.865	0.372	0.520	0.667	0.023	0.044	0.000	0.000	0.000
	32b	0.882	0.385	0.536	0.214	0.038	0.065	0.000	0.000	0.000	0.769	0.698	0.732	0.250	0.012	0.022	0.000	0.000	0.000

表-5 Autoencoderによる検出結果

モデル名	サイズ	草地									階段								
		10m			30m			60m			10m			30m			60m		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
AE(Unet)	16x16	0.581	0.684	0.628	0.656	0.747	0.698	0.694	0.442	0.540	0.449	0.460	0.455	0.325	0.609	0.424	0.577	0.542	0.559
	32x32	0.606	0.506	0.552	0.842	0.405	0.547	0.879	0.377	0.527	0.600	0.414	0.490	0.507	0.425	0.463	0.875	0.506	0.641
	64x64	0.688	0.705	0.696	0.714	0.506	0.593	0.788	0.338	0.473	0.510	0.605	0.553	0.640	0.368	0.467	0.800	0.386	0.520

Carefully examine this drone image and detect ALL trash/waste/litter items.
 Target objects (detect ALL sizes, including SMALL ones):
 - plastic bottles, - cans, - plastic bags, - styrofoam, - cardboard, - any other visible trash
 DO NOT detect:
 - stones or rocks, - natural objects (leaves, branches, soil, grass), - people or vehicles, - buildings or infrastructure
 For EACH trash item found, provide:
 - content: item type (e.g., "plastic bottle", "can"), - bbox: [x, y, width, height] in pixels (top-left origin), - confidence: 0.0 to 1.0
 Return ONLY a JSON array in exactly this format:
 [{"content": "type", "bbox": [x, y, w, h], "confidence": 0.9}, ...]
 Return [] if no trash is found.
 IMPORTANT: Pay close attention to SMALL items that are easy to miss.

図-3 ゴミの位置と種別取得用のプロンプト

比べ小規模なモデル(2b)でもF1スコアが高く、Object Detectionと同程度の精度になっている。特にPrecisionが0.8程度とObject Detectionより高い値になっている。このことから、VLMの進歩により今後も精度が向上する事が期待される。一方で、VLMを用いるためには、高精細な画像が必要で、UAVの飛行高度は低く高画質のカメラの利用が望ましい。図-4にはQwen3-vl:8bによる推論結果を示す。ペットボトルや段ボール、ゴミ袋など小ささまざまなゴミを検出しているが、ペットボトル・缶の検出精度が低い傾向にあり、位置や大きさは大きく間違える場合がある。

(3) Autoencoderによる検出結果

Autoencoderの出力はモデルのサイズ単位に評価されるため、ゴミの数を検出する事が難しい。そこで、復元率の閾値(MS-SSIM + MSE Loss)をモデルサイズ毎に定め、検証対象画像に対してのLossのヒートマップを作成する。このヒートマップに対して閾値以上の領域を矩形で抽出することで、Object Detection同様のゴミの検出が可能となり、同様の精度評価を行う。結果を表-5に示す。閾値は本稿ではMC-SSIM+MSE > 2.0e-5としている。閾値の最適化については、今後検証用の画像を撮影し、テスト

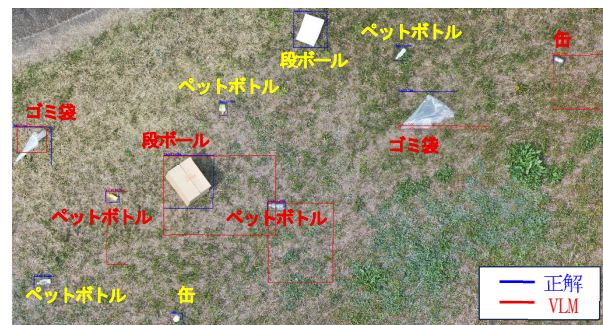


図-4 VLM(Qwen3-vl:8b)を草地 10mに適用した結果
 黄字:検出漏れのゴミの種類, 赤字:検出されたゴミの種類

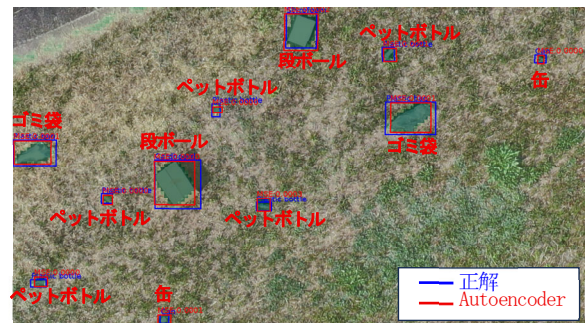


図-5 Autoencoder (16×16)を草地 30mに適用した結果
 黄字:検出漏れのゴミの種類, 赤字:検出されたゴミの種類

用の画像に対して評価を行うなど、物体検出のConfidenceの最適化同様の検証が必要となる。表-5から10mの場合には64×64のタイルにおいて高い精度でゴミを検出しており、F1スコアは草地ではObject DetectionやVLMより高い結果となった。さらに、Autoencoderでは、高度30m、60mでも高いF1スコアとなっており、高度が高い場合にはタイルのサイズが小さい方(16×16や32×32)が有効と示された。図-5に草地の高度30mの画像に適用した結果を示す。VLMとくらべ大きさや位置の再現性が高く、小さいものも良好に検出されている。一方で、石などが大量にゴミとして検出されており、全体の精度を下けている事がわかった。解決には石が含まれる範囲を教師データに入れるなどが考えられる。階段の推論で

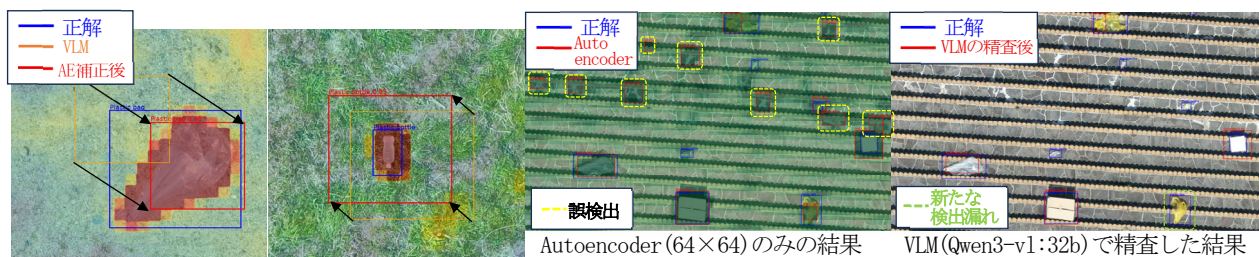


図-6 Autoencoderの復元誤差を用いてVLMの検出位置補正
矢印:VLMで推定された位置をAEで修正する移動ベクトル

図-7 Autoencoderの検出結果をVLMで精査・分類した結果
左:4(3)のAutoencoderのみの検出結果, 右:4(4)2)のVLMを
持ちいて検出結果の精査とクラスのカテゴリを行った結果

表-6 VLMの結果をAutoencoderで補正した検出結果

モデル名	サイズ	草地						階段					
		VLMのみ			VLM+Autoencoder			VLMのみ			VLM+Autoencoder		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Qwen3-vl	8b	0.822	0.385	0.536	0.771	0.346	0.478	0.865	0.372	0.520	0.667	0.140	0.231
	32b	0.822	0.385	0.536	0.560	0.654	0.604	0.769	0.698	0.732	0.673	0.407	0.507

表-7 Autoencoderの結果をVLMで補正した検出結果

モデル名	サイズ	草地						階段					
		Autoencoderのみ			VLMで精査した結果			Autoencoderのみ			VLMで精査した結果		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Auto encoder	16x16	0.581	0.684	0.628	0.731	0.695	0.713	0.449	0.460	0.455	0.930	0.459	0.615
	32x32	0.606	0.506	0.552	0.731	0.594	0.655	0.600	0.414	0.490	0.919	0.456	0.609
	64x64	0.688	0.705	0.696	0.923	0.758	0.832	0.510	0.605	0.553	0.721	0.377	0.495

は、草地と比べ誤検出が多く検出精度が低い(図-7左)。

(4) AutoencoderとVLMの複合したモデルの可能性

これまでに3種類のAIモデルをそれぞれ適用し、精度と課題について整理した。Autoencoderを用いた方法では、30mや60mの高度でも高い精度でゴミの検出が可能であることが示され、UAVを用いた河川巡視に対して有用であると考えられる。一方で、低空・高精細な画像の場合、VLMを用いた方法も良いと考えられる。しかし、どちらの方法も、誤検出(石など)や位置や大きさの誤りといった課題がある。

そこで、VLMとAutoencoderを組み合わせ、より精度の高いゴミの検出モデルの構築を考える。1)VLMによって検出されたゴミをAutoencoderで補正する方法、2)Autoencoderによって検出されたゴミをVLMで精査・分類する方法について検討を行う。ただし、前述のVLMの適用結果から高度30m、60mでは著しく精度が低下する事から、本検討は高度10mのみを対象とした。

1) VLMの結果をAutoencoderで補正する方法

VLMで検出したゴミのBounding Boxの中心座標をAutoencoderのヒートマップのピーク座標に移動させる。探索範囲はVLMの検出範囲を縦・横それぞれ±256ピクセル広げた範囲とした。図-6に位置ずれの結果と位置補正の例を示す。前述の推論結果からVLMにはQwen3-vl、Autoencoderには64×64のモデルを用いた。VLM単独による結果(表-4 Qwen3-vl抜粋)とAutoencoderを用いて補正した結果の比較を表-6に示す。草地では表中の太字で示すサイズ32bは、Autoencoderによる補正を行うことで、Recallを改善できることが確認できた。しかし、Precisionが低下し、正しく検出されている物に対しても過度な補正が行われ、補正後のNMSにより重複が除外

されたためと考えられる。階段では全体的に精度が低下する結果となっており、階段の形状、縁石などのAutoencoderの復元率の低さが原因であった。

2) Autoencoderによって検出されたゴミをVLMで精査・分類する方法

Autoencoder(64×64)の出力結果を中心に512×512ピクセルを切り出し、VLM(Qwen3-vl:32b)に入力し、前述のプロンプト(図-3)にてゴミかを判別し、ゴミでない場合には棄却する。ゴミ周辺のみを切り出すと、背景との差や段ボールや発泡スチロールなどの一様な色彩の内容を正しく理解できないため、評価範囲を切り出している。Autoencoder単独による結果(表-5 10m抜粋)とVLMを用いて補正した結果との比較を表-7と図-7に示す。Autoencoderの結果をVLMで補正する場合、Autoencoderの誤検出をVLMで補正可能であり、また、Autoencoderのゴミの判定閾値を下げる(Loss < 1.0e-5)ことで見逃したゴミについても、VLMで再評価し正しく検出できるため、全体的に精度が向上する結果となった。また、位置と範囲の検出精度は、VLM(図-4)とAutoencoder(図-5)から、Autoencoderによる検出結果を用いた方が良い事がわかる。本手法は、Object Detectionと比べても高い精度でゴミを検出可能である事が示された。

5. 推論時間

本稿で提案するゴミの検出手法を河川巡視に適用する場合、推論時間が短いことが望まれることから、検証対象の画像の推論時間を計測した。推論時間の計測には、

表-8 AIモデルの推論時間の比較

単位(秒)	適用画像	草本			階段		
		10m	30m	60m	10m	30m	60m
Object Detection	Yolo 26n	18	2	1	17	2	1
	Yolo 26x	26	3	1	21	2	1
VLM	Qwen3-vl:8b	1,764	378	348	1,127	664	551
	Qwen3-vl:32b	3,975	1,872	1,415	1,714	976	882
Auto encoder	16x16	780	78	78	609	87	87
	32x32	624	78	78	522	87	87
	64x64	546	78	78	435	87	87
VLM+AE	8b+64x64	2,470	477	380	1,799	773	594
AE+VLM	64x64+32b	763	189	135	588	208	174

CPU: AMD EPYC 9654(96core,2CPU), GPU: Nvidia H200を用いた。AIモデル毎の推論時間の比較を表-8に示す。

Object Detectionは非常に高速に動作し、高度10mの画像で20～30秒で実行可能であった。対してVLM単独で推論を行った場合、10,000×10,000ピクセル程度の画像の場合、Qwen3-vl:8bでは30分程度かかり、32bのモデルでは、草本の画像を処理するのに、1時間以上かかることが分かった。AutoencoderはVLMとObject Detectionの中間程度の時間となり、10分程度で高度10mの画像を処理可能であることが分かった。また、AE+VLMではVLM単体より短時間に解析でき、Autoencoderのみと比べ大きくは時間がかからないことが分かった。また、小型・低消費電力のエッジ向けワークステーションNvidia GB10を用いた場合でも、推論時間はH200の約2倍程度であった。

6. まとめ

本稿では、複数のゴミの検出モデルを適用し、それぞれの精度や課題を整理した。Autoencoderは高度30mや60mからの撮影でも一定の精度でゴミ検出が可能であることが示された。また、VLMは形状が鮮明な画像に対して有効な方法と考えられた。ただし、VLMのモデルによって精度が大きく変わることから、事前に適切なモデルを選定する必要がある。VLMではゴミの位置や大きさを誤ることが多く、Autoencoderを用いた補正の可能性を確認した。また、VLMをAutoencoderの検出結果の再評価に用いる事で精度が向上する事が示された。推論時間も10,000×10,000ピクセルの画像を1時間程度と短時間に処理可能であることが分かった。

本稿で提案する、教師なしモデル(Autoencoder)と事前学習済みのVLMとそれぞれを組み合わせる方法は、ゴミ検出に限らず、様々な未経験の異状・変状をAutoencoderで検出し、VLMで説明可能であると考えられ、AIの未経験の事象に対する外挿問題に対して有用な解決策となると考えられる。

謝辞：本研究は科学研究費補助金(課題番号：25K07936)の助成を受けて実施した。記して謝意を表する。

参考文献

- 1) 国土交通省: 第1回ドローンを活用した河川巡視・点検への適用検討会 資料4, 2024.
- 2) 国土交通省: 第4回ドローンを活用した河川巡視・点検への適用検討会 資料1, 2026.
- 3) 星野孝総, Namal Rathnayak, Dang Tuan Linh, Upaka Rathnayake: ドローンを用いた上流河川の3D解析と流量予測の検証, 第40回ファジィシステムシンポジウム, pp. 741-744, 2024.
- 4) 原田守啓: 河川環境モニタリングのニーズにUAVは応えられるか, RIVER RONT, Vol.84, pp.12-15, 2017.
- 5) 岡野将太, 吉田流人, 高橋悠太, 藤井純一郎, 天野匠純: 河川の不法投棄物検知におけるマルチモーダルモデルの適用検証, The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2N6-GS-10-04, 2023.
- 6) 英佑成, 吉田圭介, 藩是均, 小島崇, 西山哲: UAV写真と大規模マルチモーダルモデルを用いた河川投棄物検出手法の高度化に関する研究, 土木学会論文集, Vol. 82, No. 16, 25-16082, 2026.
- 7) 藤井純一郎, 吉田龍人, 天方匡純, 山下隆義: UAV写真測量と深層学習を活用した河道管理手法の研究, AI・データサイエンス論文集 1巻, J1号, pp.588-595, 2020.
- 8) 荒木 健, 箱石 健太, 一言 正之, 古木 宏和, 倉上 健, 与賀田 隆史, 坂本 俊二, 大石 哲也, 山口 里実: 空中写真を用いた深層学習による河道の水域抽出手法の研究, 河川技術論文集, 27巻, pp.35-40, 2021.
- 9) 下江大地, 藩是均, 吉田圭介, 西山哲, 小島崇: UAVを活用した河川巡視時のペットボトル検出モデルにおける画像生成AIの応用, 土木学会論文集, Vol. 81, No. 16, 24-16180, 2024.
- 10) Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee: Visual Instruction Tuning, arXiv, arXiv:2304.08485v2, 2023.
- 11) Qwen Team: Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al.: Qwen3-VL Technical Report, arXiv, arXiv:2511.21631v2, 2025.
- 12) Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, et al.: Ministral 3, arXiv, arXiv:2601.08584v1, 2026.
- 13) Gemma Team: Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al.: Gemma 3 Technical Report, arXiv, arXiv:2503.19786v1, 2025.
- 14) Paul Bergmann, Sindy Lowe1, Michael Fauser, David Sattlegger, Carsten Stege: Improving Unsupervised Defect Segmentation by Applying Structural Similarity To Autoencoders, arXiv, arXiv:1807.02011v3, 2019.
- 15) Hang Zhao, Orazio Gallo, Iuri Frosio, Jan Kautz: Loss Functions for Image Restoration with Neural Networks, arXiv, arXiv:1511.08861v3, 2018.
- 16) Olaf Ronneberger, Philipp Fischer, and Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv, arXiv:1505.04597v1, 2015.

(2026. 4. 2受付)